Statistical Science
1994, Vol. 9, No, 3, 416–428

Some Applications of Number-Theoretic Methods in Statistics

Kai-Tai Fang, Yuan Wang and Peter M. Bentler

Abstract. Number-theoretic methods (NTM's) are a class of techniques by which representative points of the uniform distribution on the unit cube of R^s can be generated. NTM have been widely used in numerical analysis, especially in evaluation of high-dimensional integrals. Recently, NTM's have been extended to generate representative points for many useful multivariate distributions and have been systematically applied in statistics. In this paper, we shall introduce NTM's and review their applications in statistics, such as evaluation of the expected value of a random vector, statistical inference, regression analysis, geometric probability and experimental design.

Key words and phrases: F-discrepancy, number-theoretic method, multivariate distribution, multivariate statistics, quasi–Monte Carlo method.

1. INTRODUCTION

Number-theoretic methods (NTM's) or *quasi–Monte Carlo methods* are a class of methods which represent a combination of number theory and numerical analysis. As noted by Niederreiter (1978), "The widest range of applications, and indeed the historical origin of these methods, is found in numerical integration, but related matters such as interpolation problems and the numerical solution of integral equations can also be dealt with successfully." Korobov (1959, 1989), Niederreiter (1978, 1988, 1992) and Hua and Wang (1981) give a comprehensive review in a bibliographic and historical setting.

The reader may have the following questions. What is a number-theoretic method and how is it applied to various problems in statistics? We shall illustrate that all the applications can be reduced to one key problem: how to find a set of points called an NT-net which are uniformly scattered in the s-dimensional unit cube C^s. A so-called number-theoretic method is a method by which we can generate such a set in a certain sense. Note that a "uniformly scattered" set of points generated by NTM is not a sample from the uniform distribution on C^s, the unit hypercube in R^s, by the Monte Carlo method. Figure 1a, b present the plots for the two sets of points of size 17.

Although there is a close relationship between NTM and the Monte Carlo method, it appears that only a few statisticians have directed their attention to NTM and their applications in statistics. The first applications of the NTM in statistics were naturally in evaluating probabilities and moments of a multivariate distribution (e.g., Fang and Wu, 1979, and Zhang and Fang, 1982). Fang (1980) and Wang and Fang (1981) were first to apply the NTM idea to experimental design, and they proposed a new design which is called a *uniform design*. This may be the first application of NTM in statistics except in integration. Since then many nice results were obtained on the uniform design in many areas of applications in China. Shaw (1988) gave a detailed discussion on applications of NTM to Bayesian statistics, mainly for numerical computation of posterior density and moments. Recently Wang and Fang (1990a, b, 1992), Fang and Wang (1990, 1991), Fang, Yuan and Bentler (1992) and Fang, Zhu and Bentler (1993) have systematically studied applications of NTM in statistics. Most results mentioned in this paper are from these references and the forthcoming book of Fang and Wang (1993).

Why are NTM's powerful? Let us see a simple example.

Kai-Tai Fang is Professor, Department of Mathematics, Hong Kong Baptist College, and Professor, Institute of Applied Mathematics, Academia Sinica, Beijing, China. Yuan Wang is Professor, Institute of Mathematics, Academia Sinica, Beijing, China. Peter M. Bentler is Professor, Department of Psychology, University of California, Los Angeles, California 90024-1563.

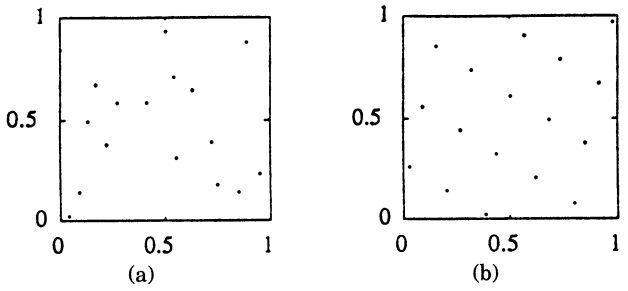

FIG. 1. *Two kinds of sets*: (a) *random numbers*; (b) *an NT-net.*

EXAMPLE 1. Suppose that random vector $\mathbf{x} \sim N_3(0, I_3)$ and we want to evaluate the probability of $\mathbf{x}$ falling in $C^3 = [0, 1]^3$, that is,

$$p = \int_0^1 \int_0^1 \int_0^1 (2\pi)^{-3/2} \exp\left\{-\tfrac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right\} \cdot dx_1\, dx_2\, dx_3 \tag{1}$$
$$= \int_{C^3} f(\mathbf{x})\, d\mathbf{x},$$

say. In fact, the probability $p = [\Phi(1) - \Phi(0)]^3 = 0.039772181953$, where $\Phi(x)$ is the cdf of the standard normal distribution. Suppose that $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a set of points in C^3. If $\{\mathbf{x}_k\}$ is uniformly scattered in C^3, we can estimate p by

$$\hat{p} = \frac{1}{n} \sum_{k=1}^{n} f(\mathbf{x}_k). \tag{2}$$

Let us choose the following three sets of points: (i) equi-lattice points

$$\left\{\frac{2i-1}{2m}, \frac{2j-1}{2m}, \frac{2k-1}{2m}, \quad i, j, k = 1, 2, \ldots, m\right\};$$

(ii) random numbers in C^3 generated by the Monte Carlo method; and (iii) points generated by the good lattice point method, one of the NTM's described in Section 2. We choose the points from the table given by Hua and Wang (1981) in (iii). Since the number of points in (i) should be of the form $n = m^3$, we choose the minimum m such that $m^3 > n$. Table 1 gives comparisons of the errors $p - \hat{p}$ obtained when using the three sets of points. In this case, the number-theoretic method is the best.

In general, if $f(\mathbf{x})$ is a continuous function of finite variation in C^s, the integral

$$p = \int_{C^s} f(\mathbf{x})\, d\mathbf{x}$$

can be estimated by $\hat{p}$ in (2), where $\{\mathbf{x}_k\}$ can be generated by one of the above three methods. It is known that under a certain condition, as $n \to \infty$, $|\hat{p} - p| \le O(n^{-1/s})$ by using equi-lattice points; $|\hat{p} - p| \le O(n^{-1/2})$ by using random numbers; and $|\hat{p} - p| \le O(n^{-1} \log^s n)$ by NTM.

In this paper we shall discuss some applications of NTM's in statistics and present some new results. The paper is organized as follows. In Sections 2 and 3 we introduce the NTM in a general sense; in particular, we shall recommend some methods of generating representative points of a multivariate distribution. In Section 4 we shall consider evaluation of $\mathbb{E}(f(\mathbf{x}))$, where $\mathbf{x}$ is a random vector with a given density function and f is a continuous function. In particular, when the domain of an integral is not a rectangle, some useful methods are suggested. The solution of many statistical problems needs theory and various algorithms in optimization. However, most current algorithms in optimization require that the objective function is unimodal and differentiable. In Section 5 we introduce a sequential number-theoretic method for optimization (SNTO), which is available for multiextremal global optimization problems, and we discuss its applications in statistics. NTM's also can be applied to statistical inference, such as tests for multinormality and for sphericity, and to robust estimation of the mean vector. We will treat these applications in Section 6. In the last two sections, we give applications of NTM in geometric probability and other topics.

TABLE 1
Errors $p - \hat{p}$ by the three sets of points

n	Equi-lattice points	n	Random numbers	n	Good lattice points
64	−2.22E − 04	35	−2.12E − 03	35	−1.16E − 04
125	−1.41E − 04	101	1.94E − 03	101	5.38E − 05
729	−4.36E − 05	597	6.35E − 05	597	−3.98E − 06
1728	−2.45E − 05	1626	−6.56E − 05	1626	6.69E − 06
5832	−1.09E − 05	5037	−3.25E − 05	5037	2.63E − 07
39304	−3.07E − 06	39029	1.34E − 05	39029	2.44E − 09

2. NUMBER-THEORETIC METHODS

In this section we shall briefly introduce NTM in terms of "statistical language." Let $F(\mathbf{x})$ be a continuous multivariate distribution in R^s and n be a given integer. We are required in many applications to find n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in R^s such that they form a good representation for $F(\mathbf{x})$. As we shall see, a set of good representative points (or rep-points for simplicity) can be generated by NTM. There are a number of measures for the closeness of the representation. The most commonly used measure is the so-called *discrepancy*. A more general concept is given as follows.

DEFINITION 1. Let $\mathcal{P} = \{\mathbf{x}_k, k = 1, \ldots, n\}$ be a set of points in R^s and $F_n(\mathbf{x})$ be its empirical distribution, that is,

$$F_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I\{\mathbf{x}_i \leq \mathbf{x}\},$$

where $I\{A\}$ is the indicator function of A, and all inequalities are understood with respect to the componentwise order of R^s. Then

$$D_F(n, \mathcal{P}) = \sup_{\mathbf{x} \in R^s} |F_n(\mathbf{x}) - F(\mathbf{x})| \tag{3}$$

is said to be the F-discrepancy of $\mathcal{P}$ with respect to $F(\mathbf{x})$.

When $F(\mathbf{x})$ is the uniform distribution on $C^s = [0, 1]^s$, denoted by $U(C^s)$, the F-discrepancy reduces to the common discrepancy in the literature (cf. Hua and Wang, 1981). In this case we shall use $D(n, \mathcal{P})$ instead of $D_F(n, \mathcal{P})$, and discrepancy instead of F-discrepancy. Obviously the F-discrepancy is a measure of representation of $\mathcal{P}$ with respect to $F(\mathbf{x})$ and is just the Kolmogorov-Smirnov distance.

DEFINITION 2. For given $F(\mathbf{x})$ and n, a set $\mathcal{P}^* = \{\mathbf{x}_k^*, k = 1, \ldots, n\}$ is called the set of optimum rep-points with respect to $F(\mathbf{x})$ if $D_F(n, \mathcal{P}^*) = \min_{\mathcal{P}} D_F(n, \mathcal{P})$, where $\mathcal{P}$ runs over all sets of n points in R^s.

The following lemma shows that a set of optimum rep-points always exist for every given continuous univariate distribution.

LEMMA 1. *Let $F(x)$ be a continuous distribution function and let $F^{-1}(x)$ be its inverse function. Then the set $\{F^{-1}((2i - 1)/(2n)), i = 1, \ldots, n\}$ with F-discrepancy $1/(2n)$ is the set of optimum rep-points with respect to $F(x)$.*

When $s > 1$ it is difficult to find the optimum rep-points even in the simple case of $U(C^s)$. Hence we want to find a set of points that has a low F-discrepancy.

An open problem in number theory states that for every given set $\mathcal{P}$ of n (≥ 2) points and every $s \geq 2$ we have

$$D(n, \mathcal{P}) \geq C(s) n^{-1} \log^{s-1} n, \tag{4}$$

where $C(s)$ is a constant depending on s. For $s = 2$ this conjecture was proved by Schmidt in 1964 [see Schmidt (1970)]. Thus if we can find a sequence of sets $\mathcal{P}_n$, where $\mathcal{P}_n$ has n elements such that the order of $D(n, \mathcal{P}_n)$ is near to the right-hand side of (4), then $\mathcal{P}_n$ can be regarded as a set of representative points of $U(C^s)$.

DEFINITION 3. Let $\mathcal{P}_n$, $n \in \mathcal{N}$, where $\mathcal{N}$ is an infinite subset of nonnegative integers, be a sequence of sets of points in R^s with a certain structure, and let $F(\mathbf{x})$ be an s-dimensional distribution function. If $D_F(n, \mathcal{P}_n) = O(n^{-1+\varepsilon})$ as $n \to \infty$, where $0 < \varepsilon < \frac{1}{2}$, the points of $\mathcal{P}_n$ are called rep-points of $F(\mathbf{x})$.

When $F(\mathbf{x})$ is the uniform distribution $U(C^s)$, the points of $\mathcal{P}_n$ are called uniformly scattered on C^s if $D(n, \mathcal{P}_n) \to 0$ as $n \to \infty$. In the literature most authors use "uniformly distributed" for our "uniformly scattered." In statistics, the words "uniformly distributed" have an exact meaning which is different from the present one. Therefore, we recommend using the word "scattered." For simplicity a set of rep-points of the uniform distribution on a bounded domain D is called an *NT-net* on D. Points of an NT-net on C^s are often called quasirandom numbers in the literature, because in some circumstances quasi-random numbers can be used instead of random numbers.

For the random number sequence $\mathcal{P}_n$ generated by the Monte Carlo method, Chung (1949) and Kiefer (1961) pointed out that

$$D(n, \mathcal{P}_n) = O\big(n^{-1/2} (\log \log n)^{1/2}\big) \tag{5}$$

with probability 1. Halton (1960) proved that for $s \geq 1$ there exist an infinite sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \}$ satisfying

$$D(n, \mathcal{P}_n) = O\big(n^{-1} (\log n)^s\big), \tag{6}$$

where $\mathcal{P}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and for $s \geq 2$ there exists a sequence $\mathcal{P}_n = \{\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nn}\}$ and

$$D(n, \mathcal{P}_n) = O\big(n^{-1} (\log n)^{s-1}\big). \tag{7}$$

His results suggest that we can find a sequence $\mathcal{P}_n$ with a lower discrepancy than that generated by the Monte Carlo method. By comparing (5) and (6) or

(7), it is now clear that NTM provide a more powerful tool than the Monte Carlo method when s is not very large. We now introduce several methods by which the sequence of NT-net on C^s can be obtained.

2.1 The glp Set

The set obtained by a so-called *good lattice point modulo n* is called the glp set, which is often used and is convenient for computations.

Let $(n; h_1, \ldots, h_s)$ be a vector of integers satisfying $1 \le h_i < n$, $h_i \ne h_j$ for $i \ne j$, and $s < n$. Let

$$(8)\quad x_{ki} = \left\{\frac{2kh_i - 1}{2n}\right\}, \quad i = 1, \ldots, s,\ k = 1, \ldots, n.$$

where $\{x\}$ is the fractional part of x. Then the set $\mathcal{P}_n = \{\mathbf{x}_k = (x_{k1}, \ldots, x_{ks}),\ k = 1, \ldots, n\}$ is called the set of *lattice points* of the *generating vector* $(n; h_1, \ldots, h_s)$. If the sequence $\mathcal{P}_n$ has discrepancy $D(n, D_n) = O(n^{-1+\varepsilon})$, the set $\mathcal{P}_n$ is called a glp set.

Korobov (1959) and Hlawka (1962) proved the existence of a sequence of glp sets. For the practical use of NTM's we need the generating vector $(n; h_1, \ldots, h_s)$ for given n. It is very heavy computational work to find the best generating vector and associated set of lattice points that has the smallest discrepancy among all possible sets of lattice points. Therefore, Korobov (1959) suggested considering $(h_1, \ldots, h_s)$ to be the form

$$(9)\quad (h_1, \ldots, h_s) = \left(1, a, a^2, \ldots, a^{s-1}\right) \pmod{n},$$

where a is an integer and $0 < a < n$. Many tables of generating vectors can be found in Hua and Wang (1981) for $1 < s < 19$, in Wang and Fang (1981) for small n and in Shaw (1988) for large s and small n.

2.2 The gp Set

Let $\gamma = (\gamma_1, \ldots, \gamma_s) \in R^s$. If $\mathcal{P}_n$ forms the first n points of

$$(10)\quad \left\{\left(\{\gamma_1 k\}, \ldots, \{\gamma_n k\}\right),\ k = 1, 2, \ldots\right\},$$

where $\{x\}$ denotes the fractional part of x, with discrepancy

$$D(n) = O\left(n^{-1+\varepsilon}\right) \quad \text{as } n \to \infty,$$

then the set $\mathcal{P}_n$ is called a gp set and γ a *good point*.

Baker (1965) and Schmidt (1970) proved the existence of the gp set. The following are some useful good points:

(a) Let $p_1, \ldots, p_s$ be the first s primes. Take

$$(11)\quad \gamma = \left(\sqrt{p_1}, \ldots, \sqrt{p_s}\right).$$

(b) Let p be a prime and $q = p^{1/(s+1)}$. Take

$$(12)\quad \gamma = \left(q, q^2, \ldots, q^s\right).$$

The reader can find other useful methods such as Halton, scrambled Halton, Haber, Hammersley, Faure and Sobol sequences in Hua and Wang (1981), Shaw (1988) and Niederreiter (1987, 1988, 1992). Shaw recommends the use of the glp set by comparison among several methods. Our experience leads to a similar conclusion, but other methods are still useful.

3. GENERATION OF REP-POINTS OF A MULTIVARIATE DISTRIBUTION

The NTM's mentioned in Section 2 only concern the generation of rep-points of the uniform distribution $U(C^s)$. However, rep-points for a given multivariate distribution are often required in many problems. The Monte Carlo method provides various techniques, such as the inverse transformation method, the compositional method, the acceptance–rejection method and the conditional distribution method for generating an observation from a given distribution. Most of these techniques can be used similarly in generating rep-points of a given distribution. However, some methods, such as acceptance–rejection, are difficult to apply in NTM.

Let $\mathbf{x}$ be a random vector in R^s, and let $F(\mathbf{x})$ be its distribution function. Suppose that $F(\mathbf{x})$ is continuous and $\mathbf{x}$ has a stochastic representation

$$(13)\quad \mathbf{x} = \mathbf{h}(\mathbf{y}),$$

where $\mathbf{y} \sim U(C^t)$, $t \le s$, and $\mathbf{h}$ is a continuous function on C^t. We want to generate rep-points for $F(\mathbf{x})$. The natural idea is as follows: we first generate an NT-net $\{\mathbf{c}_k,\ k = 1, \ldots, n\}$ on C^t, and then let $\mathbf{x}_k = \mathbf{h}(\mathbf{c}_k)$, $k = 1, \ldots, n$. Then $\{\mathbf{x}_k\}$ is a set of rep-points of $F(\mathbf{x})$. To show this idea to be true, we need to find the F-discrepancy of $\{\mathbf{x}_k\}$ with respect to $F(\mathbf{x})$. When $F(\mathbf{x})$ has independent marginals $F_i(x_i)$, that is,

$$(14)\quad F(\mathbf{x}) = F(x_1, \ldots, x_s) = \prod_{i=1}^{s} F_i(x_i),$$

the inverse transformation method suggests

$$(15)\quad \mathbf{x}_k = \left(F_1^{-1}(c_{k1}), \ldots, F_s^{-1}(c_{ks})\right)', \quad k = 1, \ldots, n,$$

where $\{\mathbf{c}_k = (c_{k1}, \ldots, c_{ks})',\ k = 1, \ldots, n\}$ is an NT-net on C^s. Then the F-discrepancy of $\{\mathbf{x}_k\}$ with respect to $F(\mathbf{x})$ equals the discrepancy of $\{\mathbf{c}_k\}$. In general, it is very difficult to find the F-discrepancy of $\{\mathbf{x}_k\}$. Let us see a simple example first.

EXAMPLE 2. Let $\mathbf{x} = (x_1, x_2)'$ be uniformly distributed on the unit disk $B = \{(x, y): x^2 + y^2 \le 1\}$, and consider the polar coordinates

$$x_1 = R\cos(2\pi\theta), \quad x_2 = R\sin(2\pi\theta), \tag{16}$$

where $(R, \theta) \in C^2$. It is easy to show that R and θ are independent with respective cdf's

$$F_R(x) = x^2, \quad 0 \le x \le 1, \qquad F_\theta(x) = x, \quad 0 \le x \le 1. \tag{17}$$

Let $\{(r_k, \theta_k)', k = 1, \ldots, n\}$ be an NT-net on C^2 with discrepancy d. Then the above idea suggests $\{\mathbf{x}_k = (\sqrt{r_k}\cos(2\pi\theta_k), \sqrt{r_k}\sin(2\pi\theta_k)), k = 1, \ldots, n\}$ to be a set of rep-points of $\mathbf{x}$. We can calculate the F-discrepancy of $\{\mathbf{x}_k\}$ even though it is not very easy. There is no analytic relation between the discrepancy of $\{(r_k, \theta_k)\}$ and the F-discrepancy of $\{\mathbf{x}_k\}$.

The more serious problem is that the F-discrepancy of $\{\mathbf{x}_k\}$ is not invariant under orthogonal transformations. It is more natural to use a measure of uniformity of $\{\mathbf{x}_k\}$ on B which is invariant under orthogonal transformations. Therefore, we need the concept of quasi-F-discrepancy. Let $\mathbf{x} \in R^s$ be a random vector with a continuous cdf $F(\mathbf{x})$ and a stochastic representation (13). Let $\mathcal{P} = \{\mathbf{c}_k, k = 1, \ldots, n\}$ be an NT-net on C^t, and let $\mathcal{P}_F = \{\mathbf{x}_k = \mathbf{h}(\mathbf{c}_k), k = 1, \ldots, n\}$. For any $\mathbf{r} \in C^t$, let

$$G_{\mathbf{r}} = \{\mathbf{x}: \mathbf{x} = \mathbf{h}(\mathbf{y}), \mathbf{y} \le \mathbf{r}\}, \tag{18}$$

and let $N(\mathbf{r}, \mathcal{P}_F)$ be the number of points in $\mathcal{P}_F$ such that $\mathbf{x}_k$ falls in $G_{\mathbf{r}}$. Then

$$D_F^*(n, \mathcal{P}_F) = \sup_{\mathbf{r} \in C^t} \left| \frac{N(\mathbf{r}, \mathcal{P}_F)}{n} - P_F(G_{\mathbf{r}}) \right| \tag{19}$$

is called the quasi-F-discrepancy of $\mathcal{P}_F$ with respect to $F(\mathbf{x})$, where $P_F(G_{\mathbf{r}})$ is the probability of $\mathbf{x}$ falling in $G_{\mathbf{r}}$. We can prove the following.

THEOREM 1. *Under the above notations we have*

$$D_F^*(n, \mathcal{P}_F) = D(n, \mathcal{P}). \tag{20}$$

For an illustration of the sense of quasi-F-discrepancy we go back to Example 2, where

$$G_{\mathbf{r}} = \{(x_1, x_2): x_1 = r\cos(2\pi\theta), x_2 = r\sin(2\pi\theta), 0 \le r \le \sqrt{r_1}, 0 \le \theta \le r_2\}$$

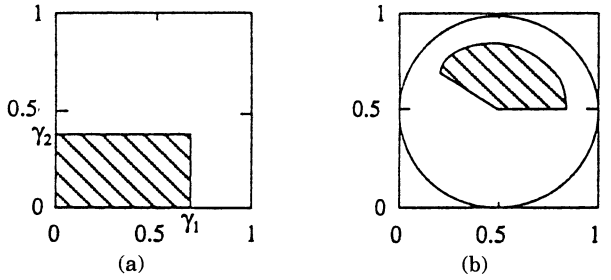

FIG. 2. *Relation of discrepancy and quasi-F-discrepancy; (a) discrepancy; (b) quasi-F-discrepancy.*

(cf. Figure 2). The areas of the rectangle $[0, r_1] \times [0, r_2]$ and the fan-shaped region $G_{\mathbf{r}}$ are $r_1 r_2$ and $\pi r_1 r_2$, respectively. The ratio of area of $[0, r_1] \times [0, r_2]$ and C^2 equals $r_1 r_2$ and the ratio of areas of $G_{\mathbf{r}}$ and B, because the latter is $\pi r_1 r_2/\pi = r_1 r_2$. Furthermore, the quasi-F-discrepancy in this case is invariant under orthogonal transformation. Therefore, the quasi-F-discrepancy can be considered a good measure of representation.

Wang and Fang (1990a) introduced F-and quasi-F-discrepancy and gave an algorithm for generating an NT-net of the uniform distribution on various domains G as well as rep-points of elliptically contoured and multivariate Liouville distributions (cf. Fang, Kotz and Ng, 1990). We shall show that an NT-net on some bounded domain G is extremely useful in simulation, experimental design and geometric probability.

4. EVALUATION OF EXPECTED VALUE OF A FUNCTION OF A RANDOM VECTOR

Evaluation of $E(g(\mathbf{x}))$, where $\mathbf{x}$ is an s-dimensional random vector with a cdf $F(\mathbf{x})$ and a pdf $p(\mathbf{x})$ on $G \subset R^s$, is often required in applied statistics. Obviously, we have

$$E(g(\mathbf{x})) = \int_G g(\mathbf{x})p(\mathbf{x})\,dv = \int_G f(\mathbf{x})\,dv \equiv I(f, G), \tag{21}$$

where $f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$ and dv is the volume element of G.

There are many deterministic quadrature formulas for computing (21) with well-behaved integrands and small s (Genz, 1991). There are many useful methods for computing (21) with $\mathbf{x}$ having a multivariate normal distribution. Unfortunately, there are few algorithms for $\mathbf{x}$ having a multivariate nonnormal distribution. If the function g fails to be regular (i.e., have continuous derivatives of moderate order) Monte Carlo integration has been recommended

in the literature. Let $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \ldots$ be an i.i.d. sequence, then by the strong law of large numbers we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) = E\big(g(\mathbf{x})\big) \tag{22}$$

with probability 1. If $E\big(g(\mathbf{x})\big)^2 < \infty$, the central limit theorem implies that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - E\big(g(\mathbf{x})\big)\right) \to_{\mathcal{L}} N\big(0, \sigma^2(g)\big) \tag{23}$$

as $n \to \infty$, where $\sigma^2(g) = \mathrm{Var}(g(\mathbf{x}))$ and "$\to_{\mathcal{L}}$" means convergence in distribution. Therefore, the rate of convergence of the Monte Carlo method is, on average, $O(1/\sqrt{n})$ and in no case worse than $O(\sqrt{\ln(\ln n)/n})$ by the law of the iterated logarithm.

NTM's can be employed for numerical evaluation of multiple integrals and are recommended by many authors. Let $\mathcal{P}_F = \{\mathbf{x}_k,\ k = 1, \ldots, n\}$ be a set of reppoints of $F(\mathbf{x})$. The expectation $E(g(\mathbf{x}))$, where $\mathbf{x} \sim F(\mathbf{x})$ can be approximated by

$$I_n(g) = \frac{1}{n}\sum_{k=1}^{n} g(\mathbf{x}_k). \tag{24}$$

This method is called the *NT-mean method*. When G is a closed and bounded domain in R^s and the volume of G, $v(G)$, can be calculated analytically, $I(f, G)$ in (21) can be approximated by

$$I_n(f, G) = \frac{1}{n} v(G) \sum_{k=1}^{n} f\big(\mathbf{y}_k\big), \tag{25}$$

where $\{\mathbf{y}_k\}$ is an NT-net on G. When G is the unit sphere, the study of distributions over G is the statistical basic of directional data (see Mardia, 1972, and Watson, 1984). If G is the simplex

$$T_s = \big\{\mathbf{x}\colon \mathbf{x} = (x_1, \ldots, x_s)',\ x_i \geq 0,\ i = 1, \ldots, s,\ x_1 + \cdots + x_s = 1\big\},$$

then the observations of distributions on G are called compositional data (see Aichison, 1986, and Fang, Kotz and Ng, 1992). The following result gives an upper bound for the error, $I_n(f, G) - I(f, G)$, when $D = C^s$.

THEOREM 2 (Koksma–Hlawka). *Suppose that $\mathcal{P}$ is a set of n points in C^s with discrepancy $D(n, \mathcal{P})$ and that $f(\mathbf{x})$ is a function of finite variation with total variation $V(f)$ over C^s. Then we have*

$$|I(f, C^s) - I_n(f, C^s)| \leq V(f) D(n, \mathcal{P}). \tag{26}$$

The case of $s = 1$ of this theorem was established by Koksma, and it was generalized by Hlawka to $s > 1$ (cf. Hua and Wang, 1981, Chapter 5).

Another upper bound for the error with bounds of derivatives of order s is given as follows. Suppose that the partial derivatives are bounded

$$\left|\frac{\partial^m f}{\partial x_{i_1} \cdots \partial x_{i_m}}\right| \leq L, \quad 1 \leq i_1 < \cdots < i_m \leq s \text{ and } 1 \leq m \leq s. \tag{27}$$

Then we can derive

$$|I(f, C^s) - I_n(f, C^s)| \leq 2^s L D(n, \mathcal{P}),$$

and consequently by Theorem 2 and the discussion in Section 2 we have

$$|I(f, C^s) - I_n(f, C^s)| \leq O\big(n^{-1+\varepsilon}\big)$$

or even

$$|I(f, C^s) - I_n(f, C^s)| \leq O\big(n^{-1}(\log n)^s\big),$$

which is better than the Monte Carlo method when s is not very large. Actually the rate of convergence $O(n^{-1}(\log n)^s)$ can still be improved, such as $O(n^{-\alpha}(\log n)^{\alpha s})$, $\alpha > 0$, or $O(n^{-1}(\log n)^{s/2})$ if the integrand is smooth enough. The above results about the rate of convergence of $I_n(f, G)$ with $G = C^s$ have been extended into the case of $G \neq C^s$ under some condition.

Since there are a number of methods for generating an NT-net on C^s with low discrepancy, we want to know which method, in general, is the best for approximating $I(f, G)$. Recently, Pagés and Xiao (1991) compared several methods including the Halton method, the scrambled Halton method, the gp set with γ in (11), on a selection of smooth periodic functions. They concluded that the gp set with $\gamma = (\sqrt{p_1}, \ldots, \sqrt{p_s})$ worked best. Unfortunately, they did not include the glp set in their comparisons. In our experience the performance of the glp set is as good as that of the gp set. Shaw (1988) considered weighted approximations to

$$I(f, \mathcal{P}, w) = \frac{1}{n}\sum_{i=1}^{n} w_i f(\mathbf{x}_i)$$

and discussed applications in Bayesian statistics.

Many variance reduction techniques in Monte Carlo methods such as stratified sampling, importance sampling, correlated sampling and the method of antithetic variates (Rubinstein, 1981) also can be

TABLE 2
Orthant probability by methods of the NT-mean and symmetrization

n	135	597	1010	2440	5037	39024
Error by NT-mean method	1.00E − 2		5.93E − 4	2.97E − 4	−1.77E − 4	3.23E − 5
Error by symmetrization method	8.59E − 4	1.17E − 4	1.85E − 5	8.00E − 7		

used in evaluating $E(g(\mathbf{x}))$ and can improve the efficiency of NTM's for integration. The following example gives an application of the method of antithetic variates in NTM's, called the symmetrization method, for integration.

EXAMPLE 3. Suppose we want to calculate the orthant probability

$$p = \int_0^\infty \int_0^\infty \int_0^\infty n_3(\mathbf{x}; \mathbf{0}, \mathbf{R})\, d\mathbf{x},$$

where $n_3(\mathbf{x}; \mathbf{0}, \mathbf{R})$ is the pdf of $N_3(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = (\rho_{ij})$, $\rho_{11} = \rho_{22} = \rho_{33} = 1$ and $\rho_{12} = \rho_{13} = \rho_{23} = 0.5$. The p-value is known and equals to 0.5 by Gupta (1963). Let $\{\mathbf{y}_k,\ k = 1, \ldots, n\}$ be an NT-net on $[0, 5]^3$. Then

$$p \approx \int_0^5 \int_0^5 \int_0^5 n_3(\mathbf{y}; \mathbf{0}, \mathbf{R})\, d\mathbf{y} \approx \frac{1}{n} \sum_{k=1}^{n} n_3(\mathbf{y}_k; \mathbf{0}, \mathbf{R}).$$

The second row of Table 2 gives some numerical results. Using a symmetrization method, the corresponding results are listed in the last row of Table 2. Obviously, the latter is better. Therefore, there is a potential for using variance reduction techniques in NTM's for integration.

5. OPTIMIZATION METHODS IN STATISTICS

There is a close relationship between optimization and statistics. There are many problems in statistics (such as maximum likelihood estimation, various estimates in regression analysis, optimal experimental design, optimal quantizer, etc.) which can be treated as optimization problems. Also, there are a number of numerical methods in optimization theory. For example, the Newton–Gauss method, Nelder and Mead simplex method, the BFGS (see Nash and Walker-Smith, 1987) or the truncated Newton method can be applied. In those methods, it is often required that the function f to be optimized is unimodal and differentiable, otherwise maybe only a local maximum (minimum) can be reached. Therefore, Horst and Tuy (1990) in their book said "The enormous practical need for solving global optimization problems coupled with a rapidly advancing computer technology has allowed one to consider problems which a year ago would have been considered computationally intractable." The book collected a number of diverse algorithms for solving a wide variety of multiextremal global optimization problems. On the other hand, in the past 20 years there has been considerable activity related to Monte Carlo simulation, including Monte Carlo optimization (cf. Rubinstein, 1986).

In this section we introduce a sequential number-theoretic method for optimization (SNTO) and its applications in statistics. We will show that it is a good addition to other optimization methods.

Let G be a closed and bounded domain in R^s, and let f be a continuous function on G. Suppose that we want to find a global maximum M of f over G, and also a point x^* of G, such that

$$M = f(\mathbf{x}^*) = \max_{\mathbf{x} \in G} f(\mathbf{x}). \tag{28}$$

An NTM for optimization requires the following steps:

1. Choose an NT-net $\mathcal{P} = \{\mathbf{x}_k,\ k = 1, \ldots, n\}$ on G.
2. Find M_n and $\mathbf{x}_n^* \in \mathcal{P}$ such that

$$M_n = f(\mathbf{x}_n^*) = \max_{1 \le i \le n} f(\mathbf{x}_i). \tag{29}$$

Then M_n and $\mathbf{x}_n^*$ are respective approximations of M and $\mathbf{x}^*$.

Fang and Wang (1990) have extended the result of Niederreiter (1983) and give the following theorems.

THEOREM 3. *Suppose that $f(\mathbf{x})$ is a continuous function defined on a closed and bounded domain G, and that $\{\mathcal{P}_{n_i}\}$, $n_1 < n_2 < \cdots$, is a sequence of sets on G which have $a_{n_i} \equiv D_F(n_i, \mathcal{P}_{n_i})$, where F is the cdf of the uniform distribution on G, such that $a_{n_i} = o(1)$ as $i \to \infty$. Let $\mathbf{x}_{n_i}^* \in \mathcal{P}_{n_i}$ be a point satisfying*

$$M_{n_i} = f(\mathbf{x}_{n_i}^*) = \max_{\mathbf{x} \in \mathcal{P}_{n_i}} f(\mathbf{x}). \tag{30}$$

Then $M_{n_i} \to M$ as $i \to \infty$. More precisely, we have $M_{n_i} = M + o(a_{n_i}^{1/s})$ and $x_{n_i}^* \to x^*$ as $i \to \infty$ if $\mathbf{x}^*$ is unique.

THEOREM 4. *Suppose that $\partial f/\partial x_i$, $i = 1, \ldots, s$, are continuous and bounded by C over G. Then*

$$M_n \le M \le M_n + sCb_n(G, \mathcal{P}_n), \tag{31}$$

where b_n is the dispersion of $\mathcal{P}_n$ on G defined by

$$b_n(G, \mathcal{P}_n) = \max_{\mathbf{x} \in G} \min_{1 \le k \le n} d(\mathbf{x}, \mathbf{x}_k), \tag{32}$$

and $d(\mathbf{x}, \mathbf{x}_k)$ is the l_2-distance between $\mathbf{x}$ and $\mathbf{x}_k$.

As a result of these theorems, an NTM for optimization can be applied to any continuous function, in particular, to a continuous multiextremal function. This is certainly a big advantage.

If we choose $\mathcal{P}_{n_i}$ with the "best" order of F-discrepancy $O(n^{-1}(\log n)^s)$, then the convergence rate of M_n to M is $O(n^{1/s} \log n)$, which is slow. Hence Niederreiter and Peart (1986) and Fang and Wang (1990) proposed independently a sequential number-theoretic method for optimization (SNTO) with a much faster convergence rate which has been successfully applied to various statistical problems.

Now we illustrate an SNTO for G being a rectangle $[\mathbf{a}, \mathbf{b}]$.

Step 0 (Initialization). Set $t = 0$, $G^{(0)} = G = [\mathbf{a}, \mathbf{b}]$, $\mathbf{a}^{(0)} = \mathbf{a}$ and $\mathbf{b}^{(0)} = \mathbf{b}$.

Step 1 (Generating an NT-net). Use an NTM to generate an NT-net of n_t points $\mathcal{P}^{(t)}$ on $G^{(t)}$.

Step 2 (Computing a new approximation). Find $\mathbf{x}^{(t)} \in \mathcal{P}^{(t)} \cup \{\mathbf{x}^{(t-1)}\}$ and $M^{(t)}$ such that $M^{(t)} = f(\mathbf{x}^{(t)}) \ge f(\mathbf{y})$, $\forall \mathbf{y} \in \mathcal{P}^{(t)} \cup \{\mathbf{x}^{(t-1)}\}$, where $\mathbf{x}^{(-1)}$ is the empty set. Then $\mathbf{x}^{(t)}$ and $M^{(t)}$ are the best approximation to $\mathbf{x}^*$ and M so far.

Step 3 (Termination criterion). Let $\mathbf{c}^{(t)} = (\mathbf{b}^{(t)} - \mathbf{a}^{(t)})/2$. If $\max \mathbf{c}^{(t)} = \max(c_1^{(t)}, \ldots, c_s^{(t)}) < \delta$, a preassigned small number, then $G^{(t)}$ is small enough; $\mathbf{x}^{(t)}$ and $M^{(t)}$ are acceptable; terminate algorithm. Otherwise, proceed to next step.

Step 4 (Contract domain). Form new domain $G^{(t+1)} = [\mathbf{a}^{(t+1)}, \mathbf{b}^{(t+1)}]$ as follows:

$$a_i^{(t+1)} = \max\left(x_i^{(t)} - rc_i^{(t)}, a_i\right) \quad \text{and}$$
$$b_i^{(t+1)} = \min\left(x_i^{(t)} + rc_i^{(t)}, b_i\right),$$

where r is a predefined *contraction rate*. Set $t = t+1$. Go to Step 1.

According to our experience, we suggest taking $n_1 > n_2 = n_3 = \cdots$ and $r = 0.5$, while Niederreiter and Peart (1986) suggested using $r_i = r^i$ for contraction ratio at the ith stage for some $0 < r < 1$.

We compared SNTO with the BFGS method, which is a quasi-Newton method available in MATLAB as a built-in function. We chose the following objective functions:

$$f_1 = -\Big[\exp\big(-(x_1 + 0.5)^2\big) + 2\exp\big(-(x_2 - 0.5)^2\big) + 4\exp\big(-(x_3 + 3)^2\big)\Big];$$

$$f_2 = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4 \quad \text{(six-hump camelback function)};$$

$$f_3 = 100(x_2 - x_1)^2 + (x_1 - 1)^2 \quad \text{(Rosenbrock function)};$$

$$f_4 = 10,000\big(x_2 - x_1^2\big)^2 + (x_1 - 1)^2 \quad \text{(scaled Rosenbrock function)};$$

$$f_5 = 100\big(x_1^2 - x_2\big)^2 + (1 - x_1)^2 + 90\big(x_4 - x_3^2\big)^2 + (1 - x_3)^2 + 10.1\big[(x_2 - 1)^2 + (x_4 - 1)^2\big] + 19.8(x_2 - 1)(x_4 - 1) \quad \text{(Wood function)};$$

$$f_6 = -\left[2\exp\left\{-\frac{1}{2}\big(x_1^2 + (x_2 - 4)^2\big)\right\} + \exp\left\{-\frac{1}{2}\left((x_1 - 4)^2 + \frac{x_2^2}{4}\right)\right\} + \exp\left\{-\frac{1}{2}\left(\frac{(x_1 + 4)^2}{4} + x_2^2\right)\right\}\right].$$

Since BFGS is a local optimization algorithm, for each objective function we calculate 100 minimizations by BFGS. The computing time is the average of computing times to attain the global minimum. Table 3 gives comparison between SNTO and BFGS, showing the percentage of times the global minimum was reached by each algorithm, and the error defined as the average of errors for those attaining the global minimum. We can see that BFGS needs less computing time than SNTO, but also that it much more frequently failed to reach the global minimum. We can expect that if SNTO can be used as a built-in function, it will be substantially faster.

Without providing a detailed comparison, we would like to say a few words that relate to another global optimization method, *simulated annealing* (SA). In a recent issue of this journal Bertsimas and Tsitsiklis (1993) gave an introduction to SA, with related discussion. Simulated annealing is a probabilistic method proposed in Kirkpatrick, Gelett and Vecchi (1983) and Cerny (1985) for finding the global minimum of a cost function defined on a finite set that may possess several local minima. "One of the great charms of SA is its extraordinary generality. Almost any optimization problem can be approached by SA, and often the coding is quite easy," said Steel (1993). The convergence rate of SA is not clear, but "the most

TABLE 3
Comparison between SNTO and BFGS

Function	Domain	Percentage attaining global minimum	Error	Average computing time (seconds)	Method
f_1	$[-2,2]\times[-1,4]$	52%	4.62E − 7	0.6123	BFGS
		100%	4.62E − 7	0.6833	SNTO
f_2	$[-3,3]\times[-2,2]$	65%	4.53E − 7	0.6328	BFGS
		100%	4.52E − 7	0.9500	SNTO
f_3	$[-2,2]^2$	97%	8.47E − 10	0.6180	BFGS
		100%	4.66E − 4	0.7333	SNTO
f_4	$[-2,2]^2$	4%	4.04E − 4	0.5170	BFGS
		100%	9.65E − 3	1.4330	SNTO
f_5	$[-2,2]^4$	88%	1.64E − 8	1.8790	BFGS
		100%	1.75E − 4	2.2300	SNTO
f_6	$[-10,7]\times[-6,7]$	20%	1.83E − 7	0.7300	BFGS
		100%	1.83E − 7	0.7500	SNTO

serious drawback lies in its very slow convergence rate" (see Ferrari, Frigessi and Schonmann, 1993). It is not easy to give a fair comparison between SNTO and SA, because there are some flexible parameters and different termination criteria in these two algorithms. Very often, a small thing in programming will cause a big difference in computing time. This is an open problem for further study.

Although we have no numerical comparison to SA, it should be clear that SNTO has certain advantages. (i) We need not calculate the derivatives or its approximation of the function f. (ii) The programming is easy and can be universally used for different problems with only a minor modification. For example, consider a general regression model

$$EY = g(\mathbf{x};\theta),$$

where $\mathbf{x} = (x_1,\ldots,x_p)$ are independent variables and $\theta = (\theta_1,\ldots,\theta_s)$ are parameters to be estimated. The first aim of regression analysis is to use a set of observations $\{Y_i, \mathbf{x}_i,\ i = 1,\ldots,N\}$ to estimate θ. This problem is often treated by the least squares method. Let

$$L(\theta) = \sum_{i=1}^{N}\left[Y_i - g(\mathbf{x}_i;\theta)\right]^2$$

Then the least squares estimator $\widehat{\theta}$ satisfies $L(\widehat{\theta}) = \min_\theta L(\theta)$, and the robust estimator $\widehat{\theta}$ satisfies $L^*(\widehat{\theta}^*) = \min_\theta L^*(\theta)$, where

$$L^*(\theta) = \sum_{i=1}^{N} h\left(Y_i - g(\mathbf{x}_i;\theta)\right)$$

and $h(\cdot)$ is a nonincreasing and nonnegative function. Usually, different methods and different computer programs are used to find $\widehat{\theta}$ and $\widehat{\theta}^*$. In contrast, we can use the same method and almost the same program to obtain $\widehat{\theta}$ and various $\widehat{\theta}^*$ by SNTO.

We have successfully applied SNTO to maximum likelihood estimation (MLE) (Fang and Yuan, 1990) and to solve a system of equations with applications in statistics (Fang and Wang, 1991).

6. APPLICATIONS OF NTM IN STATISTICAL INFERENCE

We have mentioned some applications of NTM's in statistical inference, such as in MLE and estimation of regression analysis. In this section we will consider more applications of NTM's in statistical inference.

6.1 Projection Pursuit Methods

The term "projection pursuit" (PP) was first used by Friedman and Tukey in 1974. The PP method reveals structure in the original data by offering selected low-dimensional orthogonal projections of it for inspection. Huber (1985) gave a review of developments in PP.

Let $\mathbf{X}$ be an $N \times p$ matrix of observations with p variables. For any $\mathbf{a} \in R^p$, $\mathbf{Xa}$ is an $N \times 1$ vector which is the orthogonal projection of the sample onto direction $\mathbf{a}$. Without loss of generality we always assume $\mathbf{a}'\mathbf{a} = 1$, or $\mathbf{a} \in U_p = \{\mathbf{x}: \mathbf{x}'\mathbf{x} = 1,\ \mathbf{x} \in R_p\}$. If H is a function to measure the interest of a one-

dimensional sample, then $I(\mathbf{a}) \equiv H(\mathbf{Xa})$ is called a *projection index*. We want to find $\mathbf{a}_0$ such that $I(\mathbf{a}_0) = \max_{\mathbf{a}\in U_p} I(\mathbf{a})$ [or $I(\mathbf{a}_0) = \min_{\mathbf{a}\in U_p} I(\mathbf{a})$]. For example, $I(\mathbf{a})$ is the sample variance of $\mathbf{Xa}$ in principal component analysis (PCA). There is an analytic solution to $\mathbf{a}_0$ in this case. In general, we have to use numerical optimization methods to find approximations of $\mathbf{a}_0$ and $I(\mathbf{a}_0)$. In the past statisticians had difficulties finding $\mathbf{a}_0$ and $I(\mathbf{a}_0)$ (see Malkovich and Afifi, 1973, and Rousseeuw and van Zomeren, 1990). NTM's are powerful tools for this purpose by using an NT-net on U_p. An example will be shown in the next subsection.

For orthogonal projection onto a space of dimension $q > 1$, $\mathbf{a}$ is replaced by a $p \times q$ matrix $\mathbf{A}$, and the corresponding projection index becomes $I(\mathbf{A}) = H(\mathbf{XA})$, where $\mathbf{A} \in O(p, q)$ and

$$O(p, q) = \{\mathbf{U}\text{: } \mathbf{U} \text{ is a } p \times q \text{ matrix}, \mathbf{U}'\mathbf{U} = \mathbf{I}_q\}.$$

Similarly, we need to find $\mathbf{A}_0$ and $I(\mathbf{A}_0)$. If we want to use NTM, we need an algorithm to generate an NT-net on $O(p, q)$. A simple way is based on the following fact. Let $\mathbf{Y}$ be a $p \times q$ random matrix with i.i.d. elements each having the standard normal distribution $N(0, 1)$, that is, $\mathbf{Y} \sim N_{p\times q}(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_q)$. Then $\mathbf{U} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2}$ is uniformly distributed on $O(p, q)$. With this fact, we first generate an NT-net $\mathcal{P} = \{\mathbf{c}_k, k = 1, \ldots, n\}$ on C^{pq} and obtain a set of rep-points of $\mathbf{Y} \sim N_{p\times q}(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_q)$, $\mathcal{P}_N = \{\mathbf{Y}_k(p \times q)$: $k = 1, \ldots, n\}$, by some method mentioned in Section 2. Then $\mathcal{P}_F = \{\mathbf{U}_k = \mathbf{Y}_k(\mathbf{Y}_k'\mathbf{Y}_k)^{-1/2}, k = 1, \ldots, n\}$, is an NT-net on $O(p, q)$. This algorithm is simple, but the uniformity of $\mathcal{P}_F$ on $O(p, q)$ is not good when n is small. Since there are $pq - q(q + 1)/2$ degrees of freedom in $\mathbf{A} \in O(p, q)$, the above method wastes $q(q + 1)/2$ degrees of freedom. Improvements to this are discussed in Fang and Wang (1993).

6.2 Tests for Multinormality and for Multivariate Goodness-of-Fit

Testing multinormality has received considerable attention in the past few decades. Mardia (1980), Malkovich and Afifi (1973), Gnanadesikan (1977), Cox and Small (1978), Baringhaus and Henze (1988), Csörgő (1989) and Horswell and Looney (1992), for example, constitute a large literature.

The PP algorithm discussed in the previous subsection can be applied to this problem. It is well known that a p-dimensional random vector $\mathbf{x}$ is distributed according to a multinormal distribution if and only if, for each $\mathbf{a} \in U_p$, $\mathbf{a}'\mathbf{x}$ is a univariate normal distribution. If we can find the "worst" projection direction $\mathbf{a}_0$, then the test of multinormality is equivalent to testing normality of $\mathbf{a}'_0\mathbf{x}$. Let $\{\mathbf{a}_k, k = 1, \ldots, n\}$ be an NT-net on U_p. Then the worst direction $\mathbf{a}_*$ among $\{\mathbf{a}_k\}$ is close to $\mathbf{a}_0$, and the test of multinormality is approximately equivalent to testing normality of $\mathbf{a}'_*\mathbf{x}$.

Let $Sk(\mathbf{a})$ and $Ku(\mathbf{a})$ be the sample skewness and kurtosis of $\{\mathbf{a}'\mathbf{x}_i, i = 1, \ldots, N\}$, respectively. The worst direction $\mathbf{a}_0$ can be considered as such that

$$Sk(\mathbf{a}_0) = \max_{\mathbf{a}\in U_p} |Sk(\mathbf{a})| \cong \max_k |Sk(\mathbf{a}_k)| \equiv Sk_{\max}$$

or

$$Ku(\mathbf{a}_0) = \max_{\mathbf{a}\in U_p} |Ku(\mathbf{a})| \cong \max_k |Ku(\mathbf{a}_k)| \equiv Ku_{\max}.$$

Hence, the statistics $Sk_{\max}$ and $Ku_{\max}$ can be used for testing multinormality. For a given significance level α, the rejection region is $Sk_{\max} > Sk(\alpha)$ or $Ku_{\max} > Ku(\alpha)$. The critical points of $Sk(\alpha)$ and $Ku(\alpha)$ for $\alpha = 1\%, 5\%$ with $p = 2 \sim 5$ and for various sample sizes are given by Fang, Yuan and Bentler (1992).

6.3 Test for Spherical Symmetry

Spherical distributions have been completely discussed by Fang, Kotz and Ng, (1990). Let $\mathbf{x} \in R^s$ be a random vector. It is well known that $\mathbf{x}$ has a spherical distribution if and only if, for each $\mathbf{a} \in U_s$,

$$\mathbf{a}'\mathbf{X} =_d X_1, \tag{33}$$

where "$=_d$" means that two sides of the equality have the same distribution and X_1 is the first component of $\mathbf{x}$. Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$ from an unknown underlying distribution function $G(\mathbf{x})$, we want to test the following:

$$\begin{aligned} &H_0\text{: } G(\mathbf{x}) \text{ is spherical;} \\ &H_1\text{: } G(\mathbf{x}) \text{ is not spherical.} \end{aligned} \tag{34}$$

By the characteristic (33) of spherical distributions, hypothesis (34) can be expressed as H_0: all $\mathbf{a}'\mathbf{x}$, $\mathbf{a} \in U_s$, have the same distribution. Let $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ be an NT-net on U_s. Then hypothesis (34) can be approximated by

$$H_0^*\text{: } \mathbf{a}_i'\mathbf{x},\ i = 1, \ldots, m, \text{ have the same distribution,} \tag{35}$$

if m is large. For given $0 < k < l \le m$, consider a two-sample problem:

$$\begin{aligned} &\text{I:} \quad \mathbf{a}_k'\mathbf{x}_1, \ldots, \mathbf{a}_k'\mathbf{x}_N, \\ &\text{II:} \quad \mathbf{a}_l'\mathbf{x}_1, \ldots, \mathbf{a}_l'\mathbf{x}_N. \end{aligned}$$

The following statistic is a Wilcoxon-type statistic for the two-sample problem:

$$V_N(\mathbf{a}_k, \mathbf{a}_l) = \frac{1}{N^2} \prod_{i=1}^{N} \prod_{j=1}^{N} I\{\mathbf{a}_k'\mathbf{x}_i < \mathbf{a}_l'\mathbf{x}_j\},$$

where $I\{A\}$ is the indicator function of A. The statistic

$$T_N = \min_{\substack{1 \le k,l \le m \\ k \ne l}} \{V_N(\mathbf{a}_k, \mathbf{a}_l)\} \tag{36}$$

can be used for testing hypothesis (35).

Since $V_N(\mathbf{a}_k, \mathbf{a}_l)$ is constructed by two dependent samples, it is not the traditional two-sample Wilcoxon statistic. Fang, Zhu and Bentler (1993) obtained the limiting distribution of T_N for the case of $\mathbf{a}_k$'s being orthogonal and gave some suggestions for improvement.

NTM's can be also applied in robust estimation. For example the minimum-volume ellipsoid estimator (MVE) is a popular method in robust estimation, as discussed by Rousseeuw and van Zomeren (1990). They suggested using the projection algorithm for MVE and need to find $g^* = \max_{\mathbf{a} \in U_s} g(\mathbf{a}'\mathbf{x}_1, \ldots, (\mathbf{a}'\mathbf{x}_N)$, where $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is a sample and g is a certain function. They had difficulties in evaluating g^* and suggested taking all $\mathbf{a}$ as the direction of $\mathbf{x}_i - \mathbf{m}$, $i = 1, \ldots, N$, where $\mathbf{m} = (\text{median}_j\{x_{j1}\}, \ldots, \text{median}_j\{x_{js}\})'$. Obviously, the set {direction of $\mathbf{x}_i - \mathbf{m}$} is not a set of rep-points on U_s in general. Taking an NT-net on U_s instead of {direction of $\mathbf{x}_i - \mathbf{m}$} into the projection algorithm, we can expect to get a better approximation of g^*. Furthermore the idea of projection with the α-trimmed mean can motivate a new robust estimator of the location by using NT-nets on U_s (Fang, Yuan and Bentler, 1992).

7. GEOMETRIC PROBABILITY

Many problems in geometric probability need simulation studies. Let D be a domain where a simulation is required. It is difficult to do simulation in many situations where D is not a rectangle (e.g., D is the unit sphere U_s). For illustration of the use of NTM's in geometric probability we would like to present an interesting and typical problem in geometric probability: the distribution of the life of a roller in steel production.

This problem arose from steel rolling and had no satisfactory solution since 1980 (cf. Cheng, 1983). People wish to increase the life of the roller by using a randomly rotary ball roller instead of a fixed roller. Its mathematical model can be stated as follows: Let S be a unit sphere in R^3 and be covered by independently random belts with fixed width one by one. Each belt is symmetric about a great circle of S and uniformly distributed on S. Denote by $G_h(\mathbf{x})$ the random belt on S,

$$G_h(\mathbf{x}) = \{\mathbf{a}\colon |\mathbf{a}'\mathbf{X}| \le h\}, \tag{37}$$

where $\mathbf{x}$ is the normal direction and $2h$ is thickness of the belt. Let $G_h(\mathbf{x}_1), G_h(\mathbf{x}_2), \ldots$ be a sequential sample from the population $G_h(\mathbf{x})$, where $\mathbf{x} \sim U(S)$, the uniform distribution on S. For any $\mathbf{y} \in S$ we denote by $N_M(\mathbf{y})$ the number of belts which cover $\mathbf{y}$ in the first M random belts. Given a positive integer m the life of the roller is defined by the minimum of M such that $N_M(\mathbf{y}) \ge m$ for some $\mathbf{y} \in S$. Obviously, the life of the roller, $T_m(h)$, can be expressed as

$$\begin{aligned} T_m(h) &= \min\{M\colon N_M(y) \ge m \text{ for some } \mathbf{y} \in S\} \\ &= \min\left\{M\colon \sup_{y \in S} \prod_{j=1}^{M} I_{G_h(\mathbf{x}_j)}(\mathbf{y}) \ge m\right\}, \end{aligned} \tag{38}$$

where $I_A(\cdot)$ is the indicator function of set A. We need to obtain the distribution of $T_m(h)$ and the expected value $E(T_m(h))$, and to find some way to increase the life of the roller.

We can find the limiting distribution of $T_m(h)$ (cf. Cheng et al., 1990). Unfortunately, there is big difference between the limiting distribution and the real distribution when m is not very large. For example, when $h = 0.1$ and $m = 20$, the mean and the standard deviation of the limiting distribution are $m/h = 200$ and $\sqrt{m(1-h)/h} \cong 42.4$, respectively. However, the simulation by Fang and Wei (1992) shows

$$E\big(T_{20}(0.1)\big) \approx 99.7 \quad \text{and} \quad \sigma\big(T_{20}(0.1)\big) \approx 9.8. \tag{39}$$

Therefore, simulation is more useful in practice. A simulation process was suggested to find the empirical distribution as well as the mean and variance of $T_m(h)$. For example, we generate 50 samples of size 5,000 with $m = 20$ and $h = 0.1$ and find that 50 empirical distributions are close to each other with average mean and standard deviation (39). Furthermore, we note that the longest life of the roller in 100,000 simulated observations can reach 125, which is significantly longer than the average life 99.7. Denote by $\mathbf{a}_1^*, \ldots, \mathbf{a}_{125}^*$ the corresponding normal directions of the roller with the longest life. This suggests that if we fix the normal directions at $\mathbf{a}_i = \mathbf{a}_i^*$, $i = 1, 2, \ldots$, we always have $T_m(h) = 125$ in the case $m = 20$ and $h = 0.1$. Since $\mathbf{a}_1^*, \ldots, \mathbf{a}_{125}^*$ are generated by the Monte Carlo method, we might try NTM to improve the result such that the roller has a longer life than 125. We used a glp set on S to be the normal directions and found that the longest life of the roller can be 155! This work indicates that NTM's can be

significantly better than 100,000 experiments by the Monte Carlo method. For more details, see Fang and Wei (1992).

8. CONCLUSIONS AND OTHER APPLICATIONS

We have mentioned many applications of NTM's in statistics. More applications can be found in other topics. One important area is experimental design. Suppose that there are s factors and each factor has n levels. Then the number of all possible experiments is n^s. The orthogonal array is to choose at least n^2 experiments with good representation among these n^s experiments. When n is large (e.g., $n > 7$) the number of experiments is comparatively large by the orthogonal array. For example, we met an important project with six factors, each with more than 12 levels, and had to design it within 50 experiments. Due to this requirement Fang (1980) and Wang and Fang (1981) proposed a new type of design: "uniform design" by the glp set. The uniform design has been applied satisfactorily in design of new products in the textile industry, metallurgical industry, pharmaceutics, and military industry in China. Furthermore, the idea of the uniform design can be applied in experiments with mixtures and can improve the simplex–lattice design and the simplex-centroid design proposed by Scheffé (1963) and the axial design suggested by Cornell (1981).

There are some relationships between the uniform design (UD) and the latin hypercube sampling (LHS) including its versions OA-based latin hypercube sampling (OALHS) (cf. Stein, 1987; Owen, 1992; Tang, 1993). Some comparisons among UD, LHS and OALHS and suggestion for further research can be found in Fang and Wang (1993, Section 5.6).

Based on the above discussion one can conclude that NTM's can be useful tools for statistics, in particular, for multivariate statistics. Due to lack of attention paid by statisticians, the results mentioned in this paper are preliminary. There are many open problems for further study. For example, what is the convergence rate of SNTO? How can SNTO be compared with simulated annealing and other optimization methods? How can SNTO be useful effectively with other optimization methods? It is clear that there are many potential applications of NTM's in statistics to be discovered. The purpose of this paper is to emphasize the growing importance of NTM's in statistics. On the other hand, every method has its limitation. We can not expect that NTM's will have the best performance in each field of statistics.

ACKNOWLEDGMENTS

This work was supported by a Hong Kong UPGC grant, by NSF of China and by PHS National Institute on Drug Abuse Award DA01070-17. Our special thanks are due to Professors Nancy Reid and Robert E. Kass, who gave very valuable comments that significantly improved the paper. The authors thank Professor Y. L. Tong, Dr. F. J. Hickernell, Dr. K. W. Ng and several referees for their valuable comments. The authors thank Professor R. C. Zhang and Mr. H. L. Wong for their help.

REFERENCES

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

BAKER, A. (1965). On some Diophantine inequalities involving the exponential function. *Canad. J. Math.* **17** 616–626.

BARINGHAUS, L. and HENZE, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika* **35** 339–348.

BERTSIMAS, D. and TSITSIKLIS, J. (1993). Simulated annealing. *Statist. Sci.* **8** 10–15.

CERNY, V. (1985). A thermodynamic approach to the traveling salesman problems: an efficient simulation. *J. Optim. Theory Appl.* **45** 41–51.

CHENG, P. (1983). An open problem in steel rolling. *Math. Practice Theory* **2** 79.

CHENG, P., SHI, P. D., ZHU, L. X. and WEI, G. (1990). On the life of a sphere roller. Unpublished manuscript.

CHUNG, K. L. (1949). An estimate concerning the Kolmogoroff limit distribution. *Trans. Amer. Math. Soc.* **67** 36–50.

CORNELL, J. A. (1981). *Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data*. Wiley, New York.

COX, D. R. and SMALL, N. J. H. (1978). Testing multivariate normality, *Biometrika* **65** 263–272.

CSÖRGŐ, S. (1989). Consistency of some tests for multivariate normality. *Metrika* **36** 107–116.

FANG, K. T. (1980). The uniform design: application of number theoretic methods in experimental design. *Acta Math. Appl. Sinica* **3** 363–372. (In Chinese.)

FANG, K. T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London.

FANG, K. T., KOTZ, S. and NG, K. W. (1992). On the L_1-norm distributions. In *L_1-Statistical Analysis and Related Methods* (Y. Dodge, ed.) 401–413. North-Holland, Amsterdam.

FANG, K. T. and WANG, Y. (1990). A sequential algorithm for optimization and its application to regression analysis. In *Lecture Notes in Contemporary Mathematics* (L. Yang and Y. Wang, eds.) 17–28. Science Press, Beijing.

FANG, K. T. and WANG, Y. (1991). A sequential algorithm for solving a system of nonlinear equations. *J. Comput. Math.* **9** 9–16.

FANG, K. T. and WANG, Y. (1993). *Number-Theoretic Methods in Statistics*. Chapman and Hall, London.

FANG, K. T. and WEI, G. (1992). The distribution of a special first-hit random variable. *Acta Math. Appl. Sinica* **15** 460–467.

FANG, K. T. and WU, C. Y. (1979). The extreme value problem of some probability function. *Acta Math. Appl. Sinica* **2** 132–148. (In Chinese.)

FANG, K. T. and YUAN, K. H. (1990). A unified approach to maximum likelihood estimation. *Chinese J. Appl. Probab. Statist.* **6** 412–418.

FANG, K. T., YUAN, K. H. and BENTLER, P. M. (1992). Applications of sets of points uniformly distributed on sphere to

testing multinormality and robust estimation. In *Probability and Statistics* (Z.-P. Jiang, S. J. Yan, P. Cheng and R. Wu, eds.) 56–73. World Scientific, Singapore.

FANG, K. T., ZHU, L. X. and BENTLER, P. M. (1993). A necessary test of goodness of fit for sphericity. *J. Multivariate Anal.* **45** 34–55.

FERRARI, P. A., FRIGESSI, A. and SCHONMANN, R. H. (1993). Convergence of some partially parallel Gibbs samplers with annealing. *Ann. Appl. Probab.* **3** 137–153.

GENZ, A. (1991). Subregion adaptive algorithms for multiple integrals. In *Statistical Multiple Integration* (N. Flournoy and R. Tsutakawa, eds.). Amer. Math. Soc., Providence, RI.

GNANADESIKAN, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

GUPTA, S. S. (1963). Probability integrals of multivariate normal and multivariate t. *Ann. Math. Statist.* **34** 792–828.

HALTON, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.* **2** 84–90.

HLAWKA, E. (1962). Zur angenäherten Berechnung Mehrfacher Integrale. *Monatsh. Math.* **66** 140–151.

HORST, R. and TUY, H. (1990). *Global Optimization*. Springer, Berlin.

HORSWELL, R. L. and LOONEY, S. W. (1992). A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. *J. Statist. Comput. Simulation* **42** 21–38.

HUA, L. K. and WANG, Y. (1981). *Applications of Number Theory to Numerical Analysis*. Springer, Berlin.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

KIEFER, J. (1961). On large deviations of the empirical d.f. of vector chance variables and a law of the iterated algorithm. *Pacific J. Math.* **11** 649–660.

KIRKPATRICK, S., GELETT, C. D. and VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220** 621–630.

KOROBOV, N. M. (1959). The approximate computation of multiple integrals. *Dokl. Akad. Nauk SSSR* **124** 1207–1210.

KOROBOV, N. M. (1989). *Trigonometric Sums and Their Applications*. Nauka, Moscow.

MALKOVICH, J. F. and AFIFI, A. A. (1973). On tests for multivariate normality. *J. Amer. Statist. Assoc.* **68** 176–179.

MARDIA, K. V. (1972). *Statistics of Directional Data*. Academic, New York.

MARDIA, K. V. (1980). Tests of univariate and multivariate normality. In *Handbook of Statistics, Analysis of Variance*, **1** (P. Krishnaiah, ed.) 279–320. North-Holland, Amsterdam.

NASH, J. C. and WALKER-SMITH. (1987). *Nonlinear Parameter Estimation: An Integrated System in BASIC*. Dekker, New York.

NIEDERREITER, H. (1978). Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc.* **84** 957–1041.

NIEDERREITER, H. (1983). A quasi-Monte Carlo method for the approximate computation of extreme values of a function. *Studies in Pure Math.* 523–529. Birkhäuser, Boston.

NIEDERREITER, H. (1988). Quasi-Monte Carlo methods for multidimensional numerical integration. *Internat. Ser. Numer. Math.*, Birkhäuser, Basel.

NIEDERREITER, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia.

NIEDERREITER, H. and PEART, P. (1986). Localization of search in quasi-Monte Carlo methods for global optimization. *SIAM J. Sci. Statist. Comput.* **7** 660–664.

OWEN, A. (1992). Orthogonal arrays for computer experiments, integration, and visualization. *Statist. Sinica* **2** 439–452.

PAGÉS, G. and XIAO, Y. J. (1991). Sequence with low discrepancy and pseudo-random numbers. Technical report, Laboratoire de Mathematiques et Modelisation, Paris.

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *J. Amer. Statist. Assoc.* **85** 633–651.

RUBINSTEIN, R. Y. (1981). *Simulation and The Monte Carlo Method*. Wiley, New York.

RUBINSTEIN, R. Y. (1986). *Monte Carlo Optimization, Simulation and Sensitivity of Queuing Networks*. Wiley, New York.

SCHEFFÉ, H. (1958). Experiments with mixture. *J. Roy. Statist. Soc. Ser. B* **20** 344–360.

SCHEFFÉ, H. (1963). The simplex–centroid design for experiments with mixture. *J. Roy. Statist. Soc. Ser. B* **25** 235–263.

SCHMIDT, W. M. (1970). Simultaneous approximation to algebraic numbers by rationales. *Acta Math.* **125** 189–201.

SHAW, J. E. H. (1988). A quasi-random approach to integration in Bayesian statistics. *Ann. Statist.* **16** 859–914.

STEEL, J. M. (1993). In this issue. *Statist. Sci.* **8** 1–2.

STEIN, M. (1987). Large sample properties of simulations using latin hypercube sampling. *Technometrics* **29** 143–151.

TANG, B. (1993). Othogonal array-based latin hypercubes. *J. Amer. Statist. Assoc.* **88** 1392–1397.

TONG, Y. L. (1990). *Multivariate Normal Distribution*. Springer, Berlin.

WANG, Y. and FANG, K. T. (1981). A note on uniform distribution and experimental design. *Kexue Tongba* **26** 485–489.

WANG, Y. and FANG, K. T. (1990a). Number theoretic methods in applied statistics. *Chinese Ann. Math. Ser. B* **11** 41–55.

WANG, Y. and FANG, K. T. (1990b). Number theoretic methods in applied statistics (II). *Chinese Ann. Math. Ser. B* **11** 859–914.

WANG, Y. and FANG, K. T. (1992). A sequential number-theoretic method for optimizations in statistics. In *The Development of Statistics: Recent Contributions from China* 139–156. Longman, New York.

WATSON, G. S. (1984). *Statistics on Sphere*. Wiley, New York.

ZHANG, Y. T. and FANG, K. T. (1982). *An Introduction to Multivariate Analysis*. Science Press, Beijing.