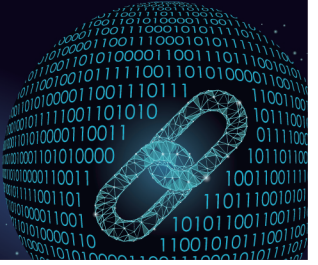




中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION



基于改进整群抽样的基效应仿真因子筛选及应用

施文, 陈奥

引用本文:

施文, 陈奥. 基于改进整群抽样的基效应仿真因子筛选及应用[J]. 控制与决策, 2024, 39(8): 2728–2736.

在线阅读 View online: <https://doi.org/10.13195/j.kzyjc.2023.0028>

您可能感兴趣的其他文章

Articles you may be interested in

基于复杂昂贵仿真的体系效能多目标优化

Complex and expensive simulation based multi-objective optimization to system-of-system effectiveness

控制与决策. 2021, 36(3): 589–598 <https://doi.org/10.13195/j.kzyjc.2019.0844>

基于16方向24邻域改进蚁群算法的移动机器人路径规划

Mobile robots path planning based on 16-directions 24-neighborhoods improved ant colony algorithm

控制与决策. 2021, 36(5): 1137–1146 <https://doi.org/10.13195/j.kzyjc.2019.0600>

一种基于改进指令整形的四旋翼吊装负载摆动抑制

Load swing suppression of quadrotor slung system based on improved command shaping

控制与决策. 2021, 36(8): 1947–1954 <https://doi.org/10.13195/j.kzyjc.2019.1733>

基于PID参数整定的线性自抗扰控制参数整定

Tuning of linear active disturbance rejection controllers based on PID tuning rules

控制与决策. 2021, 36(7): 1592–1600 <https://doi.org/10.13195/j.kzyjc.2019.1408>

基于输出协同调制的四端输出矩阵变换器开关组合优化方法

Switch combination optimization method of four-terminal output matrix converter based on output co-modulation

控制与决策. 2021, 36(12): 2991–2998 <https://doi.org/10.13195/j.kzyjc.2020.0234>

基于改进整群抽样的基效应仿真因子筛选及应用

施文, 陈奥[†]

(中南大学 商学院, 长沙 410083)

摘要: 大数据背景下, 仿真模型通常有许多因子, 仿真筛选实验(screening)就是识别出其中对响应(仿真输出或系统绩效)起最重要作用的少部分因子(亦称仿真输入或变量)的重要方法. 目前常用的筛选方法有序贯分支法(SB)与基效应法(EE). 相较于SB方法, EE以其不假设具体的仿真输入/输出数学关系(model-free)的优势在近年来不断发展并被应用于诸多领域, 然而其劣势在于计算效率. 为提高EE仿真模型的计算效率, 提出一种改进的更具一般性的整群抽样方法(简称ECS). 相较于现有整群抽样方法, ECS通过拆分矩阵的方式自动构造抽样矩阵, 利用该矩阵能够为每个因子生成数量相同且满足目标的基效应, 节省大量的仿真预算. 蒙特卡罗仿真实验表明, ECS在不损失统计效力的基础上可大大提高计算效率, 充分验证该方法的有效性.

关键词: Morris方法; 基效应; 整群抽样; 仿真实验; 因子筛选

中图分类号: TP391.9

文献标志码: A

DOI: 10.13195/j.kzyjc.2023.0028

引用格式: 施文, 陈奥. 基于改进整群抽样的基效应仿真因子筛选及应用[J]. 控制与决策, 2024, 39(8): 2728-2736.

An efficient cluster sampling generation method and its application

SHI Wen, CHEN Ao[†]

(School of Business, Central South University, Changsha 410083, China)

Abstract: In the context of big data, simulation models often involve multiple factors. The purpose of simulation screening experiments is to identify the small subset of factors (also referred to as simulation inputs or variables) that have the most significant impact on the response (simulation output or system performance). Currently, two commonly used screening methods are sequential bifurcation (SB) and elementary effects (EE). In recent years, the EE has gained traction in various fields due to its advantage of not assuming specific mathematical relationships between simulation inputs and outputs (i.e., it is model-free). However, it suffers from computational inefficiency. To address this issue and improve the computational efficiency of the EE, this paper proposes an enhanced and more versatile method called enhanced cluster sampling (ECS). Unlike existing cluster sampling methods, the ECS automatically constructs a sampling matrix by decomposing the matrix, enabling the generation of an equal number of target elementary effects for each factor. This approach significantly saves simulation budget. Monte Carlo simulation experiments demonstrate that the ECS greatly enhances computational efficiency without compromising statistical effectiveness, providing compelling evidence for the effectiveness of the proposed method.

Keywords: Morris method; elementary effect; cluster sampling; simulation experiments; factor screening

0 引言

随着信息技术的发展, 大数据逐渐成为信息化时代的特征. 大数据的广泛运用, 为海量数据的收集、存储提供了强有力的支撑. 近年来大数据更是逐渐渗透到管理科学研究的各个子领域^[1], 并应用到社会网络环境^[2]、交通^[3]、物流^[4]等多个领域. 以数据为中心、以计算为手段的决策分析与管理的范式, 正在管理的各个范畴中取代原有范式^[5-6], 仿真正是承

接这一范式, 帮助挖掘和捕捉海量数据背后信息的重要研究方法^[7], 并且随着数字孪生技术(物理实体与虚构仿真的实时互联)的应用和普及, 仿真在复杂管理决策中扮演着愈发重要的角色. 在大数据和数字孪生背景下, 仿真模型通常包含海量的实验因子(统计学术语, 也称为仿真输入、变量或参数), 而统计学中的稀疏效应定理和管理学的帕累托法则(80/20原理)都认为系统中一般只有少量因子对响应(也称为

收稿日期: 2023-01-08; 录用日期: 2023-07-04.

基金项目: 国家自然科学基金重大项目(72293574); 国家自然科学基金面上项目(71971219); 2022年湘江实验室项目(22XJ03025); 湖南省杰出青年基金项目(2022JJ10084).

责任编辑: 唐加福.

[†]通讯作者. E-mail: aochenao2001@163.com.

仿真输出或系统绩效)起到关键的作用^[8]. 因此, 识别出对响应有重要影响的因子并制定出有针对性且高效的管理决策具有重要意义^[9]. 然而, 要确定因子对模型输出的影响是具有挑战性的问题^[10]. 目前, 筛选实验(factor screening)被认为是解决该问题的重要理论方法^[11].

仿真实验领域主要有序贯分支法(sequential bifurcation, SB)和 Morris 的基效应法(elementary effect, EE)两种仿真筛选实验方法^[12]. 尽管在满足假设条件的前提下, SB 被认为是目前最具仿真实验效率(较少的仿真次数)和效力(控制两类误差)的仿真筛选实验方法^[13-14], 但并非所有仿真模型都能满足 SB 的某些假设条件^[15], 例如遗传效应不能保证对每个模型都成立^[16]. 相较于 SB, EE 不受假设条件的限制并且更加适应大数据背景下更多更广的因子水平^[15], 在近年来不断发展并被应用于众多领域. Morris^[17] 最早提出了针对确定性模型的 EE 方法, 通过基效应衡量各因子对模型输出的影响; Campolongo 等^[10]、Cropp 等^[18] 研究了交互效应的筛选问题; Sin 等^[19] 建议在使用 Morris 方法时使用基效应的非维度测量, 以确保对复杂模型中的重要因子进行可靠的筛选; 施文等^[15] 提出了基于基效应估计 Bootstrap 假设检验以及伪发现率误差控制的“后仿真分析”筛选实验设计方法, 并将 EE 方法拓展到随机模拟环境中开发了一种高效的基于 Morris 方法的框架(EMM)^[20]; Boukouvalas 等^[21] 提出了一种序贯方法增强传统 EE 方法抽样设计的计算性能, 从而识别具有重要非线性效应的因子; Fédou 等^[22] 研究了一种快速混合效应的 EE 筛选方法, 从而估计因子的交互效应. 已有研究大多聚焦于提高 EE 方法的计算性能以及对因子交互效应和非线性效应的识别能力, 对于提高计算效率的研究较少. 随着理论方面研究的不断深入, EE 方法在解决实际问题中的应用也逐步向着更广、更深的方向发展. Chang 等^[23] 利用 EE 对澳大利亚新冠疫情传播与控制模型中的因子进行敏感度分析; Diepens 等^[24] 提出了将其应用到微塑料和疏水性有机化学物质在食物网中的积累的模型中, 从而评估食物网暴露于微塑料和复杂有机混合物的潜在风险; 之后, Koks 等^[25] 将 Morris 方法应用到公路和铁路基础设施资产的全球多灾种风险分析中; 最近, Lu 等^[26] 在其流行病学建模的全局敏感度分析中应用 EE 方法进行了进一步的交互分析.

EE 方法的核心是其抽样(sampling)机制, 利用抽样矩阵(sampling matrix)获取因子组合, 进而计算出

不同因子组合下各因子的基效应, 最终实现对每个因子的主效应与交互效应的判断^[17]. 为提高 EE 方法的计算效率, Morris^[17] 提出利用整群抽样(cluster sampling)的方法生成抽样矩阵. 目前仅有 Shi 等^[20, 27] 研究了整群抽样方法. Shi 等^[20] 提出了一种整群抽样程序, 通过一个抽样矩阵为每个因子生成了指定数量的基效应; 之后, Shi 等^[27] 揭示了非整群抽样和整群抽样两种抽样方法的本质. 然而, Shi 等^[20, 27] 提出的整群抽样方法仍需仿真实验者手动选择行向量且不能保证始终为每个因子生成相同数量的基效应, 计算效率较低. 此外, Campolongo 等^[28] 通过改进 EE 的数学定义以及抽样策略降低仿真实验成本.

由上述文献可以看出, EE 方法的理论发展已取得一定成果并为解决多种实际问题提供了帮助. 然而, 提高方法的计算效率这一课题却鲜为学者关注, 整群抽样方法作为提高 EE 计算效率的重要方法始终是一个悬而未决的问题. 为此, 本文综合已有研究并提出一种改进的且具有一般性的整群抽样方法(enhanced cluster sampling, ECS). 与已有的整群抽样方法相比, ECS 完全不依赖于人工操作, 可实现通过一个整群抽样矩阵为每个因子生成相同数量的基效应, 且与给定目标之间的差距很小. 最后通过数值实验验证 ECS 的有效性, 蒙特卡罗仿真实验表明, ECS 方法在不损失统计效力的基础上可实现计算效率的提升.

1 EE

本文为 Morris 的基效应法(elementary effect, EE)中用到的整群抽样提出一种更具一般性的改进方法, 故先介绍 EE 的基本概念和步骤.

1.1 基本设定

假定仿真模型中有 k 个因子且各因子规范化后取值范围均介于 $[0, 1]$, 从而构成一个 k 维单位实验空间 $\Omega = [0, 1]^k$. 假设每一因子仅在 $[0, 1]$ 间的 p 个离散水平变化, 则实验空间 Ω 为一个 k 维 p 层的网, 其中 p 值由仿真实验者事先设定. 设 $y(\mathbf{x})$ 为运行仿真模型得到的输出, 对于实验空间 Ω 中的因子组合 $\mathbf{x} = (x_1, x_2, \dots, x_j, \dots, x_k)$, 因子 j 的基效应定义为

$$d_j(\mathbf{x}) = \frac{y(\mathbf{x} + \mathbf{e}_j \Delta) - y(\mathbf{x})}{\Delta}, \quad j = 1, 2, \dots, k. \quad (1)$$

其中: $\mathbf{e}_j$ 为 j 方向上的 k 维单位向量; Δ 为实验者事先设定的参数, 其值一般为 $1/(p-1)$ 的倍数, 通常情况下, $\Delta = p/[2(p-1)]$; $\mathbf{x}$ 为从实验空间 Ω 中随机抽样产生的因子组合, 且 $(\mathbf{x} + \mathbf{e}_j \Delta) \in \Omega$, 即 $x_j \leq 1 - \Delta$, 从而 x_j 在 $[0, 1/(p-1), \dots, 1 - \Delta]$ 中取值, 因此因子 j 能

取的水平数为 $p - \Delta(p - 1)$, 例如当 $\Delta = 1/(p - 1)$ 时, 因子 j 共可取 $p - 1$ 个水平 (即不包括 $x_j = 1$ 的水平), 其余 $k - 1$ 个因子在 $[0, 1/(p - 1), 2/(p - 1), \dots, 1]$ 中取值, 共 p 个水平, 于是这些因子共可组成 $p^{k-1}[p - \Delta(p - 1)]$ 个因子组合, 即因子 j 共有 $p^{k-1}[p - \Delta(p - 1)]$ 个基效应, 这些基效应的分布由 F_j 表示.

Morris^[17] 认为, 若 F_j 的集中趋势很明显, 则表明因子 j 对响应具有很强的主效应; 而离散程度很高则意味着因子 j 对响应的影响依赖于其他因子的值, 即因子 j 具有很强的交互效应. 因此可以根据 F_j 的总体期望和标准差 μ_j 和 σ_j 将因子划分为下述 3 类: 1) 不重要的因子; 2) 有主效应的因子; 3) 有交互 (或非线性) 效应的因子. μ_j 和 σ_j 的一个无偏估计量分别为

$$\hat{\mu}_j \triangleq \bar{d}_j = \frac{\sum_{i=1}^N d_{j,i}}{N}, \quad j = 1, 2, \dots, k; \quad (2)$$

$$\hat{\sigma}_j \triangleq S_j = \sqrt{\frac{\sum_{i=1}^N (d_{j,i} - \bar{d}_j)^2}{N - 1}}, \quad j = 1, 2, \dots, k, \quad (3)$$

其中 $d_{j,i}$ 为因子 j 的第 i 个基效应值.

1.2 非整群抽样

EE 的核心在于获得因子的基效应, 从而判断各因子所属的类别. 根据基效应的定义 (即式 (1)), 获得因子 j 的一个基效应需要两个因子组合 (即两次仿真运行), 而为了判断该因子所属的类别, 需要利用其 N 个基效应来估计 $\hat{\mu}_j$ 和 $\hat{\sigma}_j$. 进一步地, 为了判断仿真模型中的 k 个因子所属的类别, 共需 $2Nk$ 个因子组合. 显然, 当 k 较大时, 应用 EE 方法所需的仿真运行次数也将十分庞大. 为了减少运行次数 (即提高计算效率), Morris^[17] 提出利用抽样矩阵获得各因子的基效应并将抽样矩阵的经济性定义为抽样矩阵为所有因子生成的基效应数量与因子组合的数量 (即矩阵中的行) 的比率. 在此基础上, Morris 提出非整群抽样与整群抽样两种构建抽样矩阵的方法.

$$B = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}. \quad (4)$$

非整群抽样方法首先构造一个 $(k + 1) \times k$ 的 0-1 矩阵, 其第 1 行为 0 向量, 后 k 行为一个主对角线下元素全为 1 的下三角矩阵, 见式 (4), 则 ΔB 中相邻的两行可生成一个基效应. 例如 $d_1(0, 0, \dots, 0) = [y(\Delta, 0, \dots, 0) - y(0, 0, \dots, 0)]/\Delta$. 再将 B 的 N 个独

立随机形式串联起来形成矩阵 $(B_1^+; B_2^+; \dots; B_N^+)$, 则可以为每个因子生成 N 个基效应, 其中 $B_i^+ (i = 1, 2, \dots, N)$ 为矩阵 B 的随机形式. B 可以通过一个复杂的公式转换为 B_i^+ , 详见文献 [17].

1.3 整群抽样

为了提高计算效率, Morris 建议采用整群抽样矩阵 C . 与非整群抽样矩阵 B 相比, C 的主要优点是可以通过简单地加上一些额外的行到 B , 就可以为每个因子生成更多基效应 (即 C 的经济性高于 B). 例如

$$C = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

显然, 当 $k = 3$ 时, 该 C 矩阵只比式 (4) 中的 B 矩阵多 2 行就可以为每个因子额外生成 1 个基效应. 因此, C 矩阵的经济性 $1(6/6)$ 比 B 矩阵的经济性 $3/4$ 更高. 设 n 是 C 为每个因子生成的基效应数量, 则一个整群抽样矩阵 C 至多可以为每个因子生成 $n = 2^{k-1}$ 个基效应^[20]. 因此当 k 很大时, 使用一个 C 矩阵足够为每个因子生成 N 个基效应 ($N \ll 2^{k-1}$). 不过, 尽管使用一个 C 矩阵可以实现为每个因子生成足够数量的基效应, 从而达到较高的计算效率, 但与使用非整群抽样矩阵 $(B_1^+; B_2^+; \dots; B_N^+)$ 相比, 使用整群抽样矩阵 C 将导致统计效力的损失. 为了实现计算效率与统计效力的平衡, Shi 等^[20] 提出将可以为每个因子生成 n 个基效应的 $C_{(n)}$ 矩阵的 w 个随机形式 $C_{(n);w^*} (w^* = 1, 2, \dots, w)$ 串联为 $(C_{(n);1}^+; C_{(n);2}^+; \dots; C_{(n);w}^+)$, 从而为每个因子生成 $N = n \times w$ 个基效应. 对于这种整群抽样的情形, Morris 建议使用如下无偏估计量估计 $F_j (j = 1, 2, \dots, k)$ 的均值和方差:

$$\hat{\mu}_j = \frac{1}{wn} \sum_{w^*=1}^w \sum_{i=1}^n d_{j;w^*,i}, \quad (5)$$

$$\hat{\sigma}_j^2 = \frac{(p^k k! - 1)s_{j,b}^2 + p^k k!(n - 1)s_{j,w}^2}{np^k k! - 1}. \quad (6)$$

其中: $d_{j;w^*,i}$ 是 $C_{(n);w^*}$ 矩阵为因子 j 生成的第 i 个基效应, $i = 1, 2, \dots, n, w^* = 1, 2, \dots, w; p^k k!$ 为一个整群抽样矩阵 C 可以采用的随机形式的总数; $s_{j,b}^2$ 和 $s_{j,w}^2$ 分别为群间和群内均方, 即

$$s_{j,b}^2 = \frac{n \sum_{w^*=1}^w (\bar{d}_{j;w^*} - \bar{\bar{d}}_j)^2}{w - 1}, \quad (7)$$

$$s_{j,w}^2 = \frac{\sum_{w^*=1}^w \sum_{i=1}^n (d_{j;w^*,i} - \bar{d}_{j;w^*})^2}{w(n - 1)}. \quad (8)$$

这里: $\bar{d}_j = (nw)^{-1} \sum_{w^*=1}^w \sum_{i=1}^n d_{j;w^*;i}$, $\bar{d}_{j;w^*} = n^{-1} \cdot \sum_{i=1}^n d_{j;w^*;i}$.

使用整群抽样方法的难点在于: 不同于始终处于固定形式的非整群抽样矩阵, 整群抽样矩阵往往会随着仿真实验者设定的目标基效应数量变化而不断变化. 如何快速、精准地构造满足仿真实验者需求的整群抽样矩阵是一个具有挑战性的问题. Morris^[17]曾提出了构造整群抽样矩阵 C 的方法, 但该方法生成的抽样矩阵经济性较低且部分参数依赖于手动选取. Shi 等^[20]提出将包含所有因子组合的整群抽样矩阵 C 划分为下式:

$$C = \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_q \\ \vdots \\ C_k \end{bmatrix} \begin{matrix} \#1 & \#2 & \#3 & \#4 & \cdots & \#k-1 & \#k \\ \hline 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ \hline 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 1 & 1 & 1 & 1 & \cdots & 1 & 1 \end{matrix}. \quad (9)$$

其中: C_q 表示由 C 中所有包含 q 个 1 和 $k-q$ 个 0 的行向量组成的块矩阵. 基于对这些分块矩阵的研究, Shi 等^[20]总结出以下结论: 要利用块矩阵 C_q 为因子生成基效应, 则抽样矩阵 C 中必须包含块矩阵 C_{q-1} ; 并在该结论基础上提出了构造整群抽样矩阵 C 的方法.

尽管在 Morris 的方法上有所改进, 但 Shi 等提出的整群抽样方法中部分步骤仍然依赖于实验者的手动操作, 导致其不能快速且准确地构造抽样矩阵.

2 改进的整群抽样方法

尽管已有学者研究了 EE 中的整群抽样方法, 但现有整群抽样方法都在一定程度上依赖于仿真实验者的手动操作且方法的准确性缺乏理论依据, 这限制了整群抽样方法的推广应用. 为此, 本文在已有基础上提出一种改进的且具有一般性的整群抽样方法 (enhanced cluster sampling, ECS). ECS 将块矩阵 C_q 按一定规律进一步拆分为更小的子块矩阵 $C_{q,m}$, 并利用与 $C_{q,m}$ 对应的虚构子块矩阵 $C'_{q,m}$ 控制 $C_{q,m}$ 的生成, 最终实现构造 $C_{(n)}$ 的目的. 相较于现有的整群抽样方法, ECS 的优势在于: 其为每个因子生成的基效应数量完全相同, 且与实际需要生成的基效应数差值不高于 q , 而 q 通常较小. 本节将首先定义 ECS 所需的 3 个数学符号 A_s 、 $C_{q,m}$ 和 $C'_{q,m}$.

定义 1 $A_s (s = 1, 2, \dots, k)$ 表示将 k 维单位矩阵元素整体向右移动 $s-1$ 个单位 (后面的元素顺势移动至前面) 所得的矩阵, 其中 A_1 为 k 维单位矩阵.

例如, 当 $k = 4$ 时, A_1 为单位矩阵, 将 A_1 矩阵元

素向右移动 1 个单位得到 $A_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$, 将 A_1 矩

阵元素向右移动 2 个单位得到 $A_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$.

定义 2 $C_{q,m}$ 为块矩阵 C_q 中结构一致的行组成的第 m 个子块矩阵, $m = 1, 2, \dots, M_q$, 其中 M_q 指的是块矩阵 C_q 中子块矩阵 $C_{q,m}$ 的数量.

定义 2 中“结构一致的行”指的是 C_q 中某些行, 当其元素整体向右移动一定单位 (后面的元素顺势移动至前面), 所得的行向量完全一致的行的集合. 该定义本质上是找出 C_q 中 q 个数值为 1 的元素排列规

律一致的行. 例如, 当 $k = 4$ 时, 有 $C_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$.

将 C_2 矩阵的第 1 行 $[1 \ 1 \ 0 \ 0]$ 整体向右移动 1 个单位, 即前 3 个元素向右移 1 位, 末位元素移到首位, 得到 $[0 \ 1 \ 1 \ 0]$, 该向量恰与 C_2 矩阵的第 2 行相等. 同理, 第 1 行向右移动 2 位获得第 3 行, 移动 3 位获得第 4 行. 因此, 前 4 行是“结构一致的行”, 组成了一个 $C_{q,m}$. 类似地可以发现第 5、第 6 行结构也一致. 根据该定义可以将 C_2 矩阵划分为下面两个子块矩阵:

$$C_{2,1} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad C_{2,2} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \quad (10)$$

这种子块矩阵划分的好处是: 将所得的每个子块矩阵 $C_{q,m}$ 添加到 $C_{(n)} = (C_0; C_1; \dots; C_{q-1})$ 中都可以为每个因子生成相同数量的基效应. 例如添加式 (10) 中的 $C_{2,1}$ 可以为每个因子生成 2 个基效应, 添加 $C_{2,2}$ 可以为每个因子生成 1 个基效应. 本文将 1.3 节中 Shi 等^[20]提出的结论拓展到 $C_{q,m}$, 得到下述结论.

性质 1 将 C_q 矩阵的一个子块矩阵 $C_{q,m}$ 添加到 $C_{(n)} = (C_0; C_1; \dots; C_{q-1})$ 可以为每个因子生成 $\frac{q}{k} \times r$ 个基效应 (r 为 $C_{q,m}$ 的行数).

鉴于式 (10) 中 $C_{2,1}$ 和 $C_{2,2}$ 行数的区别, 本文将 $C_{q,m}$ 分为以下两种情形: 1) $C_{q,m}$ 行数为 k 的情形 (一般情形). $C_{q,m}$ 中任一行向量中的元素整体向右移动 $k'(k' \in \{1, 2, \dots, k-1\})$ 个单位 (后面的元素顺势移

动至前面),都会产生一个与另一行向量相同的行.因此,加上其自身,一共存在 k 个“结构一致”的行,从而 $C_{q,m}$ 为一个 $k \times k$ 维块矩阵.例如式(10)中的 $C_{2,1}$. 2) $C_{q,m}$ 行数小于 k 的情形(特殊情形). k' 的取值有时无法达到 $k-1$,从而导致 $C_{q,m}$ 的行数小于 k .例如:式(10)中 $C_{2,2}$ 的任一行仅能取 $k'=1$ 便得到了另外一行,但不能取 $k'=2$,因为取 $k'=2$ 时又循环到了其自身,从而 $C_{2,2}$ 的行数为2,小于 $k=4$.为了能将这两种情形合并在一起以方便处理,本文通过在其现有行的基础上有规律地添加些重复的行将情形2)中的 $C_{q,m}$ 也凑成一个 $k \times k$ 维矩阵.为区分原有的未添加重复行的 $C_{q,m}$,本文提出以下“虚构子矩阵”的概念.

定义3 $C'_{q,m}$ 为 C_q 的第 m 个 $k \times k$ 维“虚构”子矩阵,其中 $C'_{q,m}$ 要么与定义2中的 $C_{q,m}$ 一样,要么是为了凑成 $k \times k$ 方阵,而把某些重复的行也加入其中.

定义3表明,若将 $(C'_{q,1}; C'_{q,2}; \dots; C'_{q,M_q})$ 中的重复行去掉,则会得到 $C_q = (C_{q,1}; C_{q,2}; \dots; C_{q,M_q})$,故称为“虚构”.根据上述定义,本文得到以下性质.

性质2 C_q 矩阵的虚构子块矩阵 $C'_{q,m}$ 都可由 C_{q-1} 矩阵的虚构子块矩阵 $C'_{q-1,m'}$ ($m' \leq m$)加上经初等行变换的单位矩阵得到,即

$$C'_{q,m} = C'_{q-1,m'} + A_s, m' \leq M_{q-1}. \quad (11)$$

当 $q=1$ 时, m 仅能取1,从而 $C'_{1,1}$ 为 k 维单位矩阵.当 $q \geq 2$ 时,不难发现 $C'_{q,m}$ 是由 $C'_{q-1,m'} + A_s$ 构成的 $k \times k$ 矩阵,将这个虚构子矩阵去除重复行后,即可得到矩阵 C_q 真实的子块矩阵 $C_{q,m}$.例如:当 $q=2$ 时,由于 C_1 是单位矩阵,每一行仅有1个元素取值为1.结合 C_2 矩阵可以发现, C_2 中的每一行都可以由 C_1 中的行向量在某个0元素的位置上加上1得到.基于此发现,本文对 s 的取值作如下规定:设 s' 为 $C'_{q-1,m'}$ 第1行中最后一个1元素的列坐标,则式(11)中 s 的值满足 $s'+1 \leq s \leq k$.因此, C_1 矩阵加上 A_s (此处 $2 \leq s \leq k$)即可得到每行每列均包含2个值为1的元素的 $k \times k$ 矩阵,即 $C'_{2,m}$.例如,当 $k=4$ 时,有 $C'_{2,1}$

$$= C_1 + A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \text{其对应着定义2提到的}$$

“一般情形”,即 $C'_{2,1}$ 中各行均互不相等,此时 $C_{2,1} =$

$$C'_{2,1}. \text{同理有 } C'_{2,2} = C_1 + A_3 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \text{其对应}$$

着定义2提到的“特殊情形”,即 $C'_{2,2}$ 第1、第2行分别与第3、第4行重复.因此,要想得到 $C_{2,2}$,需要去除

$C'_{2,2}$ 中重复的行.进一步地对比 $C'_{2,2}$ 与 $C_{2,2}$ 发现, $C_{2,2}$ 仅有2行,而 $C'_{2,2}$ 行数为4,因此 $C_{2,2}$ 与 A_s 无法相加,即无法直接用 $C_{2,2}$ 构造子块矩阵 $C_{3,m}$,而虚构子矩阵 $C'_{2,2}$ 则恰好解决了这一问题,这是本文提出的虚构子矩阵的优势之一.若继续生成 $C'_{2,3}$,则不难发现 $C'_{2,3}$ 与 $C'_{2,1}$ 中的行完全相同,因此 C_2 仅包含两个虚构子矩阵 $C'_{2,1}$ 和 $C'_{2,2}$.

基于上述对块矩阵 C_q 内部子块矩阵的进一步分析,本文提出一种完整的ECS方法:1)先计算 $q = \arg \min_{q \in \{2,3,\dots,k\}} \left\{ \sum_{t=1}^q \binom{k-1}{t-1} \geq N \right\}$; 2)依次将 $C_0, C_1, \dots, C_{q-1}$ 添加到 $C_{(n)}$ 中,此时 $C_{(n)} = (C_0; C_1; \dots; C_{q-1})$ 可以为每个因子生成 $\sum_{t=1}^{q-1} \binom{k-1}{t-1}$ 个基效应; 3)依据性质2,以递归的方式生成 C_q 中部分子块矩阵组成 C_{q^*} 矩阵并将其添加到 $C_{(n)}$ 中,用于为每个因子生成剩下的 $N - \sum_{t=1}^{q-1} \binom{k-1}{t-1}$ 个基效应,具体算法步骤如下.

算法1 生成整群抽样矩阵 $C_{(n)}$ 的算法.

输入: 指定因子数量 k 以及要为每个因子生成的基效应数量 $N, 1 \leq N \leq 2^{k-1}$.

step 1: 计算 q .

$$q = \arg \min_{q \in \{2,3,\dots,k\}} \left\{ \sum_{t=1}^q \binom{k-1}{t-1} \geq N \right\}.$$

step 2: 构造 C_{q^*} .

1) 计算还需要为每个因子生成的基效应数量 $N - \sum_{t=1}^{q-1} \binom{k-1}{t-1}$.

2) 生成 C_{q^*} .

① 设置 $n' \leftarrow 0, m \leftarrow 1, m' \leftarrow 1, s \leftarrow s' + 1$ (s' 表示 $C'_{q-1,m'}$ 第1行中最后1个1元素的列坐标).

② 生成虚构块矩阵 $C'_{q,m} \leftarrow C'_{q-1,m'} + A_s$ ($C'_{1,1}$ 为单位矩阵).

③ 若 $s = k, m' \leftarrow m' + 1$,则 $s \leftarrow s' + 1$; 否则 $s \leftarrow s + 1$.

④ 若 $C'_{q,m}$ 与其他块矩阵重复,则去除该 $C'_{q,m}$; 否则去除 $C'_{q,m}$ 中重复的行向量后可得 $C_{q,m}$,且 $m \leftarrow m + 1$.

⑤ $C_{q^*} \leftarrow (C_{q,1}; C_{q,2}; \dots; C_{q,m})$,记录 C_{q^*} 矩阵可以为每个因子生成的基效应数量 $n' \leftarrow n' + rq/k$ (r 为 $C_{q,m}$ 的行数).

⑥ 重复②~⑤步骤直到 $n' \geq N - \sum_{t=1}^{q-1} \binom{k-1}{t-1}$.

step 3: 生成 $C_{(n)} \leftarrow (C_0; C_1; \dots; C_{q-1}; C_{q^*})$.

输出: $C_{(n)}$.

由ECS具体步骤和性质1可得如下性质.

性质3 ECS方法生成的 $C_{(n)}$ 为每个因子生成的基效应数量 n 满足 $n - N < q$.

在实际应用中 q 通常较小,所以性质3保证了 $C_{(n)}$ 为每个因子生成的基效应数量尽可能接近 N .

3 数值实验与应用

本研究首先对算法的有效性进行实验,给定因子总数 k 和目标基效应数量 N ,判断利用该算法得到的 $C_{(n)}$ 矩阵为每个因子生成的基效应数量 n 是否满足 $n - N < q$.再将其与非整群抽样方法进行对比,展示其经济性.最后进行蒙特卡罗仿真分析,应用ECS和现有方法分析因子主效应与交互效应,从而评价ECS在仿真实验中的准确性.

3.1 算法有效性检验

本节使用不同的 (k, N) 组合,比较ECS生成的 $C_{(n)}$ 为每个因子生成的基效应数量 n 与目标基效应

数量 N 的差值,检验ECS算法的有效性.本文先设定 $k \in \{4, 5, 6, 7\}$, $N = 5$,利用ECS生成 $C_{(n)}$ 矩阵,观察其为每个因子生成的基效应数量是否相同并判断是否满足 $n - N < q$,从而验证ECS的有效性.

图1结果展示了上述 (k, N) 组合下ECS生成的 $C_{(n)}$ 矩阵.可以看出:当 $k = 4$ 时,ECS给出的 $C_{(7)}$ 矩阵可以为每个因子生成 $n = 7$ 个基效应,同时此矩阵的 C_{q^*} 中 $q = 3$,满足 $n - N = 2 < q$;当 $k \in \{5, 6, 7\}$ 时,由ECS得到的 $C_{(5)}$ 矩阵可以为每个因子生成 $n = 5$ 个基效应,且这些矩阵的 C_{q^*} 中 $q = 2$,满足 $n - N = 0 < q$,即ECS给出的 $C_{(n)}$ 矩阵为每个因子生成的基效应数量相同,并且满足条件 $n - N < q$,这验证了ECS的有效性.在此基础上,本文进一步考虑当 k 和 N 取较大值时,ECS是否仍然有效,并将 $C_{(n)}$ 矩阵与利用非整群抽样方法得到的 B 矩阵相比较,验证整群抽样方法的经济性.

$$\begin{aligned}
 & \begin{matrix} \frac{C_0}{C_1} \\ \frac{C_2}{C_3^*} \end{matrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}, \quad \begin{matrix} \frac{C_0}{C_1} \\ \frac{C_2}{C_3^*} \end{matrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{matrix} \frac{C_0}{C_1} \\ \frac{C_2}{C_3^*} \end{matrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{matrix} \frac{C_0}{C_1} \\ \frac{C_2}{C_3^*} \end{matrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.
 \end{aligned}$$

$k=4$
 $k=5$
 $k=6$
 $k=7$

图1 不同 k, N 组合下的 $C_{(n)}$ 矩阵

表1展示了在 (k, N) 参数设置为较大值时,ECS生成的 $C_{(n)}$ 矩阵为每个因子生成的基效应数量 n ,矩

阵中包含 C_{q^*} 矩阵的 q 值、矩阵行数以及在当前参数设置下采用非整群抽样方法得到的 B 矩阵的行数.

表1 ECS生成的 C 矩阵和 B 矩阵部分参数

k	N	n	q	$C_{(n)}$ 的行数	$(B_1; B_2; \dots; B_N)$ 的行数
20	100	101	3	751	2 100
	500	503	4	2 911	10 500
	1 000	1 003	4	5 411	21 000
	5 000	5 000	5	21 556	105 000
	10 000	10 001	6	38 250	210 000
100	1 000	1 000	3	35 051	101 000
	2 000	2 002	3	68 451	202 000
	4 000	4 000	4	135 051	404 000
	5 000	5 003	5	168 051	505 000
200	2 000	2 000	3	140 101	402 000
	4 000	4 001	3	273 501	804 000
300	3 000	3 000	3	315 151	903 000
400	4 000	4 000	3	560 201	1 604 000
500	5 000	5 000	3	875 251	2 505 000

由表1看出:1)在较大的 (k, N) 参数设置下,ECS仍表现出较好的性能.生成的 $\mathbf{C}_{(n)}$ 矩阵尽管不能每次都精确地为每个因子生成 N 个基效应,但 n 与 N 的差值并未随着 N 的增长而迅速增加,而是仍处于一个较低的水平,始终满足 $n - N < q$,这充分验证了ECS的有效性.2)对比 $\mathbf{B}$ 矩阵可以发现,ECS方法给出的 $\mathbf{C}_{(n)}$ 矩阵的行数远小于 $(\mathbf{B}_1; \mathbf{B}_2; \dots; \mathbf{B}_N)$,体现了整群抽样方法的经济性,这对于仿真实验效率的提高具有重要意义.3)当因子数量 k 不变,而 N 不断增大时,ECS给出的 $\mathbf{C}_{(n)}$ 矩阵的行数的增速小于利用非整群抽样得到的 $\mathbf{B}$ 矩阵的行数的增速,即随着 N 的不断增大, $\mathbf{C}_{(n)}$ 矩阵的经济性在不断提高,而非整群抽样矩阵的经济性恒定且低于 $\mathbf{C}_{(n)}$ 矩阵.

由上述分析可以看出,ECS展现出良好的有效性,对于不同的 (k, N) 参数设置均可以给出满足条件的结果.此外,相较于非整群抽样,ECS体现了整群抽样的经济性,这将有助于提高仿真实验效率.

3.2 蒙特卡罗仿真分析

本节采用Morris使用的例子比较了ECS与传统EE算法之间的实验效率和效力的差异.仿真模型在因子组合 $\mathbf{x} = (x_1, x_2, \dots, x_k)$ 处的输出为

$$y = \beta_0 + \sum_{j=1}^{20} \beta_j \omega_j + \sum_{i < j}^{20} \beta_{i,j} \omega_i \omega_j + \sum_{l < i < j}^{20} \beta_{l,i,j} \omega_l \omega_i \omega_j + \sum_{s < l < i < j}^{20} \beta_{s,l,i,j} \omega_s \omega_l \omega_i \omega_j. \quad (12)$$

其中:当 $j = 3, 5, 7$ 时,采用非线性变换 $\omega_j = 2(1.1x_j / (x_j + 0.1) - 0.5)$;其他因子均采用线性变换 $\omega_j = 2(x_j - 0.5)$.对于一阶和二阶系数: $\beta_j = 20, j \in \{1, 2, \dots, 10\}$; $\beta_{i,j} = -15, i, j \in \{1, 2, \dots, 6\}$;其余 β_j 和 $\beta_{i,j}$ 从标准正态分布 $N(0, 1)$ 中独立抽样.对于三阶和四阶系数: $\beta_{l,i,j} = -10, l, i, j \in \{1, 2, \dots, 5\}$; $\beta_{s,l,i,j} = 5, s, l, i, j \in \{1, 2, 3, 4\}$;其余的 $\beta_{l,i,j}$ 、 $\beta_{s,l,i,j}$ 和 β_0 都设为0.

本文首先使用传统EE算法(即非整群抽样方法),设置为每个因子生成的基效应数量为 $N = 10\,000$,得到各因子的基效应分布 F_j 的总体均值 $\hat{\mu}_j$ 和标准差 $\hat{\sigma}_j$,并记录所需要进行的模拟运行次数作为该方法的实验效率评价标准.同时,本研究使用ECS方法,设置单个 $\mathbf{C}$ 矩阵为每个因子生成的基效应数量 n 以及 $\mathbf{C}$ 矩阵的随机形式个数 W ,得到各因子的基效应分布 F_j 的总体均值 $\hat{\mu}_j$ 和标准差 $\hat{\sigma}_j$,并记录所需要进行的模拟运行次数.之后,本研究利用现有的整群抽样方法,即Morris提出的整群抽样方法(记为EEC)以及Shi等^[20]提出的EMM方法,构造整群抽样矩阵并寻找合适的 (n, W) 组合以实现与ECS相同的模拟运行次数.需要注意的是,EEC和EMM在构造矩阵时部分依赖于手动操作.为了排除随机因素的干扰,本研究对不同的实验组合均进行1 000次独立宏重复.表2展示了本文考虑的实验具体场景.

表2中: n 为单个矩阵生成的基效应数量, W 为矩阵的随机形式数量.

表2 实验参数设置

场景	(n, W)				模拟运行次数			
	EE	ECS	EEC	EMM	EE	ECS	EEC	EMM
1	(1, 10 000)	(3, 4 500)	(3, 3 620)	(3, 4 620)	210 000	184 500	184 620	184 800
2	(1, 5 000)	(3, 2 250)	(3, 1 810)	(3, 2 310)	105 000	92 250	92 310	92 400
3	(1, 2 000)	(3, 900)	(3, 724)	(3, 924)	42 000	36 900	36 924	36 960
4	(1, 1 000)	(3, 450)	(3, 362)	(3, 462)	21 000	18 450	18 462	18 480

进一步地,本文采用Shi等^[29]提出的主效应与交互效应的平均相对误差(ARE)评价算法的统计效力,计算方法如下:

$$\text{ARE}_\mu = \frac{1}{k} \sum_{j=1}^k \left| \frac{\hat{\mu}_j - \mu_j}{\mu_j} \right|, \quad (13)$$

$$\text{ARE}_{\sigma^2} = \frac{1}{k} \sum_{j=1}^k \left(\frac{|\hat{\sigma}_j^2 - \sigma_j^2|}{\sigma_j^2} \right). \quad (14)$$

其中: μ_j 和 σ_j 分别为因子 j 的基效应对应分布 F_j 的真实均值和标准差, $\hat{\mu}_j$ 和 $\hat{\sigma}_j$ 分别为利用该方法得到各因子的基效应分布 F_j 的均值和标准差的估计值.

ARE值越大,表示该方法得到的结果与模型的真实值偏差越大,因此通过比较不同方法下的ARE

值的大小即可对这些方法的效力进行判断.本文分别对主效应和交互效应的ARE进行两样本 t 检验:

$$H_0: \text{ARE}_\mu^{\text{EE}} = \text{ARE}_\mu^{\text{ECS}} \text{ vs } H_1: \text{ARE}_\mu^{\text{EE}} \neq \text{ARE}_\mu^{\text{ECS}}; \quad (15)$$

$$H_0: \text{ARE}_{\sigma^2}^{\text{EE}} = \text{ARE}_{\sigma^2}^{\text{ECS}} \text{ vs } H_1: \text{ARE}_{\sigma^2}^{\text{EE}} \neq \text{ARE}_{\sigma^2}^{\text{ECS}}. \quad (16)$$

其中: $\text{ARE}_\mu^{\text{EE}}$ 和 $\text{ARE}_{\sigma^2}^{\text{EE}}$ 表示由传统EE算法(即非整群抽样方法)得到的主效应与交互效应的平均相对误差, $\text{ARE}_\mu^{\text{ECS}}$ 和 $\text{ARE}_{\sigma^2}^{\text{ECS}}$ 表示由ECS得到的主效应与交互效应的平均相对误差.本研究希望如果由ECS得到的ARE小于或等于由传统EE得到的ARE,则表明ECS具有不低于传统EE的统计效力.

图2展示了上述4种方法在各场景下实验所得的 ARE_μ 和 ARE_{σ^2} 值.

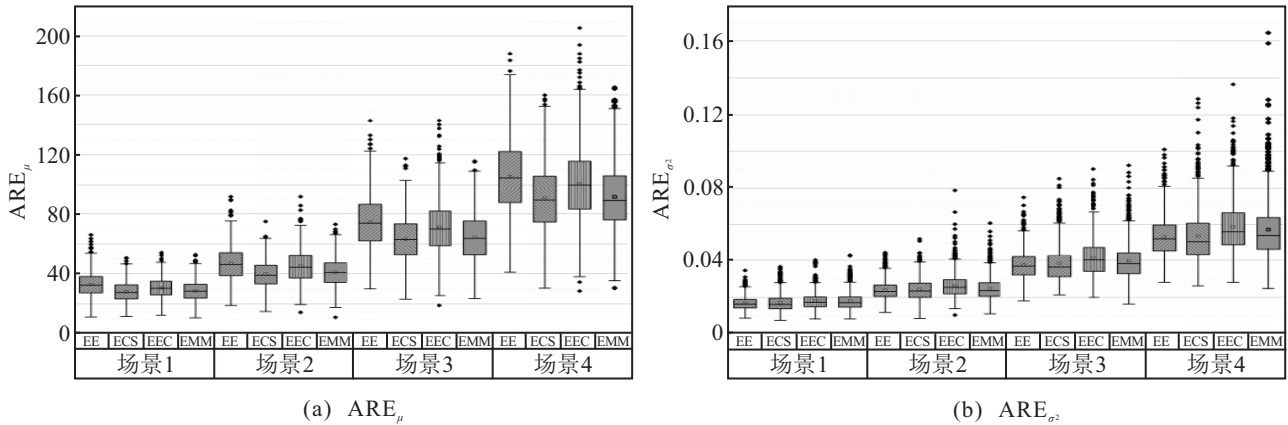


图2 4种抽样方法所得的ARE值的比较

本文发现: 1) 式(15)的假设检验拒绝 H_0 , 式(16)的假设检验接受 H_0 . 结合图2可以看出: $ARE_{\mu}^{ECS} < ARE_{\mu}^{EE}$ 且 $ARE_{\sigma^2}^{ECS}$ 接近 $ARE_{\sigma^2}^{EE}$, 这验证了ECS方法对主效应判断效力高于传统EE, 对交互效应的判断效力与传统EE无明显差别, 这与本研究预期效果一致, 验证了ECS方法的统计效力略高于传统EE. 2) 整群抽样方法的实验效率明显高于非整群抽样方法, 例如在本实验场景1中传统EE方法的模拟运行次数为210 000, 而ECS方法的模拟运行次数仅为184 500, 相较于传统EE方法减少了25 500次(约12.14%)模拟运行, 这体现了整群抽样方法在提高实验效率上的优势. 3) 与现有的整群抽样方法相比可以发现, 在相近的模拟运行次数下, ECS和EMM所展现出的ARE值接近且明显低于EEC, 展现出良好的主效应和交互效应识别能力; 同时现有方法部分依赖于手动操作, 从而不利于快速且准确地构造整群抽样矩阵, ECS则不具有上述缺陷. 简言之, 蒙特卡罗仿真实验充分验证了ECS的可行性. 与非整群抽样方法相比, ECS具有提高计算效率的优势, 且具有不低于传统EE的统计效力; 与现有的整群抽样方法相比, ECS完全不依赖于人工且展现出稳定的因子筛选能力. 需要注意的是, 虽然ECS给出了构造 $C_{(n)} (n \leq 2^{k-1})$ 的方法, 但一味地增加单个 $C_{(n)}$ 矩阵为每个因子生成的基效应数量 n 是不理智的, 正确做法应是如Shi等^[27]提出的利用 W 个 $C_{(n)}$ 矩阵的随机形式为每个因子生成 $n \times W$ 个基效应, 从而实现实验效率与效力的平衡.

4 结论

大数据背景下, 仿真模型中的因子数量急剧增加, 识别出其中对模型有重要影响的因子在近年来成为众多领域竞相研究的热门问题. 凭借着较高的实验效率和效力, Morris的基效应法在许多领域得到应用并不断发展. 然而随着数据量不断增加而急剧增多的仿真输入使得应用该方法所需的模拟运行次数

大幅增加, 提高实验效率迫在眉睫. 为提高基效应法的计算效率, 本文在已有研究的基础上提出一种改进的且更具一般性的整群抽样方法ECS. 相较于已有的整群抽样方法, 该方法保证了在较小的误差范围内为每个因子生成相同数量的基效应. 在不降低仿真实验效力的基础上大大提高了仿真实验效率, 具有重大现实意义. 未来值得研究的方向包括: 1) 提高整群抽样方法下EE对因子交互效应的识别能力; 2) 在保证不损失统计效力的前提下, 如何利用ECS尽可能地提高计算效率; 3) 利用EE直接识别现实观测数据中各因子的主效应及交互效应, 并利用ECS提高计算效率; 4) 开发更细粒度的整群抽样矩阵构造方法, 缩小 n 与 N 之间的差距.

参考文献(References)

- [1] 陈国青, 吴刚, 顾远东, 等. 管理决策情境下大数据驱动的研究和应用挑战——范式转变与研究方向[J]. 管理科学学报, 2018, 21(7): 1-10.
(Chen G Q, Wu G, Gu Y D, et al. The challenges for big data driven research and applications in the context of managerial decision-making: Paradigm shift and research directions[J]. Journal of Management Sciences in China, 2018, 21(7): 1-10.)
- [2] 徐选华, 余紫昕. 社会网络环境下基于公众行为大数据属性挖掘的大群体应急决策方法及应用[J]. 控制与决策, 2022, 37(1): 175-184.
(Xu X H, Yu Z X. A large group emergency decision making method and application based on attribute mining of public behaviour big data in social network environment[J]. Control and Decision, 2022, 37(1): 175-184.)
- [3] 郭戈, 徐涛, 韩英华, 等. 电动汽车时代的电网-交通网协同优化综述[J]. 控制与决策, 2021, 36(9): 2049-2062.
(Guo G, Xu T, Han Y H, et al. A survey of cooperative optimization of traffic-grid networks in the era of electric vehicles[J]. Control and Decision, 2021, 36(9): 2049-2062.)

- 2049-2062.)
- [4] 刘长石, 陈慧璇, 吴张. 城市物流配送的混合车辆路径规划模型与优化算法[J]. 控制与决策, 2023, 38(3): 759-768.
(Liu C S, Chen H X, Wu Z. Programming model and optimization algorithm for the mixed fleet vehicle routing problem in urban logistics distribution[J]. Control and Decision, 2023, 38(3): 759-768.)
 - [5] 徐宗本, 冯芷艳, 郭迅华, 等. 大数据驱动的管理与决策前沿课题[J]. 管理世界, 2014(11): 158-163.
(Xu Z B, Feng Z Y, Guo X H, et al. Frontier topics of management and decision-making driven by big data[J]. Management World, 2014(11): 158-163.)
 - [6] Robinson S. Simulation: The practice of model development and use[M]. Bloomsbury: Bloomsbury Publishing, 2014: 77-124.
 - [7] 闫敬, 徐龙, 曹文强, 等. 基于深度强化学习的多潜器编队控制算法设计[J]. 控制与决策, 2023, 38(5): 1457-1463.
(Yan J, Xu L, Cao W Q, et al. Design of formation control algorithm for multiple autonomous underwater vehicles based on deep reinforcement learning[J]. Control and Decision, 2023, 38(5): 1457-1463.)
 - [8] Kleijnen J P C. Design and analysis of simulation experiments[C]. International Workshop on Simulation. Cham, 2018: 3-22.
 - [9] 刘丽君, 马义中, 欧阳林寒. 高效的因子分类筛选方法及仿真应用[J]. 系统管理学报, 2021, 30(1): 3-13.
(Liu L J, Ma Y Z, Ouyang L H. An efficient factor screening and classification procedure based on multi-response sequential bifurcation and applications[J]. Journal of Systems & Management, 2021, 30(1): 3-13.)
 - [10] Campolongo F, Braddock R. The use of graph theory in the sensitivity analysis of the model output: A second order screening method[J]. Reliability Engineering & System Safety, 1999, 64(1): 1-12.
 - [11] Kleijnen J P C, Sanchez S M, Lucas T W, et al. State-of-the-art review: A user's guide to the brave new world of designing simulation experiments[J]. INFORMS Journal on Computing, 2005, 17(3): 263-289.
 - [12] Borgonovo E, Plischke E. Sensitivity analysis: A review of recent advances[J]. European Journal of Operational Research, 2016, 248(3): 869-887.
 - [13] Ankenman B E, Cheng R C H, Lewis S M. Screening for dispersion effects by sequential bifurcation[J]. ACM Transactions on Modeling and Computer Simulation, 2015, 25(1): 1-27.
 - [14] 刘丽君, 马义中, 欧阳林寒, 等. 数据污染对于序贯分支方法的影响及应对策略[J]. 系统工程理论与实践, 2020, 40(5): 1281-1292.
(Liu L J, Ma Y Z, Ouyang L H, et al. The influence of data contamination on sequential bifurcation and its counter measures[J]. Systems Engineering—Theory & Practice, 2020, 40(5): 1281-1292.)
 - [15] 施文, 冷凯君, 卿前恺. 大数据下仿真筛选实验及误差控制模型和应用[J]. 系统工程理论与实践, 2018, 38(9): 2300-2314.
(Shi W, Leng K J, Qing Q K. Simulation screening and error control for big data[J]. Systems Engineering—Theory & Practice, 2018, 38(9): 2300-2314.)
 - [16] Woods D C, Lewis S M. Design of experiments for screening[M]. Cham: Handbook of Uncertainty Quantification, 2017: 1143-1185.
 - [17] Morris M D. Factorial sampling plans for preliminary computational experiments[J]. Technometrics, 1991, 33(2): 161-174.
 - [18] Cropp R A, Braddock R D. The new Morris method: An efficient second-order screening method[J]. Reliability Engineering & System Safety, 2002, 78(1): 77-83.
 - [19] Sin G, Gernaey K V. Improving the Morris method for sensitivity analysis by scaling the elementary effects[M]. Amsterdam: Computer Aided Chemical Engineering, 2009: 925-930.
 - [20] Shi W, Chen X, Shang J. An efficient Morris method-based framework for simulation factor screening[J]. INFORMS Journal on Computing, 2019, 31(4): 745-770.
 - [21] Boukouvalas A, Gosling J P, Maruri-Aguilar H. An efficient screening method for computer experiments[J]. Technometrics, 2014, 56(4): 422-431.
 - [22] Fédou J M, Rendas M J. Extending Morris method: Identification of the interaction graph using cycle-equitable designs[J]. Journal of Statistical Computation and Simulation, 2015, 85(7): 1398-1419.
 - [23] Chang S L, Harding N, Zachreson C, et al. Modelling transmission and control of the COVID-19 pandemic in Australia[J]. Nature Communications, 2020, 11(1): 1-13.
 - [24] Diepens N J, Koelmans A A. Accumulation of plastic debris and associated contaminants in aquatic food webs[J]. Environmental Science & Technology, 2018, 52(15): 8510-8520.
 - [25] Koks E E, Rozenberg J, Zorn C, et al. A global multi-hazard risk analysis of road and railway infrastructure assets[J]. Nature Communications, 2019, 10(1): 1-11.
 - [26] Lu X F, Borgonovo E. Global sensitivity analysis in epidemiological modeling[J]. European Journal of Operational Research, 2023, 304(1): 9-24.
 - [27] Shi W, Chen X. Cluster sampling for Morris method made easy[J]. Naval Research Logistics (NRL), 2020, 68(4): 412-433.
 - [28] Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models[J]. Environmental Modelling & Software, 2007, 22(10): 1509-1518.
 - [29] Shi W, Chen X. Efficient budget allocation strategies for elementary effects method in stochastic simulation[J]. Naval Research Logistics (NRL), 2018, 65(3): 218-241.

作者简介

施文(1981—), 男, 教授, 博士生导师, 从事汽车召回管理、汽车物流与供应链管理、复杂系统建模与仿真、机器学习、大数据分析等研究, E-mail: wenshi@csu.edu.cn;

陈奥(2001—), 男, 硕士生, 从事仿真实验设计与分析、机器学习等研究, E-mail: aochenao2001@163.com.