arXiv:2304.03019v1 [math.ST] 6 Apr 2023

最优子抽样设计

亨里克·伊姆贝里格
imbergh@chalmers.se

玛丽娜·阿克塞尔松-菲斯克
marina.axelson-fisk@chalmers.se

约翰·约纳松
jonasson@chalmers.se

查尔姆斯理工大学数学科学系暨哥德堡大学，瑞典哥德堡 SE-412 96

摘要

子抽样常用于克服有限总体和海量数据集分析中的计算与经济瓶颈。现有方法的适用范围往往有限，且采用的最优性准则（例如 A-最优性）存在众所周知的缺陷，例如不具备对数据测量尺度和模型参数化方式的不变性。目前仍缺乏统一的最优子抽样设计理论。本文提出了一套适用于一般数据子抽样问题的最优设计理论，涵盖有限总体推断、参数密度估计和回归建模。该理论涵盖并推广了该领域中大多数基于不等概率抽样和逆概率加权的最优子数据选择方法。我们推导了一大类最优性准则的最优性条件，并提出了相应算法，用于在泊松抽样设计和多项式抽样设计下寻找最优抽样方案。我们提出了一类新颖的、对变换和参数化方式不变的线性最优性准则，兼具两方面优势：A-最优性的计算易处理性，以及类似于 D-最优性的不变性。在交通安全领域的一项应用中，我们展示了该方法。实验结果表明，所提出的不变线性最优性准则在计算需求降低 90–95% 的同时，达到 92–99% 的 D 效率。相比之下，A-最优性准则在两个示例中仅达到 46% 和 60% 的 D 效率。

关键词 A-最优性 · D-最优性 · L-最优性 · M估计 · 逆概率加权 · 不等概率抽样。

1 引言

考虑一个 p维参数 $\boldsymbol{\theta}_0$ 定义为

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Omega}}{\arg\min}\, \ell_0(\boldsymbol{\theta}), \tag{1}$$

即，作为参数空间$\boldsymbol{\Omega} \subset \mathbb{R}^p$上某个函数 $\ell_0(\boldsymbol{\theta})$的最小化器。我们进一步假设 $\boldsymbol{\theta}_0$ 是唯一的，并且 $\ell_0(\boldsymbol{\theta})$ 是二阶可微的，可以写成如下形式

$$\ell_0(\boldsymbol{\theta}) = \sum_{i\in\mathcal{D}} \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{v}_i), \tag{2}$$

对某个索引集 $\mathcal{D} = \{1, \ldots, N\}$ 求和，其中 $\boldsymbol{v}_i$是与成员 $i \in \mathcal{D}$ 相关联的数据向量。在这些假设下，$\boldsymbol{\theta}_0$ 也可定义为估计方程的唯一解。

$$\sum_{i\in\mathcal{D}} \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \mathbf{0}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{v}_i). \tag{3}$$

数据的形式为 $\boldsymbol{v}_i = \boldsymbol{y}_i \in \mathcal{Y}$或 $\boldsymbol{v}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathcal{Y}$，其中 $\boldsymbol{y}_i$ 是响应向量，而 $\boldsymbol{x}_i$ 是解释变量向量。我们通常不区分有无解释变量的情况，并始终将数据写作$(\boldsymbol{x}_i, \boldsymbol{y}_i)$，注意到第一项可能为空，且 $\mathcal{X}$为空集。一个

可将 (1)–(2) 解释为一个经验风险最小化问题（Vapnik，1991）。因此，我们将 $\ell_0(\boldsymbol{\theta})$称为（全数据）经验风险，并将 $\boldsymbol{\theta}_0$ 称为（全数据）经验风险最小化器（ERM）。
上述设定涵盖了统计学中广泛的推断问题、模型和估计方法，包括最大似然估计、广义线性模型（Nelder 和 Wedderburn，1972年；McCullagh 和 Nelder，1989）、拟似然方法（Wedderburn，1974年）以及某些类型的 M 估计（Stefanski 和 Boos，2002年）。第 6 节的“应用与示例”中将进一步讨论的一些具体示例包括：

i）有限总体推断：考虑一个由 N 个个体组成的有限总体，每个个体都对应一个非随机特征向量 $\boldsymbol{y}_i$。有限总体均值向量 $\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i$ 可写成形式（1）–（3），其中 $\boldsymbol{v}_i=\boldsymbol{y}_i$ 和 $\ell(\boldsymbol{\theta};\ \boldsymbol{v}_i)\ =||\boldsymbol{y}_i-\boldsymbol{\theta}||_2^2=(\boldsymbol{y}_i-\boldsymbol{\theta})^{\mathsf{T}}(\boldsymbol{y}_i-\boldsymbol{\theta})$。ii）参数密度估计：给定来自密度函数为 $f_{\boldsymbol{\theta}}(y)$ 的概率分布的独立同分布数据 $y_1,\ .....,\ y_N$，$\boldsymbol{\theta}$ 的最大似然估计可写成形式（1）–（3），其中 $\boldsymbol{v}_i=y_i$ 和 $\ell(\boldsymbol{\theta};\ \boldsymbol{v}_i)\ =-\log f_{\boldsymbol{\theta}}(y_i)$。iii）回归建模：考虑一个随机样本 $\{(\boldsymbol{x}_i,\ y_i)\}_{i=1}^{N}$，一个回归系数向量 $\boldsymbol{\theta}$，一个用于描述给定 $\boldsymbol{x}$ 时 Y 条件均值的（非线性）模型 $f_{\boldsymbol{\theta}}(\boldsymbol{x})$，以及一个可微损失函数 $l:\ \mathbb{R}^2\to\mathbb{R}_+$，使得 $l(\hat{}y,\ y)\ =0$ 当且仅当 $\hat{}y=y$。令 $\boldsymbol{v}_i=(\boldsymbol{x}_i,\ y_i)$ 且 $\ell(\boldsymbol{\theta};\ \boldsymbol{v}_i)\ =l(f_{\boldsymbol{\theta}}(\boldsymbol{x})\ y_i)$，则方程（1）–（3）定义了回归系数向量 $\boldsymbol{\theta}$ 的一个估计值。

现在考虑这样一种情况：由于经济或计算方面的限制，基于完整数据 $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ 的推断被禁止。例如，索引集可能非常庞大，以至于通过完整枚举来观测完整数据$\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ 在实践上或经济上都不可行。这是有限总体推断中的典型情况（Neyman，1938；Hansen 和Hurwitz，1943；Horvitz 和 Thompson，1952）。某些变量的测量成本可能很高，因此只能负担得起对少量实例 $i\in\mathcal{D}$进行观测，这种情况称为测量约束实验（Wang 等人，2017；Meng等人，2021；Zhang等人，2021；Imberg 等人，2022b）。另一个例子是，完整数据 $\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ 可用，但数据集的规模 N 非常大，以至于使用（1）–（2）进行估计 $\boldsymbol{\theta}$ 在计算上不可行（Ma 等人，2015；Drovandi，2017；Wang等人，2018年；Deldossi 和 Tommasi，2022；Dai，2022）。在这两种情况下，我们都可以寻找一种基于大小为 $n\ll N$ 的子集 $\mathcal{S}\subset\mathcal{D}$ 的近似解。

本文聚焦于基于不等概率抽样和逆概率加权的数据子抽样方法。具体而言，我们考虑如下形式的估计量

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}=\underset{\boldsymbol{\theta}\in\boldsymbol{\Omega}}{\arg\min}\,\hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta}),\tag{4}$$

$$\hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta})=\sum_{i\in\mathcal{S}}\frac{S_i}{\mu_i}\ell_i(\boldsymbol{\theta}),\tag{5}$$

其中，S_i 是元素 $i\in\mathcal{D}$ 被抽样机制选中的次数，μ_i 是相应的期望选中次数，而 $\mathcal{S}=\{i\in\mathcal{D}:\ S_i>0\}$ 是被选中元素构成的随机集合。可以看出，(5) 是全数据经验风险函数 (2) 的汉森-赫维茨估计量（Hansen 和 Hurwitz，1943）。因此，我们将 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 称为汉森-赫维茨经验风险最小化器。对于不放回抽样，(5) 与众所周知的 ℓ_0的 Horvitz-Thompson 估计量（$\boldsymbol{\theta}$）（Horvitz 和 Thompson，1952）一致。我们还注意到，只要对所有 $i\in\mathcal{D}$都有 $\mu_i>0$，那么 $\ell_{\boldsymbol{\mu}}(\boldsymbol{\theta})$就是 $\ell_0(\boldsymbol{\theta})$的无偏估计量；并且在一般正则条件下，$\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 是全数据参数$\boldsymbol{\theta}_0$ 的相合估计量（Binder，1983）。

一个重要的问题是，用于问题 (1)–(2) 的近似解(4)的子集 $\mathcal{S}$ used应如何选取，才能实现最佳性能。最优子抽样问题在调查抽样领域中有着悠久的传统，旨在对有限总体进行推断；参见例如 Neyman（1938）；Hájek（1959）；Cassel 等人（1976）；Brewer（1979）和 Bellhouse（1984）。然而，他们的工作主要关注有限总体标量特征的线性估计量。受现代技术发展的推动，近年来，最优子数据选择问题也重新受到关注，并扩展到更复杂的推断问题，如上文所述。相关例子包括用于大数据回归的杠杆抽样和近似数值线性代数方法（Ma 等人，2015，2020），用于二元和多项逻辑回归的最优子抽样算法（Wang 等人，2018；Yao 和Wang，2019），用于广义线性模型的最优子抽样算法（Ai等人，2021b；Zhang等人，2021；Yu 等人，2022），用于分位数回归的最优子抽样算法（Ai等人，2021a；Wang 和 Ma，2021），以及用于主动学习的最优子抽样算法（Imberg 等人，2020；Kossen 等人，2022；Zhan等人，2022）。然而，这些出版物中的大多数都具有高度算法化的视角，重点关注受限类别的模型和最优性准则。此外，许多提出的方法使用了众所周知存在缺陷的最优性准则（例如A-最优性），例如对数据测量尺度和模型参数化缺乏不变性。统一的最优子抽样设计理论仍然缺失。

我们提出了一套针对一般数据子抽样问题的最优设计理论，包括有限总体推断、参数密度估计，以及使用拟似然方法进行回归建模。我们推导了一大类最优性准则的最优性条件，包括A-、D-、E-、L- 和Kiefer的 Φ_q-最优性准则（Kiefer，1974年）。我们针对泊松抽样和多项抽样设计，提出了寻找最优抽样方案的算法。我们还从距离最小化的角度研究了最优设计，并证明了其与传统最优性准则的等价性。这自然引出了一个新型线性最优性准则类别，该类别具有良好的理论和实践性质，包括计算可处理性，以及在数据的仿射变换和模型的重新参数化下具有不变性。我们通过交通安全定义域中的一项应用，对所提出的方法和算法进行了说明。

我们首先在第 2 节简要回顾不等概率抽样和最优设计中的一些标准方法。第3节介绍数据子抽样问题的最优设计一般理论，包括寻找最优抽样方案的算法。第 4 节从最小化距离的角度讨论最优设计，并给出一些常见统计距离函数对应的最优设计。第 5 节对最优子抽样方法在实践中的实现进行讨论。第 6 节给出示例和实验。证明见附录 A。

2 预备知识

考虑一类实验 Ξ 以及相应的相合估计量$\hat{}\ \boldsymbol{\theta}_\xi, \xi \in \Xi$，其中未知参数为 $\boldsymbol{\theta}^*$。最优设计的目标是寻找一个实验 $\xi \in \Xi$，使估计量 $\boldsymbol{\theta}_\xi$的协方差$\hat{}$矩阵的某个适当函数 Φ 最小。例如， Φ 可以是其矩阵自变量的特征值之和或乘积，分别对应于 A-或 D-最优性（Atkinson和 Donev，1992），也可以是矩阵“大小”的某种其他度量。

在数据子抽样的背景下，实验由抽样设计和抽样方案 $\boldsymbol{\mu} = (\mu_1、\ldots\ldots、\ \mu_N)$ 的选择决定。对于第1节所概述的估计问题，我们希望针对某个合适的抽样设计族和目标函数 Φ： $\mathbb{R}^{p\times p} \to \mathbb{R}$，找到一个能使 $\Phi(\mathbf{Cov}(\boldsymbol{\theta}_{\boldsymbol{\mu}}))$ $\hat{}$minimises $\boldsymbol{\mu}$ 的抽样方案。一些常见的不等概率抽样设计将在第 2.1 节介绍。估计量 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 的近似协方差 $\hat{}$ 矩阵将在第2.2节给出，并在第2.3节简要回顾最优设计。

2.1 不等概率抽样设计

我们考虑这样的情形：个体元素 $i \in \mathcal{D}$ 根据不等概率抽样设计进行选择，即通过一种随机机制进行抽样，其中每个成员 $i \in \mathcal{D}$ 都有严格为正且可能各不相同的选择概率。沿用第1节中的记号，令 S_i 表示元素 $i \in \mathcal{D}$ 被抽样机制选中的次数，其中抽样可以是有放回的，也可以是无放回的；令 μ_i 表示相应的期望选择次数。我们用 n 表示子样本的期望样本量，用 $\mathcal{M}_n$ 表示$\boldsymbol{\mu} = (\mu_1、\ldots\ldots、\ \mu_N)$的相应定义域，即在指定的、具有（期望）样本量 n 的抽样设计族内，抽样方案 $\boldsymbol{\mu}$ 的可行取值集合。我们假设抽样按照以下抽样设计族之一进行：

> i) 有放回泊松抽样（PO-WR）： $S_1、\ldots、\ S_N$ 相互独立，且 $S_i \sim$服从泊松(μ_i)分布， $\mu_i > 0$。样本量 $\sum_{i\in\mathcal{D}} S_i$ 是随机的，其期望为 $\mathrm{E}[\sum_{i\in\mathcal{D}} S_i] = \sum_{i\in\mathcal{D}} \mu_i = n$。相应的定义域 $\mathcal{M}_n$由 $\boldsymbol{\mu}$ 给出：$\mathcal{M}_n = \{\boldsymbol{\mu} \in \mathbb{R}^N：\ \mu_i > 0$ 对所有 $i \in \mathcal{D}$ 和 $\sum_{i\in\mathcal{D}} \mu_i = n\}$成立。ii) 无放回泊松抽样（PO-WOR）： $S_1、\ldots、\ S_N$ 相互独立，且 $S_i \sim$服从伯努利(μ_i)分布，$\mu_i \in (0, 1]$。样本量 $\sum_{i\in\mathcal{D}} S_i$ 是随机的，其期望为 $\mathrm{E}[\sum_{i\in\mathcal{D}} S_i] = \sum_{i\in\mathcal{D}} \mu_i = n$。相应的定义域 $\mathcal{M}_n$由 $\boldsymbol{\mu}$ 给出：$\mathcal{M}_n = \{\boldsymbol{\mu} \in \mathbb{R}^N：\ \mu_i \in (0, 1]$ 对所有 $i \in \mathcal{D}$ 和 $\sum_{i\in\mathcal{D}} \mu_i = n\}$成立。iii) 多项式抽样（MULTI）：$(S_1、\ldots、\ S_N) \sim$服从多项式$(n、\ \boldsymbol{\mu}/n)$分布，$n \in \mathbb{N}、\ \mu_i > 0、\ \sum_{i\in\mathcal{D}} \mu_i = n$。抽样采用有放回方式，且样本量固定，即 $\sum_{i\in\mathcal{D}} S_i = n$。相应的定义域$\mathcal{M}_n$ 由 $\boldsymbol{\mu}$ 给出：$\mathcal{M}_n = \{\boldsymbol{\mu} \in \mathbb{R}^N：\ \mu_i > 0$ 对所有 $i \in \mathcal{D}$ 和 $\sum_{i\in\mathcal{D}} \mu_i = n\}$成立。

对于给定的样本量 n，泊松抽样设计和多项式抽样设计由均值向量 $\boldsymbol{\mu}$ 唯一确定。我们称，对于给定的样本量 n，此类抽样设计由抽样方案 $\boldsymbol{\mu}$ 索引。

还存在一些方法，可以在不放回且选择概率固定的情况下选择固定数量的元素，例如使用条件泊松抽样（Hájek，1981；Tillé，2006）。然而，这种方法在计算和分析上都难以处理，因此本文不予考虑。更多细节可参见，例如 Tillé（2006）和 Fuller（2009）。

2.2 Hansen-Hurwitz 经验风险最小化器的协方差矩阵

Binder(1983)表明，在适当的正则条件下，估计量 (4) 关于抽样机制的分布近似为高斯分布，均值

$$\mathrm{E}[\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}] = \boldsymbol{\theta}_0 + o(n^{-1/2}), \tag{6}$$

以及协方差矩阵

$$\mathbf{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}} - \boldsymbol{\theta}_0) = \boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) + o(n^{-1}), \quad \boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}. \tag{7}$$

这里， $o(n^{-1/2})$ 和 $o(n^{-1})$ 按元素解释，而 $\mathbf{H}(\boldsymbol{\theta}_0) = \frac{\partial^2 \ell_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ 是全数据经验风险函数 (2) 在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处的海森矩阵。

$$\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \mathbf{Cov}\left(\nabla_{\boldsymbol{\theta}} \hat{\ell}_{\boldsymbol{\mu}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) = \sum_{i,j \in \mathcal{D}} \frac{\mathrm{Cov}(S_i, S_j)}{\mu_i \mu_j} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_j(\boldsymbol{\theta}_0)^\mathsf{T} \tag{8}$$

$\hat{}$ 是相对于抽样机制的梯度 $\nabla_{\boldsymbol{\theta}} \ell_{\boldsymbol{\mu}}(\boldsymbol{\theta})$ 的协方差矩阵，在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处求值，且 $\boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})$。更多详情请参阅 Binder (1983) 和 Fuller (2009)。

根据第 2.1 节所述抽样设计的性质，矩阵 $\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ 可简化为

$$\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0) = \begin{cases} \sum_{i \in \mathcal{D}} \mu_i^{-1} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}, & \text{for PO-WR and MULTI designs, and} \\ \sum_{i \in \mathcal{D}} (\mu_i^{-1} - 1)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}, & \text{for PO-WOR.} \end{cases} \tag{9}$$

例如，参见 Tillé（2006）。为得到上述多项式抽样设计的结果，我们还使用了（3）。

2.3 最优设计

对于未知参数 $\boldsymbol{\theta}^*$，考虑一类实验 Ξ 以及相应的相合估计量 $\hat{}\,\boldsymbol{\theta}_\xi$， $\xi \in \Xi$，其协方差矩阵 $\boldsymbol{\Gamma}_\xi$不相同。理想情况下，我们希望找到一个实验 $\xi^* \in \Xi$，使得对于所有 $\xi \in \Xi$， $\boldsymbol{\Gamma}_\xi - \boldsymbol{\Gamma}_{\xi^*}$ 都是半正定的。然而，通常无法实现这种普遍最优性。因此，我们转而考虑定义在实对称半正定 $p \times p$ 矩阵集合上的函数 $\Phi\colon \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$，并寻求其最小化器 $\xi^* \in \Xi$。为了使 Φ 成为有意义的最优性度量，我们要求该函数关于 Loewner序单调，即

$$\Phi(\mathbf{U}) \geq \Phi(\mathbf{V}) \text{ for all } \mathbf{U}, \mathbf{V} \in \boldsymbol{S}_+^{p\times p} \text{ such that } \mathbf{U} \geq \mathbf{V}, \tag{10}$$

其中 $\mathbf{U} \geq \mathbf{V}$ 表示$\mathbf{U} - \mathbf{V}$ 是半正定的（Pukelsheim，1993）。

表 1 定义并总结了一些常用的最优性准则，包括 D-最优性准则（使协方差矩阵的行列式最小）、E-最优性准则（使协方差矩阵的最大特征值最小）和 L-最优性准则（使一组线性组合$\mathbf{L}^\mathsf{T}\hat{}\,\boldsymbol{\theta}_\xi$的平均方差最小）。L-最优性准则的两个重要特例是 A-最优性准则（使平均方差最小）和 c-最优性准则（使线性组合$\mathbf{c}^\mathsf{T}\hat{}\,\boldsymbol{\theta}_\xi$的方差最小）；当$\mathbf{L} = \mathbf{I}_{p\times p}$ 以及$\mathbf{L} = \mathbf{c}$ 时，分别可得到这两个准则，其中 $p \times 1$ 为向量$\mathbf{c}$（Silvey，1980；Atkinson 和 Donev，1992）。表 1 还列出了Φ_q-和 $\Phi_{q,\mathrm{A}}$-最优性准则，它们涵盖了表中的所有其他最优性准则。具体而言，当 $q = 0$时， Φ_q-最优性与 D-最优性一致；当 $q = 1$时，与 A-最优性一致；当 $q = \infty$ 时，则与 E-最优性一致（Kiefer，1974）。因此， Φ_q-最优性可用于在 A-、D-和 E-最优性之间进行插值。

A-、D-和 E-最优性准则有如下简单的几何解释。考虑随机集合$\mathcal{C}(\hat{}\,\boldsymbol{\theta}_\xi) := \{\boldsymbol{\theta} \in \mathbb{R}^p : (\boldsymbol{\theta} - \hat{}\,\boldsymbol{\theta}_\xi)^\mathsf{T}\boldsymbol{\Gamma}_\xi^{-1}(\boldsymbol{\theta} - \hat{}\,\boldsymbol{\theta}_\xi) \leq \chi^2_{p,\,\alpha}\}$，其中 $\chi^2_{p,\,\alpha}$ 是具有 p 个自由度的 χ^2 分布的 α 分位数。对于近似服从正态分布的估计量 $\boldsymbol{\theta}_\xi$，这定义了 $\boldsymbol{\theta}^*$ 在 $\mathbb{R}^p$ 中的一个（近似） $100 \times (1-\alpha)\%$ 椭球置信集。D-最优性在实验类别 Ξ中使该置信椭球的体积最小。E-最优性使其最长轴的长度最小，而 A-最优性则使包围该置信椭球的最小边界框（平行六面体）的对角线长度最小（Pronzato 和 Pázman，2013）。

另一种常用的最优性准则是 V-最优性准则，它使设计空间 $\mathcal{X}$ 上相对于某个测度 $\nu(\boldsymbol{x})$ 的平均预测方差最小化（Welch，1984）。这是一种线性最优性准则，因此可由针对矩阵 $\mathbf{L}$ 的 L-最优性准则涵盖，其中 $\mathbf{L}\mathbf{L}^\mathsf{T} = \int_{\mathcal{X}} \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T} d\nu(\boldsymbol{x})$（表 1）（Atkinson 和 Donev，1992）。在数据子抽样问题中，测度 $\nu(\boldsymbol{x})$ 的一个自然选择是关于$\{\boldsymbol{x}_i\}_{i\in\mathcal{D}}$的经验测度。

最优设计通常具有一项理想性质：在对数据进行非奇异仿射变换以及对模型重新参数化时保持不变。也就是说，最优设计及其所得估计量的统计性质不应取决于参数化方式的选择，也不应取决于建模前数据的缩放或编码。变换和参数化不变的最常见例子是 D-最优性准则。相比之下，A-最优性和 E-最优性准则对参数化方式或数据的变化较为敏感，因此不具备这种不变性（Atkinson 和 Donev，1992）。L-最优设计是否具有参数化和变换不变性，取决于 L-最优性准则的系数矩阵 $\mathbf{L}$ 是否针对问题的参数化方式和数据缩放进行了调整。第4.3节将讨论一些变换和参数化不变的线性最优性准则。

如需进一步了解详情，请参阅 Silvey(1980)、Atkinson和 Donev (1992)、Pukelsheim (1993) 以及 Pronzato 和 Pázman (2013)。

表 1. 最优设计中一些常见最优性准则的定义。 Φ 是定义在实对称半正定 $p \times p$ 矩阵集合上的实值函数， $\mathbf{\Gamma}$ 为估计量 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$的 $p \times p$ 协方差矩阵，$\lambda_{\max}(\mathbf{\Gamma})$为 $\mathbf{\Gamma}$的最大特征值，$\mathbf{c}$ 为非零 $p \times 1$ 向量，$\mathbf{A} = \mathbf{L}$ 为列向量为$\mathbf{a}_1, \ldots, \mathbf{a}_m$的非零 $p \times m$ 矩阵，而$\mathbf{I}_{p\times p}$为 $p \times p$ 单位矩阵。 $\mathcal{X}$是预测变量 $\boldsymbol{x}$ 的可能取值集合，而 $\boldsymbol{\varphi}: \mathcal{X} \to \mathbb{R}^p$ 是数据的特征映射。

最优性准则	**描述**	**目标函数** $\Phi(\mathbf{\Gamma})$
A-最优性	最小化平均方差， 最小化协方差矩阵的迹， 最小化特征值之和。	$\frac{1}{p}\sum_{i=1}^{p} \mathrm{Var}(\hat{\theta}_i) = \frac{1}{p}\mathrm{tr}(\mathbf{\Gamma})$
c-最优性	最小化线性组合的方差 或对比$\mathbf{c}^\mathsf{T}\hat{\boldsymbol{\theta}}$ 。	$\mathrm{Var}(\mathbf{c}^\mathsf{T}\hat{\boldsymbol{\theta}}) = \mathbf{c}^\mathsf{T}\mathbf{\Gamma}\mathbf{c} = \mathrm{tr}(\mathbf{\Gamma}\mathbf{c}\mathbf{c}^\mathsf{T})$
D-最优性	最小化广义方差， 最小化协方差矩阵的行列式， 最小化特征值的乘积。	$\det(\mathbf{\Gamma})^{1/p}$ 或 $\log\det(\mathbf{\Gamma})$
$\mathrm{D_A}$-最优性	最小化子集的广义方差 参数的线性组合集合， 组合，或对比$\mathbf{A}^\mathsf{T}\boldsymbol{\theta}$ 。	$\det(\mathbf{A}^\mathsf{T}\mathbf{\Gamma}\mathbf{A})$
E-最优性	最小化最大特征值， 最小化沿该方向的方差 最大的不确定性。	$\lambda_{\max}(\mathbf{\Gamma})$
L-最优性	最小化一组数据的平均方差 线性组合或对比的$\mathbf{L}^\mathsf{T}\boldsymbol{\theta}$ 。 $\mathbf{L} = \boldsymbol{c}$ $\Leftrightarrow$ c-最优性 $\mathbf{L} = \mathbf{I}_{p\times p}$ $\Leftrightarrow$A-最优性	$\frac{1}{m}\sum_{i=1}^{m}$ 方差$(\mathbf{a}_i^\mathsf{T}\hat{\boldsymbol{\theta}}) = \frac{1}{m}\mathrm{tr}(\mathbf{\Gamma}\mathbf{L}\mathbf{L}^\mathsf{T})$
V-最优性	通过最小化平均预测方差 相对于 $\mathcal{X}$ 上的测度 $d\nu(\boldsymbol{x})$，假设 一个线性模型 $\hat{y} = \boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T}\hat{\boldsymbol{\theta}}$。	$\int_{\mathcal{X}} \mathrm{Var}(\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T}\hat{\boldsymbol{\theta}})d\nu(\boldsymbol{x}) =$ $\mathrm{tr}\left(\mathbf{\Gamma}\int_{\mathcal{X}}\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\mathsf{T}d\nu(\boldsymbol{x})\right)$
$\Phi_{q,\mathbf{A}}$-最优性 $q \in [0,\infty]$	Φ_0 $\Leftrightarrow$ D-最优性 $\Phi_{0,\ \mathbf{A}}$ $\Leftrightarrow \mathrm{D_A}$-最优性 Φ_1 $\Leftrightarrow$A-最优性 $\Phi_{1,\ \mathbf{L}}$ $\Leftrightarrow$L-最优性 Φ_∞ $\Leftrightarrow$ E-最优性	$\Phi_{q,\mathbf{A}}(\mathbf{\Gamma}) = \frac{1}{m}\mathrm{tr}[(\mathbf{A}^\mathsf{T}\mathbf{\Gamma}\mathbf{A})^q]^{1/q}, q \in (0,\infty)$ $\Phi_{0,\mathbf{A}}(\mathbf{\Gamma}) = \lim_{q\downarrow 0}\Phi_{q,\mathbf{A}}(\mathbf{\Gamma})$ $\Phi_{\infty,\mathbf{A}}(\mathbf{\Gamma}) = \lim_{q\uparrow\infty}\Phi_{q,\mathbf{A}}(\mathbf{\Gamma})$ $\Phi_q(\mathbf{\Gamma}) = \Phi_{q,\ \mathbf{A}}(\mathbf{\Gamma})$，$\mathbf{A} = \mathbf{I}_{p\times p}$

3 最优子抽样设计

本节在可微性假设下，针对一类广义最优性准则介绍最优抽样方案。 Φ-最优性定义于第3.1节，该节还介绍了三个重要引理。泊松和多项式抽样设计的最优性准则见第3.2节，而寻找最优抽样方案的算法见第3.3节。

˄ 首先，我们注意到，如式 (7) 所示，估计量 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 的近似协方差矩阵 $\mathbf{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$ 通常取决于完整数据 $\{(\boldsymbol{x}_i,\ \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$和全数据参数 $\boldsymbol{\theta}_0$。显然，如果在设计阶段能够获得这些信息，就不需要进行子抽样。然而，这是最优设计中的一个普遍问题，并非我们所研究设置中特有的问题，因此并不构成本文所提出理论的重大局限。本节和第4节中，我们将假设能够获得这些信息，同时牢记由此得到的理论最优设计通常无法在实践中找到。

我们将第 5 节留作讨论如何在实践中实施最优子抽样设计。

在全文中，我们假设满足正则条件，使得式（7）成立，并且 $\mathbf{H}(\boldsymbol{\theta}_0)$ 是满秩的。除非另有说明，否则所有向量均假设为列向量。令 $||\mathbf{u}||_2^2 = \mathbf{u}^\intercal\mathbf{u}$ 表示向量 $\mathbf{u}$ 的欧几里得范数。另请回忆，$\psi_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta})$。

3.1 最优性准则

通过一个 o 最优抽样方案 $\boldsymbol{\mu}^*$，我们指的是以下内容。 翼:

定义 1(Φ最优性)。考虑一个函数 $\Phi: \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$，它关于 *Loewner*序是单调的，即满足(10) 。再考虑一族由抽样方案 $\boldsymbol{\mu}$索引的不等概率抽样设计（例如 *PO-WR*、*PO-WOR* 或 *MULTI*）。令期望样本量E $[\sum_{i\in\mathcal{D}} S_i] = n$ 固定，并令 $\mathcal{M}_n$ 表示 $\boldsymbol{\mu}$的相应定义域。若抽样方案 $\boldsymbol{\mu}^*$ 满足以下条件，则称其为 Φ-最优:

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu}\in\mathcal{M}_n}{\arg\min}\, \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)),$$

ˆ其中 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 是 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$的近似协方差矩阵，如(7)所示。

寻找 Φ-最优抽样方案可归结为一个非线性且可能为非凸的约束优化问题，其定义在 $\mathbb{R}^N$中的一个 $(N-1)$ 维超平面上。当 N 较小时，可以使用数值优化方法求解该问题，但对于大型数据集，这通常并不可行。因此，我们需要最优设计理论，以便在 N 较大时，能够据此设计高效算法来寻找最优抽样方案。为使问题易于处理，我们将仅考虑在最优值 $\boldsymbol{\mu}^*$的邻域内关于$\boldsymbol{\mu}$ 可微的最优性准则 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$。下面给出三个重要引理。

引理 1（链式法则）。考虑一个函数 $\Phi: \boldsymbol{S}_+^{p\times p} \to \mathbb{R}^p$，并假设 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 在某个点 $\boldsymbol{\mu}^*$的邻域内可微。则 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 μ_i 的偏导数为

$$\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = \operatorname{tr}\left(\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))\frac{\partial\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial\mu_i}\right), \tag{11}$$

其中 $\boldsymbol{\phi}(\mathbf{U}) = \frac{\partial\Phi}{(\mathbf{U})\,\partial\mathbf{U}}$ 是 $p\times p$ 矩阵导数，表示 Φ 相对于其矩阵参数的导数；而 $\frac{\partial\boldsymbol{\Gamma}}{(\boldsymbol{\mu};\boldsymbol{\theta}_0)\,\partial\mu_i}$是 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 相对于 μ_i 的逐元素导数。

进一步假设

i) $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 随 $\mu_1,\ldots,\mu_N$按洛文纳序单调递减，即对于满足 $\boldsymbol{\mu}_1 \le \boldsymbol{\mu}_2$（逐元素）的每一对向量 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2 \in \mathbb{R}^N_{>0}$，$\boldsymbol{\Gamma}(\boldsymbol{\mu}_1;\boldsymbol{\theta}_0) - \boldsymbol{\Gamma}(\boldsymbol{\mu}_2;\boldsymbol{\theta}_0)$ 均为半正定，且

ii) Φ 在*Loewner*序下是单调的，即(10) 成立。

然后，矩阵 $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 是半正定的，并且存在一个实矩阵 $\mathbf{L}(\boldsymbol{\mu}$<style id='18'>; $\boldsymbol{\theta}_0)$，使得$\mathbf{L}(\boldsymbol{\mu}$<style id='28'>; $\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu}$; $\boldsymbol{\theta}_0)^\intercal = \boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu}$<style id='50'>; $\boldsymbol{\theta}_0))$。

引理 1 的第一部分由矩阵微分演算中的链式法则以及 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 的对称性得出，第二部分则由 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 和 Φ 的单调性假设得出。矩阵 $\mathbf{L}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$ 例如可以通过求 $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 的矩阵平方根获得；当 $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 满秩时，也可以通过乔列斯基分解获得。引理 2 中给出了一些示例。

引理 2 ($\boldsymbol{\mu}$-可微 Φ-最优性准则)。考虑一个 *PO-WR*、*PO-WOR* 或 *MULTI* 设计，并假设$\mathbf{H}(\boldsymbol{\theta}_0)$ 满秩。设$\mathbf{c}$ 为一个非零的 $p\times 1$ 向量，$\mathbf{L}$ 为一个非零的 $p\times m$ 矩阵，$\lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 为 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$的最大特征值，并且 $\boldsymbol{v}_{\boldsymbol{\mu}}$为相应的特征向量。令 $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 按照引理 *1* 中的方式定义。则以下结论成立:

a) $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$关于 $\boldsymbol{\mu}$ 且 $\frac{\partial\boldsymbol{\Gamma}(\boldsymbol{\mu}\,\boldsymbol{\theta}_0}{\partial\mu_i} = -\mu_i^{-2}$*;) H*$(\boldsymbol{\theta}_0)^{-1}\psi_i(\boldsymbol{\theta}_0)\psi_i(\boldsymbol{\theta}_0)^\intercal\mathbf{H}(\boldsymbol{\theta}_0)^{-1}$ 可微，前提是 $\mu_i > 0$*. b) D*-最优性目标函数 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)) = \log\det(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 且 $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}$ 可微，前提是 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$为满秩*.c) E*-最优性目标函数 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 且$\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \boldsymbol{v}_{\boldsymbol{\mu}}\boldsymbol{v}_{\boldsymbol{\mu}}^\intercal$ 可微，前提是 $\lambda_{\max}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 的重数为 *1.*

d) L-最优性的目标函数 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \mathrm{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}\mathbf{L}^\mathsf{T})$ 关于 $\boldsymbol{\mu}$ 可微，且$\phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \mathbf{L}\mathbf{L}^\mathsf{T}$。特别是，这对 *A*-最优性（$\mathbf{L} = \mathbf{I}_{p\times p}$）和 *c*-最优性（$\mathbf{L} = \mathbf{c}$）均成立。

e) Φ_q最优性目标函数 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = tr(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^q)^{1/q}$ 关于 $\boldsymbol{\mu}$ 在 $q \in (0,\infty)$ 上可微，并且 $\phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = tr(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^q)^{1/q-1}\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{q-1}$，前提是 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 满秩。

结合引理 1 和引理 2 的结果，我们得到如下结论：

引理 3 （偏导数 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$）。考虑一个 *PO-WR*、*PO-WOR* 或 *MULTI* 设计。还考虑一个函数 Φ： $\boldsymbol{S}_+^{p\times p} \to \mathbb{R}$，使得 Φ 关于 *Loewner*序是单调的。假设 $\mathbf{H}(\boldsymbol{\theta}_0)$ 是满秩的，并且 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 在某点 $\boldsymbol{\mu}^*$ 的邻域内可微。令 $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 如引理 *1* 所定义。则

$$\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = -\mu_i^{-2}\left|\left|\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|_2^2.$$

3.2 最优性条件

利用引理1–3的结果，我们在命题1中给出了在可微性假设下，关于 a Φ-最优性准则的泊松和多项抽样设计的最优性条件。

命题1 (Φ-最优性条件)。考虑大小为（期望）n的 *PO-WR*、*PO-WOR* 或 *MULTI* 设计族。另考虑函数 Φ：$\boldsymbol{S}_+^{p\times p} \to \mathbb{R}$，使得 Φ 关于*Loewner*序是单调的。假设 $\mathbf{H}(\boldsymbol{\theta}_0)$ 为满秩，并且 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 在某个点 $\boldsymbol{\mu}^*$的邻域内可微。令$\mathbf{L}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$ 按照引理 *1* 定义，并且

$$c_i = \left|\left|\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|_2^2. \tag{12}$$

则有：

a) $\boldsymbol{\mu}^*$ 是 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 的驻点，对于大小为 n 的 *PO-WR* 或 *MULTI* 设计，如果

$$\mu_i^* = n\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}}\sqrt{c_j}} \quad \textit{for all } i \in \mathcal{D}. \tag{13}$$

b) $\boldsymbol{\mu}^*$ 是 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 对于大小为 n 的 *PO-WOR* 设计，如果

$$\mu_i^* \le 1 \qquad \textit{for all } i \in \mathcal{D}, \tag{14a}$$

$$\mu_i^* = (n - n_\mathcal{E})\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}\setminus\mathcal{E}}\sqrt{c_j}} \quad \textit{for all } i \in \mathcal{D}\setminus\mathcal{E}, \tag{14b}$$

$$\sqrt{c_i} \ge \sqrt{c_j}/\mu_j^* \qquad \textit{for all } i \in \mathcal{E} \textit{ and } j \in \mathcal{D}\setminus\mathcal{E}, \tag{14c}$$

其中 $\mathcal{E} = \{i \in \mathcal{D}:\ \mu_i^* = 1\}$和 $n_\mathcal{E} = |\mathcal{E}|$。

因此，如果 $\boldsymbol{\mu}^*$ 根据 *a)* 或 *b)* 满足最优性条件，且 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 是凸函数，那么 $\boldsymbol{\mu}^*$ 就是 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 的全局极小值点。

注意，根据引理 1，只要目标函数在 $\boldsymbol{\mu}^*$ 处可微，命题1中的矩阵 $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ 就存在。不过，它未必唯一，而且可能同时依赖于 $\boldsymbol{\mu}^*$ 和 $\boldsymbol{\theta}_0$。一些例子见引理2。对于线性最优性准则，矩阵 $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ 不依赖于 $\boldsymbol{\mu}^*$，但可能依赖于全数据参数 $\boldsymbol{\theta}_0$；更多讨论和例子见第4.3节。

命题1的结果可分别通过 a) 中的拉格朗日乘数法和 b) 中的 Karush-Kuhn-Tucker 条件，由引理 3 得出。我们将在命题 2 中证明，D-和 L-最优性准则关于 $\boldsymbol{\mu}$ 是凸的，因此可以据此推导出全局最优性。

命题 2 （D-和 L-最优性准则的凸性）。考虑期望大小为 n 的 *PO-WR*、*PO-WOR* 或多项式抽样设计族。假设$\mathbf{H}(\boldsymbol{\theta}_0)$ 满秩。那么

a) L-最优性准则关于 $\boldsymbol{\mu}$ 是凸的。

进一步假设 $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$（在(8) 中定义）对每个 $\boldsymbol{\mu} \in \mathcal{M}_n$ 都是正定的。那么

b) D-最优性准则关于 $\boldsymbol{\mu}$是（对数）凸的。

命题 2 中的第一个假设是确保 $\mathbf{H}(\boldsymbol{\theta}_0)$ 的逆矩阵存在，并且近似协方差矩阵 $\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ 有明确定义所必需的。对于 D-最优性准则，我们还需要 $\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ 满秩；如果满足关于 $\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ 的附加假设，这一点便可得到保证。我们注意到，这更多是关于模型和数据的假设，而非关于抽样设计的假设。此外，在大多数情况下，命题 2 中的两个假设都成立。然而，在（多变量）回归分析中，如果模型矩阵 $\mathbf{X}$ 或响应矩阵 $\mathbf{Y}$ （即行分别为 $\boldsymbol{x}_i^\mathsf{T}$ 和 $\boldsymbol{y}_i^\mathsf{T}$ 的矩阵）存在列线性相关，就会出现违反这些假设的情况。另一个例子是存在完全分离的逻辑回归，即结果可由预测变量线性分离。$\mathbf{H}(\boldsymbol{\theta}_0)$ 也可能是满秩的，而 $\mathbf{V}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$ 在 $\mathcal{M}_n$ 上却是秩亏的。在这种情况下，L-最优性准则仍有明确定义，而 D-最优性准则则没有。对此类问题有多种解决方法，例如从数据中移除冗余列，使用岭惩罚来避免海森矩阵秩亏（Hastie，2020），或者使用所谓的 $\mathrm{D_A}$-最优性将 D-最优性准则限制在参数的一个子集上（表 1）（Sibson，1974）。不过，通过仔细构建模型，通常可以避免大多数此类情况。

即使目标函数是凸函数，也可能不存在可行的全局最优解，因为定义域 $\mathcal{M}_n$ 并非闭集。对于 L-最优性准则，如果式 (12) 中的 c_i 对某个 $i \in \mathcal{D}$等于零，就会出现这种情况。在这种情况下，目标函数不依赖于相应的 μ_i，且关于 μ_i 的偏导数等于零。最优选择是相应地令 $\mu_i = 0$，但这不是一个可行解。无论选择何种$\mu_i > 0$，总可以通过减小 μ_i，并将重新获得的概率质量最优地分配给 $\mathcal{D}$中的其余元素，来改善目标函数的值。通过施加额外限制，确保对所有 $i \in \mathcal{D}$都有 $\mu_i \geq \mu_{\min}$，且满足某个 $\mu_{\min} \in (0,\ n/N)$，即可保证存在可行的全局最优解。第 5 节提出了一种替代方案，无需明确指定下界 $\mu_{\min}$，但仍能确保存在满足 $\mu_i > 0$的可行解。

3.3 最优抽样方案

本小节介绍用于求解最优抽样方案的算法。首先考虑关于 $p \times m$ 矩阵 $\mathbf{L}$ 的线性最优性准则。在这种情况下，PO-WR 和 MULTI 设计的最优抽样方案有闭式解，由式 (12)–(13) 给出，其中 $\mathbf{L}(\boldsymbol{\mu}^*; \boldsymbol{\theta}_0) = \mathbf{L}$，前提是相应的 $c_i > 0$对所有 $i \in \mathcal{D}$均成立。具体而言，A-最优性对应于 $\mathbf{L}(\boldsymbol{\mu}^*; \boldsymbol{\theta}_0) = \mathbf{I}_{p \times p}$，而 c-最优性对应于 $\mathbf{L}(\boldsymbol{\mu}^*; \boldsymbol{\theta}_0) = \mathbf{c}$。对于 PO-WOR，可能需要进行简单调整，以确保得到满足 $\mu_i \leq 1$ 的可行解（算法 1）。

Alg法 1. L-最　p优抽样　pling 方案：泊松抽样和多项式抽样　　pling designs.

输入：索引集 $\mathcal{D}$、（期望）样本量 n、非零 $p \times m$ 矩阵 $\mathbf{L}$、海森矩阵 $\mathbf{H}(\boldsymbol{\theta}_0)$、梯度 $\{\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\}_{i \in \mathcal{D}}$、抽样设计族（PO-WR、PO-WOR 或 MULTI）。

1:令 $c_i = ||\mathbf{L}^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2$ **对所有** $i \in \mathcal{D}$。**2:如果任意** $c_i = 0$ **则 3:停止。可行解不存在。 4: 否则 5:令** $\mu_i^* = n \frac{\sqrt{c_i}}{\sum_{j \in \mathcal{D}} \sqrt{c_j}}$ **对所有** $i \in \mathcal{D}$。**6: 如果 PO-WOR 则7: 当任意** $\mu_i^* > 1$ **时8:令** $\mathcal{E} = \{i \in \mathcal{D} : \mu_i^* \geq 1\}$**和** $n_\mathcal{E} = |\mathcal{E}|$。**9: 令** $\mu_i^* = \begin{cases} 1 & \text{如果 } i \in \mathcal{E}, \\ (n - n_\mathcal{E}) \frac{\sqrt{c_i}}{\sum_{j \in \mathcal{D} \setminus \mathcal{E}} \sqrt{c_j}} & \text{如果 } i \in \mathcal{D} \setminus \mathcal{E}. \end{cases}$**10: 结束当循环 11: 结束如果 12: 返回最优抽样方案** $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_N^*)$。 **13: 结束如果**

利用命题1和算法1的结果，我们在算法2中提出了一种用于寻找非线性最优性准则下最优抽样方案的迭代算法。该算法以初始抽样方案作为输入，并通过将目标函数局部近似为线性最优性准则来求解一系列凸优化问题。当两次连续迭代之间目标函数的相对改进小于某个预先指定的容差水平 ϵ （例如，$\epsilon = 10^{-3}$）时，算法因收敛而终止。如果目标函数的值在迭代之间增加，算法也可能因发散而终止。如果该算法

收敛时，它会收敛到由算法 1 定义的函数 $h(u)$的不动点：$\mathbf{L} = \mathbf{L}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)$，根据命题1，该不动点是 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0))$ 的驻点。对于L-最优性，该方法是精确的，并在单次迭代内终止。对于L-最优性之外的情形，该算法不一定收敛；即使收敛，除非问题是凸的，否则也不一定收敛到全局最优解。该算法对于非线性最优性准则的性能将在第 6 节中进行评估。

算法2。 不动点迭代。

输入：索引集 $\mathcal{D}$、（期望）样本量 n、最优性准则 Φ、海森矩阵$\mathbf{H}(\boldsymbol{\theta}_0)$、梯度 $\{\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\}_{i\in\mathcal{D}}$、初始抽样方案 $\boldsymbol{\mu}_0$、抽样设计族（PO-WR、PO-WOR或 MULTI）、最大迭代次数T、容差参数 $\epsilon > 0$。

1: 对t = 1、...、**T执行 2:令** $\mathbf{L}_t$ **为一个矩阵，使得**$\mathbf{L}_t\mathbf{L}_t^\intercal = \phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}_{t-1}; \boldsymbol{\theta}_0))$。 **3:令** $c_i = \|\mathbf{L}_t^\intercal \mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\|_2^2$ **对所有** $i \in \mathcal{D}$ **成立。4:如果存在任何** $c_i = 0$**，则 5:停止。在迭代过程中遇到不可行解。6:否则 7:根据算法1，关于**$\mathbf{L} = \mathbf{L}_t$ **找到L 最优抽样方案** $\boldsymbol{\mu}_t$ **。 8: 如果目标函数的值增加，则 9: 停止。算法发散。 10: 否则，如果目标函数的相对改进** $< \epsilon$**，则11: 算法收敛。返回** $\boldsymbol{\mu}^* = \boldsymbol{\mu}_t$。 **12: 结束 if 13: 结束 if 14: 结束 for**

4 从距离最小化的角度看最优子抽样设计

回顾第1节中介绍的数据子抽样的总体目标：为原本难以处理的问题（1）–（2）寻找一个近似解。因此，在此背景下，最优设计的一个自然目标是，对于某个合适的统计距离函数d：$\boldsymbol{\Omega} \to \mathbb{R}_+$，最小化估计量 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 与全数据参数 $\boldsymbol{\theta}_0$之间的预期ˆ ˆ距离 $\mathrm{E}[d(\boldsymbol{\theta}_{\boldsymbol{\mu}})]$ 。在第4.1节中，我们定义了一类用于最小化预期距离的最优性准则，并讨论其与传统最优性准则的关系。第4.2节介绍了一些具体示例，并在第4.3节讨论不变性性质。

4.1 d-最优性

考虑一个统计距离函数 $d(\boldsymbol{\theta})$，满足 $d(\boldsymbol{\theta}) \geq 0$ 对所有 $\boldsymbol{\theta} \in \boldsymbol{\Omega}$均成立，且仅当 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$时等号成立。为便于解析和计算，我们还要求该距离函数二阶可微，并令$\mathbf{H}_d(\boldsymbol{\theta}) = \frac{\partial^2 d(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\intercal}$ 表示 $d(\boldsymbol{\theta})$的海森矩阵。我们有如下结果：

ˆ 引理 *4*（ $d(\boldsymbol{\theta})$ 的泰勒展开）。令 $\boldsymbol{\theta}_0$ 和 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 分别根据 *(1)–(2)* 及 *(4)–(5)* 定义。假设 *(6)–(7)* ˆ 成立，并且对于某个 $\delta > 0$， $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 具有有界的 $2 + \delta$ 阶矩。考虑函数 $d : \boldsymbol{\Omega} \to \mathbb{R}_+$，使得 $d(\boldsymbol{\theta}) = 0$ 当且仅当 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$。假设 $d(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0$ 的某个邻域内二阶可微，且 $H_d(\boldsymbol{\theta}_0)$ 非零。则

$$\mathrm{E}[d(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})] = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0)\right) + o(n^{-1}).$$

引理 4 的结果由在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处对 $d(\boldsymbol{\theta})$ 进行泰勒展开以及二次型的性质得到。基于这一结果，我们定义如下的一类期望距离最小化最优性准则：

定义 2 (*d*-最优性)。 考虑一个函数 d： $\boldsymbol{\Omega} \to \mathbb{R}_+$，它满足引理 *4* 的条件。再考虑一个由抽样方案$\boldsymbol{\mu}$索引的不等概率抽样设计族（例如 *PO-WR*、*PO-WOR* 或 *MULTI*）。设期望样本量 $\mathrm{E}[\sum_{i\in\mathcal{D}} S_i] = n$ 固定，并令 $\mathcal{M}_n$ 表示相应的 $\boldsymbol{\mu}$ 的定义域。若抽样方案 $\boldsymbol{\mu}^*$ 相对于统计距离函数 $d(\boldsymbol{\theta})$ 满足以下条件，则称其为 *d*-最优：

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu}\in\mathcal{M}_n}{\arg\min}\,\mathrm{tr}\left(\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0)\right).$$

我们将这一最优性准则称为距离的 d-最优性，须注意不要将其与第2.3节介绍的 D-最优性准则混淆。我们认识到，d-最优性准则是一个线性最优性准则，其中 $\mathbf{L}\mathbf{L}^\intercal = \mathbf{H}_d(\boldsymbol{\theta}_0)$。事实上，有如下等价性结果：

命题3 （d-最优性与 Φ-最优性之间的等价关系）。

a) 考虑函数 $d: \boldsymbol{\Omega} \to \mathbb{R}_+$，它满足引理 *4* 的条件，并记$\mathbf{H}_d(\boldsymbol{\theta})$ 为$d(\boldsymbol{\theta})$的海森矩阵。假设抽样方案 $\boldsymbol{\mu}^*$ 关于距离函数 $d(\boldsymbol{\theta})$是 d-最优的。那么，存在一个实矩阵$\mathbf{L}$ 使得$\mathbf{L}\mathbf{L}^\intercal = \mathbf{H}_d(\boldsymbol{\theta}_0)$，且关于L是 L-最优的。

b) 设 $\Phi: \boldsymbol{S}_+^{p\times p} \to \mathbb{R}$ 和$\mathbf{L}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$ 如引理 *1* 中所定义，并假设 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$在其最优参数 $\boldsymbol{\mu}^*$的某个邻域内可微。那么，$\boldsymbol{\mu}^*$ 关于距离函数 $d(\boldsymbol{\theta}) = ||\mathbf{L}(\boldsymbol{\mu}^*;\ \boldsymbol{\theta}_0)^\intercal(\boldsymbol{\theta}-\boldsymbol{\theta}_0)||_2^2$是 d-最优的。

命题3可由定义和命题1的最优性条件立即推出。根据这一结果，ˆ 任意 Φ-最优性准则都可以视为：针对某一特定的距离函数，最小化估计量 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 与全数据参数 $\boldsymbol{\theta}_0$ 之间的期望距离。例如，A-最优性等价于 d-最优性，其中$d(\boldsymbol{\theta}) = ||\boldsymbol{\theta}-\boldsymbol{\theta}_0||_2^2$。除了线性最优性准则之外，所诱导的距离函数可能是隐式的，并且取决于 Φ-最优抽样方案 $\boldsymbol{\mu}^*$。例如，E-最优性等价于 d-最优性，其中 $d(\boldsymbol{\theta}) = ||\boldsymbol{v}_{\boldsymbol{\mu}^*}^\intercal(\boldsymbol{\theta}-\boldsymbol{\theta}_0)||_2^2$，其中 $\boldsymbol{v}_{\boldsymbol{\mu}^*}$ 是对应于 $\boldsymbol{\Gamma}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ 的最大特征值的特征向量，而 $\boldsymbol{\mu}^*$ 是相应的 E-最优抽样方案。在这种情况下，只有在已知 E-最优抽样方案的情况下，才能计算 d-最优性准则的距离函数。

4.2 一些距离最小化设计

接下来我们将展示如何利用d-最优性推导出一类新颖的线性最优性准则，这类准则具有良好的理论性质，包括变换不变性和参数化不变性。考虑以下自然产生于数据子抽样应用、且在统计学中常见的统计距离函数：

i) 经验风险距离：由于 $\boldsymbol{\theta}_0$ 被定义为全数据经验风险 (2) 的最小化器，因此我们可以通过经验风险的取值来度量参数值 $\boldsymbol{\theta}$ 与全数据参数 $\boldsymbol{\theta}_0$ 之间的距离。我们将 $\boldsymbol{\theta}$ 与 $\boldsymbol{\theta}_0$ 之间的经验风险距离定义为 $d_{\mathrm{ER}}(\boldsymbol{\theta}) = \ell_0(\boldsymbol{\theta}) - \ell_0(\boldsymbol{\theta}_0)$。 ii) 库尔贝克-莱布勒发散：考虑一个随机向量 $\boldsymbol{Y}$，其概率密度函数为 $f_{\boldsymbol{\theta}}(\boldsymbol{y})$，累积分布函数为 $F_{\boldsymbol{\theta}}(\boldsymbol{y})$。令 $\mathcal{Y}$ 表示 $\boldsymbol{Y}$的定义域。 $f_{\boldsymbol{\theta}}$ 与 $f_{\boldsymbol{\theta}_0}$ 之间的库尔贝克-莱布勒发散定义为 KL ($f_{\boldsymbol{\theta}_0}||f_{\boldsymbol{\theta}}) = \int_{\mathcal{Y}} \log\ ^{f_{\boldsymbol{\theta}_0}(\boldsymbol{y})}\ _{f_{\boldsymbol{\theta}}(\boldsymbol{y})}\ dF_{\boldsymbol{\theta}_0}(\boldsymbol{y})$。为纳入协变量，我们将 $\boldsymbol{\theta}$ 与 $\boldsymbol{\theta}_0$ 之间的库尔贝克-莱布勒距离定义为 $d_{\mathrm{KL}}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{D}} \int_{\mathcal{Y}} \log\ ^{f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)}\ _{f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)}\ dF_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)$。 iii) 马氏距离：考虑定义在 $\mathbb{R}^p$ 上的一个概率分布，其均值向量为 $\boldsymbol{\gamma}$，协方差矩阵为$\boldsymbol{\Sigma}$。点 $\boldsymbol{\theta} \in \mathbb{R}^p$ 与均值 $\boldsymbol{\gamma}$之间的马氏距离为 $\sqrt{(\boldsymbol{\theta}-\boldsymbol{\gamma})^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\gamma})}$。我们将 $\boldsymbol{\theta}$ 与 $\boldsymbol{\theta}_0$ 之间相对于实对称正定离散矩阵 $\boldsymbol{\Sigma}$ 的平方马氏距离定义为 $d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta}) = (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\intercal\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$。

对于马氏距离，离散矩阵 $\boldsymbol{\Sigma}$ 有四种自然的选择：

ˆ iii.a) $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$， $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 的近似协方差矩阵。

iii.b) $\boldsymbol{\Sigma} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}$，对于参数模型而言，这是对 $\boldsymbol{\theta}_0$协方差矩阵的估计，其中 $\boldsymbol{\theta}_0$被视为某个潜在超总体参数 $\boldsymbol{\theta}^*$ 的估计量。在此情况下，$\mathbf{H}(\boldsymbol{\theta}_0)$也称为观测费Fisher 信息矩阵，通常记作$\mathbf{I}(\boldsymbol{\theta}_0)$（Efron 和 Hinkley，1978）。

iii.c) $\boldsymbol{\Sigma} = \tilde{\mathbf{H}}(\boldsymbol{\theta}_0)^{-1}$，其中 $\tilde{\mathbf{H}}(\boldsymbol{\theta}_0)$ 对参数模型 $f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$定义为 $\tilde{\mathbf{H}}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{y}\sim f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x})}[\mathbf{H}(\boldsymbol{\theta}_0)]$。在$\tilde{}$这种情况下，$\mathbf{H}(\boldsymbol{\theta}_0)$也称为期望 Fisher 信息矩阵，通常记作 $\mathcal{I}(\boldsymbol{\theta}_0)$（Efron 和 Hinkley，1978）。

iii.d) $\boldsymbol{\Sigma} = \mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}$，其中

$$\mathbf{V}(\boldsymbol{\theta}_0) = \sum_{i\in\mathcal{D}} \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\intercal, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}). \tag{15}$$

这种对矩阵 $\boldsymbol{\Sigma}$ 的选择，对应于 $\boldsymbol{\theta}_0$ 的协方差矩阵的“稳健估计量”或“夹心估计量”；在半参数模型或可能设定错误的参数模型下，将其视为某个潜在超总体参数 $\boldsymbol{\theta}^*$ 的估计量（Stefanski 和 Boos，2002年）。

我们相应地定义 d_{ER}-、 d_{KL}-和 $d_{\boldsymbol{\Sigma}}$-最优性，即将距离函数取为下标所指示的距离函数的d-最优性。我们还将 d_{I}-、 $d_{\mathcal{I}}$-和 d_{S}-最优性定义为离散矩阵 $\boldsymbol{\Sigma}$ 分别取iii.b)（观测信息矩阵的逆）、iii.c)（期望信息矩阵的逆）和 iii.d)（夹心方差估计量）时的 $d_{\boldsymbol{\Sigma}}$-最优性。

请注意， d_{KL}-最优性和 $d_{\mathcal{I}}$-最优性仅针对参数模型定义，而 d_{ER}-、 d_{I}-和 d_{S}-最优性也适用于半参数方法和分布无关方法，包括有限总体特征的估计。对于回归问题，基于经验风险距离的d-最优性准则（即 d_{ER}-最优性）与V-最优性准则（表 1，第2.3节）密切相关。事实上，当 $\nu(\boldsymbol{x})$是 $\{\boldsymbol{x}_i\}_{i\in\mathcal{D}}$上的经验测度时，这两个最优性准则对于普通最小二乘回归是等价的。

ˆ 马氏距离与 $\boldsymbol{\Sigma}=\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$是通过考察 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 作为全数据参数 $\boldsymbol{\theta}_0$的估计量时的不确定性而得到的。相比之下，我们在上文iii.b)–iii.d)中引入离散矩阵的动机来自超总体视角，在这种视角下， $\boldsymbol{\theta}_0$ 被视为某个潜在参数 $\boldsymbol{\theta}^*$ 的估计量（参见 Hartley 和 Sielken，1975）。随后，离散矩阵 $\boldsymbol{\Sigma}$ 的不同选择自然源于与全数据参数 $\boldsymbol{\theta}_0$ 相关的不同不确定性度量（参见 Stefanski 和 Boos，2002年）。不过，我们强调，此处采用超总体视角纯粹是修辞性的。即使无意进行超总体推断，所得距离函数也同样有效。命题4及第4.3节将进一步阐明这些特定距离函数和离散矩阵选择的意义。

命题4(d_{ER}- d_{KL}-、 $d_{\boldsymbol{\Sigma}}$-最优性以及与 L-最优性的等价性)。

a) 关于经验风险距离的 d-最优性等价于关于满足 $\mathbf{L}\mathbf{L}^{\mathsf{T}}=\mathbf{H}(\boldsymbol{\theta}_0)$ 的 $p\times p$矩阵$\mathbf{L}$ 的 L-最优性。

b) 关于马氏距离的 d-最优性等价于关于满足 $\mathbf{L}\mathbf{L}^{\mathsf{T}}=\boldsymbol{\Sigma}^{-1}$ 的 $p\times p$ 矩阵$\mathbf{L}$ 的 L-最优性。

c) 考虑一个密度函数为 $f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$、累积分布函数为$F_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$的参数统计模型。令 $\boldsymbol{\theta}_0$ 由 *(1)–(2)* 定义，其中 $\ell_i(\boldsymbol{\theta})=-\log f_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}_i)$。假设对所有$i\in\mathcal{D}$ 以及邻域或 $\boldsymbol{\theta}_0$中的所有参数值 $\boldsymbol{\theta}$，以下条件均成立： $d_{\mathrm{KL}}(\boldsymbol{\theta})$ 有限，$\ell_i(\boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$二次连续可微，且 $\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)$ 的所有一阶和二阶导数在 L_1中关于测度 $dF_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)$ 有界。那么，关于库尔贝克-莱布勒距离的 d-最优性等价于关于满足 $LL^{\mathsf{T}}=\tilde{}H(\boldsymbol{\theta}_0)$ 的 $p\times p$ 矩阵 L 的 L-最优性。

命题4的结果可由命题3立即推出。注意，对于 c)，我们需要对模型$f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})$作出一些条件限制，以便能够交换积分与微分的顺序。

根据命题4a)和 c)，我们观察到， d_{ER}-最优性与 d_{I}-最优性是等价的（取 $\boldsymbol{\Sigma}=\mathrm{H}(\boldsymbol{\theta}_0)^{-1}$）。对于 d_{KL}-最优性与 $d_{\mathcal{I}}$-最优性，结论同样成立（取 $\boldsymbol{\Sigma}=\tilde{}\mathrm{H}(\boldsymbol{\theta}_0)^{-1}$）。我们还注意到，对于许多模型，包括指数族和具有典型连接函数的广义线性模型，观测信息矩阵 $\mathrm{I}(\boldsymbol{\theta}_0)=\mathrm{H}(\boldsymbol{\theta}_0)\tilde{}$与期望信息矩阵 $\mathcal{I}(\boldsymbol{\theta}_0)=\mathrm{H}(\boldsymbol{\theta}_0)$相等，因此这四种最优性准则也是等价的（例如参见 McCullagh 和 Nelder，1989）。对于正确设定的参数模型，它们还与 d_{S}-最优性渐近等价（当 $N\to\infty$时），因为在这种情况下， $N\mathrm{H}(\boldsymbol{\theta}_0)^{-1}$、 $N\tilde{}\mathrm{H}(\boldsymbol{\theta}_0)^{-1}$ 和 $N\mathrm{H}(\boldsymbol{\theta}_0)^{-1}\mathrm{V}(\boldsymbol{\theta}_0)\mathrm{H}(\boldsymbol{\theta}_0)^{-1}$都收敛于同一个极限（例如参见 Stefanski 和 Boos，2002年）。

上述最优性准则经过适当的变量变换后，也与 A-最优性有关。例如，考虑一个线性回归模型，并假设模型矩阵 $\mathbf{X}$ （即行向量为 $\boldsymbol{x}_i^{\mathsf{T}}$ 的矩阵）的列相互正交。那么， d_{ER}-和 d_{KL}-最优性准则等价于 A-最优性，因为在这种情况下$\mathbf{H}(\boldsymbol{\theta}_0)=\tilde{}\mathbf{H}(\boldsymbol{\theta}_0)\propto\mathbf{X}^{\mathsf{T}}\mathbf{X}=\mathbf{I}_{p\times p}$。在非正交的情况下， d_{ER}-和 d_{KL}-最优性准则取决于模型的参数化方式、数据的尺度以及变量之间的相关性，这些因素通过海森矩阵 $\mathbf{H}(\boldsymbol{\theta}_0)$ 起作用。因此，可以实现数据在非奇异仿射变换下以及模型重新参数化下的不变性（见第4.3节）。从几何角度看，A-最优性准则使估计量 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 与全数据参数 $\boldsymbol{\theta}_0$ 之间的期望欧氏距离ˆ 最小（命题3，第4.1节）。 d_{ER}-和 d_{KL}-最优性准则则使相对于模型空间自然几何结构的期望距离最小。

最后，我们考察d-最优性与D-最优性之间的关系。当距离函数取为带有离散矩阵 $\boldsymbol{\Sigma}=\boldsymbol{\Gamma}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$的平方马氏距离 $d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$时，这两个准则是一致的，其中 $\boldsymbol{\mu}^*$ 是D-最优抽样方案（见命题3和命题4b)）。特别地，D-最优性等价于L-最优性，其中$\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$按照式（8）定义，并且$\mathbf{L}=\mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^{-1/2}$。然而，这一结果并不十分实用，因为L-最优性准则的系数矩阵取决于D-最优抽样方案 $\boldsymbol{\mu}^*$。An

与 D-最优性密切相关的最优性准则是 L-最优性，即 $\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^{-1/2}$，其中由 (15) 给出的 $\mathbf{V}(\boldsymbol{\theta}_0)$ 不依赖于 $\boldsymbol{\mu}$。根据命题4b)，这等价于 d_{S}-最优性。
我们指出，系数矩阵 $\mathbf{L}$ 取决于全数据海森矩阵$\mathbf{H}(\boldsymbol{\theta}_0)$和参数 $\boldsymbol{\theta}_0$ 并不构成限制，因为所有最优设计无论如何都取决于未知的全数据特征。处理这一问题的方法将在第 5 节中讨论。

4.3 不变性

除了具有令人称道的几何和统计解释外，上一节介绍的期望距离最小化最优性准则还具备两个理想性质：计算上的可处理性和参数化不变性。事实上，由于属于线性最优性准则这一类，d_{ER}、d_{KL} 和 d_{S}最优性准则根据算法 1均具有最优抽样方案的简单解。下文将确立这些最优性准则及其相应最优抽样方案的不变性性质。
考虑重新参数化 $\boldsymbol{g}$： $\boldsymbol{\theta} \mapsto \boldsymbol{\eta}$，其中 $\boldsymbol{g}$ 是参数空间上的一一可微映射。在这种变换下，全数据经验风险最小化器 $\boldsymbol{\theta}_0$ 具有等变性，即诱导经验风险 $\ell_0^*(\boldsymbol{\eta}) := \sum_{i\in\mathcal{D}} \ell_i(\boldsymbol{g}^{-1}(\boldsymbol{\eta}))$ 的最小化器由 $\boldsymbol{\eta}_0 = \boldsymbol{g}(\boldsymbol{\theta}_0)$给出（参见 Casella 和 Berger，2001）。$\hat{}$ 通过类似论证，$\boldsymbol{\eta}_0$ 的 Hansen-Hurwitz 经验风险最小化器由 $\hat{}\boldsymbol{\eta}_{\boldsymbol{\mu}} = \boldsymbol{g}(\boldsymbol{\theta}_{\boldsymbol{\mu}})$ 给出。对诱导经验风险$\ell_0^*(\boldsymbol{\eta})$求导，并利用式（7），我们得到 $\hat{}\boldsymbol{\eta}_{\boldsymbol{\mu}}$ 的协方差矩阵：

$$\mathbf{Cov}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\mu}} - \boldsymbol{\eta}_0) = \boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0) + o(n^{-1}), \quad \boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0) = \mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{\mathsf{T}}, \tag{16}$$

其中 $\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta})$ 是 $\boldsymbol{g}$ 的雅可比矩阵，即其各行为 $\nabla_{\boldsymbol{\theta}} g_i(\boldsymbol{\theta})^{\mathsf{T}}$ 的矩阵。若对于重新参数化 $\boldsymbol{g}$： $\boldsymbol{\theta} \mapsto \boldsymbol{\eta}$， $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 和 $\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 的最优抽样方案相同，则称最优性准则在该重新参数化下具有不变性。命题 5 证明了 d_{ER}-、d_{KL}-和 d_{S}-最优性准则的不变性。

命题5（参数化不变性）。 设$\mathbf{V}(\boldsymbol{\theta}_0)$ 如(15)中所定义，并假设$\mathbf{H}(\boldsymbol{\theta}_0)$ 和$\mathbf{V}(\boldsymbol{\theta}_0)$ 均为满秩。则， d_{ER}-和 d_{S}-最优性准则在重新参数化 $\boldsymbol{g}$： $\boldsymbol{\theta} \mapsto \boldsymbol{\eta}$下保持不变，其中 $\boldsymbol{g}$ 是参数空间上的一一可微映射。在命题*4c*）的假设下， d_{KL}-最优性准则也具有同样性质。

对于数据的非奇异仿射变换不变性，也可以得到类似的结果。事实上，在许多情况下，对数据的变换会在参数空间上诱导出一个变换，并且该变换满足命题 5 中关于变换 $\boldsymbol{g}$ 的条件。然而，需要注意确保应用变换后经验风险函数仍然有定义，并且该变换产生一个在数学上等价的模型。在这种情况下，在大多数实际情形中，变换不变性和参数化不变性的概念是可以互换的。不过也存在例外情况：对数据的变换使海森矩阵$\mathbf{H}(\boldsymbol{\theta}_0)$保持不变。在这种情况下， d_{ER}和 d_{KL}最优性准则不再对数据的仿射变换保持不变。我们在第6.4节中给出了这样一个例子。需要指出的是，即使在此类情况下，D-和 d_{S}-最优性准则仍然对数据的仿射变换保持不变。

5 实际实施

到目前为止，我们假设完整数据 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ 和全数据参数 $\boldsymbol{\theta}_0$ 已知。然而，如果在设计阶段能够获得此类信息，那么一开始就不需要进行子抽样。本节介绍一种最优子抽样的实用方法。在第 5.1 节中，我们引入预期协方差矩阵（参见Isaki和Fuller,1982 年），将其作为未知协方差矩阵 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 的替代量，用于优化。在抽样过程中逐步获取优化所需信息的序贯最优设计和多阶段抽样程序将在第 5.2 节中讨论。

5.1 辅助变量辅助子抽样设计

除数据 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ 外，我们现在假设存在一组辅助变量 $\{\boldsymbol{z}_i\}_{i\in\mathcal{D}}$，这些变量对所有成员而言均可先验地 获得 $i \in \mathcal{D}$。根据具体情境，辅助变量可能包括 $\boldsymbol{x}_i$ 中的部分变量和/或 $\boldsymbol{y}_i$ 中的部分变量。例如，考虑一项病例对照研究，旨在调查某些暴露变量对已知二元结局的影响。在这种情况下，辅助变量包含（标量）结果 y_i，还可能包含某些解释变量或其代理变量（参见 Imberg 等人，2022a）。主动学习中则会遇到相反的情况（Settles，2012）。在这种情况下，所有预测变量向量 $\boldsymbol{x}_i$ 均已知，但只能观测到子集 $\mathcal{S} \subset \mathcal{D}$ 的结果变量 y_i，因此 $\boldsymbol{z}_i = \boldsymbol{x}_i$（参见 Bach，2007；Wang 等人，2017；Meng

等人，2021；Zhang 等人，2021；Imberg 等人，2022b）。在极端情况下，甚至可以获取全数据$\{(\boldsymbol{x}_i,\ \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$，但利用这些信息计算 $\boldsymbol{\theta}_0$ 的计算量可能过大，难以实际实现（例如，参见 Ma 等人，2015；Drovandi 等人，2017；Wang 等人，2018；Deldossi 和 Tommasi，2022）。实践中可能遇到介于这两种极端情况之间的任何情形。辅助变量可能与未观测研究变量弱相关、强相关，甚至完全相关。相关性越强，最优抽样的潜在收益就越大。

第3.3节中介绍的用于寻找最优抽样方案的算法，需要全数据海森矩阵$\mathbf{H}(\boldsymbol{\theta})$和梯度 $\boldsymbol{\psi}_i(\boldsymbol{\theta})$ 的信息，并且是在全数据参数 $\boldsymbol{\theta}_0$ 处计算的。此外，如果模型中包含解释变量 $\boldsymbol{x}_i$，海森矩阵还取决于这些解释变量，有时也取决于结果变量 $\boldsymbol{y}_i$。同样，梯度同时取决于结果变量和解释变量。为处理这一问题，我们引入一组随机变量 $\{(\mathbf{X}_i,\mathbf{Y}_i)\}_{i\in\mathcal{D}}$，用于描述我们对数据 $\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i\in\mathcal{D}}$ 未知值的不确定性。对于也包含在 $\boldsymbol{z}_i$ 中的任何变量，我们可以将一个退化的（确定性）分布与 $(\mathbf{X}_i,\mathbf{Y}_i)$ 的相应~分量相关联，并以 $\boldsymbol{z}_i$ 为条件。我们还假设已有全数据参数 $\boldsymbol{\theta}_0$ 的初步估计 $\tilde{\boldsymbol{\theta}}_0$，以及一个辅助模型 $f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})$，用于描述在给定辅助变量 $\boldsymbol{z}_i$ 的条件下随机变量 $(\mathbf{X}_i,\mathbf{Y}_i)$ 的条件分布。此类信息可能来自领域知识、先前研究、先导样本，或这些来源的组合。在第 5.2 节中，我们将讨论如何在子抽样过程中逐步获取此类信息。下面我们将预期协方差矩阵定义为在针对未知量的辅助模型协助下进行优化的目标。

定义3（预期协方差）。考虑一个数据三元组 $\{(\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{z}_i)\}_{i\in\mathcal{D}}$，其中 $(\mathbf{X}_i,\mathbf{Y}_i)$ 是一个随机向量~，并且对所有 $i\in\mathcal{D}$，$\boldsymbol{z}_i$ 均为已知。此外，考虑全数据参数 $\boldsymbol{\theta}_0$的一个初步估计 $\tilde{\boldsymbol{\theta}}_0$，以及一个模型$f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})$，用于描述给定辅助变量 $\boldsymbol{z}_i$时（$\mathbf{X}_i$、$\mathbf{Y}_i$）的条件分布。$\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 的预期协方差矩阵^定义为

$$\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\mu};\tilde{\boldsymbol{\theta}}_0)=\mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}[\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)]_{\boldsymbol{\theta}_0=\tilde{\boldsymbol{\theta}}_0}\,.$$

定义3中的预期协方差矩阵，是我们根据可用辅助信息对实际未知协方差矩阵 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$所作的预测。我们沿用 Isaki and Fuller (1982) 的术语，称之为预期而非期望，以强调上述定义中涉及的期望是一种假设性构造，通常不同于数据生成机制下的期望。

现在，可以将第3节和第4节中的所有结果重新表述为相对于预期协方差矩阵~ ~$\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$的 Φ-最优性，而非相对于近似协方差矩阵 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$的最优性。在对模型 $f(\boldsymbol{x},\ \boldsymbol{y}|\boldsymbol{z})$作出较弱假设、从而允许我们交换积分与微分的次序后，唯一的变化是：算法2中的系数 c_i 由其相应的期望替代。

$$\tilde{c}_i:=\mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}\left[C_i|\{\boldsymbol{z}_i\}_{i\in\mathcal{D}}\right],\quad C_i=\left|\left|\mathbf{L}_t^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|^2_{2,\boldsymbol{\theta}_0=\tilde{\boldsymbol{\theta}}_0},\tag{17}$$

其中 C_i 是一个函数。关于随机变量 $\{(\mathbf{X}_i,\ \mathbf{Y}_i)\}_{i\in\mathcal{D}}$，以及$\mathbf{L}_t$ 一个矩阵，使得 $\mathbf{L}_t\mathbf{L}_t^{\intercal}=\boldsymbol{\phi}(\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\mu}_{t-1};\ \tilde{\boldsymbol{\theta}}_0)$).

我们注意到，(17) 中的 C_i 是一个正随机变量，这意味着只要 $C_i>0$ 的概率为正，$\tilde{c}_i>0$ 就成立。只要 $(\mathbf{X}_i,\ \mathbf{Y}_i)$ 中未纳入 $\boldsymbol{z}_i$ 的各分量的协方差矩阵对所有 i 都是满秩的，这一条件就能得到满足。因此，通过考虑辅助分布下的预期协方差，并适当承认未知量的不确定性，我们便能有效避免所提出的算法（算法 1 和 2）收敛到不可行解的情况。

5.2 序贯最优设计

上一节引入的预期协方差使我们进一步接近实用的最优子抽样框架。借助这一概念，可以使用第3.3节的方法寻找最优抽样方案，其中系数的未知值 c_i 被辅助模型~$f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})$ 下的期望（17）以及初步参数估计值 $\boldsymbol{\theta}_0$ 所替代。然而，在大多数情况下，即使是这些信息，也无法在观测到任何数据之前获得。可以使用序贯最优设计来处理这一问题。因此，子抽样分多个阶段进行，在前面抽样阶段获取的信息可用于制定后续阶段的最优抽样方案。我们承认，已有许多体现这一思想的算法和方法（例如参见 Bach, 2007；Wang等人,2018年；Imberg 等人, 2020；Ai等人, 2021b）。算法 3给出了一般程序。

算法 3 中的抽样阶段数 K 可以从单个阶段（包含 n 个观测）到 n 个阶段（每个子样本中只有一个观测）不等。例如，在线性回归中，如果已知解释变量 $\boldsymbol{x}_i$，就无需进行序贯子抽样。这是因为在这种情况下，(17) 是预测变量 $\boldsymbol{x}_i$（已知）、海森矩阵 $\mathbf{H}(\boldsymbol{\theta}_0)$（仅取决于预测变量 $\boldsymbol{x}_i$）以及残差二阶矩的函数。参见

算法 3. K 阶段子抽样过程。

输入：索引集 $\mathcal{D}$、最优性准则 Φ、抽样设计族（PO-WR、PO-WOR 或 MULTI）、抽样阶段数 K、批次大小 $\{n_k\}_{k=1}^{K}$。

1: **对于** $k = 1, 2, \ldots, K$ **do**
2: 计算（最优）抽样方案。
3: 随机选择一个大小为 n_k的子样本。
4: 估计目标参数 $\boldsymbol{\theta}_0$。
5: 更新辅助模型 $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$。
6: 评估性能/精度。
7: 如果达到足够的精度，则停止；否则继续。
8: **结束循环**

Ma 等人（2020）针对这一情境给出了多种最优性准则及相应的最优抽样方案。与此相反，主动学习方法会采用大量抽样阶段，通常每个阶段只抽取一个观测，以便在抽样过程中获得最大的灵活性（Bach, 2007；Imberg 等人, 2020；Kossen 等人, 2022；Zhan 等人, 2022）。大数据中的子抽样方法通常依赖两个抽样阶段：先进行初始简单随机抽样，再进行最优不等概率抽样（Wang 等人, 2018年；Ai等人, 2021b；Wang 和 Ma, 2021）。

针对 $\boldsymbol{\theta}_0$ 在 k 个抽样阶段之后的估计量可定义为

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}^{(k)} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} \hat{\ell}_{\boldsymbol{\mu}}^{(k)}(\boldsymbol{\theta}), \tag{18}$$

$$\hat{\ell}_{\boldsymbol{\mu}}^{(k)}(\boldsymbol{\theta}) = m_k^{-1} \sum_{j=1}^{k} n_j \hat{\ell}_{\boldsymbol{\mu},j}(\boldsymbol{\theta}), \quad \hat{\ell}_{\boldsymbol{\mu},j}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{D}} S_{ji} w_{ji} \ell_i(\boldsymbol{\theta}),$$

其中，S_{ji} 是实例 $i \in \mathcal{D}$ 在第 j阶段由抽样机制选中的次数，μ_{ji} 是相应的预期选中次数，$m_k = n_1 + \ldots + n_k$ 是经过 k 个阶段后的累积样本量，而$\hat{}w_{ji} = 1/\mu_{ji}$。此处$\ell_{\boldsymbol{\mu},j}(\boldsymbol{\theta})$是从第 j阶段获得的样本中计算出的全数据经验风险 $\ell_0(\boldsymbol{\theta})$的无偏汉森-赫维茨估计量，$\hat{}\ell_{\boldsymbol{\mu}}^{(k)}(\boldsymbol{\theta})$则是根据前 k 个子样本计算出的合并估计量。

所得估计量(18)的性质已在一些特定情形下得到研究，并已证明，在适当的正则条件下，估计量$\hat{}\boldsymbol{\theta}_{\boldsymbol{\mu}}^{(k)}$ 渐近正态分布且对于 $\boldsymbol{\theta}_0$相合。有关抽样阶段数 $K = 2$时广义线性模型和拟似然方法的结果，参见例如 Ai等人 (2021b) 和Yu等人 (2022)。Imberg 等人(2022b)建立了有限总体向量特征的估计量的渐近性质，其中子样本量 n_k 有界且抽样阶段数$K \to \infty$。结合鞅极限定理（Hall和 Heyde，1980）与调查抽样中估计方程估计量的渐近理论（Binder，1983），当批次大小 n_k有界且 $K \to \infty$ 时，(18) 的相合性和渐近正态性也可以推导出来（参见Zhang等人，2021）。我们推测，当抽样阶段数 K 有界且子样本量 n_k 趋于无穷，同时 $N \to \infty$ 和 $n_k/N \to \gamma_k \in$ （0,1）时，也存在类似结果。然而，对这一问题的全面处理是未来研究的课题。

6 应用与示例

已有大量文献展示了最优子抽样的优势，例如可参见第 1 节中的参考文献。我们不再为这些已有充分说服力的结果提供更多证据。相反，本节将通过示例说明所介绍的方法，并从计算方面和估计量效率的角度比较数据子抽样的不同最优性准则。

我们考虑一个应用案例：为高级驾驶辅助系统的虚拟安全评估生成场景。第 6.1 节简要介绍该应用的背景、数据描述和问题表述。参数密度估计的示例、说明和结果见第 6.2 节，回归建模见第 6.3 节，有限总体推断见第 6.4 节。

6.1 材料与方法

背景 道路交通伤害是全球范围内导致死亡的主要原因之一（世界卫生组织，2018年）。为降低这些风险，人们不断开发应对措施，例如高级驾驶辅助系统。一种方法

是在此类系统上市前，通过虚拟仿真对其进行评估（Anderson 等人，2013；Seyedi 等人，2021）。由于此类评估是在虚拟而非实体测试环境中进行的，因此比传统测试台更具成本效益。然而，这也带来了巨大的计算负荷。通过子抽样可以大幅降低计算需求（Mullins 等人，2018年；Imberg 等人，2022；Sun 等人，2022）。

数据集 我们的数据集包含 44,220 条观测数据，这些数据由 44 起重建的真实追尾碰撞事件经过多种变体生成。这些变体通过改变后车驾驶员的驾驶行为生成，涉及注视行为（自某一特定时间锚点之后偏离道路的注视时长）和制动特征（制动期间的最大减速度）。对于每种变体，我们都在虚拟环境和仿真软件中设置了相应场景，并借此模拟整个事件过程。仿真的结果包括是否发生碰撞，以及发生碰撞时的撞击速度。因此，数据集中的每条观测都代表一个合成事件，描述在特定条件变体下，原始碰撞事件可能发生的情况。每个场景还分别在两种“处理条件”下运行：一种是配备自动紧急制动（AEB）系统的场景，另一种是不配备 AEB 的基线手动驾驶场景。

数据集包含以下变量：

- 输入变量：案例标识符（分类变量，共 44 个水平，对应 44 起原始追尾碰撞）、偏离道路注视时长（67 个水平，0–6.6 秒）以及制动期间最大减速度（15 个水平，3.3–10.3 米/秒^2）。
- 直接结果变量：碰撞指标（发生碰撞时为 1，否则为 0），以及自动紧急制动系统和基线手动驾驶场景下的撞击速度。
- 计算的结果变量：配备自动紧急制动系统时的受伤风险，以及在基线人工驾驶场景下的受伤风险、撞击速度降低、受伤风险降低，以及与基线人工驾驶相比，配备自动紧急制动系统时的碰撞避免指标。

每个观测还关联一个观测权重 $w_i > 0$，用于描述特定输入参数配置（即偏离道路注视时长和制动期间最大减速度）在现实生活中出现的概率。更多详情请参阅 Imberg et al. (2022b)。

目标特征我们关注以下内容：

i) 基线场景下的碰撞速度分布，仅限于会导致碰撞的输入值子集。ii) 基线场景下的碰撞速度响应面，作为偏离道路注视时长和最大减速度的函数。iii) 与基线人工驾驶相比，启用自动紧急制动后的平均撞击速度降低幅度、平均受伤风险降低幅度和碰撞避免率，仅限于基线场景中会发生碰撞的变化子集。

数据集的特征，包括自动紧急制动相较于基线人工驾驶的基准撞击速度分布、碰撞速度响应面和安全效益分布，均以图形方式呈现在附录 B 的图 S1 和图 S2 中。

与实践中的常见情况一样，我们假设运行所有感兴趣的模拟在实际中不可行，因此不可避免地需要进行子抽样。在这种情况下，数据集中所有实例的输入变量（即案例标识符、偏离道路注视时长和制动期间最大减速度）以及场景概率均可先验地 获取。因此，这些就是我们的辅助变量。其余变量只能针对一个子集观测到，推断将基于该子集进行。为简便起见，对于问题 i)（第 6.2 节）和 iii)（第 6.4 节），我们仅考虑在基线场景中发生碰撞的模拟。因此，在相应评估中排除未导致碰撞的 4299 条观测。

性能评估 我们针对上述应用和推断问题，从计算时间和统计效率两个方面评估所提出的最优子抽样方法的性能。此外，对于非线性最优性准则，我们还评估不动点迭代算法（算法2，第3.3节）收敛所需的迭代次数，即找到最优抽样方案所需的时间。对于抽样方案 $\hat{}\boldsymbol{\mu}$，估计量 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 相对于准则 Φ 的统计效率，通过相对 Φ效率来衡量

$$\Phi\text{-eff}(\boldsymbol{\mu}) = \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}^*; \boldsymbol{\theta}_0))/\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu}; \boldsymbol{\theta}_0)),$$

其中， $\boldsymbol{\mu}^*$ 是 Φ最优抽样方案（Atkinson 和 Donev，1992；Pukelsheim，1993）。相对 Φ效率衡量抽样方案 $\boldsymbol{\mu}$ 在准则 Φ下为 $\boldsymbol{\theta}_0$ 提供最大信息量的程度。其倒数表示：相对于 Φ最优性准则，要达到与最优设计相同的性能水平，所需样本量的相对增幅。相对效率通过近似协方差矩阵的表达式（7）进行解析计算。

考虑以下最优性准则：A-、c-、D- 和 E-最优性， $q = 0.5, q = 5$ 下的 Φ_q-最优性，以及 $q = 10$, d_{ER}-最优性（即采用 $\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)^{1/2}$ 的 L-最优性；对于本次评估中的所有模型，这也等价于 d_{KL}-最优性），以及 d_{S}-最优性（即采用 $\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^{-1/2}$ 的 L-最优性）。更多详细信息和定义请参见第2.3节和 4.2。对于 D-最优性准则，使用目标函数的非对数版本（$\det(\boldsymbol{\Gamma})^{1/p}$）（表 1，第2.3节）。

所有算法和评估均使用 R 语言及其统计计算环境实现，版本为 4.2.3（R Core Team，2023）。所有计算均在运行 Windows 11 的台式机上使用单核完成，该计算机配备 2.1 GHz Intel i7 处理器。子样本量设为全数据量的 1%。线性最优性准则的抽样方案按照算法 1 计算；非线性最优性准则的抽样方案按照算法2计算，容差参数为 $\epsilon = 0.001$。假设已知完整数据 $\{(\boldsymbol{x}_i,\ \boldsymbol{y}_i)\}_{i\in\mathcal{D}}$ 和全数据参数 $\boldsymbol{\theta}_0$，因此可以找到理论上的最优抽样方案。数据集和 R 代码可在 https://github.com/imbhe/OSD 在线获取。

本文给出了 PO-WR 和多项式抽样设计的结果，两者得到相同的解析结果。采用类似方法，也可以得到 PO-WOR 的相应结果。

6.2 参数密度估计

首先，我们考察基线场景下撞击速度的分布，如附录 B 中的图 S1 所示。

模型 假设撞击速度服从对数正态分布，其参数为 (η, σ)，分别表示对数撞击速度的均值和标准差。全数据参数 $\boldsymbol{\theta}_0 = (\eta_0, \sigma_0)^{\mathsf{T}}$ 定义为

$$\boldsymbol{\theta}_0 = \underset{\eta\in\mathbb{R},\sigma\in\mathbb{R}_{>0}}{\arg\min} \frac{1}{2}\sum_{i=1}^{N} w_i \left(\frac{(\log y_i - \eta)^2}{\sigma^2} + \log\sigma^2\right), \tag{19}$$

其中， w_i 是先验地已知的观测权重， y_i 是场景 $i \in \mathcal{D}$中的撞击速度。不失一般性，我们假设观测权重已归一化，使得 $\sum_{i=1}^{N} w_i = 1$。

最优抽样方案 作为一个说明性示例，我们考虑 c-最优性准则，其中 $\mathbf{c} = (1,0)^{\mathsf{T}}$，即最小化位置参数 η_0估计值的方差。由于这里的最优性准则是线性的，因此可以根据算法 1 并结合以下条件求得最优抽样方案

$$\mathbf{L} = (1,0)^{\mathsf{T}}, \quad \mathbf{H}(\boldsymbol{\theta}_0) = \frac{1}{\sigma_0^2}\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) = -w_i\left(\frac{\log y_i - \eta_0}{\sigma_0^2}, \frac{(\log y_i - \eta_0)^2}{\sigma_0^3} - \frac{1}{\sigma_0}\right)^{\mathsf{T}},$$

and

$$c_i = \left|\left|\mathbf{L}^{\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|_2^2 \propto w_i^2(\log y_i - \eta_0)^2. \tag{20}$$

[{"id": 0, "output": "为了针对 η_0 的估计量预期方差找到最优抽样方案，我们将 y_i 替换为随机变量 Y_i，并在对 Y_i 所作的假设模型下，计算式 (20) 的相应期望。如果假设 $\log Y_i$ 的均值为 $\hat{}y_i$、方差为 σ_i^2，则可得到"}]

$$\tilde{c}_i \propto \sqrt{\mathrm{E}_{Y_i}[w_i^2(\log Y_i - \eta_0)^2]} = w_i\sqrt{(\hat{y}_i - \eta_0)^2 + \sigma_i^2},$$

在实践中，该值可能会在 $\tilde{\eta}_0$ 的 η_0初步估计处进行评估。预测值 $\hat{}y_i$ 和离散参数 σ_i^2可以建模为观测到的辅助变量（即案例标识符、偏离道路注视时长和制动期间最大减速度）的函数，并从先导样本中进行估计，或使用序贯子抽样方法进行估计（算法 3）。只要所有 $w_i, \sigma_i > 0$，所得抽样方案就能保证产生严格为正的抽样概率。

结果 各种最优性准则下的计算时间、收敛所需的迭代次数以及相对效率列于表2中。利用算法2，通过四次不动点迭代得到了D-最优抽样方案。$\Phi_{0.5}$-、 Φ_5-最优抽样方案分别经过三次和25次迭代得到。由于目标函数的非凸性，无法找到E-最优抽样方案。寻找最优抽样方案的计算时间从线性最优性准则的0.10秒到D-最优性准则的0.96秒以及 Φ_5-最优性准则的4.95秒不等。 d_{ER}-和d_{S}-最优性准则的最优抽样方案达到了97–99%的 A 效率、96–99%的D 效率以及92–94%的 Φ_5-效率。A 最优抽样方案具有相似的性能。

表2。 使用各种最优性准则通过最优子抽样估计对数正态模型（19）的性能度量。各列显示寻找最优抽样方案所需的不动点迭代次数和执行时间，以及相对于其他最优性准则的相对效率。

最优性准则	迭代次数	时间（秒）	A-效率	$\mathbf{c}_{(1,0)}$-效	$\mathbf{c}_{(0,1)}$-效	D-效	d_{ER}-效	Φ_5-效
A		0.11	**1.00**	0.90	0.76	0.99	0.99	0.96
$\mathbf{c}, \mathbf{c} = (1,0)^\mathsf{T}$		0.10	0.14	**1.00**	0.04	0.24	0.10	0.09
$\mathbf{c}, \mathbf{c} = (0,1)^\mathsf{T}$		0.10	0.27	0.16	**1.00**	0.47	0.34	0.19
D	4	0.96	0.98	0.86	0.79	**1.00**	0.99	0.91
d_{ER}		0.10	0.99	0.84	0.83	0.99	**1.00**	0.92
d_{S}		0.10	0.97	0.84	0.79	0.96	0.98	0.94
E	发散	-						
$\Phi_{0.5}$	3	0.78	>0.99	0.88	0.78	>0.99	0.99	0.94
Φ_5	25	4.95	0.95	0.91	0.65	0.90	0.91	**1.00**
Φ_{10}	发散	-						

6.3 回归建模

接下来，我们将考察基准撞击速度作为场景生成的输入变量（即越野瞥视持续时间和制动期间最大减速度）的函数时的分布。

我们首先注意到，随着偏离道路注视时长水平的增加以及减速度水平的降低，撞击速度单调增加。因此，由同一次原始追尾碰撞生成的变异，其撞击速度存在一个上界；当变异的偏离道路注视时长达到最大值且减速度水平达到最小值时，可达到该上界。我们假设这一最大撞击速度是已知的，例如可通过运行相应的虚拟仿真来观测。随后，可以将撞击速度表示为相对于该特定案例的最大撞击速度的值，其取值位于通常范围 $[0,1]$ 内。注意，在这种情况下，解释变量是先验地已知的，而结果（即相对撞击速度）只有在运行相应的虚拟仿真后才能观测到。

模型 对于单位区间上的响应变量，一种简单的模型是准二项逻辑回归模型，其全数据参数 $\boldsymbol{\theta}_0$ 定义为

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min} -\sum_{i=1}^{N} y_i \log p_i(\boldsymbol{\theta}) + (1-y_i)\log(1-p_i(\boldsymbol{\theta})), \quad p_i(\boldsymbol{\theta}) = (1+\exp(-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}))^{-1}, \tag{21}$$

其中， $\boldsymbol{x}_i$ 是与实例 i相关的特征向量， $\boldsymbol{\theta}$ 是回归系数向量。解释变量包括原始追尾碰撞事件的案例标识符（包含 44 个类别，以 44 个二元变量进行哑变量编码）、偏离道路注视时长、制动期间最大减速度，以及所有三重交互作用。对于这 44 个案例中的每一个，碰撞速度响应面由 4 个参数描述：一个截距参数，以及分别对应于偏离道路注视时长、减速度水平和两者交互作用的三个斜率参数。联合参数向量 $\boldsymbol{\theta}$ 的维度为 $44 \times 4 = 176$。请注意，在本例中，我们未将观测权重 w_i 纳入经验风险函数，因为这些权重是解释变量的函数，因此在此情境下可以忽略。附录 B 的图 S2 展示了其中三个案例的观测碰撞速度响应面和预测碰撞速度响应面。

最优抽样方案 为便于说明，我们考虑 d_{ER}-最优性准则。由于这是线性最优性准则，因此可以根据算法 1 找到最优抽样方案，并且

$$\mathbf{L} = \mathbf{H}(\boldsymbol{\theta}_0)^{1/2}, \quad \mathbf{H}(\boldsymbol{\theta}_0) = \mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X}, \quad \boldsymbol{\psi}_i(\boldsymbol{\theta}_0) = -(y_i - p_i(\boldsymbol{\theta}_0))\boldsymbol{x}_i,$$

其中 $\mathbf{W}(\boldsymbol{\theta})$是具有条目 $p_i(\boldsymbol{\theta})(1-p_i(\boldsymbol{\theta}))$的对角矩阵，而$\mathbf{X}$ 是行向量为 $\boldsymbol{x}_i^\mathsf{T}$ 的矩阵，一个 d

$$c_i = ||\mathbf{L}^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2 = (y_i - p_i(\boldsymbol{\theta}_0))^2\boldsymbol{x}_i^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X})^{-1}\boldsymbol{x}_i. \tag{22}$$

为了找到关于预期协方差矩阵的 d_{ER}-最优抽样方案，我们将（22）中的 y_i 替换为随机变量 Y_i，并根据给定已知解释变量 $\boldsymbol{x}_i$时 Y_i 的模型，计算相应的期望。例如，我们可以假设 Y_i 的均值为 $p(\boldsymbol{\theta}_0)$，方差为 $p_i(\boldsymbol{\theta}_0)(1-p_i(\boldsymbol{\theta}_0))$。由此可得

$$\tilde{c}_i = \sqrt{\mathrm{E}_{Y_i}[(Y_i - p_i(\boldsymbol{\theta}_0))^2\boldsymbol{x}_i^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X})^{-1}\boldsymbol{x}_i]} = \sqrt{h_{ii}(\boldsymbol{\theta}_0)}, \tag{23}$$

其中 $h_{ii}(\boldsymbol{\theta}_0)$ 是“帽子矩阵”或投影矩阵的 i^{th} 对角元素

$$\mathbf{W}(\boldsymbol{\theta}_0)^{1/2}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}(\boldsymbol{\theta}_0)^{1/2}$$

（见 Hoaglin 和 Welsch，1978；Pregibon，1981）。为了考虑数据点 $(\boldsymbol{x}_i, y_i)$ 对其自身预测的影响，适宜将 Y_i 的方差按因子 $1-h_{ii}(\boldsymbol{\theta}_0)$ 进行折减，从而得到

$$\tilde{c}_i = \sqrt{h_{ii}(\boldsymbol{\theta}_0)(1-h_{ii}(\boldsymbol{\theta}_0))}$$

˜而不是（23）（参见 Ma 等人，2020）。在实践中，我们可以在由先导样本获得的初步估计 $\boldsymbol{\theta}_0$ 处，或使用序贯子抽样方法（算法 3）估计的初步估计处，计算 $\tilde{c}_i$。只要预测值 $p_i(\boldsymbol{\theta}_0)$与 0 和 1 均保持正距离，所得抽样方案就能˜保证产生严格为正的抽样概率。

结果 表 3 展示了针对各种最优性准则，找到最优抽样方案所需的计算时间、相对效率和迭代次数。对于 D-最优性，经过五次不动点迭代找到了最优抽样方案；对于 $\Phi_{0.5}$-最优性，经过四次迭代找到了最优抽样方案；而对于 Φ_5-、 Φ_{10}-和 E-最优性准则，则未能找到最优抽样方案。找到 L 最优抽样方案所需的计算时间比非线性 D 最优性准则少 95%。 d_{ER}-和 d_{S}-最优方案达到了 40–47%的 A 效率和 92–96%的 D 效率。A 最优抽样方案的 D 效率仅为 60%。 $\Phi_{0.5}$-最优抽样方案在 A-最优性和 D-最优性之间进行插值，实现了 92%的 A 效率和 80%的 D 效率。

表 3. 使用不同最优性准则的最优子抽样估计准二项逻辑回归模型 (21) 的性能指标。各列显示寻找最优抽样方案所需的不动点迭代次数和执行时间，以及相对于其他最优性准则的相对效率。将模型拟合到完整数据集的计算时间为 8.46 秒。

最优性准则	迭代次数	时间（秒）	A-效率	D-效率	d_{ER}-效率	d_{S}-效率	$\Phi_{0.5}$-效率
A	5	1.16	**1.00**	0.60	0.47	0.42	0.93
D		27.69	0.49	**1.00**	0.89	0.94	0.77
d_{ER}		1.11	0.47	0.92	**1.00**	0.91	0.71
d_{S}		1.12	0.40	0.96	0.92	**1.00**	0.67
E	已发散	-					
$\Phi_{0.5}$	4	20.37	0.92	0.80	0.68	0.65	**1.00**
Φ_5	已发散	-					
Φ_{10}	已发散	-					

6.4 有限总体推断

最后，我们考察自动紧急制动系统相较于基线手动驾驶场景可能带来的安全效益。对于场景 $i \in \mathcal{D}$，令 $\boldsymbol{y}_i = (y_{i1},\ y_{i2},\ y_{i3})^\mathsf{T}$，其中 y_{1i} 表示碰撞速度降低，y_{i2} 表示受伤风险降低，y_{i3} 表示自动紧急制动系统相较于基线人工驾驶的二元碰撞避免指标。这些特征的分布如附录 B 中的图 S1 所示。

模型 我们关注平均碰撞速度降低、平均受伤风险降低和碰撞避免率，它们由总向量给出

$$\boldsymbol{t}_{\boldsymbol{y}} = \sum_{i=1}^{N} w_i\boldsymbol{y}_i,$$

其中，观测权重 w_i 经过归一化，使得 $\sum_{i=1}^{N} w_i = 1$。这也可以表示为

$$\boldsymbol{t}_{\boldsymbol{y}} = \boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\mathbb{R}^3}{\arg\min}\,\frac{1}{2}\sum_{i=1}^{N} w_i||\boldsymbol{y}_i - \boldsymbol{\theta}||_2^2. \tag{24}$$

最优抽样方案 举例来说，考虑 d_{S}-最优性准则。由于这是一个线性最优性准则，可以根据算法 1 找到最优抽样方案，其中

$$\mathbf{L}=\mathbf{H}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^{-1/2},\quad \mathbf{H}(\boldsymbol{\theta}_0)=\mathbf{I}_{3\times 3},\quad \mathbf{V}(\boldsymbol{\theta}_0)=\sum_{i\in\mathcal{D}}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^T,\quad \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)=-w_i(\boldsymbol{y}_i-\boldsymbol{\theta}_0),$$

and

$$c_i=||\mathbf{L}^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2=w_i^2(\boldsymbol{y}_i-\boldsymbol{\theta}_0)^{\intercal}\mathbf{V}(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta}_0). \tag{25}$$

为了找到关于预期协方差矩阵的L 最优抽样方案，我们引入一个随机ˆ向量 $\mathbf{Y}_i$，将 $\mathbf{Y}_i$ 替换为 $\boldsymbol{y}_i$ 在（25）中，并计算其期望。因此，令 ˆ$\boldsymbol{y}_i$ 和 $\boldsymbol{\Sigma}_i$ 分别表示$\mathbf{Y}_i$的均值向量和协方差矩阵。根据二次型的性质（Mathai 和 Provost，1992），我们得到

$$\tilde{c}_i=\sqrt{\mathrm{E}\left[w_i^2(\boldsymbol{y}_i-\boldsymbol{\theta}_0)^{\intercal}\mathbf{V}(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta}_0)\right]}=w_i\sqrt{\left[(\hat{\boldsymbol{y}}_i-\boldsymbol{\theta}_0)^{\intercal}\mathbf{V}(\boldsymbol{\theta}_0)^{-1}(\hat{\boldsymbol{y}}_i-\boldsymbol{\theta}_0)+\mathrm{tr}(\mathbf{V}(\boldsymbol{\theta}_0)^{-1}\hat{\boldsymbol{\Sigma}}_i)\right]}. \tag{26}$$

˜为在实践中实施最优抽样，我们在从先导样本获得的初步估计 $\boldsymbol{\theta}_0$ 处计算 (26)。ˆ 预测值 ˆ$\boldsymbol{y}_i$ 和离散矩阵 $\boldsymbol{\Sigma}_i$ 可建模为观测到的辅助变量（即案例标识符、偏离道路注视时长和制动期间最大减速度）的函数，并通过序贯子抽样方法（算法 3）迭代更新。只要所有 $\boldsymbol{\Sigma}_i$ 均为满秩，由 (26) 得出的抽样方案就能保证产生ˆ严格为正的抽样概率。

结果 关于计算时间、收敛所需的迭代次数以及各种最优性准则的相对效率，结果见表 4。D-最优性准则通过四次不动点迭代找到最优抽样方案，而其他非线性最优性准则则通过两次迭代找到方案。计算时间从线性最优性准则的 0.10 秒到 D-最优性准则的 0.94 秒不等。 d_{ER}-最优抽样方案的 A 效率为 100%，D 效率为 46%，E 效率为 >99%。事实上，在这种情况下， d_{ER}-最优性准则与 A-最优性相同。相比之下， d_{S}–最优抽样方案的 A 效率为 73%，D 效率为 98%，E 效率为 72%。 $\Phi_{0.5}$-准则的 A 效率为 99%，D 效率为 58%，E 效率为 99%。A-和 E-最优性准则主要受平均碰撞速度降低的影响，因为该指标的测量尺度比受伤风险降低和避免碰撞的测量尺度大几个数量级（图 S1，附录 B）。

表 4. 采用不同最优性准则进行最优子抽样，以估计有限总体均值向量 (24) 的性能指标。各列分别显示寻找最优抽样方案所需的不动点迭代次数和执行时间，以及相对于其他最优性准则的相对效率。

最优性准则	迭代次数	时间（秒）	A-效率	$\mathbf{c}^{(1,0,0)}$-效率	$\mathbf{c}^{(0,1,0)}$-效率	$\mathbf{c}^{(0,0,1)}$-效率	D-效率	E-效率
A		0.10	**1.00**	>0.99	0.36	0.25	0.46	>0.99
$\mathbf{c}, \mathbf{c}=(1,0,0)^{\intercal}$		0.10	0.98	**1.00**	0.05	0.04	0.20	>0.99
$\mathbf{c}, \mathbf{c}=(0,1,0)^{\intercal}$		0.10	0.12	0.12	**1.00**	0.11	0.22	0.12
$\mathbf{c}, \mathbf{c}=(0,0,1)^{\intercal}$		0.10	0.41	0.41	0.50	**1.00**	0.70	0.41
D	4	0.94	0.65	0.65	0.76	0.82	**1.00**	0.65
d_{ER}		0.10	**1.00**	>0.99	0.36	0.25	0.46	>0.99
d_{S}		0.10	0.73	0.72	0.77	0.77	0.98	0.72
E	2	0.70	0.99	>0.99	0.07	0.06	0.22	**1.00**
$\Phi_{0.5}$	2	0.57	0.99	0.99	0.46	0.38	0.58	0.99
Φ_5	2	0.58	0.99	>0.99	0.07	0.06	0.22	**1.00**
Φ_{10}	2	0.57	0.99	>0.99	0.07	0.06	0.22	**1.00**

7 讨论

我们提出了一套针对一般类别的估计量、抽样设计和最优性准则的最优子抽样设计理论。尽管所提出的最优性条件适用于任何可微目标函数，但用于寻找最优抽样方案的算法最适用于凸函数。未来的研究可以包括开发处理非凸最优性准则的方法，例如 E-最优性和 G-最优性（Kiefer 和 Wolfowitz，1960年；Kiefer，1974年）。

从应用角度来看，我们认为，所提出的不变线性最优性准则（即 d_{ER}、 d_{KL}和d_{S}-最优性）在计算效率与统计效率之间实现了良好的折中。非线性最优性准则需要迭代过程以及计算成本高昂的协方差矩阵计算，这限制了它们在计算复杂度是主要考量的问题和应用中的可用性。鼓励开展进一步研究，以评估这些方法在实践中及其他应用中的性能。
序贯子抽样是将最优子抽样方法付诸实践的一种可行方案。迄今为止，仅在有限的情形下对这类序贯子抽样方法所得到的估计量的理论性质进行了严格研究。未来有必要在这一方向上开展更多研究。

致谢

我们感谢沃尔沃汽车公司的 Malin Svärd 和 Simon Lundell 允许我们在实验中使用他们的数据。

参考文献

Ai, M.、Wang, F.、Yu, J. 和 Zhang, H.（2021a）。大规模分位数回归的最优子抽样。复杂性杂志，62:101512。Ai, M.、Yu, J.、Zhang, H. 和 Wang, H.（2021b）。大数据回归的最优子抽样算法。《统计学报》。Anderson, R.、Doecke, S.、Mackenzie, J. 和 Ponte, G.（2013）。基于深入碰撞重建与仿真的自动紧急制动潜在益处。载于第*23*届国际车辆安全增强会议论文集。Atkinson, A. C. 和 Donev, A. N.（1992）。最优试验设计。克拉伦登出版社，牛津。Bach, F. R.（2007）。错误设定的广义线性模型的主动学习。载于神经信息处理系统进展*19*。Bellhouse, D. R.（1984）。调查抽样中的最优设计综述。加拿大统计学杂志，12(1):53–65。Binder, D. A.（1983）。复杂抽样调查中渐近正态估计量的方差。国际统计评论，51(3):279–292。Boyd, S. 和 Vandenberghe, L.（2004）。凸优化。剑桥大学出版社，剑桥。Brewer, K. R. W.（1979）。一类适用于大规模调查的稳健抽样设计。美国统计学会会刊，74(368):911–915。Casella, G. 和 Berger, R.（2001）。统计推断。Duxbury，太平洋丛林市。Cassel, C. M.、Särndal, C. E. 和 Wretman, J. H.（1976）。有限总体的广义差值估计和广义回归估计的一些结果。《生物统计学》，63(3):615–620。Dai, W.、Song, Y. 和 Wang, D.（2022）。基于最小能量准则的回归问题子抽样方法。《技术计量学》。在线提前发表。https://doi.org/10.1080/00401706.2022.2127915。Deldossi, L. 和 Tommasi, C.（2022）。大型数据集的最优设计子抽样。质量技术杂志，54(1):93–101。Drovandi, C. C.、Holmes, C. C.、McGree, J. M.、Mengersen, K.、Richardson, S. 和 Ryan, E. G.（2017）。大数据分析实验设计原理。统计科学，32(3):385–404。Efron, B. 和 Hinkley, D. V.（1978）。评估最大似然估计量的准确性：观测费希尔信息与期望费希尔信息。《生物统计学》，65(3):457–482。Fuller, W. A.（2009）。抽样统计学。威利出版社，霍博肯。Hájek, J.（1981）。有限总体抽样。马塞尔·德克尔出版社，纽约。Hall, P. 和 Heyde, C.（1980）。鞅极限定理及其应用。学术出版社，纽约。Hansen, M. H. 和 Hurwitz, W. N.（1943）。有限总体抽样理论。数学统计年刊，14(4):333–362。Hartley, H. O. 和 Sielken, R. L.（1975）。有限总体抽样的“超总体观点”。生物统计学，31(2):411–422。Hastie, T.（2020）。岭正则化：数据科学中的一个基本概念。《技术计量学》，62(4):426–433。

Hoaglin, D. C. 和 Welsch, R. E.（1978）。回归与方差分析中的帽子矩阵。美国统计学家，32(1):17–22。Horn, R. 和 Johnson, C.（1990）。矩阵分析。剑桥大学出版社，剑桥。Horvitz, D. G. 和 Thompson, D. J.（1952）。有限总体不放回抽样的一般化。美国统计学会会刊，47(260):663–685。Hájek, J.（1959）。概率抽样中的最优策略及其他问题。数学杂志，84(4):387–423。Imberg, H.、Jonasson, J. 和 Axelson-Fisk, M.（2020）。无偏主动学习中的最优抽样。载于第*23*届国际人工智能与统计学会议论文集。Imberg, H.、Lisovskaja, V.、Selpi 和 Nerman, O.（2022a）。两阶段抽样设计的优化及其在自然驾驶研究中的应用。*IEEE*智能交通系统汇刊，23(4):3575–3588。Imberg, H.、Yang, X.、Flannagan, C. 和 Bärgman, J.（2022b）。主动抽样：一种利用机器学习、针对有限总体推断的最优子样本框架。arXiv:2212.10024 [stat.ME]。Isaki, C. T. 和 Fuller, W. A.（1982）。回归超总体模型下的调查设计。美国统计学会会刊，77(377):89–96。Kiefer, J.（1974）。最优设计的一般等价性理论（近似理论）。统计年刊，2(5):849–879。Kiefer, J. 和 Wolfowitz, J.（1960）。两个极值问题的等价性。加拿大数学杂志，12:363–366。Kossen, J.、Farquhar, S.、Gal, Y. 和 Rainforth, T.（2022）。主动替代估计量：一种高效利用标签进行模型评估的主动学习方法。载于神经信息处理系统进展。Kullback, S. 和 Leibler, R. A.（1951）。关于信息与充分性。数学统计年刊，22(1):79–86。Ma, P.、Mahoney, M. W. 和 Yu, B.（2015）。算法杠杆的统计学视角。《机器学习研究杂志》，16:861–911。Ma, P.、Zhang, X.、Xing, X.、Ma, J. 和 Mahoney, M. W.（2020）。随机数值线性代数算法中抽样估计量的渐近分析。载于第*23*届国际人工智能与统计学会议论文集。Mathai, A. 和 Provost, S.（1992）。随机变量的二次型。马塞尔·德克尔出版社，纽约。McCullagh, P. 和 Nelder, J. A.（1989）。广义线性模型。CRC出版社，博卡拉顿。Meng, C.、Xie, R.、Mandal, A.、Zhang, X.、Zhong, W. 和 Ma, P.（2021）。LowCon：错误设定线性模型中的一种基于设计的子抽样方法。《计算与图形统计学杂志》，30(3):694–708。Mullins, G. E.、Stankiewicz, P. G.、Hawthorne, R. C. 和 Gupta, S. K.（2018）。为自动驾驶车辆的测试与评估自适应生成具有挑战性的场景。系统与软件杂志，137:197–215。Nelder, J. A. 和 Wedderburn, R. W. M.（1972）。广义线性模型。英国皇家统计学会杂志。*A*辑（综合），135(3):370–384。Neyman, J.（1938）。对人类总体抽样理论的贡献。美国统计学会会刊，33(201):101–116。Petersen, K. B. 和 Pedersen, M. S.（2012）。矩阵烹饪书。Pregibon, D.（1981）。逻辑回归诊断。统计年刊，9(4):705–724。Pronzato, L. 和 Pázman, A.（2013）。非线性模型中的试验设计。施普林格，纽约。Pukelsheim, F.（1993）。最优试验设计。威利出版社，纽约。R Core Team（2023）。*R*：统计计算语言与环境。R统计计算基金会，维也纳。Settles, B.（2012）。主动学习。人工智能与机器学习综合讲座，6(1):1–114。Seyedi, M.、Koloushani, M.、Jung, S. 和 Vanli, A.（2021）。基于真实世界碰撞仿真的前向碰撞规避辅助系统安全评估与参数研究。高级交通杂志。在线提前发表。https://doi.org/10.1155/2021/4430730。Sibson, R.（1974）。D_A-最优性与对偶性。载于《统计学进展》第*2*卷：第*9*届欧洲统计学家会议论文集。

Silvey, S.（1980）。最优设计。Chapman &Hall，伦敦。Stefanski,L.A. 和 Boos, D. D.（2002年）。M估计的微积分。美国统计学家，56(1)：29–38。Sun, J.、Zhou, H.、Xi,H.、Zhang,H.和Tian,Y.（2022）。面向自动驾驶车辆安全性评估的自适应实验设计。*IEEE*智能交通系统汇刊，23(9)：14497–14508。Tillé, Y.（2006）。抽样算法。施普林格，纽约。Vapnik, V.（1991）。学习理论中的风险最小化原理。在第*4*届神经信息处理系统国际会议论文集中。Wang,H.和 Ma, Y.（2018年）。大数据下的分位数回归最优子抽样。《生物统计学》，108(1)：99–112。Wang,H.、Zhu,R. 和 Ma, P.（2018年）。大样本逻辑回归的最优子抽样。美国统计学会会刊，113(522)：829–844。Wang, Y.、Yu,A.W.和 Singh,A.（2017）。关于测量约束回归模型中计算上可处理的实验选择。《机器学习研究杂志》，18(143)：1–41。Wedderburn,R.W. M.（1974年）。拟似然函数、广义线性模型和高斯-牛顿法。《生物统计学》，61(3)：439–447。Welch, W. J.（1984）。用于响应估计的计算机辅助实验设计。《技术计量学》，26(3)：217–224。世界卫生组织（2018）。《*2018*年全球道路安全状况报告》。https://www.who.int/publications/i/item/9789241565684。Yao,Y.和Wang, H.（2019）。Softmax回归的最优子抽样。《统计学论文》，60：585–599。Yu, J.、Wang, H.、Ai,M.和Zhang,H.（2022）。海量数据下极大拟似然估计量的最优分布式子抽样。美国统计学会会刊，117(537)：265–276。Zhan, X.、Wang,Y.和 Chan, A. B.（2022）。主动学习过程的渐近最优性。在第三十八届人工智能不确定性会议论文集中。Zhang, T.、Ning,Y.和Ruppert, D.（2021）。测量约束下广义线性模型的最优抽样。《计算与图形统计学杂志》，30(1)：106–114。

A 证明

A.1 引理 1 的证明

根据 矩阵微分演算中的链式法则（Petersen 和 Pedersen，2012 年），我们有 t

$$\frac{\partial \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i} = \operatorname{tr}\left(\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))^{\mathsf{T}} \frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}\right),$$

其中 $\boldsymbol{\phi}(\mathbf{U}) = \frac{\partial \Phi}{\partial \mathbf{U}}$ 是 Φ 关于其矩阵自变量的 $p \times p$ 矩阵导数，而 $\frac{\partial \boldsymbol{\Gamma}}{\partial \mu_i}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 是 $\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 关于 μ_i 的逐元素导数。由于 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 是对称的，因此 $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 也必定是对称的，由此证明了式（11）。

根据假设，$\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 随着 μ_i 的增加，按照洛文纳序单调递减，这意味着 $\frac{\partial \boldsymbol{\Gamma}}{\partial \mu_i}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$是负半定的。因此，存在一个实矩阵 $\mathbf{U}$，使得 $\mathbf{U}\mathbf{U}^{\mathsf{T}} = -\frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}$。此外，$\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 按 Loewner序单调变化，因此是关于 μ_i 的单调递减函数，所以必有

$$\frac{\partial \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial \mu_i} \leq 0,$$

由上可知，这等价于

$$\operatorname{tr}\left(\mathbf{U}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))\mathbf{U}\right) \geq 0, \quad \mathbf{U}\mathbf{U}^{\mathsf{T}} = -\frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial \mu_i}.$$

当且仅当 $\boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 是半正定矩阵时，对于每个$\mathbf{U}$以及矩阵 $\frac{\partial \boldsymbol{\Gamma}}{\partial \mu_i}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 的每个可能取值，这个不等式都成立。因此，存在一个实$\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 矩阵，使得$\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\mathsf{T}} = \boldsymbol{\phi}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$。

A.2 引理 2 的证明

a）的证明 由 (7) 和 (9) 可得

$$\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)=\begin{cases}\sum_{i\in\mathcal{D}}\mu_i^{-1}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}, & \text{for PO-WR or MULTI designs, and}\\ \sum_{i\in\mathcal{D}}(\mu_i^{-1}-1)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}, & \text{for PO-WOR.}\end{cases}$$

对 μ_i求导，得到

$$\frac{\partial\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)}{\partial\mu_i}=-\mu_i^{-2}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}.$$

b)–e) 的证明 由矩阵微分演算中的以下规则可得（Petersen 和 Pedersen，2012 ）：

b) $\frac{\partial}{\partial U}\log\det(U)=U^{-1}$，其中 U 为满秩矩阵。c) $\frac{\partial\lambda_{\max}(U)}{\partial U}=vv^{\intercal}$，其中 v 是对应于 U 的最大特征值的特征向量，且该特征向量唯一。d) $\frac{\partial}{\partial U}\mathrm{tr}(UA)=A^{\intercal}$。e) $\frac{\partial}{\partial U}\mathrm{tr}(U^q)=q(U^{q-1})^{\intercal}$，因此 $\frac{\partial}{\partial U}\mathrm{tr}(U^{q\,1/q}{}_1)\,\mathrm{tr}(U^q)^{1/q-1}\frac{\partial}{\partial U}\mathrm{tr}(U^p)=\mathrm{tr}(U^q)^{1/q-1}(U^{q-1})^{\intercal}$，其中 U 为满秩矩阵。最后，由于 $\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 的对称性，得出最终结果。

A.3 引理 3 的证明

结合引理 1 和 2 的结果，我们注意到，对于PO-WR、PO-WOR 和 MULTI设计，$\Phi(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 μ_i 的偏导数（只要其存在）为

$$\begin{aligned}\frac{\partial\Phi(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i}&=-\mathrm{tr}\left(\boldsymbol{\phi}(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))\mu_i^{-2}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\right)\\&=-\mu_i^{-2}\mathrm{tr}(\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0))\\&=-\mu_i^{-2}\left|\left|\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|_2^2.\end{aligned}$$

第二个等式由迹的循环性质以及$\mathbf{L}(\boldsymbol{\mu},\boldsymbol{\theta}_0)$的定义得到，第三个等式则是因为注意到括号内的表达式是一个标量，并且等于向量$\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)$的欧几里得范数的平方。

A.4 引理 4 的证明

通过在 $\boldsymbol{\theta}_0$ 附近进行二阶泰勒展开，我们有

$$d(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})=d(\boldsymbol{\theta}_0)+\nabla d(\boldsymbol{\theta})^{\intercal}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)+\frac{1}{2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}_d(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)+o_p(||(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)||_2^2),$$

其中前两项根据 $d(\boldsymbol{\theta})$ 的定义为零，且 $\mathbf{H}_d(\boldsymbol{\theta}_0)=\frac{\partial^2 d(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\intercal}}$。根据关于 $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ 的假设，我们有 $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0=o_p(n^{-1/2})$ 和 $\mathrm{E}[|\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0|^{2+\delta}]<\infty$（逐元素）成立，其中某个 $\delta>0$。根据有界收敛定理，这意味着余项满足 $\mathrm{E}[o_p(||(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)||_2^2)]=o(n^{-1})$。此外，我们还有

$$\begin{aligned}\mathrm{E}\left[(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)^{\intercal}\mathbf{H}_d(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)\right]&=\mathrm{tr}\left(\mathbf{H}_d(\boldsymbol{\theta}_0)\mathbf{Cov}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0)\right)+\mathrm{E}[\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0]^{\intercal}\mathbf{H}_d(\boldsymbol{\theta}_0)\mathrm{E}[\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}-\boldsymbol{\theta}_0]\\&=\mathrm{tr}\left(\mathbf{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0)\right)+o(n^{-1}),\end{aligned}$$

其中，第一个等号由二次型的性质（Mathai 和 Provost，1992）得出，第二个等号则由关于 $\boldsymbol{\theta}_{\boldsymbol{\mu}}$ 的假设（6）–（7）以及迹的循环性质得出。

A.5 命题1的证明

首先注意，根据引理 1，只要目标函数在 $\boldsymbol{\mu}^*$ 处可微，矩阵 $\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)$ 就存在。因此，系数 c_i 为正，平方根 $\sqrt{c_i}$ 为实数，且最优性条件 (13) 和 (14a)–(14c) 定义良好。

a) 的证明 考虑函数 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$，并满足约束条件 $\sum_{i\in\mathcal{D}}\mu_i = n$，且对所有 $i\in\mathcal{D}$， $\mu_i > 0$ 均成立。根据拉格朗日乘子法（Boyd 和 Vandenberghe，2004）， $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 的约束驻点可通过求其拉格朗日函数的驻点得到:

$$\Lambda(\boldsymbol{\mu},\alpha) = \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) + \alpha g(\boldsymbol{\mu}), \quad g(\boldsymbol{\mu}) = \sum_{i\in\mathcal{D}}\mu_i - n.$$

分别对 $\boldsymbol{\mu}$ 和 α求导，我们得到方程组

$$\nabla\Lambda(\boldsymbol{\mu},\alpha) = \mathbf{0} \quad \Leftrightarrow \quad \begin{cases} g(\boldsymbol{\mu}) = 0 \\ -\nabla_{\boldsymbol{\mu}}\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \alpha\nabla g(\boldsymbol{\mu}). \end{cases}$$

现在，根据引理 3 和 c_i 的定义，有 $\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = -c_i/\mu_i^2$，以及 $\frac{\partial g(\boldsymbol{\mu})}{\partial\mu_i} = 1$。因此，驻点满足方程组 $\alpha = c_1/\mu_1^2 = \ldots = c_N/\mu_N^2$，对所有 $i\in\mathcal{D}$ 均成立。要使 $\boldsymbol{\mu}^*$ 为 Φ-最优，我们必须有 $\mu_i^* \propto \sqrt{c_i}$、 $\mu_i^* > 0$ $\sum_{i\in\mathcal{D}}\mu_i = n$，并且，因此

$$\mu_i^* = n\frac{\sqrt{c_i}}{\sum_{j\in\mathcal{D}}\sqrt{c_j}} \text{ for all } i\in\mathcal{D}.$$

b) 的证明 考虑满足约束 $\sum_{i\in\mathcal{D}}\mu_i = n$ 和 $0 < \mu_i \le 1$（对所有 $i\in\mathcal{D}$）的函数 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$。再考虑拉格朗日函数。

$$\Lambda(\boldsymbol{\mu},\alpha,\boldsymbol{\beta}) = \Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) + \alpha g(\boldsymbol{\mu}) + \sum_{i\in\mathcal{D}}\beta_i h_i(\boldsymbol{\mu}),$$

其中 $g(\boldsymbol{\mu}) = \sum_{i\in\mathcal{D}}\mu_i - n$ 和 $h_i(\boldsymbol{\mu}) = \mu_i - 1$。 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 的约束驻点刻画为 d 作为 K 的解 Arush-Kuhn-Tucker 条件（Boyd 和 Vandenberghe，2004）：

- 平稳性: $-\nabla_{\boldsymbol{\mu}}\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \alpha\nabla g(\boldsymbol{\mu}) + \sum_{i\in\mathcal{D}}\beta_i\nabla h_i(\boldsymbol{\mu})$.
- 原始可行性: $g(\boldsymbol{\mu}) = 0$，且 $h_i(\boldsymbol{\mu}) \le 0$ 对所有 $i\in\mathcal{D}$。
- 对偶可行性： $\beta_i \ge 0$ 对于所有 $i\in\mathcal{D}$。
- 互补松弛： $\beta_i h_i(\boldsymbol{\mu}) = 0$对所有 $i\in\mathcal{D}$。

首先注意到 $\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = -c_i/\mu_i^2$ 根据引理 3 以及 c_i, $\frac{\partial g(\boldsymbol{\mu})}{\partial\mu_i} = 1$的定义，当 $i = j$ 时为 $\frac{\partial h_i(\boldsymbol{\mu})}{\partial\mu_j} = 1$，否则为 0。考虑一个抽样方案 $\boldsymbol{\mu}\in\mathcal{M}_n$，并令 $\mathcal{E} = \{i\in\mathcal{D}: \mu_i = 1\}$ 和 $n_{\mathcal{E}} = |\mathcal{E}|$。为了满足 Karush-Kuhn-Tucker 条件，我们必须有

i) $\mu_i \le 1$ and $\sum_{i\in\mathcal{D}\setminus\mathcal{E}}\mu_i = n - n_{\mathcal{E}}$, by the primal feasibility condition,

ii) $\beta_i = 0$若 $\mu_i < 1$，根据互补松弛条件,

iii) $c_i/\mu_i^2 = \alpha + \beta_i$ 根据平稳性条件，由上述内容可知,

$$c_i = \begin{cases} \alpha+\beta_i & \text{if } \mu_i = 1 \\ \alpha\mu_i^2 & \text{if } \mu_i < 1 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \beta_i = c_i - \alpha & \text{if } \mu_i = 1 \\ \alpha = c_i/\mu_i^2 & \text{if } \mu_i < 1, \end{cases}$$

iv) $\beta_i \ge 0$ 根据对偶可行性条件，由上述可知， $\beta_i = c_i - \alpha = c_i - c_j/\mu_j^2 \ge 0$ 对于$i\in\mathcal{E}$ 以及 $j\in\mathcal{D}\setminus\mathcal{E}$。

条件（14a）由 i) 得出，（14b）由 i) 和 iii) 得出，而（14c）由 iv) 得出。

A.6 命题 2 的证明

首先注意到， $\boldsymbol{\mu}$ 的定义域 $\mathcal{M}_n$ 是凸的。因此，通过证明 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$的海森矩阵在 $\mathcal{M}_n$ 上是半正定的，结论即可由二阶导数检验得出。

a) 的证明 由引理 3 可得,

$$\frac{\partial\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} = -\mu_i^{-2}\left|\left|\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|_2^2.$$

对于 L-最优性准则，矩阵 $\mathbf{L}(\boldsymbol{\mu};\boldsymbol{\theta}_0) = \mathbf{L}$ 不依赖于 $\boldsymbol{\mu}$。二阶偏导数为

$$\frac{\partial^2\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i\partial\mu_j} = \begin{cases} 2\mu_i^{-3}\left|\left|\mathbf{L}^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\right|\right|_2^2 \geq 0 & \text{if } i=j, \\ 0 & \text{if } i\neq j. \end{cases}$$

对于所有 $\mu_i > 0$，该矩阵都是对角矩阵，且其对角元素非负，因此在 $\mathcal{M}_n$ 上半正定。

证明 b) 的证明 我们证明 $\det(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 是对数凸的，即 $\log\det(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 是凸函数 .
首先注意到，$\log\det(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)) = \log\det(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))) - 2\log\det(\mathbf{H}(\boldsymbol{\theta}_0))$，其中 $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ 由 (9) 给出，且$\mathbf{H}(\boldsymbol{\theta}_0)$ 不依赖于 $\boldsymbol{\mu}$。因此，只需证明 $\log\det(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 关于 $\boldsymbol{\mu}$ 是凸函数。我们通过证明 $\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 的海森矩阵可分解为两个半正定矩阵的 Hadamard积，从而得到所需结果；因此该海森矩阵是半正定的（Horn 和 Johnson，1990）。

考虑 first 一种PO-WR 或多项式抽样设计。$\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))$ 的偏导数为 by

$$\begin{aligned}
\frac{\partial\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i} &= -\mu_i^{-2}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0), \\
\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i^2} &= 2\mu_i^{-3}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0) - \mu_i^{-4}(\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0))^2, \\
\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i\partial\mu_j} &= -(\mu_i\mu_j)^{-2}(\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_j(\boldsymbol{\theta}_0))^2, \quad i\neq j.
\end{aligned}$$

这些结果类似于利用链式法则（11）以及以下矩阵求导规则所作的引理 3 的证明（Petersen和Pedersen，2012 年）：

$$\frac{\partial\mathbf{a}^\mathsf{T}\mathbf{X}\mathbf{a}}{\partial\mathbf{X}} = \mathbf{a}\mathbf{a}^\mathsf{T}, \quad \frac{\partial\log\det(\mathbf{Y})}{\partial x} = \operatorname{tr}\left(\mathbf{Y}^{-1}\frac{\partial\mathbf{Y}}{\partial x}\right), \quad \text{and} \quad \frac{\partial\mathbf{Y}^{-1}}{\partial x} = -\mathbf{Y}^{-1}\frac{\partial\mathbf{Y}}{\partial x}\mathbf{Y}^{-1}.$$

令 $\boldsymbol{u}_i = \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)/\sqrt{\mu_i}$，且 $\mathbf{U}$ 为行向量为 $\boldsymbol{u}_i^\mathsf{T}$ 的矩阵。另设 $\mathbf{A} = \mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}$，且 a_{ij} 为 $\mathbf{A}$ 的元素。注意以下几点:

- $\mathbf{U}^\mathsf{T}\mathbf{U} = \sum_{i\in\mathcal{D}}\mu_i^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T} = \mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$,
- $a_{ij} = \boldsymbol{u}_i^\mathsf{T}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\boldsymbol{u}_j = (\mu_i\mu_j)^{-1/2}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_j(\boldsymbol{\theta}_0)$,
- $\mathbf{A}$ is an idempotent matrix, i.e., $\mathbf{A}^2 = \mathbf{A}$, which implies that $a_{ii} = \sum_j a_{ij}^2$,
- $a_{ii} = \boldsymbol{u}_i^\mathsf{T}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\boldsymbol{u}_i = \boldsymbol{u}_i^\mathsf{T}\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)^{-1}\boldsymbol{u}_i > 0$, since $\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ by assumption is positive definite.

现在我们可以写出

$$\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i\partial\mu_j} = \begin{cases} \mu_i^{-2}(2a_{ii} - a_{ii}^2) & \text{if } i=j, \\ (\mu_i\mu_j)^{-1}a_{ij}^2 & \text{if } i\neq j. \end{cases}$$

我们注意到海森矩阵 $\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^\mathsf{T}}$ 是 Hadamard积 $\mathbf{M}\otimes\mathbf{B}$，由一个秩一矩阵 $\mathbf{M} = \mathbf{m}\mathbf{m}^\mathsf{T}$ 和$\mathbf{m} = (\mu_1^{-1},\ldots,\mu_N^{-1})^\mathsf{T}$，以及一个对称矩阵 $\mathbf{B}$ 组成，其元素为

$$b_{ij} = \begin{cases} 2a_{ii} - a_{ii}^2 & \text{if } i=j, \\ a_{ij}^2 & \text{if } i\neq j. \end{cases}$$

矩阵 $\mathbf{M}$ 的特征值为 $\mathbf{m}^\mathsf{T}\mathbf{m}$ 和 0，因此是半正定的。矩阵$\mathbf{B}$ 是对角占优且具有正元素，因为 $a_{ii} = \sum_j a_{ij}^2$，$b_{ii} = 2a_{ii} - a_{ii}^2 = a_{ii} + \sum_{j\neq i}a_{ij}^2$，并且 $a_{ii} > 0$ 意味着

$$b_{ii} > a_{ii} > 0, \quad \text{and } b_{ii} > \sum_{j\neq i}a_{ij}^2 = \sum_{j\neq i}b_{ij}.$$

因此，$\mathbf{B}$ 是正定的（Horn 和 Johnson，1990）。由此可得，对于 PO-WR 和多项式抽样设计，海森矩阵 $\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^\mathsf{T}}$ 在 $\mathcal{M}_n$ 上是半正定的。

剩下要证明 PO-WOR 的凸性。首先注意，函数 log det $(\mathbf{V}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 根据关于$\mathbf{V}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)$的假设，在 $\mathcal{M}_n$上可微且连续。根据连续性，只需证明海森矩阵在 $\mathcal{M}_n$的内部是半正定的。因此，考虑满足对所有 i均有$\mu_i \leq 1$的点 $\boldsymbol{\mu} \in \mathcal{M}_n$。令 $\boldsymbol{u}_i = \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\sqrt{1-\mu_i}/\sqrt{\mu_i}$，并令$\mathbf{U}$ 为行向量为 $\boldsymbol{u}_i^\mathsf{T}$的矩阵。再令$\mathbf{A}\ = \mathbf{U}(\mathbf{U}^\mathsf{T}\mathbf{U})^{-1}\mathbf{U}^\mathsf{T}$，且 a_{ij} 为$\mathbf{A}$的元素。与上文类似，现在可以写成

$$\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\mu_i\partial\mu_j} = \begin{cases}\mu_i^{-2}(1-\mu_i)^{-1}(2a_{ii}-a_{ii}^2) & \text{if } i=j,\\ (\mu_i\mu_j)^{-1}(1-\mu_i)^{-1/2}(1-\mu_j)^{-1/2}a_{ij}^2 & \text{if } i\neq j.\end{cases}$$

我们将海森矩阵 $\frac{\partial^2\Phi(\mathbf{V}(\boldsymbol{\mu};\boldsymbol{\theta}_0))}{\partial\boldsymbol{\mu}\partial\boldsymbol{\mu}^\mathsf{T}}$ 识别为 Hadamard积 $\mathbf{M}\otimes\mathbf{B}$，其中一个因子是秩一矩阵 $\mathbf{M}\ = \mathbf{m}\mathbf{m}^\mathsf{T}$，其中$\mathbf{m}\ = (\mu_1^{-1}(1-\mu_1)^{-1/2},\ldots,\mu_N^{-1}(1-\mu_N)^{-1/2})^\mathsf{T}$，另一个因子是对称矩阵 $\mathbf{B}$，其元素为

$$b_{ij} = \begin{cases}2a_{ii}-a_{ii}^2 & \text{if } i=j,\\ a_{ij}^2 & \text{if } i\neq j.\end{cases}$$

证明的其余部分　完全仿照 PO-WR 和多项式抽样设计的证明即可得出　s.

A.7 命题3的证明

a) 的证明 首先注意到，海森矩阵 $\mathbf{H}_d(\boldsymbol{\theta})$ 在 $\boldsymbol{\theta}=\boldsymbol{\theta}_0$处半正定，因为 $\boldsymbol{\theta}_0$ 是 $d(\boldsymbol{\theta})$ 的全局极小值点。因此，存在矩阵 $\mathbf{L}$ 使得 $\mathbf{L}\mathbf{L}^\mathsf{T}\ = \mathbf{H}_d(\boldsymbol{\theta})$。由于根据假设，$\mathbf{H}_d(\boldsymbol{\theta})$ 非零，因此 $\mathbf{L}$ 非零。<style id='d-最优抽样方案 $\boldsymbol{\mu}^*$ 定义为函数 tr($\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)$ $\mathbf{H}_d(\boldsymbol{\theta}_0)$) =tr($\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}\mathbf{L}^\mathsf{T}$), 的最小化器；根据定义，这等价于关于满足 $\mathbf{L}\mathbf{L}^\mathsf{T} = \mathbf{H}_d(\boldsymbol{\theta})$.的矩阵 $\mathbf{L}$ 的 L-最优性。

b) 的证明 假设 $\boldsymbol{\mu}^*$ 是 $\Phi(\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0))$ 的最小化器，并令 $d(\boldsymbol{\theta})\ = \frac{1}{2}||\mathbf{L}(\boldsymbol{\mu}^*;\ \boldsymbol{\theta}_0)^\mathsf{T}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)||_2^2$，其海森矩阵为 $\mathbf{H}_d(\boldsymbol{\theta}_0)$ $= \mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}$。根据命题 1，$\Phi$- 最优抽样方案 $\boldsymbol{\mu}^*$ 必须满足最优性条件 (13) 或 (14a)–(14c)，其中

$$\begin{aligned}c_i &= ||\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)||_2^2\\ &= \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0)\\ &= \boldsymbol{\psi}_i(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\mathbf{H}_d(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\psi}_i(\boldsymbol{\theta}_0).\end{aligned}$$

这与 d-最优性准则的最优性条件相同。此外，根据命题 2a) 和 3a)，d-最优性准则关于 $\boldsymbol{\mu}$ 是凸的，因此 $\boldsymbol{\mu}^*$ 必定是 d-最优性准则的全局极小值点。现在，对 tr($\boldsymbol{\Gamma}(\boldsymbol{\mu};\ \boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0)$) 求最小值，等价于对任意常数 $k>0$，对 tr($k\boldsymbol{\Gamma}$ $(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{H}_d(\boldsymbol{\theta}_0)$) 求最小值。因此，相对于距离函数 $d(\boldsymbol{\theta})\ = ||\mathbf{L}(\boldsymbol{\mu}^*;\boldsymbol{\theta}_0)^\mathsf{T}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)||_2^2$，$\boldsymbol{\mu}^*$ 也是 d-最优的。

命题4的证明

结果来自　命题3a)，因为 $d_{\mathrm{ER}}(\boldsymbol{\theta})$、$d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ 和 d_{KL} 的海森矩阵由　y

$$\begin{aligned}\frac{\partial^2 d_{\mathrm{ER}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} &= \frac{\partial^2\ell_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} = \mathbf{H}(\boldsymbol{\theta}),\\ \frac{\partial^2 d_{\boldsymbol{\Sigma}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = \boldsymbol{\Sigma}^{-1},\ \text{and}\\ \frac{\partial^2 d_{\mathrm{KL}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}\sum_{i\in\mathcal{D}}\int_{\mathcal{Y}}\log\frac{f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)}{f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)}dF_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)\\ &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}\sum_{i\in\mathcal{D}}\int_{\mathcal{Y}}\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)dF_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)\\ &= -\sum_{i\in\mathcal{D}}\int_{\mathcal{Y}}\frac{\partial^2\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}dF_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}_i)\\ &= \mathrm{E}_{\boldsymbol{y}\sim f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x})}\left[-\sum_{i\in\mathcal{D}}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}\log f_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}_i)\right]\\ &= \mathrm{E}_{\boldsymbol{y}\sim f_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x})}[\mathbf{H}(\boldsymbol{\theta})] = \widetilde{\mathbf{H}}(\boldsymbol{\theta}).\end{aligned}$$

对于库尔贝克-莱布勒距离的海森矩阵，我们使用了莱布尼茨积分法则来改变积分与微分的顺序（参见库尔贝克和莱布勒，1951 年）。

A.9 命题 5 的证明

考虑一个一一对应的可微映射 $\boldsymbol{g}:\ \boldsymbol{\theta}\mapsto\boldsymbol{\eta}$。将诱导的经验风险记为 $\ell_0^*(\boldsymbol{\eta})\ =\sum_{i\in\mathcal{D}}\ell_i(\boldsymbol{g}^{-1}(\boldsymbol{\eta}))$，其最小化器为 $\boldsymbol{\eta}_0=\boldsymbol{g}(\boldsymbol{\theta}_0)$。根据链式法则， $\ell_0^*(\boldsymbol{\eta})$ 在 $\boldsymbol{\eta}=\boldsymbol{\eta}_0$ 处的海森矩阵为

$$\begin{aligned}\mathbf{H}_{\ell_0^*}(\boldsymbol{\eta}_0)=\left.\frac{\partial^2\ell_0^*(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^\mathsf{T}}\right|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}&=\mathbf{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{g}^{-1}(\boldsymbol{\eta}_0))^\mathsf{T}\mathbf{H}_{\ell_0}(\boldsymbol{g}^{-1}(\boldsymbol{\eta}_0))\mathbf{J}_{\boldsymbol{g}^{-1}}(\boldsymbol{g}^{-1}(\boldsymbol{\eta}_0))\\&=\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}.\end{aligned}$$

这里还用到了以下事实：根据 $\boldsymbol{\theta}_0$ 作为 $\ell_0(\boldsymbol{\theta})$的最小化器的定义， $\nabla_{\boldsymbol{\theta}}\ell_0(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}=\mathbf{0}$。

现在假设，对于 $\boldsymbol{\theta}_0$ 和 $\boldsymbol{\eta}_0$， $\boldsymbol{\mu}^*$ 和 $\tilde{\boldsymbol{\mu}}^*$ 分别是 d_{ER}-最优的。后者是指， $\tilde{\boldsymbol{\mu}}^*$ 使$\boldsymbol{\eta}_{\boldsymbol{\mu}}=\boldsymbol{g}(\boldsymbol{\theta}_{\boldsymbol{\mu}})$与 $\boldsymbol{\eta}_0=\boldsymbol{g}(\boldsymbol{\theta}_0)$之间的期望距离最小，其中该期望是相对于诱导的经验风险距离 $d_{\mathrm{ER}}^*(\boldsymbol{\eta})\ =\ell_0^*(\boldsymbol{\eta})\ -\ \ell_0^*(\boldsymbol{\eta}_0)$而言的。根据命题4， $\boldsymbol{\mu}^*$ 相对于满足 $\mathbf{L}$ 的矩阵$\mathbf{L}$ 是 L-最优的，其中$\mathbf{L}\mathbf{L}^\mathsf{T}=\mathbf{H}(\boldsymbol{\theta}_0)$。类似地， $\tilde{\boldsymbol{\mu}}^{*\sim}$相对于满足以下条件的矩阵$\mathbf{L}$ 是 L-最优的：

$$\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\mathsf{T}=\mathbf{H}_{\ell_0^*}(\boldsymbol{\eta}_0)=\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}\mathbf{H}(\boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}=\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}\mathbf{L}\mathbf{L}^\mathsf{T}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}.\tag{27}$$

现在， $\boldsymbol{\mu}^*$ 是该函数的最小化器

$$\begin{aligned}\operatorname{tr}(\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{L}\mathbf{L}^\mathsf{T})&=\operatorname{tr}(\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)\boldsymbol{\Gamma}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}\mathbf{L}\mathbf{L}^\mathsf{T})\\&=\operatorname{tr}(\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}\mathbf{L}\mathbf{L}^\mathsf{T}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1})\\&=\operatorname{tr}(\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\mathsf{T}).\end{aligned}$$

第一个等式是通过两次插入单位矩阵 $\mathbf{I}_{p\times p}\ =\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-1}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)\ =\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^\mathsf{T}\mathbf{J}_{\boldsymbol{g}}(\boldsymbol{\theta}_0)^{-\mathsf{T}}$ 得到的；第二个等式由式（16）和迹的循环性质得到；第三个等式由式（27）得到。但是， $\tilde{\boldsymbol{\mu}}^*$ 也是 $\operatorname{tr}(\boldsymbol{\Gamma}_{\boldsymbol{g}}(\boldsymbol{\mu};\boldsymbol{\theta}_0)\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\mathsf{T})$ 的一个最小化器。由于 L-最优性准则关于 $\boldsymbol{\mu}$ 是凸的，最优解是唯一的，因此必须有 $\boldsymbol{\mu}^*=\tilde{\boldsymbol{\mu}}^*$。所以， d_{ER}-最优性准则在重新参数化 $\boldsymbol{g}:\ \boldsymbol{\theta}\mapsto\boldsymbol{\eta}$下是不变的。

d_S-最优性和 d_{KL}-最优性的结果可类似地得到。

B 补充图

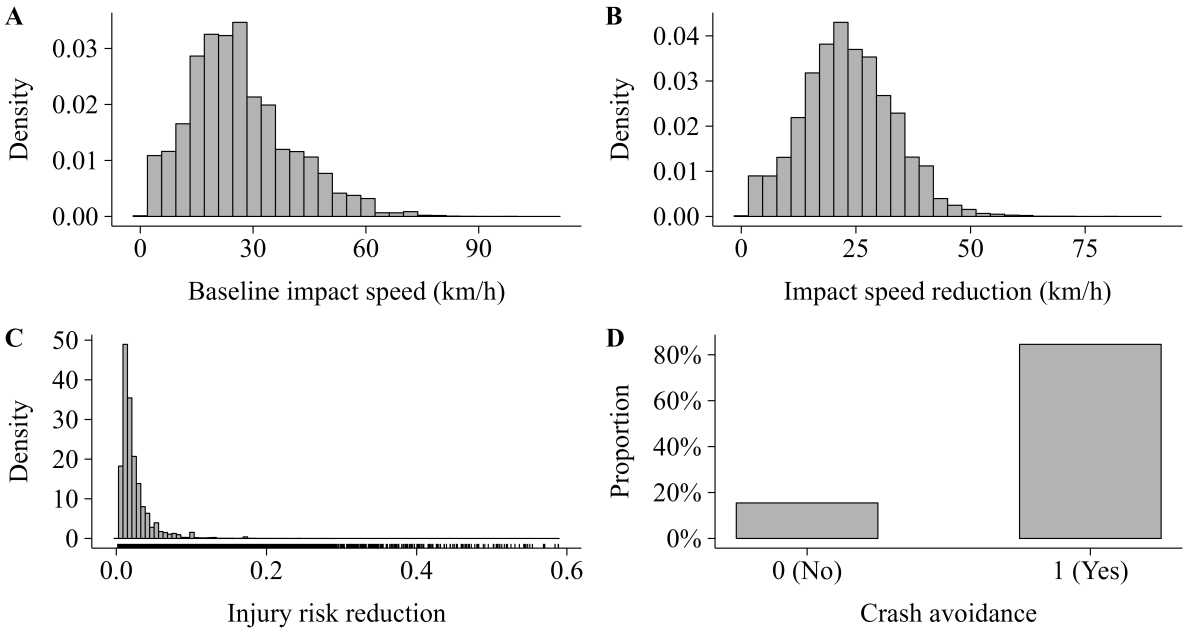


图 S1。 第 6 节所考虑的车辆安全数据集的特征。**A**：基线人工驾驶场景下的撞击速度分布。**B–D**：与基线手动驾驶场景相比，配备自动紧急制动系统时，碰撞速度降低、受伤风险降低和碰撞避免率的分布。

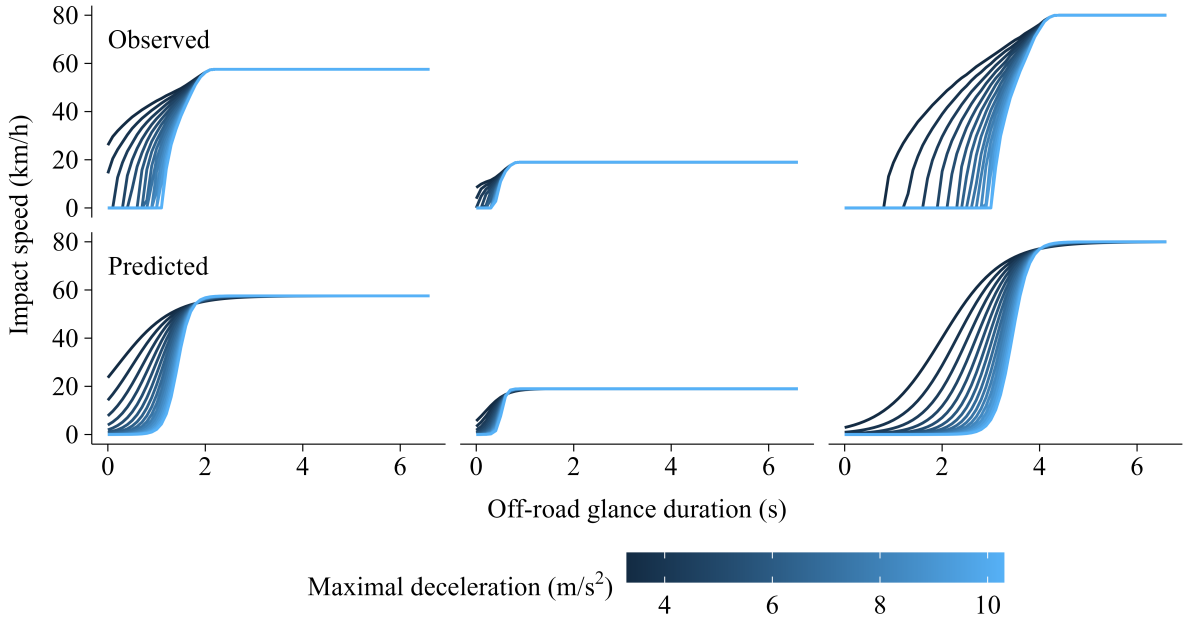


图 S2。 对于三个重建的追尾碰撞的反事实变化，碰撞速度响应面是偏离道路注视时长和制动期间最大减速度的函数。**上图**：观测到的碰撞速度。**下图**：使用准二项逻辑回归模型（21）预测的碰撞速度。响应值已从模型范围 $[0,1]$ 映射至原始范围 $[0, y_{\max,k}]$，其中 $y_{\max,\ k}$ 是由案例 k、 $k=1$，……，44 生成的变化中的最大可能碰撞速度。