FM- 代表点

周永道^{1*}, 方开泰^{2,3}

1. 南开大学统计研究院, 天津 300071;

2. 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 519085;

3. 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: ydzhou@nankai.edu.cn, ktfang@uic.edu.hk

收稿日期: 2017-08-10; 接受日期: 2018-03-21; 网络出版日期: 2019-04-24; * 通信作者

国家自然科学基金 (批准号: 11471229)、中国科学院数学与系统科学研究院科研基金和珠海市优势学科基金资助项目

摘要 本文提出一种寻找连续性随机变量代表点的新准则: FM- 准则. 该准则在前 $n-1$ 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之间差异的 L_2 - 范数. 本文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.

关键词 F - 偏差 均方误差 矩 代表点

MSC (2010) 主题分类 62E10

1 引言

在实际应用中, 经常要求用一个离散分布来近似某个连续概率分布. 例如, Fang 和 He^[1] 用代表点方法寻找服装的最优尺寸. Flury^[2]、Flury 和 Tarpey^[3] 用主成分点决定瑞士军队的防毒面具的最优尺寸和形状, 以及函数型数据的最优子集等.

给定某连续型随机变量 X 的概率密度函数 $f(x)$ 和相应的分布函数 $F(x)$, 我们寻找一个离散型随机变量 ξ , 使其分布函数 $G(x)$ 能尽可能多地保留 $F(x)$ 的性质. 寻找 ξ 的一个基本想法是, 选择一些代表点, 并对每个点赋以一定的权重. 文献中有许多构造方法, 如 Monte Carlo 方法、括号 - 中位数法^[4]、推广的 Pearson-Tukey 法^[5]、数论方法^[6] 和均方误差法^[1,2,7]. Monte Carlo 方法考虑一个随机样本 $\{\xi_1, \dots, \xi_n\}$, 并且每个点有相同的权重; 我们把相应的代表点记为 MC-RP (Monte Carlo-representative point). 括号 - 中位数法是把连续分布函数 $F(x)$ 分成一些相同的区间, 并对每个区间取其括号 - 中位数为代表点. 推广的 Pearson-Tukey 法选择三个点作为代表点, 即中位数、0.05 和 0.95 分位数, 且其权重分别为 0.63、0.185 和 0.185. 对于一维情形, 数论方法选择 $\{\xi_i = F^{-1}((2i-1)/(2n)), i = 1, \dots, n\}$ 作为 n 个代表点, 其中 $F^{-1}(x)$ 为 $F(x)$ 的逆函数, 记相应的代表点为 NTM-RP (number-theoretic

英文引用格式: Zhou Y D, Fang K T. FM-criterion for representative points (in Chinese). Sci Sin Math, 2019, 49: 1009–1020, doi: 10.1360/SCM-2017-0529

method-representative point). 从某种意义来讲, 数论方法是一维时的括号 - 中位数法的一种特殊情形. 均方误差法是在均方误差意义下, 选择最优的 n 个点使其尽可能好地逼近连续分布.

然而, 这些准则并没有考虑代表点的样本矩与总体分布的总体矩之间的差异. Jiang 等^[8] 证明了反正弦分布的一个有趣的性质, 即由数论方法得到的 n 个点的离散分布 ξ 的前 $n-1$ 阶矩恰好都等于反正弦分布相应的总体矩. 然而, 这个性质并不能推广至任意分布. 根据该性质, 本文提出一种新准则用于选择离散分布 ξ , 即限定 ξ 的前 $n-1$ 阶矩都等于连续分布 X 相应矩的条件下, 最小化 ξ 的分布函数 $F_n(x)$ 与 X 的分布函数 $F(x)$ 之间差异的 L_2 - 范数. 称该新准则为 FM- 准则, 因为其包括修改的 F - 偏差以及矩相等的要求. 在该准则下得到的代表点称为 FM- 代表点. 我们将证明该准则对均匀分布以及其他分布是有意义的. 然而, 该准则的约束条件对有些分布而言太过苛刻. 例如, 对于正态分布而言, 当 $n \leq 4$ 时, 其约束条件能满足, 而当 $n \geq 5$ 时不能满足. 因此, 对于这些情形, 有必要把 FM- 准则推广至伪 FM- 准则, 其要求代表点的样本矩近似于总体矩, 从而减弱了约束要求.

本文余下部分安排如下: 第 2 节给出一些记号、代表点的一些概念以及 FM- 准则的定义. 第 3 节给出在 FM- 准则下的一些分布的最优离散近似 ξ . 第 4 节把 FM- 准则推广至伪 FM- 准则, 并给出标准正态分布的代表点. 第 5 节比较在不同准则下不同类型的代表点的性能. 第 6 节给出结论.

2 FM- 准则

设 $F(\mathbf{x})$ 为 $\mathbb{R}^s$ 中连续随机向量 $\mathbf{X} = (X_1, \dots, X_s)^T$ 的分布函数, $\mathcal{P} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ 为 $\mathbb{R}^s$ 中的点集. 设离散随机变量 ξ 以 $\mathcal{P}$ 为支撑, 并对每点赋以相同的权重 $1/n$. 定义 $\mathcal{P}$ 的 F - 偏差为

$$D_F(\mathcal{P}) = \sup_{\mathbf{x} \in \mathbb{R}^s} |F_n(\mathbf{x}) - F(\mathbf{x})|, \quad (2.1)$$

其中 $F_n(\mathbf{x})$ 为 $\mathcal{P}$ 的分布函数. F - 偏差可以衡量 $\mathcal{P}$ 对于 $F(\mathbf{x})$ 的近似程度. $D_F(\mathcal{P})$ 恰好是对于 $F(\mathbf{x})$ 的拟合优度检验中的 Kolmogorov-Smirnov 统计量. 若分布函数 $F(\mathbf{x})$ 是 s 维单位立方体 $[0, 1]^s$ 上均匀分布的分布函数, 则 F - 偏差变为星偏差这一特殊的均匀性测度.

另一常用于搜索代表点的准则是均方误差准则. 设 s 维随机向量 $\mathbf{X}$ 的密度函数为 $f(\mathbf{x})$, 其支撑为 $D \subset \mathbb{R}^s$. 定义点集 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ 在 D 中关于 $\mathbf{X}$ 的均方误差为

$$\begin{aligned} \text{MSE}(\mathbf{x}_1, \dots, \mathbf{x}_n) &= \mathbb{E} \left(\min_{i=1, \dots, n} \|\mathbf{X} - \mathbf{x}_i\|^2 \right) = \int_D \min_{i=1, \dots, n} \|\mathbf{x} - \mathbf{x}_i\|^2 f(\mathbf{x}) d\mathbf{x} \\ &= \sum_{i=1}^n \int_{D_i} \|\mathbf{x} - \mathbf{x}_i\|^2 f(\mathbf{x}) d\mathbf{x}, \end{aligned} \quad (2.2)$$

其中 $\|\cdot\|$ 为 L_2 - 范数, $D_i = \{\mathbf{x} \in D : \|\mathbf{x} - \mathbf{x}_i\| < \|\mathbf{x} - \mathbf{x}_j\|, i \neq j\}$, 且 $\{D_1, \dots, D_k\}$ 为试验区域 D 的一个剖分. 具有最小的 MSE 值的点集 $\mathcal{P}$ 称为 MSE- 代表点, 记为 MSE-RP (mean square error-representative point). 对于 MSE-RP 中第 i 个点 $\mathbf{x}_i$, 相应的权重为 $p_i = \int_{D_i} f(\mathbf{x}) d\mathbf{x}$, $i = 1, \dots, n$. 不同点的权重可能是不同的.

F - 偏差考虑 $\mathcal{P}$ 的经验分布函数与总体分布函数之间的差异, 而 MSE 准则是基于 X 与 $\mathcal{P}$ 之间的 L_2 - 距离. 这两个准则都没有考虑 X 与 $\mathcal{P}$ 之间矩的差异. 由 X 的矩母函数 $\mathbb{E}(e^{tX})$ 展开的 Maclaurin 序列可知, 随机变量的矩具有很重要的信息. 因此, 我们可以合理地要求 $\mathcal{P}$ 的前 $n-1$ 阶矩分别等于或接近 X 的相应总体矩. Jiang 等^[8] 证明了该要求是可以满足的.

例 2.1 设 X 服从标准反正弦分布, 其密度函数为 $f(x) = 1/(\pi\sqrt{x(1-x)})$ ($0 < x < 1$) 且分布函数为 $F(x) = 2/\pi \arcsin(\sqrt{x})$, $0 < x < 1$. 设离散随机变量 ξ 的支撑为 $\mathcal{P} = \{F^{-1}((2i-1)/(2n)) = \sin^2((2i-1)\pi/(4n)), i = 1, \dots, n\}$, 且每点的权重都等于 $1/n$, 则 ξ 的前 $n-1$ 阶矩都等于 X 的相应矩, 即 $E(\xi^m) = E(X^m) = \prod_{j=1}^m \frac{2j-1}{2j}$, $m = 1, \dots, n-1$.

从例 2.1 可知, 对一些一维情形, 我们可以限制 ξ 的前 $n-1$ 阶矩等于 X 的相应矩. 另外, (2.1) 的 F -偏差可以衡量 ξ 与 X 的分布函数之间的逼近程度. 因此, 我们提出 FM-准则这一新准则来寻找 X 的代表点, 称其为 FM-代表点, 记为 FM-RP (FM-representative point). FM-准则包括对分布以及矩的信息要求. 对于一个 s 维随机变量 $\mathbf{X}$, 设其前 $n-1$ 矩存在, 考虑下面的优化问题:

$$\begin{aligned} \min O_F &= \int_{\mathbb{R}^s} |F_n(\mathbf{x}) - F(\mathbf{x})|^2 d\mathbf{x} \\ \text{s.t. } E(\xi_j^m) &= E(\mathbf{X}_j^m), \quad m = 1, \dots, n-1, \quad j = 1, \dots, s, \end{aligned} \quad (2.3)$$

其中 $F_n(\mathbf{x})$ 和 $F(\mathbf{x})$ 分别是 ξ 和 $\mathbf{X}$ 的分布函数, ξ_j 和 $\mathbf{X}_j$ 分别是 ξ 和 $\mathbf{X}$ 的第 j 个元素. FM-准则的目标函数是 $F_n(\mathbf{x})$ 和 $F(\mathbf{x})$ 之间差异的 L_2 -范数, 其与 (2.1) 的 F -偏差中的 L_∞ -范数相比, 更容易计算. 对于一维随机变量 X , 问题 (2.3) 变为

$$\begin{aligned} \min O_F &= \int_{\mathbb{R}} |F_n(x) - F(x)|^2 dx \\ \text{s.t. } E(\xi^m) &= E(X^m), \quad m = 1, \dots, n-1. \end{aligned} \quad (2.4)$$

在下一节中, 我们将说明很多一维分布存在 FM-RP.

3 常见分布的 FM-RP

本节将考虑一些常见分布的 FM-RP, 如均匀分布和幂分布, 尤其是前者在统计模拟中普遍应用. 为了比较, 我们也给出均匀分布的均方误差代表点 MSE-RP.

首先, 考虑 $[0, 1]$ 上的均匀分布 $X \sim U(0, 1)$ 的 FM-RP. X 的分布为 $F(x) = x$, $0 \leq x \leq 1$, 则 FM-准则的目标函数为

$$O_F = \int_{-\infty}^{\infty} |F_n(x) - F(x)|^2 dx = \int_0^1 |F_n(x) - x|^2 dx = \frac{1}{3} + \frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n^2} \sum_{i=1}^n (2i-1)x_i,$$

其中 $E(X^k) = \int_0^1 x^k dx = 1/(k+1)$. 对等式右边删除固定项 $1/3$ 后, 对剩余两项都乘以 n^2 , $U(0, 1)$ 的 FM-RP 为下面优化问题的解:

$$\begin{aligned} \min n \sum_{i=1}^n x_i^2 - \sum_{i=1}^n (2i-1)x_i \\ \text{s.t. } (x_1, \dots, x_n) \in [0, 1]^n, \quad \frac{1}{n} \sum_{i=1}^n x_i^k = \frac{1}{k+1}, \quad k = 1, \dots, n-1, \end{aligned} \quad (3.1)$$

其中 $(x_1, \dots, x_n) \in [0, 1]^n$ 表示每个分量 $x_i \in [0, 1]$. 记

$$\Phi_{\text{uniform}} = \left\{ \mathbf{x} : \mathbf{x} = (x_1, \dots, x_n) \in [0, 1]^n, \sum_{i=1}^n x_i^k = \frac{n}{k+1}, k = 1, \dots, n-1 \right\}. \quad (3.2)$$

为优化问题 (3.1) 的解, 我们首先证明约束条件 (3.1) 是有意义的, 即 (3.2) 中的集合 Φ_{uniform} 非空.

命题 3.1 均匀分布 $U(0,1)$ 存在 FM-RP, 即集合 Φ_{uniform} 非空.

证明 对每个 k , 记

$$\Phi_k = \left\{ \mathbf{x} : \mathbf{x} \in [0,1]^n, \sum_{i=1}^n x_i^k = \frac{n}{k+1} \right\}, \quad (3.3)$$

则在几何上 Φ_k 为高维球体, 且 $\Phi_{\text{uniform}} = \Phi_1 \cap \Phi_2 \cap \cdots \cap \Phi_{n-1}$. 记 $\mathbf{0} = (0, \dots, 0)$ 为原点. 记原点 $\mathbf{0}$ 和 Φ_k 之间的最小和最大 L_2 - 距离分别为 $d_{k,\min}$ 和 $d_{k,\max}$, 则 $d_{k,\min} = \min_{\mathbf{x} \in \Phi_k} \|\mathbf{x}\|^2$, $d_{k,\max} = \max_{\mathbf{x} \in \Phi_k} \|\mathbf{x}\|^2$, 其中 $\|\cdot\|$ 为 L_2 - 范数.

当 $k = 2$ 时, $d_{k,\min} = d_{k,\max} = n/3$. 当 $k \neq 2$ 时, 为了获得 $d_{k,\min}$, 记 Lagrange 乘子为

$$L = \sum_{i=1}^n x_i^2 + \lambda \left(\sum_{i=1}^n x_i^k - \frac{n}{k+1} \right).$$

令 $\partial L / \partial x_i = 0$ 且 $\partial L / \partial \lambda = 0$, 则

$$\begin{cases} x_i(2 + k\lambda x_i^{k-2}) = 0, & i = 1, \dots, n, \\ \sum_{i=1}^n x_i^k = \frac{n}{k+1}. \end{cases} \quad (3.4)$$

当 $k = 1$ 时, $x_1 = \cdots = x_n = 1/2$ 且 $d_{k,\min} = n/4$. 当 $k \geq 3$ 时, 由 (3.4) 可知, $x_i = 0$ 或 $x_i = (-\frac{2}{k\lambda})^{1/(k-2)}$, $i = 1, \dots, n$. 假设在 $\{x_1, \dots, x_n\}$ 中存在 m 个 $x_i \neq 0$, 记 $x_{i_1} = \cdots = x_{i_m} = (-\frac{2}{k\lambda})^{1/(k-2)}$, 则 $x_{i_{m+1}} = \cdots = x_{i_n} = 0$, 其中 $\{i_1, \dots, i_n\}$ 为 $\{1, \dots, n\}$ 的一个置换. 根据条件 $\sum_{i=1}^n x_i^k = \frac{n}{k+1}$ 可知, $x_{i_1} = \cdots = x_{i_m} = (\frac{n}{m(k+1)})^{1/k}$, $x_{i_{m+1}} = \cdots = x_{i_n} = 0$. 由于 $x_i \in [0,1]$, $m \geq \lceil \frac{n}{k+1} \rceil$ 且 $\sum_{i=1}^n x_i^2 = m^{1-2/k} (\frac{n}{k+1})^{2/k}$, 其为 m 的增函数, 因此, $d_{k,\min} = m_k^{1-2/k} (\frac{n}{k+1})^{2/k}$, 其中 $m_k = \lceil \frac{n}{k+1} \rceil$, 且 $d_{k,\max} = n(\frac{1}{k+1})^{2/k}$, 即对于任意 $\mathbf{x} \in \Phi_k$ 和 $k \geq 3$, 有

$$\|\mathbf{x}\|^2 \in \left[m_k^{1-\frac{2}{k}} \left(\frac{n}{k+1} \right)^{\frac{2}{k}}, n \left(\frac{1}{k+1} \right)^{\frac{2}{k}} \right] \doteq [a_k, b_k]. \quad (3.5)$$

当 $k = 1$ 时, 寻找最大距离 $d_{1,\max}$ 等价于一个目标函数为 $\sum_{i=1}^n x_i^2$ 且约束条件为 $\sum_{i=1}^n x_i = n/2$ ($x_i \in [0,1]$) 的优化问题. 该问题的解为 $x_{i_1} = \cdots = x_{i_r} = 1$, $x_{i_{r+1}} = n/2 - r$, $x_{i_{r+2}} = \cdots = x_{i_n} = 0$, 其中 $r = \lfloor \frac{n}{2} \rfloor$. 则 $d_{1,\max} = \lfloor \frac{n}{2} \rfloor + (\frac{n}{2} - \lfloor \frac{n}{2} \rfloor)^2$, 且对于任意 $\mathbf{x} \in \Phi_1$, 有 $\|\mathbf{x}\|^2 \in [\frac{n}{4}, \lfloor \frac{n}{2} \rfloor + (\frac{n}{2} - \lfloor \frac{n}{2} \rfloor)^2] \doteq [a_1, b_1]$. 易知 $d_{2,\min} = d_{2,\max} = \frac{n}{3} \in [a_k, b_k]$, 其中 $k \neq 2$; 且当 $k \geq 2$ 时, $[a_1, b_1] \cap [a_k, b_k] \neq \emptyset$, 当 $3 \leq k_1 < k_2 \leq n-1$ 时, $[a_{k_1}, b_{k_1}] \subset [a_{k_2}, b_{k_2}]$. 因此, 对于这 $n-1$ 个连续的凸曲面 Φ_k , $k = 1, \dots, n-1$, 至少存在一点 $\mathbf{x} = (x_1, \dots, x_n)$ 使得 $\mathbf{x} \in \Phi_1 \cap \cdots \cap \Phi_{n-1}$. 结论得证. $\square$

通过求解优化问题 (3.1), 我们可以获得均匀分布 $U(0,1)$ 的 n 个点的 FM-RP. 表 1 给出 $n = 2, \dots, 10$ 的 FM-RP. 从中可知, 对于任意 n , FM-RP 关于中心点 0.5 对称, 且当 n 为奇数时, 0.5 为其中一个 FM-RP. 定义 $\varepsilon = \sum_{k=1}^{n-1} |\frac{1}{n} \sum_{i=1}^n x_i^k - \frac{1}{k+1}|$ 为前 $n-1$ 阶样本矩向量和总体矩向量之间差异的 L_1 - 范数. 表 1 的最后一列显示, ε 值非常小, 其不等于 0 的原因在于求解优化问题 (3.1) 时存在搜索精度问题. 本质上, ε 值应该为 0.

由 FM-RP 的定义可知, 一个均匀分布的支撑从 $[0,1]$ 变到 $[0,a]$ ($a > 0$) 时, 相应的 FM-RP 也会存在, 且具有下面的性质.

表 1 均匀分布的 n 个点的 FM-RP

n	FM-RP										ε
2	0.2500	0.7500									0
3	0.1464	0.5000	0.8536								2.19E-13
4	0.1127	0.3709	0.6291	0.8873							1.22E-13
5	0.0838	0.3127	0.5000	0.6873	0.9162						2.91E-14
6	0.0708	0.2561	0.4205	0.5795	0.7439	0.9292					9.77E-15
7	0.0581	0.2352	0.3380	0.5000	0.6620	0.7648	0.9419				5.55E-17
8	0.0514	0.2001	0.3070	0.4302	0.5698	0.6930	0.7999	0.9486			1.39E-16
9	0.0442	0.1995	0.2356	0.4160	0.5000	0.5840	0.7644	0.8005	0.9558		8.33E-17
10	0.0401	0.1708	0.2263	0.3570	0.4615	0.5385	0.6430	0.7737	0.8292	0.9599	1.11E-16

命题 3.2 若 $\{x_1, \dots, x_n\}$ 为均匀分布 $U(0, 1)$ 的 n 个 FM-RP, 则 $\{ax_1, \dots, ax_n\}$ 为均匀分布 $U(0, a)$ 的 n 个 FM-RP, 其中 $a > 0$.

然而对于一般的均匀分布, 并没有这种线性性. 例如, 设 $\{x_1, \dots, x_n\}$ 为均匀分布 $U(0, 1)$ 的 n 个 FM-RP, 则 $\{a + (b - a)x_1, \dots, a + (b - a)x_n\}$ 可能不是均匀分布 $U(a, b)$ 的 n 个 FM-RP. 另一方面, 为了方便比较, 我们也给出均匀分布的 MSE-RP.

命题 3.3 均匀分布 $U(0, 1)$ 的 n 个 MSE-RP 为 $\{\frac{1}{2n}, \frac{3}{2n}, \dots, \frac{2n-1}{2n}\}$, 且每个点的权重都是 $1/n$.

证明 从 (2.2) 的 MSE 准则的定义可知, $U(0, 1)$ 的 n 个点 $\{x_1, \dots, x_n\}$ 的 MSE 值为

$$L = \int_0^{(x_1+x_2)/2} (x-x_1)^2 dx + \sum_{i=2}^{n-1} \int_{(x_{i-1}+x_i)/2}^{(x_i+x_{i+1})/2} (x-x_i)^2 dx + \int_{(x_{n-1}+x_n)/2}^1 (x-x_n)^2 dx. \quad (3.6)$$

令 $\frac{\partial L}{\partial x_1} = 0$, 可得 $\int_0^{(x_1+x_2)/2} 2(x-x_1)dx = 0$, 则 $x_2 = 3x_1$. 类似地, 令 $\frac{\partial L}{\partial x_i} = 0$, 可得

$$x_{i+1} = (2i+1)x_1, \quad i = 2, \dots, n-1. \quad (3.7)$$

令 $\frac{\partial L}{\partial x_n} = 0$, 由 (3.7) 可得

$$4n(n-1)x_1^2 - 2(2n-1)x_1 + 1 = 0,$$

则 $x_1 = 1/(2n)$ 且 $x_i = (2i-1)/(2n)$. 易知, 每个点的权重都等于 $1/n$. 结论得证. $\square$

对于统计模拟中很重要的均匀分布 $U(0, 1)$, MSE-RP 与数论方法得到的代表点完全相同. 然而, 由表 1 知, 当 $n=2$ 时, FM-RP 和 MSE-RP 是相同的; 而当 $n>2$ 时, FM-RP 和 MSE-RP 是不同的.

下面考虑幂分布, 其为参数是 $1/2$ 和 1 的特殊 Beta 分布, 密度函数和分布函数分别为 $f(x) = 1/2x^{-1/2}$ ($0 < x < 1$) 和 $F(x) = x^{1/2}$ ($0 < x < 1$). 对于任意 n 个点 $\{x_1, \dots, x_n\}$, 求 FM-RP 的目标函数为

$$O_F = \int_{-\infty}^{\infty} |F_n(x) - F(x)|^2 dx = \frac{1}{6} + \frac{3}{4n} \sum_{i=1}^n x_i^{3/2} - \frac{1}{n^2} \sum_{i=1}^n (2i-1)x_i, \quad (3.8)$$

且幂分布的第 k 个矩为 $E(X^k) = 1/(2k+1)$, $k \geq 1$. 删除 (3.8) 中的固定项, 可得优化问题

$$\begin{aligned} \min & \frac{3}{4n} \sum_{i=1}^n x_i^{3/2} - \frac{1}{n^2} \sum_{i=1}^n (2i-1)x_i \\ \text{s.t. } & (x_1, \dots, x_n) \in [0, 1]^n, \quad \sum_{i=1}^n x_i^k = \frac{n}{2k+1}, \quad k = 1, \dots, n-1. \end{aligned} \quad (3.9)$$

记

$$\Phi_{\text{power}} = \left\{ \mathbf{x} : \mathbf{x} = (x_1, \dots, x_n) \in [0, 1]^n, \sum_{i=1}^n x_i^k = \frac{n}{2k+1}, k = 1, \dots, n-1 \right\}. \quad (3.10)$$

类似于命题 3.1 的证明过程, 我们也可以证明 (3.9) 中的约束条件是有意义的, 即集合 Φ_{power} 非空. 我们忽略其证明.

命题 3.4 幂分布的 FM-RP 存在, 即 (3.10) 的集合 Φ_{power} 非空.

由命题 3.1 和 3.4 可知, 我们可以得到均匀分布和幂分布等很多分布的 FM-RP, 然而下节将说明对于有些分布而言, 相应的 FM-RP 可能并不存在.

4 伪 FM-RP

对于很多分布而言, (2.4) 的条件可能不满足. 例如, 当 $n \geq 5$ 时, 下面将说明标准正态分布不存在 FM-RP. 对于标准正态分布而言, (2.4) 的约束条件意味着

$$\frac{1}{n} \sum_{i=1}^n x_i^k = c_k = \begin{cases} 0, & k \text{ 为奇数,} \\ (k-1)!!, & k \text{ 为偶数,} \end{cases} \quad k = 1, \dots, n-1, \quad (4.1)$$

其中 $(k-1)!! = (k-1)(k-3)\cdots 3 \cdot 1$. 当 $n = 5$ 时, 该约束条件变为

$$\begin{cases} x_1 + x_2 + x_3 + x_4 + x_5 = 0, \\ x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 = 5, \\ x_1^3 + x_2^3 + x_3^3 + x_4^3 + x_5^3 = 0, \\ x_1^4 + x_2^4 + x_3^4 + x_4^4 + x_5^4 = 15. \end{cases} \quad (4.2)$$

易知, 该方程组并不存在实数解, 即约束 (2.4) 无实数解. 当 $n > 5$ 时, 该约束也没有实数解. 因此, 我们有必要弱化约束 (2.4).

对于 s 维随机变量 $\mathbf{X}$, 可以弱化约束 (2.3) 如下:

$$\begin{aligned} \min \quad & O_F = \int_{\mathbb{R}^s} |F_n(\mathbf{x}) - F(\mathbf{x})|^2 d\mathbf{x} \\ \text{s.t.} \quad & \sum_{j=1}^s \sum_{k=1}^{n-1} |\mathbb{E}(\xi_j^k) - \mathbb{E}(X_j^k)| < \varepsilon, \end{aligned} \quad (4.3)$$

其中 ε 为给定门限值. 问题 (4.3) 是一个具有非线性约束条件的非线性优化问题. 类似地, 若点集 $\mathcal{P} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ 是 (4.3) 的解, 则称该点集为伪 FM 代表点, 记为 QFM-RP (quasi FM-representative point). 显然, 不同的 ε 会导致不同的 QFM-RP. 若 ε 趋于 ∞ , 约束条件相当于无约束条件, 相应的 QFM-RP 变为最小化目标函数 $\int_{\mathbb{R}^s} |F_n(\mathbf{x}) - F(\mathbf{x})|^2 d\mathbf{x}$ 的解. 当 $\varepsilon = 0$ 时, QFM-RP 变为 FM-RP. 由于 ξ 与 $\mathbf{X}$ 之间的矩相差越小, $\mathbf{X}$ 的代表点的性能越好, 我们可能会选择尽可能小的门限 ε 使得约束 (4.3) 有意义. 然而, 我们将说明 ε 尽可能小时, 可能会导致一些不合理现象, 因此, 我们需要考虑另外一个准则用于确定门限 ε .

记 $\varepsilon_{\min}$ 为下面优化问题的解:

$$\min \varepsilon = \sum_{j=1}^s \sum_{k=1}^{n-1} \left| \frac{1}{n} \sum_{i=1}^n x_{ij}^k - E(X_j^k) \right|. \quad (4.4)$$

对于给定的 s 维分布, 并不容易获得 $\varepsilon_{\min}$. 例如, 对于一维分布, 目标函数 (4.4) 变为

$$\min \varepsilon = \sum_{k=1}^{n-1} \left| \frac{1}{n} \sum_{i=1}^n x_i^k - E(X^k) \right|. \quad (4.5)$$

在很多情形下, $\varepsilon_{\min}$ 通常存在正的下界.

例 4.1 考虑标准正态分布 $N(0, 1)$. 根据 $N(0, 1)$ 的密度函数的对称性, 可以假设其 n 个 QFM-RP 也关于原点 0 对称, 即 $x_i = -x_{n+1-i}$, $i = 1, \dots, n$, 且当 n 为奇数时, $x_{(n+1)/2} = 0$. 则对于任意奇数 k , 有

$$\frac{1}{n} \sum_{i=1}^n x_i^k = 0.$$

设 $y_i = x_i^2$, 则 (4.3) 的约束变为

$$\sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \left| \frac{2}{n} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} y_i^k - (2k-1)!! \right| \leq \varepsilon, \quad y_i \geq 0, \quad i = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor, \quad (4.6)$$

其中 $\lfloor \cdot \rfloor$ 表示向下取整. 当 $n = 5$ 时, 约束 (4.6) 变为

$$\left| \frac{2}{5}(y_1 + y_2) - 1 \right| + \left| \frac{2}{5}(y_1^2 + y_2^2) - 3 \right| \leq \varepsilon, \quad y_1, y_2 \geq 0.$$

在约束 $y_1, y_2 \geq 0$ 下, 直线 $(y_1 + y_2) = 5/2$ 与圆 $(y_1^2 + y_2^2) = 15/2$ 的最小距离为 $\sqrt{15/2} - 5/2$. 因此, ε 不能小于 $\sqrt{6/5} - 1$, 即

$$\varepsilon_{\min} \geq \sqrt{\frac{6}{5}} - 1.$$

记 $\phi_\varepsilon = (\varepsilon_1, \dots, \varepsilon_{n-1})$, $\varepsilon_k = |\frac{1}{n} \sum_{i=1}^n x_i^k - E(X^k)|$, $k = 1, \dots, n-1$, 为点集 $\{x_1, \dots, x_n\}$ 关于随机变量 X 的误差向量. 给定 X 的分布, 总体矩 $E(X^k)$ 可相应求出, 然而 $\varepsilon_1, \dots, \varepsilon_{n-1}$ 的取值依赖于点集 $\{x_1, \dots, x_n\}$, 且最优的 n 个点并不容易得到. 该问题可换成下面的方式求解: 给定 n 维实空间 $\mathbb{R}^n$ 中的一个点 $\mathbf{x}^0 = (x_1^0, \dots, x_n^0)$, 分别计算 $\mathbf{x}^0$ 与各曲面 $\frac{1}{n} \sum_{i=1}^n x_i^k = E(X^k)$ ($i = 1, \dots, n-1$) 之间的距离, 则可得相应的 ε_k . 此问题转化为寻找最优的 $\mathbf{x}^0$. 换句话说, 问题 (4.5) 可认为 $\mathbb{R}^n$ 中的一个非线性优化问题.

设一维随机变量 X 的分布函数为 $F(x)$, 一个确定 (4.3) 中的门限值 ε , 以及相应近似解的经验方法如下所示:

(1) 令 $F(x)$ 的分位数 $\mathbf{x}_0 = \{x_{i0} = F^{-1}((2i-1)/(2n)), i = 1, \dots, n\}$ 为初值.

(2) 计算误差

$$\varepsilon_0 = \sum_{k=1}^{n-1} \left| \frac{1}{n} \sum_{i=1}^n x_{i0}^k - E(X^k) \right|.$$

(3) 给定合适的门限 $\varepsilon = \delta \varepsilon_0$, $\delta_0 \leq \delta \leq 1$, 求解问题 (4.3), 使得误差向量 $\phi_\varepsilon = (\varepsilon_1, \dots, \varepsilon_{n-1})$ 在字典序意义下达到最小.

在步骤 (1) 中, Fang 和 Wang^[6] 证明了 n 个点的点集 $\mathbf{x}_0$ 恰好是在 F - 偏差意义下的最优解, 因此易得步骤 (2) 中的 ε_0 . 步骤 (3) 中, 取 δ_0 为使问题 (4.3) 有意义的最小值, 即 $\delta_0 = \varepsilon_{\min}/\varepsilon_0$. 显然, 当 $\delta = 1$ 时, 问题 (4.3) 的解必然存在, 因为 $\mathbf{x}_0$ 就满足其约束条件. 而 δ_0 值可由一些优化方法确定. 例如, Matlab 内嵌函数 “fmincon” 可用于求解优化问题 (4.3), 且能显示解是否存在, 因此可由该函数确定 δ_0 值. 步骤 (3) 中的字典序意味着, 对于最优的 ϕ_ε 和任意误差向量 $\phi_{\varepsilon'} = (\varepsilon'_1, \dots, \varepsilon'_{n-1})$, 存在一个 $k \in \{1, \dots, n-1\}$ 使得 $\varepsilon_1 = \varepsilon'_1, \dots, \varepsilon_{k-1} = \varepsilon'_{k-1}$ 且 $\varepsilon_k < \varepsilon'_k$. 与 X 的代表点的样本高阶矩相比, 我们更愿意其低阶矩逼近相应的总体矩. 该要求与 FM-RP 的要求是一致的, 因为对于 FM-RP 而言, $\phi_\varepsilon = (0, \dots, 0)$, 其恰好达到字典序意义下的最优.

为了考察 δ 值的影响, 我们重新考虑标准正态分布的 $n = 5$ 个代表点. 对于不同的 δ , 得到相应的 QFM-RP, $\varepsilon_1, \dots, \varepsilon_4$, $\varepsilon = \sum_{k=1}^4 \varepsilon_k$, 以及目标函数值 O_F . 由 Matlab 软件可知 $\delta_0 = 0.045$, 即当 $\delta < \delta_0$ 时, 优化问题 (4.3) 无解. 表 2 给出不同 δ 的 QFM-RPs. 易知, 随着 δ 值的增大, 差异值 ε 增大, 而目标函数值 O_F 减少. 因此, 其中存在 ε 与 O_F 之间的一个平衡. 我们选择 $\delta \in [\delta_0, 1]$ 使得误差向量 $\phi_\varepsilon = (\varepsilon_1, \dots, \varepsilon_{n-1})$ 达到字典序意义下的最优. 表 2 显示, 当 $n = 5$ 时, $\delta = 0.4699$, 相应的 $\varepsilon_2 = 0$. 换句话说, 最优的 δ 不等于 δ_0 , 否则会出现不合理现象, 即目标函数值 O_F 太大, 而且 5 个点中有 3 个点等于或非常接近 0.

对于标准正态分布而言, 当 $n = 2$ 时, 对任意的点集 $\{-x, x\}$, $x \in \mathbb{R}$, 都有 $\varepsilon = 0$; 当 $n = 3$ 时, 对任意的点集 $\{(-x - \sqrt{6 - 3x^2})/2, x, (-x + \sqrt{6 - 3x^2})/2\}$, $|x| \leq \sqrt{2}$, 都有 $\varepsilon = 0$, 因为其前两阶样本矩都等于总体矩. 当 $n = 4$ 时, 对任意的点集 $\{x, -x, -\sqrt{2 - x^2}, \sqrt{2 - x^2}\}$, $|x| \leq \sqrt{2}$, 都有 $\varepsilon = 0$. 因此, 当 $n = 2, 3, 4$ 时, 都可以找到点集使得 $\varepsilon = 0$. 此时, 我们自然在这些点集中选择最优的 x 使得目标函数 (4.3) 达到最小, 相应的 QFM-RP 变为 FM-RP. 当 $n = 5$ 时, 例 4.1 说明 FM-RP 不存在, 且当 $n > 5$ 时, FM-RP 也不存在. 为了比较, 我们给出 $n = 3, 4, \dots, 10$ 时的 FM-RP 或 QFM-RP, 以及相应的 NTM-RP 和 MSE-RP. 数论方法代表点 NTM-RP 为 $\Phi^{-1}((i - 0.5)/n)$, $i = 1, \dots, n$, 其中 $\Phi(\cdot)$ 为标准正态分布的分布函数. Fang 等^[9] 给出标准正态分布的 MSE-RP. 由于 QFM-RP、NTM-RP 和 MSE-RP 都是关于原点对称, 我们仅列出 $\lfloor n/2 \rfloor$ 个正的代表点, 具体见表 3. 当 n 为偶数时, 这 n 个代

表 2 标准正态分布 5 个代表点时 δ 值的影响

δ	QFM-RPs				ε_1	ε_2	ε_3	ε_4	ε	O_F	
0.0450	-1.6549	0.0003	0	-0.0003	1.6549	0	0.0954	0	0.0000	0.0954	0.0632
0.1000	-1.6548	-0.1518	0	0.1518	1.6548	0	0.1046	0	0.0000	0.1046	0.0426
0.2000	-1.6529	-0.4364	0	0.4364	1.6529	0	0.1690	0	0.0000	0.1690	0.0233
0.3000	-1.5695	-0.5198	0	0.5198	1.5695	0	0.0934	0	0.5438	0.6372	0.0191
0.4000	-1.5242	-0.5199	0	0.5199	1.5242	0	0.0374	0	0.8121	0.8495	0.0177
0.4699	-1.4918	-0.5240	0	0.5240	1.4918	0	0.0000	0	0.9889	0.9889	0.0168
0.5000	-1.4795	-0.5337	0	0.5337	1.4795	0	0.0106	0	1.0512	1.0618	0.0165
0.6000	-1.4451	-0.5327	0	0.5327	1.4451	0	0.0511	0	1.2232	1.2743	0.0157
0.7000	-1.4086	-0.5314	0	0.5314	1.4086	0	0.0933	0	1.3932	1.4866	0.0150
0.8000	-1.3695	-0.5297	0	0.5297	1.3695	0	0.1376	0	1.5615	1.6991	0.0145
0.9000	-1.3273	-0.5275	0	0.5275	1.3273	0	0.1840	0	1.7274	1.9114	0.0141
1.0000	-1.2816	-0.5244	0	0.5244	1.2816	0	0.2331	0	1.8908	2.1238	0.0139

表 3 标准正态分布的 QFM-RP 以及与 NTM-RP 和 MSE-RP 之间的比较

n	代表点类型					ε_2	ε_4	ε_6	ε_8	δ	O_F	
3	FM-RP	1.224745				0					0.045	
	NTM-RP	0.967422				0.376					0.035	
	MSE-RP	1.224006				0.190					0.045	
4	FM-RP	0.345228	1.371429			0					0.026	
	NTM-RP	0.318639	1.150350			0.288					0.021	
	MSE-RP	0.452780	1.510418			0.243					0.036	
5	QFM-RP	0.524035	1.491773			0	0.989			0.470	0.017	
	NTM-RP	0.524400	1.281552			0.233	1.891				0.014	
	MSE-RP	0.764568	1.724147			0.249	0.223				0.033	
6	QFM-RP	0.209220	0.676358	1.580749		0	0.848			0.472	0.012	
	NTM-RP	0.210428	0.674490	1.382994		0.196	1.711				0.010	
	MSE-RP	0.317716	1.000106	1.893595		0.538	1.467				0.032	
7	QFM-RP	0.365780	0.793642	1.654188		0	0.742	9.074		0.710	0.009	
	NTM-RP	0.366106	0.791639	1.465234		0.169	1.566	12.102			0.008	
	MSE-RP	0.560577	1.188147	2.033369		0.510	1.885	3.612			0.033	
8	QFM-RP	0.157159	0.488714	0.890331	1.715742	0	0.662	8.495		0.695	0.007	
	NTM-RP	0.157311	0.488776	0.887147	1.534121	0.149	1.446	11.616			0.006	
	MSE-RP	0.245094	0.756005	1.343909	2.151946	0.725	2.936	9.659			0.033	
9	QFM-RP	0.282241	0.589621	0.969394	1.770017	0	0.594	7.972	83.414	0.851	0.005	
	NTM-RP	0.282216	0.589456	0.967422	1.593219	0.133	1.345	11.174	95.601		0.005	
	MSE-RP	0.443639	0.918796	1.476392	2.254664	0.687	3.242	13.049	30.505		0.034	
10	QFM-RP	0.125663	0.385330	0.674630	1.038552	1.817144	0	0.541	7.529	80.944	0.839	0.004
	NTM-RP	0.125661	0.385320	0.674490	1.036433	1.644854	0.120	1.259	10.772	94.009		0.004
	MSE-RP	0.199623	0.609858	1.057825	1.591340	2.345096	0.858	4.137	18.928	70.125		0.035

表点为 $\{-x_{\lfloor n/2 \rfloor}, \dots, -x_1, x_1, \dots, x_{\lfloor n/2 \rfloor}\}$; 而当 n 为奇数时, 代表点为 $\{-x_{\lfloor n/2 \rfloor}, \dots, -x_1, 0, x_1, \dots, x_{\lfloor n/2 \rfloor}\}$. 因此, 这三种代表点的奇数阶样本矩都等于 0, 且当 k 为奇数时, $\varepsilon_k = 0$. 表 3 也给出这三种代表点的偶数阶矩相应的误差, 以及目标函数值 O_F . 对每个 n , QFM-RP 的 ε_k 值在字典序意义下都优于 NTM-RP 和 MSE-RP. 例如, 当 $n = 6$ 时, QFM-RP 的 ε_2 等于 0, 而 NTM-RP 和 MSE-RP 的 ε_2 分别等于 0.196 和 0.538; 当 $n = 3, 4$ 时, NTM-RP 的二阶矩并不等于总体二阶矩, 而 QFM-RP 可以克服该缺点. 此外, MSE-RP 的目标函数值 O_F 大于其他两种代表点.

5 仿真

本节比较均匀分布 $U(0, 1)$ 和幂分布相应的 FM-RP、MSE-RP、NTM-RP 和 MC-RP 的性能. FM-RP 是通过求解非线性优化问题 (2.4) 的一些优化算法而得到的, 如内点法等 (参见文献 [10]). 对于 MC-RP, 我们重复 1,001 次随机抽样, 分别估计总体均值得到 1,001 个估计误差, 并选择误差恰好为中位数对应的这 n 个点作为 MC-RP. 对每个分布 n 个点的 FM-RP、NTM-RP 和 MC-RP, 考虑其支撑

的每个点具有相同的权重 $1/n$. 对于 MSE-RP $\{x_1, \dots, x_n\}$ 的每个点 x_i , 其权重为

$$p_i = F\left(\frac{x_{i-1} + x_i}{2}\right) - F\left(\frac{x_i + x_{i+1}}{2}\right),$$

其中 $x_0 = -\infty$, $x_{n+1} = \infty$. 我们给出每个分布的每种代表点的均值、方差和峰度, 并得到与总体分布相应的数字特征之间的偏差.

对于均匀分布 $U(0, 1)$, 总体均值、方差和峰度这三个数字特征分别为 0.5、 $1/12$ 和 1.8. 根据命题 3.3, 均匀分布的 MSE-RP 与 NTM-RP 是一样的. 对于幂分布, 这三个数字特征分别为 $1/3$ 、 $4/45$ 和 $(2/27 - 2/35)/(4/45)^2$, 其 NTM-RP 为 $\{(\frac{i-0.5}{n})^2, i = 1, \dots, n\}$, 而 MSE-RP 由 Fang 和 Wang^[6] 提出的 SNT0 (sequential number-theoretic optimization) 方法而得到. 表 4 给出当 $n = 5, 10, 15, 20, 25$ 时, 这两个分布的 4 种代表点的样本均值、方差和峰度与总体均值、方差和峰度之间的差异. 由于均匀分布的 MSE-RP 与 NTM-RP 相同, 故仅列出 MSE-RP 的结果. 根据 FM-RP 的约束条件, 这三个数字特征的差异应该为 0, 然而由于 FM-RP 是通过优化算法得到的, 其存在计算误差问题, 则这些偏差可能不等于 0. 从表中知, MSE-RP 和 MC-RP 的偏差比 FM-RP 的大.

此外, 对每个分布的每种代表点, 考虑用 bootstrap 方法重复 1,000 次. 对每个 bootstrap 样本, 得到其均值、标准差、峰度以及第 10、25、50、75 和 90 个百分位数与真值之间的差异. 对于均匀分布 $U(0, 1)$, 第 10、25、50、75 和 90 个百分位数的真值分别为 0.1、0.25、0.5、0.75 和 0.9, 而幂分布的第 10、25、50、75 和 90 个百分位数的真值分别为 0.01、0.0625、0.25、0.5625 和 0.81. 表 5 和 6 分别给出 $n = 8, 10, 15, 20, 25$ 时, 均匀分布 $U(0, 1)$ 和幂分布的不同代表点的 bootstrap 样本的这些数字特征值与真值之间的偏差. 从表 5 和 6 可知, 当 $n = 8$ 时, FM-RP 的 bootstrap 样本的性质通常比 MSE-RP 的 bootstrap 样本的更好; 当 n 更大时, 这两类代表点的性质差不多. FM-RP 比 MC-RP 更优, 因为 FM-RP 通常有更小的偏差.

表 4 均匀分布 $U(0, 1)$ 和幂分布不同代表点的均值、方差和峰度与总体相应数字特征之间的偏差

n	数字特征	均匀分布 $U(0, 1)$			幂分布			
		FM	MSE	MC	FM	MSE	NTM	MC
5	均值	0	0	0.0056	0	-0.0068	-0.0033	0.1721
	方差	0	-0.0033	-0.0355	0	-0.0030	-0.0044	-0.0249
	峰度	0	-0.1000	-0.5085	0	-0.1727	-0.2489	-0.8520
10	均值	0	0	-0.0022	-4.8E-05	-0.0027	-0.0008	0.1626
	方差	0	-0.0008	0.0456	0.0002	-0.0011	-0.0011	-0.0355
	峰度	0	-0.0242	-0.4189	0.0030	-0.0441	-0.0638	0.3081
15	均值	-5.9E-06	0	-0.0002	-8.2E-05	-0.0013	-0.0004	0.1704
	方差	-4.4E-06	-0.0004	0.0092	0.0004	-0.0004	-0.0005	0.0075
	峰度	-0.0039	-0.0107	-0.1194	0.0072	-0.0218	-0.0285	-0.5714
20	均值	-1.7E-05	0	-0.0014	2E-05	-0.0007	-0.0002	0.1656
	方差	2.7E-05	-0.0002	0.0048	9E-05	-0.0003	-0.0003	0.0030
	峰度	-0.0002	-0.0060	-0.2464	0.0044	-0.0132	-0.0160	-0.4662
25	均值	-2.6E-05	0	0.0023	2.9E-05	-0.0001	-0.0001	0.1615
	方差	5.3E-05	-0.0001	-0.0095	-4.5E-05	-4.1E-05	-0.0002	-0.0025
	峰度	0.0009	-0.0038	-0.0276	0.0030	0.0025	-0.0103	-0.2969

表 5 均匀分布的不同代表点的 bootstrap 样本的数字特征与真值之间的偏差

n	代表点	数字特征			百分位数				
		均值	标准差	峰度	10	25	50	75	90
8	FM	0.0004	-0.0041	0.1789	0.0487	0.0246	0.0009	-0.0263	-0.0461
	MSE	-0.0026	-0.0070	0.2016	0.0498	0.0267	-0.0046	-0.0313	-0.0518
	MC	-0.0015	0.0274	0.2675	-0.0017	-0.0007	0.0206	-0.0084	-0.0318
10	FM	0.0009	-0.0047	0.1823	0.0415	0.0217	0.0013	-0.0238	-0.0404
	MSE	-0.0025	-0.0048	0.1764	0.0377	0.0170	-0.0023	-0.0274	-0.0420
	MC	-0.0025	-0.0668	0.2479	0.1081	0.0837	-0.0005	-0.0722	-0.1287
15	FM	-0.0004	-0.0026	0.1465	0.0239	0.0169	0.0044	-0.0201	-0.0297
	MSE	0.0025	-0.0014	0.1342	0.0255	0.0156	0.0042	-0.0100	-0.0199
	MC	0.0020	0.0267	-0.1174	0.0180	-0.0513	0.0255	0.0176	-0.0127
20	FM	-0.0028	-0.0014	0.1187	0.0182	0.0100	-0.0017	-0.0167	-0.0239
	MSE	0.0018	-0.0015	0.1161	0.0215	0.0138	0.0010	-0.0081	-0.0167
	MC	-0.0065	-0.0056	0.1997	-0.0011	-0.0145	0.0519	-0.0549	-0.0707
25	FM	0.0017	-0.0001	0.1088	0.0121	0.0128	0.0067	-0.0084	-0.0169
	MSE	0.0037	-0.0015	0.0920	0.0182	0.0133	0.0028	-0.0060	-0.0116
	MC	-0.0010	-0.0063	0.0728	0.0263	0.0163	0.0070	-0.0173	-0.0240

表 6 幂分布的不同代表点的 bootstrap 样本的数字特征与真值之间的偏差

n	代表点	数字特征			百分位数				
		均值	标准差	峰度	10	25	50	75	90
8	FM	0.0058	-0.0047	0.0411	0.0412	0.0417	0.0358	-0.0118	-0.0698
	MSE	-0.0113	-0.0114	0.0900	0.0235	0.0308	0.0137	-0.0308	-0.0831
	NTM	0.0006	-0.0096	0.0066	0.0275	0.0411	0.0288	-0.0132	-0.0724
	MC	0.0014	-0.0086	-0.0599	0.0165	0.0366	0.0530	-0.0142	-0.0898
10	FM	-0.0011	-0.0063	0.1289	0.0266	0.0242	0.0237	-0.0195	-0.0615
	MSE	-0.0118	-0.0093	0.1772	0.0184	0.0218	0.0054	-0.0311	-0.0697
	NTM	0.0007	-0.0070	0.0838	0.0225	0.0315	0.0225	-0.0126	-0.0574
	MC	-0.0037	-0.0137	-0.2925	0.0094	0.0169	0.0479	0.0024	-0.1085
15	FM	0.0027	-0.0027	0.1166	0.0073	0.0231	0.0233	-0.0106	-0.0342
	MSE	0.0011	-0.0043	0.1460	0.0130	0.0221	0.0163	-0.0092	-0.0379
	NTM	-0.0028	-0.0057	0.1746	0.0124	0.0202	0.0101	-0.0172	-0.0431
	MC	-0.0062	0.0223	0.4904	0.0165	0.0184	-0.0395	-0.0077	0.0267
20	FM	0.0004	-0.0039	0.1570	0.0049	0.0199	0.0152	-0.0099	-0.0324
	MSE	0.0003	-0.0014	0.0927	0.0087	0.0137	0.0124	-0.0065	-0.0259
	NTM	0.0018	-0.0026	0.1162	0.0104	0.0163	0.0154	-0.0037	-0.0283
	MC	-0.0001	-0.0076	-0.0083	0.0117	0.0286	0.0055	-0.0037	-0.0421
25	FM	-0.0018	-0.0033	0.1458	0.0045	0.0146	0.0098	-0.0127	-0.0261
	MSE	-0.0014	-0.0029	0.1183	0.0069	0.0121	0.0085	-0.0094	-0.0252
	NTM	-0.0018	-0.0039	0.1106	0.0071	0.0125	0.0089	-0.0087	-0.0273
	MC	-0.0031	-0.0038	0.2209	0.0225	0.0405	-0.0291	-0.0067	-0.0080

6 结论

本文提出的 FM 准则用于搜索给定分布的代表点, 该准则考虑在前 $n - 1$ 阶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和总体分布函数之间的 L_2 - 距离. 这种约束可以用来搜索比其他代表点类型更优的代表点, 因为这种约束可使代表点更加贴近真实的分布. 我们证明了 FM- 准则对于很多分布是有意义的, 如均匀分布和幂分布. 对于不能满足约束条件的分布, 我们把该准则推广至伪 FM 准则, 并获得相应的伪 FM 代表点 QFM-RP. 这种代表点能很好地刻画真实的分布函数.

致谢 感谢主编和审稿人对该文提出的宝贵意见, 使文章得以提高.

参考文献

- 1 Fang K T, He S. The problem of selecting a given number of representative points in a normal population and a generalized mill's ratio. Technical Report. Palo Alto: Department of Statistics, Stanford University, 1982
- 2 Flury B. Principal points. *Biometrika*, 1990, 77: 33–41
- 3 Flury B, Tarpey T. Representing a large collection of curves: A case for principal points. *Amer Statist*, 1993, 47: 304–306
- 4 Clemen R C. Making Hard Decision: An Introduction to Decision Analysis, 2nd ed. New York: Duxbury Press, 1997
- 5 Keefer D L, Bodily S E. Three-point approximations for continuous random variables. *Manage Sci*, 1983, 29: 595–609
- 6 Fang K T, Wang Y. Number-Theoretic Methods in Statistics. London: Chapman and Hall, 1994
- 7 Cox D R. Note on grouping. *J Amer Statist Assoc*, 1957, 52: 543–547
- 8 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
- 9 Fang K T, Zhou M, Wang W J. Applications of the representative points in statistical simulations. *Sci China Math*, 2014, 57: 2609–2620
- 10 Boyd S, Vandenberghe L. Convex Optimization. Cambridge: Cambridge University Press, 2004

FM-criterion for representative points

Yongdao Zhou & Kaitai Fang

Abstract In this paper, a new criterion, FM-criterion, is proposed to choose representative points of a given distribution. Such a criterion is to minimize the L_2 -norm of the difference between the empirical distribution and the given distribution under the constraint that the first $n - 1$ sample moments equal the population moments. It is shown that the FM-criterion is meaningful for many distributions. This criterion is also extended to quasi-FM-criterion if the constraint is not satisfied. The empirical study shows that the corresponding representative points are better than other types of representative points.

Keywords F -discrepancy, mean square error, moments, representative points

MSC(2010) 62E10

doi: 10.1360/SCM-2017-0529