中国科学: 数学 2020 年 第 50 卷 第 9 期: 1149 ~ 1168
SCIENTIA SINICA Mathematica
论 文

统计分布的代表点集及其应用

献给越民义教授 100 华诞

方开泰[1,2*], 贺平[1], 杨骏[1]

1. 北京师范大学 - 香港浸会大学联合国际学院理工科技学部, 珠海 519087;
2. 中国科学院数学与系统科学研究院应用数学研究所, 北京 100190
E-mail: ktfang@uic.edu.hk, heping@uic.edu.hk, jyang5037@gmail.com

收稿日期: 2019-10-15; 接受日期: 2020-03-16; 网络出版日期: 2020-09-14; * 通信作者
珠海市优势学科基金 (批准号: R1050) 和北京师范大学 - 香港浸会大学联合国际学院校内科研基金 (批准号: R201712, R201810, R201912 和 R202010) 资助项目

摘要 用一个离散统计分布来近似一个连续的统计分布 (一维或多维) 一直是统计学研究的核心内容. 显然这个离散统计分布的支撑点集必须有代表性, 故称它们为代表点集, 或简称代表点. 选择代表点可以有不同的考虑, 本文回顾并比较 4 类近似离散统计分布: 随机样本 (独立同分布)、修改的 Monte Carlo 方法、数论方法的样本 (伪 Monte Carlo 方法) 及在最小平方误准则下的代表点集和相应的统计分布. 其中修改的 Monte Carlo 方法是本文新提出的. 本文比较 4 类方法在密度估计和重采样的统计推断中的表现, 其中有一类是改进的自助法. 本文对最小平方误准则下的代表点的性质和数值算法进行了详细回顾, 并且得到一些新结果, 例如, 随机样本的最小平方误准则的统计分布、椭球等高分布代表点的几何结构以及椭球等高分布代表点和主成分的关系.

关键词 统计分布代表点 伪 Monte Carlo 方法 统计推断 正态分布 椭球等高分布 主成分和主成分点

MSC (2010) 主题分类 65C05, 65C50

1 统计分布的代表点

统计分布是统计学用来建模的重要工具, 一个随机变量 X 的分布函数定义为 $F(x)=\mathrm{P}(X\leqslant x)$. 离散的随机变量 X 的分布, 常常表示为

$$\begin{array}{c|c} X & x_1\ x_2\ \cdots\ x_n \\ \hline p & p_1\ p_2\ \cdots\ p_n \end{array}. \tag{1.1}$$

这里 $x_1,\ldots,x_n$ 称为 X 的支撑点并且 $\mathrm{P}(X=x_i)=p_i>0$, $i=1,\ldots,n$. 连续的统计分布有分布密度函数 $p(x)$, 它是一条曲线. 统计学是通过样本来研究总体的科学和艺术. 如果已知总体的分布类型, 需

英文引用格式: Fang K T, He P, Yang J. Sets of representative points of statistical distributions and their applications (in Chinese). Sci Sin Math, 2020, 50: 1149–1168, doi: 10.1360/SSM-2019-0251

要用样本来估计分布的未知参数. 如果不知道总体的分布类型, 可以用样本来估计分布的分布密度函数, 称为密度估计.

由于分布函数是一条曲线, 为了应用方便, 统计学建议用一些特征数来表达分布的个性, 如平均值、方差、偏态系数和峰度系数等. 如果使用者不知道总体的分布类型, 可以通过样本的经验分布函数

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I\{x_i\leqslant x\} \tag{1.2}$$

来近似总体分布 $F(x)$, 上式中 $I\{A\}$ 为事件 A 的示性函数, 即

$$I\{A\}=\begin{cases}1, & \text{若 } A \text{ 成立},\\ 0, & \text{若 } A \text{ 不成立}.\end{cases}$$

$F_n(x)$ 和 $F(x)$ 的 L_2- 距离为

$$D(F_n,F)=\left[\int_{-\infty}^{\infty}(F_n(x)-F(x))^2dx\right]^{1/2}, \tag{1.3}$$

概率论证明当 $n\to\infty$ 时 $F_n(x)\to F(x)$. 如果从 $D(F_n,F)$ 角度考虑, 那么可以选择其他的离散的随机变量 Y,

$$\begin{array}{c|c} Y & y_1\ y_2\ \cdots\ y_k \\ \hline p & p_1\ p_2\ \cdots\ p_k \end{array}, \tag{1.4}$$

用它的分布函数来近似总体分布 $F(x)$, 这里 $y_1,\ldots,y_k$ 为 Y 的支撑点, 相应的概率为 $\mathrm{P}(Y=y_i)=p_i$, $i=1,\ldots,k$. 我们可以将 $\{y_1,\ldots,y_k\}$ 看成一个 “样本” (不是随机样本), 希望它的分布函数 $F_Y(x)$ 接近总体分布 $F(x)$, 但是 $y_1,\ldots,y_k$ 不互相独立. 上面的考虑可以延伸到多元统计分布, 随机向量 $\boldsymbol{x}=(X_1,\ldots,X_p)\sim F(\boldsymbol{x})$, 它的离散化随机向量 $\boldsymbol{\xi}$ 有支撑点 $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k$, 其概率分布为 $\mathrm{P}(\boldsymbol{\xi}=\boldsymbol{y}_i)=p_i$, $i=1,\ldots,k$. $\boldsymbol{\xi}$ 的分布函数表示为 $F_{\boldsymbol{\xi}}$, 或者 F_{NTM}. 这个思路在信息工程界有许多研究, 如文献 [1–3]. Fang 等 [4] 讨论了 3 类构造 Y 的方法, 本文增加一类修改的 Monte Carlo 方法.

(1) Monte Carlo 方法

随机抽样又称为 Monte Carlo (MC) 方法, k 个随机样本可以看作总体分布的代表点, 定义一个随机向量 $\boldsymbol{\xi}_{\mathrm{MC}}$, 它的支撑点是 k 个随机样本 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k$, 相应的概率为 $1/k$.

(2) 修改的 Monte Carlo 方法

Monte Carlo 方法由于完全基于随机抽样, 有时候效果欠佳. 本文建议用随机样本附加不等概率的近似分布, 称为修改的 Monte Carlo (revised Monte Carlo, RMC) 方法. 令 $x_1\leqslant\cdots\leqslant x_k$ 是 k 个随机样本, 定义一个随机向量 $\boldsymbol{\xi}_{\mathrm{RMC}}$ 使得 $\mathrm{P}(\boldsymbol{\xi}_{\mathrm{RMC}}=x_i)=p_i^{(\mathrm{RMC})}$. 对一维随机变量, 其支撑点 $x_1,\ldots,x_k$ 的概率为

$$\begin{aligned} p_1^{(\mathrm{RMC})}&=\int_{-\infty}^{\frac{x_1+x_2}{2}}f(x)dx,\quad p_2^{(\mathrm{RMC})}=\int_{\frac{x_1+x_2}{2}}^{\frac{x_2+x_3}{2}}f(x)dx,\\ &\vdots\\ p_i^{(\mathrm{RMC})}&=\int_{\frac{x_{i-1}+x_i}{2}}^{\frac{x_i+x_{i+1}}{2}}f(x)dx,\quad \ldots,\quad p_n^{(\mathrm{RMC})}=\int_{\frac{x_{n-1}+x_n}{2}}^{\infty}f(x)dx. \end{aligned} \tag{1.5}$$

对多维随机向量, 可以用本文介绍的 NTLBG (number-theoretic Linde-Buzo-Gray) 算法获得.

(3) 数论方法 (伪 Monte Carlo 方法)

数论方法 (number-theoretic methods, NTM) 的产生是基于多元数值积分的需要, Korobov [5] 创造了在高维矩形上生成均匀布点的方法, 其方法是建立在数论的基础上, 数学家们称这个方法为数论方法 (参见文献 [6]). 后来统计学家们发现数论方法生成均匀布点的方法和 Monte Carlo 方法有许多共同点, 故称它为伪 Monte Carlo (quasi-Monte Carlo, QMC) 方法 (参见文献 [7]). 数论方法需要定义一个均匀性量度, 早期均匀性测度用星偏差

$$D(F,F_{\boldsymbol{\xi}})=\int|F(\boldsymbol{x})-F_{\boldsymbol{\xi}}(\boldsymbol{x})|d\boldsymbol{x} \tag{1.6}$$

表示, 这里 $F_{\boldsymbol{\xi}}$ 是 $\boldsymbol{\xi}$ 的分布函数. 其他的均匀性测度有 L_2- 中心化偏差 (centered L_2-discrepancy, CD)、L_2- 可卷偏差 (wrap-around L_2-discrepancy, WD) 和 L_2- 混合偏差 (mixture L_2-discrepancy, MD) (参见文献 [8]). 如果 $F(x)$ 是区间 $(0,1)$ 上的均匀分布 $U(0,1)$, 使上述的偏差达到最小的的点集为 $\{\frac{2j-1}{2k},j=1,\ldots,k\}$, 则在这个点集上的均匀分布定义为 $U(0,1)$ 的 $Y_{\rm NTM}$. 如果总体分布 $F(x)$ 是一元连续的统计分布, 由 Monte Carlo 逆变换法, $F(x)$ 的代表点为

$$y_j=F^{-1}\left(\frac{2j-1}{2k}\right),\quad j=1,\ldots,k, \tag{1.7}$$

这里 $F^{-1}(x)$ 为 $y=F(x)$ 的反函数. 这时, $F(x)$ 的数论方法的 $Y_{\rm NTM}$ 是在 y_i $(i=1,\ldots,k)$ 上的均匀分布. 显然, y_j 是 $F(x)$ 的 $\frac{2j-1}{2k}$- 分位点, 详见文献 [9].

(4) 最小均方误代表点的方法

最小均方误代表点的方法来源于实际的需要. 如果成年男子的身高遵从正态分布 $N(\mu,\sigma^2)$, 从服装设计的需要选择 k 个服装的号码, 同时考虑每个服装号码代表的人数, 如何来确定服装号码呢? 方开泰和贺曙东 [10] 提出了最小均方误 (mean squared error, MSE), 它定义为

$$\mathrm{MSE}(\boldsymbol{b})=\mathrm{MSE}(b_1,\ldots,b_k)=\int_{-\infty}^{\infty}\min_{1\leqslant j\leqslant k}(x-b_j)^2p(x)dx, \tag{1.8}$$

这里 $p(x)$ 是总体分布的密度函数. Cox [11] 是第一个提出均方误准则的人, 目的是用少数的点代表正态分布使得均方误最小. Bofinger [12] 从多元统计分析的典型相关的角度也提出了同样的代表点准则. *IEEE Transaction on Information Theory* 在 1982 年 3 月出版了一个关于这个方向的研究专辑, 主要用于信息传送, 他们用 "quantizer" 表示数字转换器或编码器. 使 $\mathrm{MSE}(b)$ 达极小的 $\{b_1,\ldots,b_n\}$ 作为 $Y_{\rm MSE}$ 的支撑点称为分布 $F(x)$ 的 MSE 代表点. 相应的概率分布为 $\mathrm{P}(Y_{\rm MSE}=b_i)=p_i$, $i=1,\ldots,k$, 这里

$$\begin{aligned}
p_1&=\int_{-\infty}^{(b_1+b_2)/2}p(x)dx=\int_{a_0}^{a_1}p(x)dx,\\
p_i&=\int_{(b_{i-1}+b_i)/2}^{(b_i+b_{i+1})/2}p(x)dx=\int_{a_i}^{a_{i+1}}p(x)dx,\quad i=2,\ldots,k-1,\\
p_k&=\int_{(b_{k-1}+b_k)/2}^{\infty}p(x)dx=\int_{a_{k-1}}^{a_k}p(x)dx,
\end{aligned}$$

其中 $a_0=-\infty$, $a_i=(b_i+b_{i-1})/2$, $i=2,\ldots,k-1$, $a_k=\infty$.

本文仅考虑连续的统计分布, 回顾统计分布代表点的发展并且介绍一些新的研究结果. 第 2 节介绍一元和多元统计分布代表点的算法. 第 3 节介绍椭球等高分布的一些基本的知识, 方便后面的讨论.

第 4 节比较 3 类代表点构造的近似分布在统计推断中的表现, 其中包括密度估计、重采样和 MSE 的统计分布, 以及反正弦分布的一个有趣性质. 第 5 节讨论椭球等高分布代表点的性质和它与主成分分析的联系. 文中不少结果是新的.

2 统计分布代表点的算法

首先考虑一元统计分布代表点的算法. 容易证明, 当 $k=1$ 时, 唯一的 MSE 代表点正好是 $F(x)$ 的平均值 $\mu=\mathrm{E}(X)$. 如果 $k=2$ 并且统计分布是关于原点对称的, 即 $p(x)=p(-x)$, 两个 MSE 代表点为 $\mu-\mathrm{E}(|X-\mu|)$ 和 $\mu+\mathrm{E}(|X-\mu|)$. 如果 b_i $(i=1,\dots,k)$ 是 $F(x)$ 的代表点, 那么 $u+vb_i$ $(i=1,\dots,k)$ 是 $u+vX$ 的代表点 (参见文献 [10,13]). 当 $k>2$ 时, 求 MSE 代表点需要数值算法.

一元标准正态分布代表点的第一个算法由 Max [1] 提出, 方开泰和贺曙东 [10] 给出了更系统的算法并且证明了算法的收敛性. (1.8) 可以表示为

$$\begin{aligned}\mathrm{MSE}(b_1,\dots,b_k)&=\int_{-\infty}^{\infty}\min_{1\leqslant i\leqslant k}(x-b_i)^2\varphi(x)dx\\&=\int_{-\infty}^{(b_1+b_2)/2}(x-b_1)^2\varphi(x)dx+\int_{(b_1+b_2)/2}^{(b_2+b_3)/2}(x-b_2)^2\varphi(x)dx\\&\quad+\cdots+\int_{(b_{k-1}+b_k)/2}^{\infty}(x-b_k)^2\varphi(x)dx,\end{aligned}$$

这里 $\varphi(x)$ 是标准正态分布的分布密度函数, 它是偶函数, 即 $\varphi(x)=\varphi(-x)$, 因而 $-b_1=b_k,-b_2=b_{k-1},\dots,-b_i=b_{k-i+1}$. 因此, 变量 b_i 的数目减少到 m 个, $m=k/2$ 或 $m=(k-1)/2$, 依赖于 k 是偶数或奇数. 当 k 为偶数时, 记 $0<b_1,\ b_2<\cdots<b_m$, 对 $\mathrm{MSE}(b_1,\dots,b_m)$ 求偏导数得方程组

$$\begin{cases}\varphi(0)-\varphi\left(\dfrac{b_1+b_2}{2}\right)=b_1\left[\Phi\left(\dfrac{b_1+b_2}{2}\right)-\Phi(0)\right],\\ \varphi\left(\dfrac{b_1+b_2}{2}\right)-\varphi\left(\dfrac{b_2+b_3}{2}\right)=b_2\left[\Phi\left(\dfrac{b_2+b_3}{2}\right)-\Phi\left(\dfrac{b_1+b_2}{2}\right)\right],\\ \quad\vdots\\ \varphi\left(\dfrac{b_{m-2}+b_{m-1}}{2}\right)-\varphi\left(\dfrac{b_{m-1}+b_m}{2}\right)=b_{m-1}\left[\Phi\left(\dfrac{b_{m-1}+b_m}{2}\right)-\Phi\left(\dfrac{b_{m-2}+b_{m-1}}{2}\right)\right],\\ \varphi\left(\dfrac{b_{m-1}+b_m}{2}\right)=b_m\left[1-\Phi\left(\dfrac{b_{m-1}+b_m}{2}\right)\right].\end{cases}\tag{2.1}$$

上述的方程式可以分成 3 类, 方开泰和贺曙东 [10] 给出了下列的算法:

步骤 1 给一个初始值 $b_1=b_1^{(1)}$, 从方程组 (2.1) 的第 1 个方程解得 $b_2^{(1)}$.

步骤 2 给定当前的 $b_1^{(1)}$ 和 $b_2^{(1)}$, 从方程组 (2.1) 的第 2 个方程解得 $b_3^{(1)}$.

步骤 3 使用类似的方法可以从前 $m-1$ 个方程式解得 $b_4^{(1)},\dots,b_m^{(1)}$.

步骤 4 利用方程组 (2.1) 的最后一个方程, 从步骤 3 的 $b_{m-1}^{(1)}$ 解得 $b_m^{(1*)}$.

步骤 5 如果 $|b_m^{(1)}-b_m^{(1*)}|<\varepsilon$, 输出方程组 (2.1) 的近似解 $b_1^{(1)},\dots,b_m^{(1)}$; 否则, 变化 $b_1^{(1)}$ 为 $b_1^{(2)}$, 然后重复上面步骤 N 次, 直至 $|b_m^{(N)}-b_m^{(N*)}|<\varepsilon$, 这里 ε 是一个预先给定的小的正数.

利用上面的算法, 方开泰和贺曙东 [10] 给出了 $k\leqslant 31$ 标准正态分布的代表点及相应的近似分布. 利用方开泰和贺曙东的算法, 傅洪海 [14,15] 和费荣昌 [16] 给出了 Gamma 分布、Weibull 分布与 Pearson

分布族的代表点及相应的近似分布. 方程组 (2.1) 可以用其他方法来求数值解, 例如, 用序贯数论优化 (sequential number-theoretic for optimization, SNTO) 方法, 详见文献 [17]. 费荣昌 [18] 利用类似的方程组 (2.1), 证明了分布的代表点的一些重要性质.

Sharma [19] 指出关于零点对称的分布其代表点不一定关于零点对称, 并且给出了相关的例子. 幸运的是标准正态分布的代表点是关于零点对称的. 对给定 k 关于 MSE 代表点是否唯一, Fleischer [20] 给出了详细讨论, 并且给出了一个容易实行的充分条件, 称为对数凹性检验 (log-concavity test), 即如果 X 的分布密度函数满足

$$\frac{\partial^2 \log p(x)}{\partial x^2} < 0,$$

则 X 的 MSE 代表点对任何正整数 k 唯一. 标准正态分布的密度函数满足这个条件, 它的 MSE 代表点唯一.

一元分布的 MSE 代表点的定义容易延伸到多元统计分布的情况, 但是寻找 MSE 代表点的算法方面有一定的困难. 令 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 是一个连续的随机向量, 有分布函数 $F(\boldsymbol{x})$ 和分布密度函数 $p(\boldsymbol{x}) = p(x_1, \ldots, x_p)$. 定义 $\mathbb{R}^p$ 中的点 $\boldsymbol{x}$ 到 $\mathbb{R}^p$ 中的点集 $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ 的平方距离为

$$d^2(\boldsymbol{x} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k) = \min_{1 \leqslant i \leqslant k} ((\boldsymbol{x} - \boldsymbol{z}_i)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{z}_i)). \tag{2.2}$$

上式左边是向量 $\boldsymbol{x}$ 到 $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k \in \mathbb{R}^p\}$ 的距离. 如果 $\mathbb{R}^p$ 中的点集 $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ 能最小化下列的平方误:

$$\mathrm{MSE}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) = \mathrm{E}(d^2(\boldsymbol{x} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k)) = \min_{\substack{\boldsymbol{z}_j \in \mathbb{R}^p \\ 1 \leqslant j \leqslant k}} \mathrm{E}(d^2(\boldsymbol{x} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)), \tag{2.3}$$

则称 $\boldsymbol{\xi} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ 为 $\boldsymbol{x} \sim F(\boldsymbol{x})$ 的 MSE 代表点集. Flury [13] 发现椭球等高分布 MSE 代表点的性质与主成分分析有联系, 故称 $\boldsymbol{\xi}$ 为主成分点集. 由于主成分点集使用的范围有局限, 我们更喜欢使用代表点集这个名称.

令 $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ 是 $\mathbb{R}^p$ 的点集,

$$S_j = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{y}_j)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}_j) \leqslant (\boldsymbol{x} - \boldsymbol{y}_i)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}_i)\}, \quad j = 1, \ldots, k. \tag{2.4}$$

$\{S_j\}$ 是一组凸多边形, 称为 Voronoi 分割 (Voronoi partitions), 详情参见文献 [3]. 如果 $\mathrm{E}(\boldsymbol{x} \mid x \in S_j) = \boldsymbol{y}_j$, $j = 1, \ldots, k$, 则称代表点 $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ 是自相容的.

由于 (2.3) 中需要在多个凸多边形 S_j 上积分, 在数值计算上有相当的复杂性. 所以, Linde 等 [2] 提出了一个迭代算法 (LBG (Linde-Buzo-Gray) 算法), 其思想和聚类分析中的 k- 均值算法类似. 他们应用这个算法到多元正态分布、多元 Laplace 分布、多元 Gamma 分布和多元均匀分布.

LBG 算法 令 $\boldsymbol{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ 是一个连续的随机向量, 其分布函数为 $F(\boldsymbol{x})$, 分布密度为 $p(\boldsymbol{x})$. 使用 Monte Carlo 方法从总体分布 $F(\boldsymbol{x})$ 抽 N 个样本 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ ($N \gg k$) 作为训练样本. $F(\boldsymbol{x})$ 的 k- 水平的数字转换器 $Q = \{\mathcal{Y}, \mathcal{S}\}$ 利用下列步骤获得. 令 $t = 0$.

步骤 1 给出初始向量集 $\mathcal{Y}_t = \{\boldsymbol{y}_{t1}, \ldots, \boldsymbol{y}_{tk}\}$, 使用最近距离的方法获得训练样本的划分 $\mathcal{S}_t = \{\mathcal{S}_{t1}, \ldots, \mathcal{S}_{tk}\}$, 其中

$$\mathcal{S}_{ti} = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{y}_{ti}\| \leqslant \|\boldsymbol{x} - \boldsymbol{y}_{tj}\|, j \neq i\}, \quad \boldsymbol{x} \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \quad i = 1, \ldots, k.$$

步骤 2 计算训练样本在 $\mathcal{S}_{ti}$ $(i=1,\ldots,k)$ 中的样本平均值, 表示为 $\mathcal{Y}_{t+1}=\{\boldsymbol{y}_{t+1,1},\ldots,\boldsymbol{y}_{t+1,k}\}$.

步骤 3 如果 $\mathcal{Y}_{t+1}=\mathcal{Y}_t$ 输出 $\mathcal{Y}_t$ 作为 $F(\boldsymbol{x})$ 的代表点; 否则 $t:=t+1$, 重复上面的步骤.

从 LBG 算法的输出, 可以获得 $Q=\{\mathcal{Y},\mathcal{S}\}$:

(i) $\boldsymbol{y}_{\mathrm{MSE}}$ 的支撑点集 $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k\}$.

(ii) 用 (2.4) 获得 $\mathbb{R}^p$ 空间的一个划分 $\mathcal{S}=\{S_1,\ldots,S_k\}$.

(iii) 定义映射 $\mathbb{R}^p\to\mathcal{S}$: 如果 $\boldsymbol{x}\in S_i$, 令 $Q(\boldsymbol{x})=\boldsymbol{y}_i$. 令 n_j 为训练样本落在 S_j 中的数目, 那么 $\hat{p}_j=n_j/N$ 作为 $\mathrm{P}(\boldsymbol{y}_{\mathrm{MSE}}=\boldsymbol{y}_j)$ 的估计.

(iv) 最优的数字转换器 $Q(\cdot)$ 选择 $\{\mathcal{Y},\mathcal{S}\}$ 使得

$$\mathrm{MSE}(Q)=\mathrm{E}\|\boldsymbol{x}-Q(\boldsymbol{x})\|$$

达到最小.

(v) 如果 $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k\}$ 是 $\mathbb{R}^p$ 空间的点集,

$\mathrm{E}(\boldsymbol{x}\mid x\in S_j)=\boldsymbol{y}_j$, $j=1,\ldots,k$, 我们称代表点 $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k\}$ 是自相容的.

LBG 算法强烈依赖初始点集, 并且收敛速度的阶为 $O_p(\frac{1}{\sqrt{n}})$, 可以提高. Fang 等 [21] 建议使用数论方法取初始点集和 N 个训练样本, 可以改善 LBG 算法的输出质量和收敛速度. 修改以后的算法称为 NTLBG 算法, 应用这个算法, 他们给出了某些椭球等高分布子类的分布代表点. 如何使用数论方法产生椭球等高分布成为 NTLBG 算法的关键.

3 椭球等高分布族的数论方法代表点

椭球等高分布族 (椭球对称分布族) 是非常有用的多元统计分布族, 它包含多元正态分布、多元指数分布、多元 t 分布、多元 F 分布、多元均匀分布 (在多维矩形、多维球体、多维球面、多维单纯形上的均匀分布等)、多元 Beta 分布和多元 Gamma 分布等. 本文中 "$\boldsymbol{u}\overset{d}{=}\boldsymbol{v}$" 表示随机向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 有同样的分布.

椭球等高分布族的标准型是球对称分布. 如果随机向量 $\boldsymbol{y}=(Y_1,\ldots,Y_p)^{\mathrm{T}}$ 对一切 p 维正交矩阵 $\boldsymbol{P}$ 满足条件

$$\boldsymbol{y}\overset{d}{=}\boldsymbol{P}\boldsymbol{y}, \tag{3.1}$$

则称 $\boldsymbol{x}$ 遵从球对称分布, 它的特征函数必有形式 $\phi(\boldsymbol{t}^{\mathrm{T}}\boldsymbol{t})$. 记为 $\boldsymbol{y}\sim S_p(\phi)$. 球对称分布不一定有密度函数, 例如, 在单位球面的多元均匀分布不存在密度函数. 如果 $\boldsymbol{y}$ 遵从球对称分布并且存在分布密度, 后者必有形式 $g(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})$. 如果 $\boldsymbol{y}\sim S_p(\phi)$ 并且 $\mathrm{P}(\boldsymbol{y}=\boldsymbol{0})=0$, 那么 $\|\boldsymbol{y}\|\overset{d}{=}R$ 和 $\boldsymbol{y}/\|\boldsymbol{y}\|\overset{d}{=}\boldsymbol{u}^{(p)}$ 独立, 这里随机向量 $\boldsymbol{u}^{(p)}$ 在 $\mathbb{R}^p$ 的单位球面 S^p 上均匀分布, 并且 $\boldsymbol{y}\overset{d}{=}R\boldsymbol{u}^{(p)}$. 称为球对称分布的随机表示.

如果随机向量 $\boldsymbol{x}\in\mathbb{R}^p$ 有随机表示

$$\boldsymbol{x}\overset{d}{=}\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\boldsymbol{y}\overset{d}{=}\boldsymbol{\mu}+R\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}^{(p)}, \tag{3.2}$$

这里 $\boldsymbol{\mu}\in\mathbb{R}^p$, $\boldsymbol{\Sigma}$ 是 p 维正定矩阵, $\boldsymbol{\Sigma}^{1/2}$ 是 $\boldsymbol{\Sigma}$ 的正定平方根矩阵, R 和 $\boldsymbol{u}^{(p)}$ 独立. 称 $\boldsymbol{x}$ 遵从椭球对称分布, 或椭球等高分布 (elliptically contoured distribution, ECD), 记为 $\boldsymbol{x}\sim\mathrm{ECD}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\phi)$, 这里 $R\boldsymbol{u}^{(p)}$ 的特征函数为 ϕ. 注意, $\boldsymbol{\Sigma}$ 不一定是 $\boldsymbol{x}$ 的协方差矩阵, 但是正比例于协方差矩阵. 为简单起见, 本文视 $\boldsymbol{\Sigma}$ 为协方差矩阵.

椭球等高分布 $\boldsymbol{x}\sim \mathrm{ECD}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\phi)$ 存在分布密度的充分必要条件是 R 存在分布密度 (记为 $f(t)$). 这时, $\boldsymbol{x}$ 的分布密度有形式

$$|\boldsymbol{\Sigma}|^{1/2}g((\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})),$$

可以写为 $\boldsymbol{x}\sim \mathrm{ECD}_p(\boldsymbol{\mu},\boldsymbol{\Sigma},g)$. 因为它的密度等高的几何形状是椭球, 因而产生椭球等高分布的名称. 函数 g 和 f 有下列关系:

$$f(r)=\frac{2\pi^{p/2}}{\Gamma(p/2)}r^{p-1}g(r^2). \tag{3.3}$$

椭球等高分布的理论可以参见文献 [22]. 我们可以利用椭球等高分布的随机表示 (3.2) 来生成它的随机样本或数论方法样本.

假设随机向量 $\boldsymbol{x}\sim F(\boldsymbol{x})$, $\boldsymbol{x}\in\mathbb{R}^p$ 有随机表示

$$\boldsymbol{x}=\boldsymbol{h}(\boldsymbol{y}),\quad \boldsymbol{y}\sim U(C^t),\quad t\leqslant p,$$

这里随机向量 $\boldsymbol{y}$ 在单位立方体 C^t 上均匀分布, $\boldsymbol{h}$ 是 C^t 上的连续函数. 利用数论方法 (如好格子方法 [9]) 在 C^t 上产生均匀分散的点集 $\{\mathbf{c}_i, i=1,\ldots,n\}$, 令

$$\boldsymbol{x}_i=\boldsymbol{h}(\mathbf{c}_i),\quad i=1,\ldots,n, \tag{3.4}$$

那么 $\boldsymbol{x}_i$ $(i=1,\ldots,k)$ 是 $\boldsymbol{x}$ 的数论方法代表点. 这个算法称为 NTSR (number-theoretic stochastic representation) 算法.

利用随机表示 (3.2), 我们可以获得下列产生椭球等高分布数论方法代表点的 NTSR 算法.

步骤 1 在 C^p 上生成均匀分散的点集 $\{\mathbf{c}_i=(c_{i1},\ldots,c_{ip}), i=1,\ldots,k\}$.

步骤 2 记 $F_R(r)$ 为 R 的分布函数, 它的逆函数为 F_R^{-1}, 计算 $r_i=F_R^{-1}(c_{ip})$, $i=1,\ldots,k$.

步骤 3 用 NTSR 算法产生在 $\mathbb{R}^p$ 的单位球面 S^p 上均匀分布的数论方法代表点 $\{\boldsymbol{u}_i, i=1,\ldots,k\}$, 计算是基于 $\{\mathbf{c}_i=(c_{i1},\ldots,c_{i,p-1}), i=1,\ldots,k\}$.

步骤 4 输出椭球等高分布数论方法代表点 $\{\boldsymbol{y}_i=\boldsymbol{\mu}+r_i\boldsymbol{\Sigma}^{1/2}\boldsymbol{u}_i, i=1,\ldots,k\}$.

对于给定的 p、$\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$, 不同的椭球等高分布的相异仅仅是不同的 R 分布. 关键的困难是如何产生 $\boldsymbol{u}^{(p)}$ 的随机样本或数论方法样本, Fang 和 Wang [9] 给出了详细介绍. 表 1 列出椭球等高分布经常使用的子类的分布密度函数格式. 本文将讨论多元正态分布、Kotz 型、Pearson II 型和 Pearson VII 型分布.

表 1 椭球等高分布的一些子类

子类	$g(\boldsymbol{x})$ 在 $\mathbb{R}^p$ 的分布密度函数
Kotz 型	$g(\boldsymbol{x})=c(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{N-1}\exp[-r(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^s]$, $r,s>0$, $2N+p>2$
多元正态分布	$g(\boldsymbol{x})=c\exp(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})$
Pearson VII 型	$g(\boldsymbol{x})=c(1+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}/s)^{-N}$, $N>p/2$, $s>0$
多元 t	$g(\boldsymbol{x})=c(1+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}/s)^{-(p+q)/2}$, $q>0$ 是一个整数, $s>0$
多元 Cauchy	$g(\boldsymbol{x})=c(1+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-(p+1)/2}$
Pearson II 型	$g(\boldsymbol{x})=c(1-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^q$, $q>0$
Logistics	$g(\boldsymbol{x})=c\,\exp(-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})/[1+\exp(-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})]^2$
混合型 (scale mixture)	$g(\boldsymbol{x})=c\int_0^\infty t^{-p/2}\exp(-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}/2t)dG(t)$, $G(t)$ 是分布函数

4 统计分布代表点的比较和应用

第 1 节介绍了 4 类统计分布代表点, 本节给出 4 类统计分布代表点在密度估计和统计推断中的比较. 通过样本来推断总体是统计学的基本目的之一. 为了方便讨论, 本节主要考虑一元统计分布. 当知道总体的分布类型时, 需要通过样本来估计总体的未知参数, 如分布的均值和方差. 如果不知道总体分布的类型, 核密度估计的方法可帮助我们用样本或代表点来直接估计总体的分布密度函数.

4.1 密度估计

在信息传递过程中, 输出方在不同的时间发送了 k 个序列数, 输入方希望从收到的数据获得尽量多的信息, 密度估计的方法提供了有力的工具. 核密度估计方法由 Rosenblatt [23] 和 Parzen [24] 提出. 设 $x_1,\dots,x_n$ 是来自总体的一组样本, 总体有密度函数 $p(x)$. 核密度估计建议用泛函

$$\hat{p}_h(x)=\frac{1}{n}\sum_{i=1}^{n}k_h(x-x_i)=\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{x-x_i}{h}\right) \tag{4.1}$$

作为 $p(x)$ 的一个估计, 这里 $k(\cdot)$ 为核函数, $k_h(y)=\frac{1}{h}k(x/h)$, h 为窗宽. 最常用的核函数为标准正态分布密度函数, 即 $k(x)=\phi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2}$, 本文就采用这个核函数. 如果我们掌握的信息不是一组随机样本, 而是总体的一个代表点集 $y_1,\dots,y_k$, 其概率分布由 (1.4) 给出, 这时核密度估计的公式可推广为

$$\hat{p}_h(x)=\sum_{i=1}^{k}k_h(x-y_i)p_i=\frac{1}{h}\sum_{i=1}^{k}k\left(\frac{x-y_i}{h}\right)p_i. \tag{4.2}$$

窗宽 h 的选择很重要, 这方面虽然有一些理论探讨, 但在实际中, 一些统计软件如 MATLAB 可以提供最好的 h. Fang 等 [4] 在正态分布的模型下, 比较了 3 种方法的密度估计, 当 $k=30$ 时, 这 3 个核密度的估计 $\hat{p}_h(x)$ 与总体分布密度 $p(x)$ 的 L_2- 距离分别为 MC: 0.8725、QMC: 0.0468 和 MSE: 0.0033, 以 MSE 方法最好. 由于随机样本的随机性, MC 方法的表现欠佳. 我们建议修改为: 把随机样本 $x_1,\dots,x_k$ 视为总体近似分布的支撑点, 相应的概率用 (1.5) 计算. 当代表点个数 $k=25,28,31$ 时, 表 2 给出了 4 个核密度的估计 $\hat{p}_h(x)$ 与总体分布密度 $p(x)$ 的 L_2- 距离. 图 1 展示了当 $k=25$ 时 4 类代表点方法对标准正态分布的核密度估计的效果, 从此图可以看出 MSE 代表点的估计效果最好, 与表 2 的结果一致.

他们的结论可以延伸到其他的统计分布吗? 本文考虑混合正态分布的模型, 其分布密度是两个正态分布的加权和

$$p(x)=\alpha n_1(x;\mu_1,\sigma_1^2)+(1-\alpha)n_2(x;\mu_2,\sigma_2^2), \tag{4.3}$$

这里 $n_i(x;\mu_i,\sigma_i^2)$ $(i=1,2)$ 是两个正态分布密度函数, $0<\alpha<1$. 如果两个正态分布分别取为 $N(-8,3^2)$ 和 $N(3,4^2)$, $\alpha=0.6$, 比较相应的混合正态分布的密度估计.

表 2 4 个核密度的估计 $\hat{p}_h(x)$ 与总体分布密度 $p(x)$ 的 L_2- 距离

k	MC	RMC	QMC	MSE
25	1.6027	0.4536	0.3499	0.0984
28	1.8084	0.4293	0.3283	0.0872
31	2.3160	0.4025	0.3145	0.0771

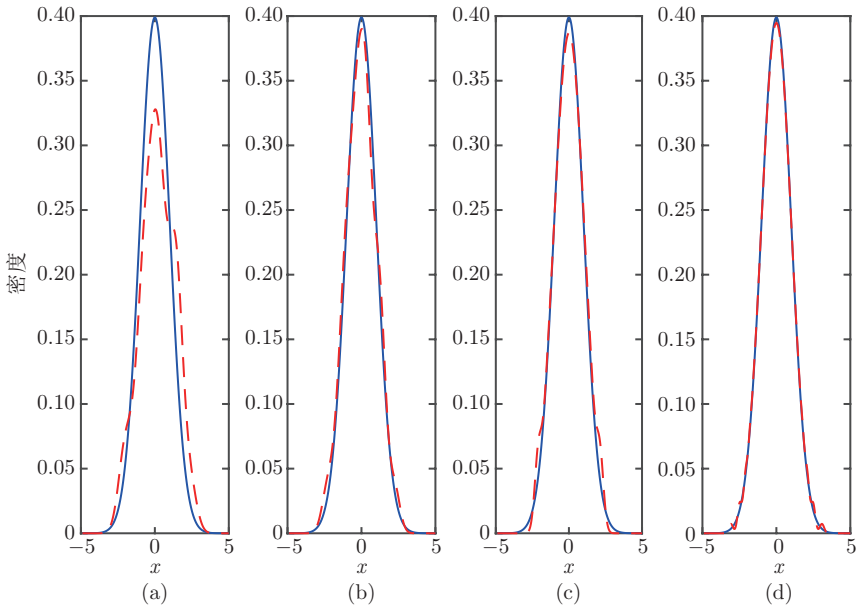

图 1 (网络版彩图) 4 类代表点方法对标准正态分布的核密度估计, $k=25$, 其中实线表示总体分布密度, 虚线表示核密度估计. (a) MC ($h=0.45$); (b) RMC ($h=0.35$); (c) QMC ($h=0.25$); (d) MSE ($h=0.15$)

当 k 分别等于 25、28 和 31 时, 4 个核密度的估计 $\hat{p}_h(x)$ 与总体分布密度 $p(x)$ 的 L_2- 距离列于表 3. RMC 方法在混合正态分布模型下进行密度估计比 MC 方法有更好的表现, 但其效果不及 QMC 方法和 MSE 方法, 而 MSE 方法仍然最好. 图 2 展示了当 $k=31$ 时 4 类代表点对混合正态分布核密度估计的表现, 如图所示, MSE 最接近原总体分布.

4.2 4 类代表点的统计推断

考虑随机变量 $X\sim F(x)$, 有密度函数 $p(x)$, 存在如下前 4 阶矩均值 (μ)、方差 (Var)、偏态系数 (Sk) 和峰度系数 (Ku):

$$\mu=\mathrm{E}(X),\quad \sigma^2=\mathrm{Var}(X),\quad \mathrm{Sk}(X)=\frac{\mathrm{E}(X-\mu)^3}{\sigma^3},\quad \mathrm{Ku}(X)=\frac{\mathrm{E}(X-\mu)^4}{\sigma^4}-3.$$

令 Y 是 X 的近似离散随机变量, 它的分布由 (1.4) 给出, 则 Y 的上述 4 个统计量为

$$\begin{aligned}&\mathrm{E}(Y)=\sum_{i=1}^{k}y_ip_i\equiv\mu_y,\quad \mathrm{Var}(Y)=\sum_{i=1}^{k}(y_i-\mu_y)^2p_i\equiv\sigma_Y^2,\\&\mathrm{Sk}(Y)=\frac{1}{\sigma_Y^3}\sum_{i=1}^{k}(y_i-\mu_y)^3p_i,\quad \mathrm{Ku}(Y)=\frac{1}{\sigma_Y^4}\sum_{i=1}^{k}(y_i-\mu_y)^4p_i-3.\end{aligned}\tag{4.4}$$

由标准正态分布的性质可知 $\mathrm{E}(Z)=0$, $\sigma^2=1$, $\mathrm{Sk}(Z)=\mathrm{Ku}(Z)=0$. Fang 等 [4] 在正态分布的模型下,

表 3 混合正态分布模型下 4 个核密度的估计 $\hat{p}_h(x)$ 与总体分布密度 $p(x)$ 的 L_2- 距离

k	MC	RMC	QMC	MSE
25	0.4447	0.2588	0.0884	0.0330
28	0.3382	0.1783	0.0818	0.0285
31	0.3544	0.0980	0.0745	0.0246

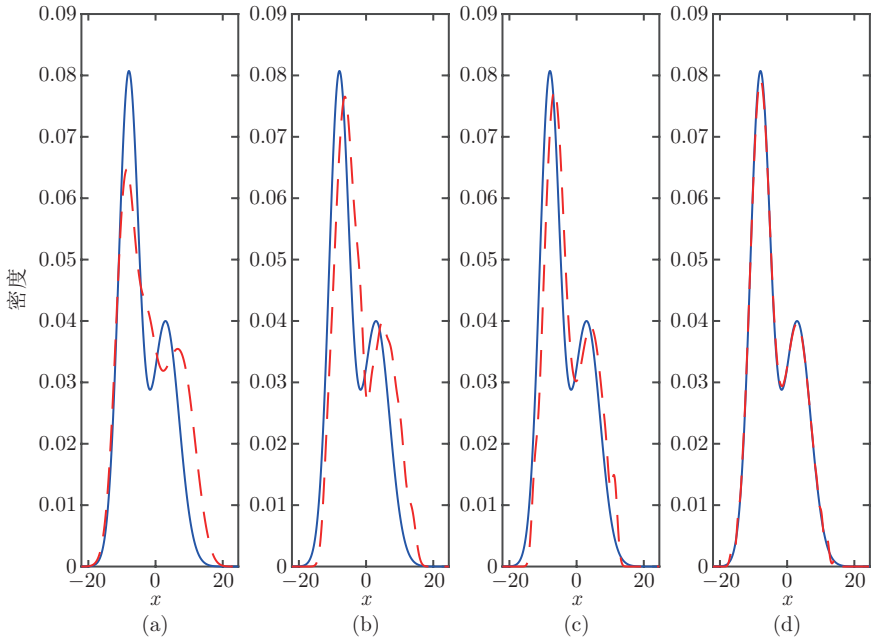

图 2 (网络版彩图) 4 类代表点方法对混合正态分布的核密度估计, $k=31$, 其中实线表示总体分布密度, 虚线表示核密度估计 (a) MC ($h=2.25$); (b) RMC ($h=1.20$); (c) QMC ($h=0.90$); (d) MSE ($h=0.60$)

比较了 4 种方法对上述 4 个统计量的估计, 他们取 k 分别等于 5、10、15、20、25、28 和 31, 从计算看出 MC 方法最差, 而大部分情形, MSE 方法最好. 表 4 给出了上面 4 种方法对标准正态分布的 4 个统计量估计的误差, 我们看到 RMC 的表现比 MC 好, 对均值和偏态系数的估计, QMC 比 RMC 好, 但是对方差和峰度系数的估计, RMC 比 QMC 好, MSE 在这次比较中最好.

表 4 4 种方法对标准正态分布的 4 个统计量估计的误差

统计量	产生近似分布的方法	$k=10$	$k=20$	$k=25$	$k=28$	$k=31$
均值	MC	−0.0011	0.4190	−0.0313	−0.2046	0.2905
	RMC	−0.0193	0.1167	−0.0030	−0.0178	0.0010
	QMC	3.6554E−14	1.3578E−13	1.6279E−13	1.6985E−13	1.6900E−13
	MSE	−1.2050E−12	1.8745E−13	−8.3149E−14	−4.8168E−14	−4.0390E−14
方差	MC	−0.7141	−0.2152	0.2901	0.0516	0.2048
	RMC	−0.4301	−0.3547	−0.0334	−0.0916	−0.0375
	QMC	−0.1202	−0.0614	−0.0494	−0.0443	−0.0401
	MSE	−0.0229	−0.0062	−0.0040	−0.0032	−0.0027
偏态系数	MC	−0.2824	0.3751	−0.1164	−0.2676	−0.3717
	RMC	−0.1335	0.6400	−0.0089	−0.1534	−0.0224
	QMC	2.5017E−13	1.2973E−12	1.6946E−12	1.8323E−12	1.8807E−12
	MSE	−2.2351E−11	5.1300E−12	−2.5291E−12	−1.5429E12	1.3553E−12
峰度系数	MC	0.1864	−1.1284	−0.5257	−0.8999	−0.2767
	RMC	−1.1980	−0.5361	−0.2360	−0.4398	−0.2259
	QMC	−0.7512	−0.4943	−0.4291	−0.3988	−0.3732
	MSE	−0.2023	−0.0627	−0.0421	−0.0343	−0.0284

考虑随机变量 X 服从两个正态分布 $N(\mu_i,\sigma_i^2)$ $(i=1,2)$ 组成的混合正态分布, 其权重为 $\alpha_1=\alpha$ 和 $\alpha_2=1-\alpha$, 则 X 的上述 4 种统计量分别为

$$\begin{aligned}&\mathrm{E}(X)=\sum_{i=1}^{2}\alpha_i\mu_i\equiv\mu,\quad \mathrm{Var}(X)=\sum_{i=1}^{2}\alpha_i(\sigma_i^2+\mu_i^2)-\mu^2\equiv\sigma^2,\\&\mathrm{Sk}(X)=\frac{1}{\sigma^3}\sum_{i=1}^{2}\alpha_i(\mu_i-\mu)(3\sigma_i^2+(\mu_i-\mu)^2),\\&\mathrm{Ku}(X)=\frac{1}{\sigma^4}\sum_{i=1}^{2}\alpha_i(3\sigma_i^4+6(\mu_i-\mu)^2\sigma_i^2+(\mu_i-\mu)^4).\end{aligned}\tag{4.5}$$

由 (4.5) 可以得到 4 种统计量的理论值. 令 Y 为 X 的近似离散随机变量, 则 Y 的上述 4 种统计量可以用 (4.4) 计算. 在模型 (4.3) 下, 我们仍然以两个正态分布 $N(-8,3^2)$ 和 $N(3,4^2)$ 组成的混合正态分布作为例子, 比较 4 种方法对上述 4 种统计量的估计, 取 k 分别等于 10、20、25、28 和 31, 其结果列于表 5. 在大部分情形下, 可以看出 MC 方法最差, MSE 方法最好.

Fang 等 [4] 进一步比较 3 种方法在重采样中的性能. 重采样是从一个近似的总体 Y (1.4) 抽样, 记重采样样本大小为 N, 从重采样样本得到有关的参数的估计. 例如, 估计 Y 的均值和方差作为 X 的均值和方差的估计. 众所周知, 自助法是一种重采样的方法, 是 Efron [25] 于 1979 年提出来的, 非常广泛地应用于估计总体的参数. 因为 QMC 和 MSE 的代表点集不包含随机性, 好像不能用于随机模拟. Fang 等 [4] 是第一个利用重采样这个平台将 QMC 和 MSE 的代表点集用于随机模拟.

对于混合正态分布, 我们做相同的试验, 结果列于表 6–8. MC 代表点集的重采样估计方差 ($k=31$, $N=2{,}000$, $5{,}000$) 时有好的表现; RMC 代表点集估计均值 ($k=25$, $N=1{,}000$; $k=31$, $N=2{,}000$, $5{,}000$) 和峰度 ($k=28$, $N=10{,}000$; $k=31$, $N=1{,}000$, $2{,}000$, $5{,}000$, $10{,}000$) 时表现最佳; QMC 代表

表 5 混合正态分布的几种近似分布估计参数的误差

统计量	产生近似分布的方法	$k=10$	$k=20$	$k=25$	$k=28$	$k=31$
均值	MC	−1.5194	−1.2916	−0.8254	−0.2367	0.2974
	RMC	−0.2706	−0.0714	−0.0461	0.0036	0.0110
	QMC	−0.0172	−0.0079	−0.0061	−0.0054	−0.0048
	MSE	5.6222E−13	1.6342E−13	1.5232E−13	1.4122E−13	6.6613E−15
方差	MC	−17.1440	−13.5439	−13.3645	−4.3663	4.3873
	RMC	−8.2007	−5.0780	−2.9665	−1.4037	−1.2487
	QMC	−2.6334	−1.2827	−1.0187	−0.9063	−0.8160
	MSE	−0.6344	−0.1709	−0.1112	−0.0893	−0.0732
偏态系数	MC	0.2416	0.4112	0.8121	0.0264	−0.0484
	RMC	−0.1602	−0.0645	0.2289	0.0182	0.0109
	QMC	−0.0380	−0.0207	−0.0171	−0.0156	−0.0143
	MSE	−3.6945E−4	−1.7203E−4	−1.1813E−4	−9.6992E−5	−8.1012E−5
峰度系数	MC	0.2521	0.3165	2.4727	0.0516	-0.3770
	RMC	−0.6202	−0.4362	0.5333	−0.1172	−0.1500
	QMC	−0.2925	−0.1768	−0.1497	−0.1375	−0.1273
	MSE	−0.0888	−0.0265	−0.0176	−0.0143	−0.0118

点集估计均值 ($k=25$, $N=2{,}000$; $k=28$, $N=1{,}000$, $10{,}000$; $k=31$, $N=1{,}000$) 和峰度 ($k=25$, $N=1{,}000$, $2{,}000$) 时获得最精确的估计. 其余 32 种情形, 都以 MSE 代表点组成的近似总体为最优. 在上述的 3 个表中, 黑体表示同类中最好的. 表 9 统计 4 种方法估计 4 种统计量偏差最小的次数, MSE 代表点组成的近似总体在混合正态分布模型中仍然有明显的优势.

表 6　对混合正态分布使用 $k=25$ 个代表点, 用 4 种近似分布的重采样对 4 个参数估计的误差

统计量	产生近似分布的方法	$N=1{,}000$	$N=2{,}000$	$N=5{,}000$	$N=10{,}000$
均值	MC	−0.8166	−0.8123	−0.8210	−0.8312
	RMC	−0.0087	−0.0516	−0.0360	−0.0422
	QMC	−0.0489	−0.0361	−0.0213	0.0122
	MSE	−0.0493	−0.0503	**0.0142**	−0.0081
方差	MC	−13.8159	−14.2544	−14.5410	−14.4436
	RMC	−3.7956	−4.2640	−4.4867	−4.5306
	QMC	−2.5794	−2.5570	−2.7826	−2.4634
	MSE	−1.8174	−1.8934	−1.7205	−1.7869
偏态系数	MC	0.6068	0.5830	0.5766	0.5837
	RMC	0.1707	0.1552	0.1406	0.1386
	QMC	−0.0213	−0.0093	−0.0232	−0.0264
	MSE	−0.0251	−0.0182	−0.0155	−0.0186
峰度系数	MC	1.7145	1.6689	1.6419	1.6556
	RMC	0.3957	0.3662	0.3363	0.3251
	QMC	−0.0756	−0.0391	−0.0643	−0.0762
	MSE	−0.0048	**0.0177**	**0.0012**	**0.0149**

4.3　MSE 的分布

如果我们用 MSE 准则来比较 MC 和 MSE 方法, 需要考虑如下意义的 MSE 分布. 假设随机向量 $\boldsymbol{x}\sim F(\boldsymbol{x})$, $(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k)$ 是 $\boldsymbol{x}$ 的一个样本. MC 方法用这个样本当作代表点, 相应的 MSE 是一个随机变量, 记为 $(\mathrm{MSE}(\boldsymbol{x})\mid \boldsymbol{y}_1,\ldots,\boldsymbol{y}_k)$, 它的统计分布提供了 MC 方法的随机性.

考虑一元正态分布 $X\sim N(\mu,\sigma^2), k=1$ 的情形. 令 y 是 $N(\mu,\sigma^2)$ 的一个样本, 并且 X 和 y 独立. 有

$$\begin{aligned}(\mathrm{MSE}(X)\mid Y=y)&=\mathrm{E}_X[(X-y)^2]=\mathrm{E}_X[(X-\mu+\mu-y)^2]\\&=\mathrm{E}_X(X-\mu)^2+(\mu-y)^2=\sigma^2+(\mu-y)^2.\end{aligned}$$

因为 $y\sim N(\mu,\sigma^2)$, 因而 $(y-\mu)\sim N(0,\sigma^2)$, $\frac{1}{\sigma^2}(y-\mu)^2\sim\chi_1^2$ (自由度为 1 的卡方分布). 因此, $(\mathrm{MSE}(X)\mid Y=y)$ 有随机表示

$$(\mathrm{MSE}(X)\mid Y=y)\overset{d}{=}\sigma^2(1+W),\quad \text{这里 } W\sim\chi_1^2.$$

进一步考虑 $\boldsymbol{x}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $k=1$ 的情形. 令 $\boldsymbol{y}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 与 $\boldsymbol{x}$ 独立. 有

$$(\mathrm{MSE}(\boldsymbol{x})\mid \boldsymbol{y}=\boldsymbol{y}_0)=\mathrm{E}_{\boldsymbol{x}}[(\boldsymbol{x}-\boldsymbol{y}_0)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}_0)]$$

表 7 对混合正态分布使用 $k=28$ 个代表点, 用 4 种近似分布的重采样对 4 个参数估计的误差

统计量	产生近似分布的方法	$N=1{,}000$	$N=2{,}000$	$N=5{,}000$	$N=10{,}000$
均值	MC	−0.0554	−0.0732	−0.1378	−0.2379
	RMC	−0.0335	−0.0251	−0.0186	0.0032
	QMC	**0.0102**	0.0211	0.0144	**0.0028**
	MSE	−0.0205	−0.0142	−0.0123	−0.0247
方差	MC	−4.6482	−4.7448	−5.0219	−5.6802
	RMC	−4.3472	−4.1379	−3.5968	−2.8893
	QMC	−2.4380	−2.4518	−2.4522	−2.1757
	MSE	−1.8123	−1.7014	−1.6184	−1.7492
偏态系数	MC	0.1227	0.1083	0.0678	−0.0136
	RMC	0.1255	0.1111	0.0729	**0.0014**
	QMC	−0.0252	−0.0246	−0.0255	−0.0226
	MSE	−0.0137	−0.0153	−0.0154	−0.0076
峰度系数	MC	0.2984	0.2702	0.1878	0.0389
	RMC	0.2889	0.2483	0.1275	−0.0707
	QMC	−0.0749	−0.0720	−0.0728	−0.0677
	MSE	**0.0213**	**0.0176**	**0.0123**	**0.0226**

表 8 对混合正态分布使用 $k=31$ 个代表点, 用 4 种近似分布的重采样对 4 个参数估计的误差

统计量	产生近似分布的方法	$N=1{,}000$	$N=2{,}000$	$N=5{,}000$	$N=10{,}000$
均值	MC	0.0280	0.0657	0.1587	0.2882
	RMC	0.0044	**0.0054**	**0.0047**	0.0257
	QMC	**0.0033**	0.0144	0.0068	0.0038
	MSE	−0.0293	−0.0281	−0.0233	−0.0012
方差	MC	−2.2671	−1.6709	**0.0514**	2.9883
	RMC	−2.8255	−2.7934	−2.8144	−2.5487
	QMC	−2.1605	−2.1353	−2.1949	−2.0730
	MSE	−1.7192	−1.7166	−1.6455	−1.3404
偏态系数	MC	−0.0022	−0.0090	−0.0274	−0.0511
	RMC	**0.0015**	**0.0022**	**0.0011**	**0.0028**
	QMC	−0.0217	−0.0246	−0.0239	−0.0240
	MSE	−0.0086	−0.0069	−0.0103	−0.0089
峰度系数	MC	−0.0908	−0.1144	−0.1809	−0.2873
	RMC	−0.0728	−0.0742	−0.0777	−0.0883
	QMC	−0.0658	−0.0669	−0.0700	−0.0692
	MSE	**0.0207**	**0.0279**	**0.0201**	**0.0182**

$$
\begin{aligned}
&= \mathrm{E}_{\boldsymbol{x}}[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})] + (\boldsymbol{\mu}-\boldsymbol{y}_0)^{\mathrm{T}}(\boldsymbol{\mu}-\boldsymbol{y}_0)\\
&= \mathrm{tr}(\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\boldsymbol{y}_0)^{\mathrm{T}}(\boldsymbol{\mu}-\boldsymbol{y}_0).
\end{aligned}
$$

表 9　参数估计中表现最好的次数

	MC	RMC	QMC	MSE
均值	0	3	4	5
方差	2	0	0	10
偏态系数	0	5	2	5
峰度系数	0	0	0	12
总计	2	8	6	32

如果 $\boldsymbol{\Sigma}=\sigma^2\boldsymbol{I}_p$, 那么 $(\boldsymbol{y}-\boldsymbol{\mu})\sim N_p(\boldsymbol{0},\sigma^2\boldsymbol{I}_p)$ 因而 $\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu})\sim\chi^2_p$ (自由度为 p 的卡方分布). 因此, $(\mathrm{MSE}(\boldsymbol{x})\mid\boldsymbol{y})\overset{d}{=}\sigma^2(p+W_p)$, 这里 $W\sim\chi^2_p$. 研究 $k=2$, $X\sim N(\mu,\sigma^2)$ 的情形, 令 y_1 和 y_2 是 $N(\mu,\sigma^2)$ 的随机样本. 已知 $Y_1=y_1$ 和 $Y_2=y_2$ 的条件时, X 的 MSE 为

$$(\mathrm{MSE}(X)\mid Y_1=y_1,Y_2=y_2)=\mathrm{E}_X[\min\{(X-y_1)^2,(X-y_2)^2\}].$$

当固定 $X=x$ 时, 有 $(y_1-x)^2\sim N(\mu-x,\sigma^2)$ 和 $(y_1-x)^2\sim\chi^2_1((\mu-x)^2)$, 其中 $\chi^2_1((\mu-x)^2)$ 为自由度为 1 的非中心卡方分布. 类似地, $(y_2-x)^2\sim\chi^2_1((\mu-x)^2)$. 那么, 对给定 $X=x$, $\min\{(y_1-x)^2,(y_2-x)^2\}$ 的分布可以从

$$\mathrm{P}(\min\{(y_1-x)^2,(y_2-x)^2\}>w)=\mathrm{P}((y_1-x)^2>w)\mathrm{P}((y_2-x)^2>w)=(1-G_{W_x}(w))^2$$

导出, 这里 $G_{W_x}(w)$ 是 $W_x\sim\chi^2_1((\mu-x)^2)$ 的分布函数.

对于 $k>2$, $\boldsymbol{x}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 或其他椭球等高分布的情形, $(\mathrm{MSE}(\boldsymbol{x})\mid\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k)$ 的分布更为复杂. 因此, 寻找它的近似分布更加实用. 通过筛选, 我们发现广义极值分布是一个很好的选择. 广义极值分布的密度函数为

$$f(x\mid\kappa,\mu,\sigma)=\frac{1}{\sigma}\exp\left(-\left(1+\kappa\frac{x-\mu}{\sigma}\right)^{-1/\kappa}\right)\left(1+\kappa\frac{x-\mu}{\sigma}\right)^{-1-\frac{1}{\kappa}},\quad \text{如果 } 1+\kappa\frac{x-\mu}{\sigma}>0,$$

这里 κ 是形状参数, σ 是刻度参数, μ 是位置参数. 广义极值分布分为 3 类:

类 I: Gumbel 分布, $\kappa=0$, 定义域 $-\infty<x<\infty$;

类 II: Frechet 分布, $\kappa>0$, 定义域 $0<x<\infty$;

类 III: Weibull 型, $\kappa<0$, 定义域 $-\infty<x<0$.

例如 $\boldsymbol{x}\sim N_p(\boldsymbol{0},\boldsymbol{\Sigma})$ 的情形, 其中 $\boldsymbol{\Sigma}$ 的主对角线元素为 1, 非对角线元素为 ρ. 考虑 $p=2$, k 分别等于 2 和 5, ρ 分别等于 0、0.3、0.6 和 0.9, 以及 $p=3$, $k=5$, ρ 分别等于 0、0.3、0.6 和 0.9 的情形, 它们用广义极值分布的拟合均比较理想. 图 3 给出了用广义极值分布对 $p=3$, $k=5$, $\rho=0.9$ 情形的拟合直方图, 其中训练样本量 $N=2{,}000$. 从我们的实验可以看出, 对椭球等高分布族的不同子类、高维或低维、不同的 ρ 值, 用广义极值分布拟合效果都不错. MSE 代表点达到最小的 MSE 值, 但是 MC 代表点的 MSE 变化很大. 这解释了为什么自助法在重采样的表现不稳定.

4.4　反正弦分布的一个有趣性质

对大部分统计分布, MSE 代表点比 QMC 代表点在统计模拟中表现要好一些. 但是我们发现, 对反正弦分布有不同的结论. 反正弦分布是 Beta 分布的特例, 它的分布函数为

$$F(x)=\frac{2}{\pi}\arcsin(\sqrt{x}),\quad 0\leqslant x\leqslant 1,$$

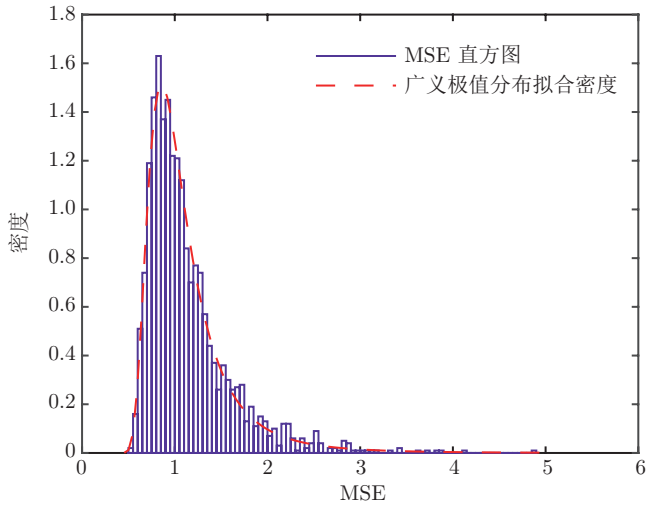

图 3　(网络版彩图) 广义极值分布的拟合直方图 ($p = 3$, $k = 5$, $\rho = 0.9$)

所以称为反正弦分布. 一般而言, 连续的统计分布函数 $F(x)$ 与它的离散近似 $F_Y(x)$ 有彼此接近的低阶矩, 但是值一般不同. Jiang 等 [26] 发现反正弦分布的 k 个 QMC 代表点为

$$F^{-1}\left(\frac{2i-1}{2k}\right) = \sin^2\left(\frac{(2i-1)\pi}{4k}\right), \quad i = 1, \ldots, k,$$

并且证明它的离散近似分布 $F_Y(x)$ 和原来的反正弦分布有相同的前 $k-1$ 阶矩, 直观看来这似乎是不可能的性质.

是否只有反正弦分布具有这个性质? 周永道和方开泰 [27] 推断出可以有许多其他分布具有这个性质, 并且建议了一个新的分布代表点的准则, 称为 FM- 准则. 该准则在前 $k-1$ 个样本矩等于相应的总体矩的约束条件下最小化经验分布函数与总体分布函数之间的 L_2- 距离. 我们证明该准则对很多分布都是有意义的. 当约束条件不满足时, 该准则被推广至伪 FM- 准则. 一些例子表明 FM- 代表点比其他类型的代表点更优.

5　椭球等高分布代表点的一些性质

本节假设 p 维随机向量 $\boldsymbol{x} \sim F(\boldsymbol{x})$, $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$, $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Flury [13] 和他的合作者发现多维统计分布的 MSE 代表点与 $\boldsymbol{\Sigma}$ 的主成分有一些联系, 并且称多维统计分布的 MSE 代表点为主成分点 (principal points). 他们特别关注椭球等高分布的 MSE 代表点, 这时 $\boldsymbol{x} \sim \mathrm{ECD}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, 椭球等高分布的定义在第 3 节已经介绍. 这里列举相关的一些结论.

(i) 假设 $\boldsymbol{x} \sim \mathrm{ECD}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, $k = 2$, 那么 $\boldsymbol{x}$ 的两个主成分点为 $\boldsymbol{y}_1 = \boldsymbol{\mu} + \gamma_1\boldsymbol{\beta}$, $\boldsymbol{y}_2 = \boldsymbol{\mu} + \gamma_2\boldsymbol{\beta}$, 这里 $\boldsymbol{\beta} \in \mathbb{R}^p$ 是 $\boldsymbol{\Sigma}$ 的第一主成分单位化向量 ($\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = 1$), γ_1 和 γ_2 是随机变量 $\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\mu})$ 的两个 MSE 代表点.

(ii) 假设 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ 是 p 维随机向量 $\boldsymbol{x}$ 的 MSE 代表点, 它们一定是自相容的, 并且 $\boldsymbol{x}$ 的均值向量 $\mathrm{E}(\boldsymbol{x})$ 落在 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ 的凸包 [28].

(iii) 假设随机向量 $\boldsymbol{x} \sim \mathrm{ECD}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, 如果 $\boldsymbol{x}$ 的 k 个自相容的点集张成的线性子空间 (记为 $\mathcal{Q}$) 的秩 q 满足 $q < \min(k, p)$, 那么这个线性子空间可以由 $\boldsymbol{\Sigma}$ 的最大的 q 个主成分张成. 这个线性子空间也可以用 $\boldsymbol{x}$ 的 k 个代表点张成.

(iv) 如果 $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k$ 是 p 维随机向量 $\boldsymbol{x}$ 的 MSE 代表点, 那么 $\boldsymbol{a}+b\boldsymbol{H}\boldsymbol{y}_j$ $(j=1,\ldots,k)$ 是随机向量 $\boldsymbol{a}+b\boldsymbol{H}\boldsymbol{x}$ 的 MSE 代表点, 这里 $\boldsymbol{a}\in\mathbb{R}^p$, $b\in\mathbb{R}$, $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}=\boldsymbol{I}_p$.

Flury [13] 对二元正态分布的协方差矩阵是对角阵的情形给出了 k 分别等于 2、3、4 和 5 的 MSE 代表点, 并且提出了一些需要解决的问题. 费荣昌 [18] 对二元正态分布给出了 $2\leqslant k\leqslant 12$ 的 MSE 代表点, 他将二元积分分解为两个一元积分, 并且对代表点的结构做了假定. 我们用 NTLBG 算法检验他的结果, 发现他获得的 MSE 代表点是高质量的. NTLBG 算法是 Fang 等 [21] 提出的, 该算法可以数值计算椭球等高分布一些子类的 MSE 代表点, 代表点数任意. 但是他们没有进一步研究椭球等高分布代表点的性质. NTLBG 算法是研究多元统计分布, 特别是椭球等高分布代表点的强有力的工具. 利用这个工具, 我们研究了下列问题.

(1) 不同椭球等高分布子类 MSE 代表点的差别

从椭球等高分布的随机表示 (3.2), 对相同的 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$, 不同椭球等高分布子类的差别仅在于 R 的分布. 图 4 给出了椭球等高分布 4 个子类 R 的分布密度. 我们看到, 正态分布的 R 分布比较分散, 有长尾; Kotz 型 R 分布的质量主要集中在 0.25 到 1.0 之间; Pearson II 型和 Pearson VII 型的 R 分布比较接近, 质量主要集中在 0 到 1 之间. 从椭球等高分布的随机表示, 对相同的 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$, 不同椭球等高分布子类的 MSE 代表点应该有某些相似关系, 通过不同的 R 分布来缩小或放大. 所以, 我们可以通过多元正态分布的分布图案, 来想象其他椭球等高分布子类 MSE 代表点的图案.

(2) 多元正态分布代表点与主成分的联系

假定 $\boldsymbol{x}\sim N_p(\boldsymbol{0},\boldsymbol{\Sigma}_\rho)$, 这里

$$\boldsymbol{\Sigma}_\rho=\begin{bmatrix}1&\rho&\cdots&\rho\\ \rho&1&\cdots&\rho\\ \vdots&\vdots&\ddots&\vdots\\ \rho&\rho&\cdots&1\end{bmatrix}. \tag{5.1}$$

令 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$ 为 $\boldsymbol{\Sigma}_\rho$ 的特征值, $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p$ 为相应的标准特征向量. 表 10 列出 2 和 5 维正态分布代表点的 MSE 值. 因为当 $p=5$, $\rho\leqslant 0.3$ 时, $\boldsymbol{\Sigma}_\rho$ 不是正定矩阵, 相应的多元正态分布不存在, 表

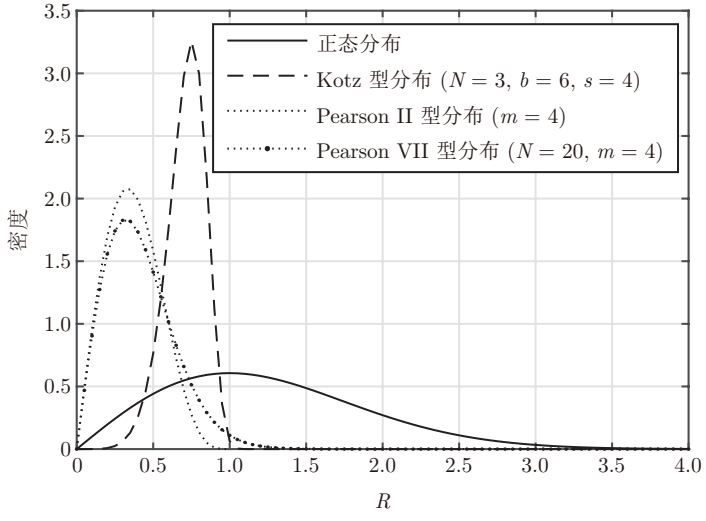

图 4　椭球等高分布 4 个子类 R 的分布密度

中有一部分空白. 多元统计分析定义第一主成分的贡献率为 $C_1=\lambda_1/\mathrm{tr}(\boldsymbol{\Sigma})$. 从表 10 可以清楚地看到:

(i) 当 k 增加时, MSE 减小;

(ii) 当 p 增加时, MSE 增加;

(iii) 当 $|\rho|$ 增加时, MSE 减小.

为了研究多元正态分布代表点与主成分的联系, 首先考虑 $p=2,\boldsymbol{\Sigma}=\mathrm{diag}(\sigma^2,1)\equiv\boldsymbol{\Sigma}_\sigma$. 当 $k=2$ 时, $\boldsymbol{x}$ 的两个代表点在 $\boldsymbol{\Sigma}$ 的第一主成分上. 当 $k=3$ 时, 我们发现当 $\sigma<\sigma_0(3)=1.5416$ 时, 3 个代表点构成一个等腰三角形; 但是当 $\sigma\geqslant 1.5416$ 时, 3 个代表点全落在第一主成分上. 这个现象告诉我们, 当 $\sigma\geqslant 1.5416$ 时, 仅用一个主成分可以取得 $\boldsymbol{x}$ 的主要信息, 其贡献率为 $C_1=\frac{\sigma^2}{1+\sigma^2}$. 当 $\sigma=1.5416$ 时, 相应的贡献率为 $C_1=0.7038$. 在 $k=4$ 的情形, 当 $\sigma<\sigma_0(4)=2.1509$ 时, 4 个代表点构成一个矩形; 当 $\sigma\geqslant\sigma_0(4)=2.1509$ 时, 4 个代表点全落在第一主成分上, 这时第一主成分的贡献率为 $C_1=0.8223$. 在 $k=5$ 的情形, 5 个代表点全落在第一主成分上的 σ 的临界值 $\sigma_0(5)=2.6147$, 相应的第一主成分的贡献率为 $C_1=0.8724$. 表 11 列出 $3\leqslant k\leqslant 5$ 相应的结果, 包括当 $\sigma=\sigma_0(k)$ 时代表点的 MSE 值.

现在考虑 $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_\rho$ 的情形, 这时 p 个特征值为 $\lambda_1=1+\rho(p-1)$, $\lambda_2=\cdots=\lambda_p=1-\rho$. 当 $\rho\geqslant 0$ 时, 最大的特征值为 λ_1, 第一主成分的贡献率为 $C_1=\lambda_1/p$. 但是对于 $\rho<0$, $\lambda_1=1+\rho(p-1)$ 可能是负数 ($\rho\leqslant 1/(1+p)$), 这时 $\boldsymbol{\Sigma}_\rho$ 不是正定矩阵, 没有资格作为协方差矩阵. 类似于上面的研究, 对于 $p=2$, k 分别等于 3、4 和 5 的情形, 我们研究当 ρ 增加时, k 个代表点位置的改变. 记 $\rho_0(k)$ 为 ρ 的临界值, 当 $\rho\geqslant\rho_0(k)$ 时, k 个代表点全落在第一主成分上. 例如, $k=3$ 时, 当 $\rho<\rho_0(3)=0.4078$ 时, 3 个代表点构成一个等腰三角形; 但是当 $\rho\geqslant 0.4078$ 时, 3 个代表点全落在第一主成分上, 这时 3 个代表点的 MSE $=0.8600$, 第一主成分的贡献率为 $C_1=(1+\rho_0)/2=(1+0.4078)/2=0.7039$, 详情见表 12.

椭球等高分布代表点的几何结构正在研究中, 例如, $p>4$, $k=3$ 时, 4 个代表点是否分为两对, 分别落在第一和第二主成分上; 不同椭球等高分布子类的主成分的几何结构之间的关系等. 图 5 给出上面讨论的图表示.

表 10 2 和 5 维正态分布代表点的 MSE 值

	$p=2$					$p=5$				
ρ	$k=2$	$k=3$	$k=4$	$k=8$	$k=11$	$k=2$	$k=3$	$k=4$	$k=8$	$k=11$
−0.2	1.2361	0.9124	0.6975	0.3890	0.2954	4.2361	3.7104	3.3037	2.4247	2.1565
−0.1	1.2997	0.9227	0.7167	0.3966	0.3006	4.2997	3.8181	3.4455	2.6392	2.3935
0	1.3634	0.9257	0.7268	0.4024	0.3031	4.3634	3.9252	3.5866	2.7919	2.4624
0.1	1.2997	0.9226	0.7167	0.3966	0.3006	4.1087	3.7460	3.4603	2.7145	2.4145
0.2	1.2361	0.9124	0.6975	0.3890	0.2954	3.8541	3.5295	3.2399	2.5745	2.3066
0.3	1.1724	0.8938	0.6720	0.3799	0.2872	3.5994	3.2182	2.9692	2.3980	2.1405
0.4	1.1087	0.8632	0.6414	0.3680	0.2753	3.3448	2.8942	2.6906	2.1719	1.9481
0.5	1.0451	0.7853	0.6062	0.3419	0.2585	3.0901	2.5702	2.3529	1.9289	1.7239
0.6	0.9814	0.7043	0.5672	0.3107	0.2364	2.8355	2.2462	1.9999	1.6389	1.4735
0.7	0.9177	0.6233	0.5248	0.2748	0.2102	2.5808	1.9222	1.6469	1.3081	1.1839
0.8	0.8541	0.5424	0.4115	0.2312	0.1731	2.3262	1.5983	1.2940	0.9511	0.8624
0.9	0.7904	0.4614	0.3233	0.1667	0.1251	2.0715	1.2743	0.9410	0.5590	0.4889

表 11　二元正态分布 σ 的临界值和对应的 C_1

k	$\sigma_0(k)$	C_1	MSE
3	1.5416	0.7038	1.4520
4	2.1509	0.8223	1.5436
5	2.6147	0.8724	1.5468

表 12　二元正态分布 ρ 的临界值和对应的 C_1

k	$\rho_0(k)$	C_1	MSE
3	0.4078	0.7039	0.8600
4	0.6445	0.8222	0.5487
5	0.7448	0.8724	0.3947

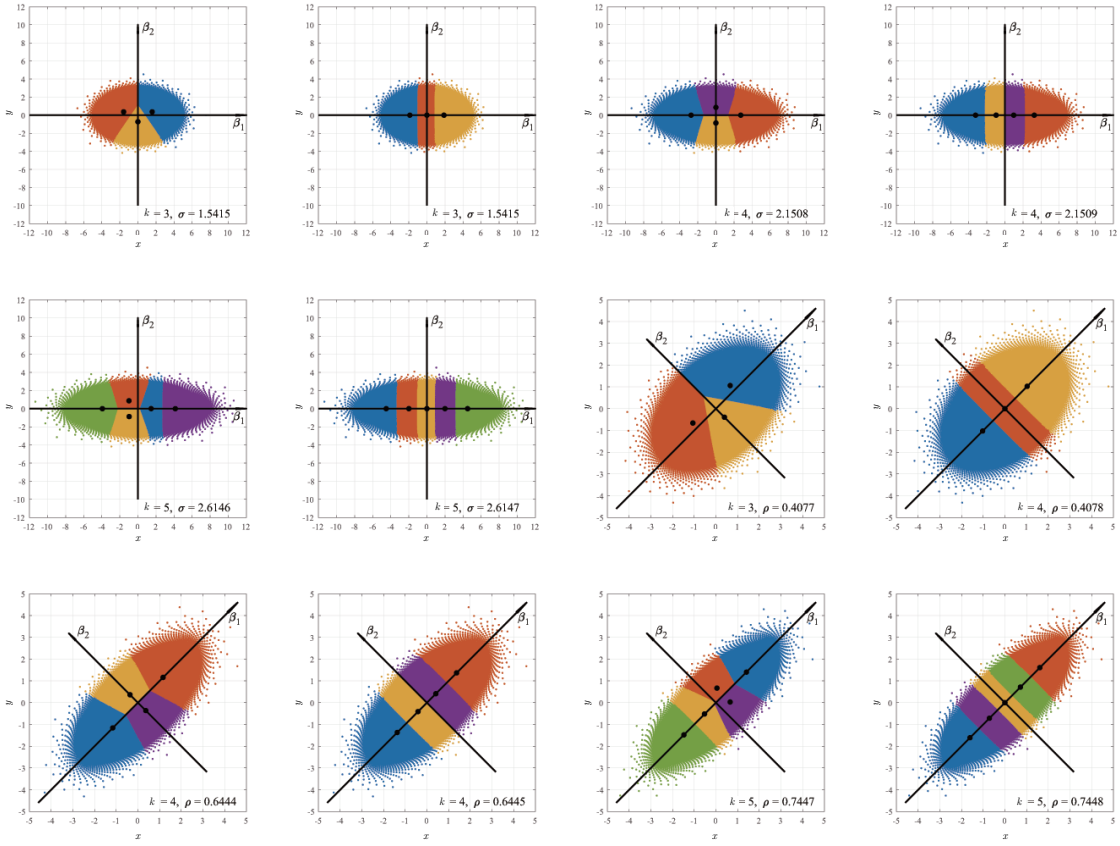

图 5　二元正态分布 σ 和 ρ 临界值对应的代表点集

6　进一步研究

统计分布代表点集的概念是基于信息传送和多元统计应用等多方面的背景而产生的一个研究方向, 已有六十多年历史. 搜索不同统计分布 (一元、多元、对称的、不对称的和混合分布) 的代表点集产生了许多理论研究和算法探讨. 当代表点数 k 比较小时, 对一元和多元统计分布已获得一些理论结

果; 当 k 稍大时, 获得代表点集主要依赖于数值计算. 本文对一元分布发展了一些有效的算法, 并且探讨了解的存在性、唯一性和对称性 (仅对对称分布) 等. 但是寻求多元分布代表点有相当的复杂性, 当前流行的 LBG 算法和 NTLBG 算法以 k- 均值算法为基础, 所以是局部最优的, 不能保证全局最优. 对椭球等高分布, 因为它有清晰的随机表示, NTSR 算法是一个强有力的工具. 利用这些有效的算法, 我们可以方便地估计和研究椭球等高分布的代表点集, 本文是一些初步结果, 更多的正在研究中. 代表点集在统计推断中的应用有很大的潜力, 使用代表点集的思想, 本文改进了传统的自助法, 大大提高了自助法在统计推断中的效力. 代表点集在多元统计推断中的应用到目前为止没有系统研究, 是进一步研究的重要对象.

致谢 本文第一作者方开泰是越民义先生的第一个研究生, 他感谢导师当年的培养和教导, 本文的第二作者贺平是方开泰的博士研究生, 本文的第三作者是贺平的研究生, 四代师生薪火相传. 三位作者祝越民义先生百岁生日快乐, 身体健康!

参考文献

1 Max J. Quantizing for minimum distortion. IEEE Trans Inform Theory, 1960, 6: 7–12
2 Linde Y, Buzo A, Gray R. An algorithm for vector quantizer design. IEEE Trans Commun, 1980, 28: 84–95
3 Graf S, Luschgy H. Foundations of Quantization for Probability Distributions. Berlin-Heidelberg: Springer-Verlag, 2007
4 Fang K T, Zhou M, Wang W J. Applications of the representative points in statistical simulations. Sci China Math, 2014, 57: 2609–2620
5 Korobov N M. The approximation of multiple integrals. Dokl Akad Nauk SSSR, 1959, 124: 1207–1210
6 Hua L K, Wang Y. Applications of Number Theory to Numerical Analysis. Berlin: Springer; Beijing: Science Press, 1981
7 Niederreiter H. Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM, 1992
8 Fang K T, Liu M Q, Qin H, et al. Theory and Application of Uniform Experimental Designs. Beijing: Science Press; New York: Springer, 2018
9 Fang K T, Wang Y. Number-Theoretic Methods in Statistics. London: Chapman & Hall, 1994
10 方开泰, 贺曙东. 在正态总体中如何选择给定数目的代表点的问题. 应用数学学报, 1984, 7: 293–306
11 Cox D R. Note on grouping. J Amer Statist Assoc, 1957, 52: 543–547
12 Bofinger E. Maximizing the correlation of grouped observations. J Amer Statist Assoc, 1970, 65: 1632–1638
13 Flury B A. Principal points. Biometrika, 1990, 77: 33–41
14 傅洪海. 在 Γ 分布的总体中如何选取给定数目的代表点问题. 中国矿业学院学报, 1985, 4: 107–117
15 傅洪海. 威布尔分布总体的最优代表点问题. 无锡轻工业学院学报, 1993, 22: 78–83
16 费荣昌. 在 Pearson 分布族总体中选取代表点的问题. 无锡轻工业学院学报, 1990, 9: 71–78
17 Fang K T, Wang Y. A sequential algorithm for solving a system of nonlinear equations. J Comput Math, 1991, 9: 9–16
18 费荣昌. 在二元正态总体中如何选取代表点的问题. 无锡轻工业学院学报, 1991, 10: 74–85
19 Sharma D K. Design of absolutely optimal quantizers for a wide class of distortion measures. IEEE Trans Inform Theory, 1978, 24: 693–702
20 Fleischer P E. Sufficient conditions for achieving minimum distortion in a quantizer. IEEE Int Conv Rec, 1964, 104–111
21 Fang K T, Yuan K H, Bentler P M. Applications of number-theoretic methods to quantizers of elliptically contoured distributions. Multivariate Analysis and Its Applications. IMS Lecture Notes-Monograph Series. Hayward: Institute of Mathematical Statistics, 1994, 237–251
22 Fang K T, Kotz S, Ng K W. Symmetric Multivariate and Related Distributions. London-New York: Chapman and Hall, 1990
23 Rosenblatt M. Remarks on some nonparametric estimates of a density function. Ann Math Statist, 1956, 27: 832–837
24 Parzen E. On estimation of a probability density function and mode. Ann Math Statist, 1962, 33: 1065–1076
25 Efron B. Bootstrap methods: Another look at the Jackknife. Ann Statist, 1979, 7: 1–26
26 Jiang J J, He P, Fang K T. An interesting property of the arcsine distribution and its applications. Statist Probab Lett, 2015, 105: 88–95
27 周永道, 方开泰. FM- 代表点. 中国科学: 数学, 2019, 49: 1009–1020

28 Tarpey T, Li L, Flury B D. Principal points and self-consistent points of elliptical distributions. Ann Statist, 1995, 23: 103–112

Sets of representative points of statistical distributions and their applications

Kaitai Fang, Ping He & Jun Yang

Abstract An important issue of how to use a discrete distribution to approximate a given continuous statistical distribution has been studied in statistics. Obviously, the support points of this discrete distribution must have a good representative in a certain sense. The set of these support points are called representative points (RPs). There are different considerations for representation. This paper reviews four approaches: Monte Carlo (i.i.d.), revised Monte Carlo, number-theoretic methods, and the mean squared error criterion. The revised Monte Carlo method is new. We compare the performance of resampling by these four methods in density estimation and statistical inference, one of which is a revised bootstrap method. The paper pays more attention to the properties of MSE (mean squared error) representative points and algorithms for their generation. Some new results are obtained, for example, the distribution of MSE, the geometric pattern of RPs of elliptical distributions and relationships between the RPs and principal components.

Keywords **representative points of statistical distributions, quasi-Monte Carlo method, statistical inference, normal distribution, elliptically contoured distribution, principal components and principal points**

MSC(2010) **65C05, 65C50**

doi: 10.1360/SSM-2019-0251