超大数据集下装袋估计量的渐近性质

Yuan Gao[a], Riquan Zhang[a*] Hansheng Wang[b]

[a]统计学院与 *KLATASDS-MOE*，华东师范大学，中国上海；[b]北京大学光华管理学院，中国北京

arXiv:2304.06278v1 [math.ST] 13 Apr 2023

摘要

装袋法是大规模统计分析的一种有效方法，尤其是在计算资源非常有限时。本文研究海量数据集下 M-估计问题中装袋估计量的渐近性质。我们从理论上证明，在适当条件下，所得估计量是一致的且渐近正态。结果表明，只要装袋子样本量和子样本数量足够大，装袋估计量就能达到最优统计效率。此外，我们推导出一种方差估计量，可用于有效的渐近推断。所有理论结果均通过大量模拟研究进一步验证。最后，我们将装袋法应用于美国航空公司数据集，以展示其实际用途。

关键词：子抽样、装袋法、 M-估计量、渐近分析

*通讯作者。电子邮箱：rqzhang@stat.ecnu.edu.cn

1 引言

由于信息时代数据的爆炸式增长，海量数据集经常出现在各个领域。例如，社交媒体推文、网上购物记录、监控摄像视频和在线搜索查询。它们每天都会产生数量巨大的数据（Fan et al.，2020）。然而，海量数据的规模也使相应的统计估计和推断变得极具挑战性。数据通常可能过于庞大，无法完整地加载到计算机内存中。这意味着直接计算基于完整样本的经典估计量会十分困难，甚至不可行。因此，如何在提供统计保证的同时有效分析这些海量数据，是一个极其重要的问题。

过去十年间，人们提出了许多方法来应对大规模统计计算问题。这些方法大致可分为两大类。第一类是分布式计算方法，其基本思想是分而治之，即将复杂问题分解为若干小问题，再由多台机器并行处理。分布式计算已被证实在许多大规模统计问题中非常有用，例如分位数回归（Volgushev 等，2019；Chen 等，2020；Pan 等，2021）、 M估计（Zhang 等，2013；Huang and Huo，2019；Jordan 等，2019）、变量选择与特征筛选（Zhu 等，2021a；Li 等，2020）、主成分分析（Fan 等，2019；Chen 等，2021），以及非参数和半参数回归（Chang 等，2017；Lian 等，2019）。第二类方法是子抽样方法。其核心思想是根据从完整样本中精心选取的子样本计算估计量。因此，选择相对较小的子样本可以大幅降低相应的计算成本。为了充分利用包含信息的观测值，人们设计了多种非均匀抽样策略，包括基于杠杆分数的子抽样（Drineas 等，2012；Ma 等，2015、2020）和最优子抽样（Wang 等，2018、2019；Wang and Ma，2021）。

当可用计算资源十分有限时，这些子抽样方法尤其有用。不过，这些方法有两点局限值得一提。首先，需要仔细设定每个观测值的子抽样概率。通常，这会导致抽样成本达到 $O(N)$ 的量级，其中 N 是完整样本的样本量。其次，所得估计量仅基于一个或两个相对较小的子样本计算。因此，与基于完整样本的估计量相比，其统计效率通常较低。为解决这些问题，自助聚合方法（Breiman，1996；Bühlmann and Yu，2002），即装袋法，提供了一种有效方案。装袋法的基本思想是，基于通过有放回简单随机抽样生成的多个子样本计算估计量，随后将这些估计量聚合为一个更稳定的估计量。近年来，类似策略已被用于解决大规模统计问题。例如，Wu 等（2021）提出了两种基于子抽样的矩估计量，并研究了它们的渐近性质。Zhu 等（2021b）提出了一种基于子抽样的特征筛选程序，并建立了相应的筛选一致性。然而，对于海量数据集上的一般 M 估计问题，装袋估计量的渐近性质似乎仍鲜为人知。为此，我们致力于从理论上研究超大数据集下的装袋 M 估计量。需要指出的是，我们的工作与现有文献密切相关，但也有明显区别。例如，Wager（2014）研究了基于树的装袋估计量（即随机森林）的渐近性质，而我们关注的是参数回归模型。再如，Kleiner 等人（2014）提出了用于分布式估计的小型自助法集成（BLB），其中使用加权子样本。相比之下，我们使用未加权子样本，因此我们的方法更易于实现和研究。近期，Zou 等（2021）研究了一个类似问题，但其子样本是通过无放回简单随机抽样获得的。与之不同，我们的子样本通过有放回抽样获得，因此抽样成本大幅降低。当数据集规模超大、因而必须存放在硬盘上时，这一点尤其明显。

对于 这一问题的精彩讨论，请参阅Wu 等人（2021） .

为实施装袋法，我们采用简单随机有放回抽样生成子样本。值得指出的是，这种抽样策略不仅简单易于实际操作，而且省去了指定子抽样概率的成本。获得一个装袋子样本后，我们便可计算感兴趣参数对应的装袋子样本估计量。随后，对这些子样本估计量取平均，得到最终的装袋估计量。我们从理论上研究了所得估计量的统计性质，并证明了其一致性和渐近正态性。此外，我们还表明，在温和的条件下，装袋估计量在渐近意义上可以与基于完整样本的估计量同样有效。我们还为装袋估计量构造了方差估计量，以便进行有效的渐近推断。我们开展了大量模拟实验，以验证理论结果。

本文其余部分安排如下。第2 节介绍问题设定和装袋估计量的详细内容，并给出渐近分析。第 3 节介绍模拟研究和一个真实数据示例。第 4 节总结全文。所有技术证明均包含在附录中。

2 方法论

2.1 模型与符号 обозначations

令 $Z_i = (X_i^\top,\ Y_i)^\top \in \mathbb{R}^{p+1}$ 表示从第 i个受试者收集的观测值，其中N 为完整样本的样本量。此外， $Y_i \in \mathbb{R}$ 是我们关注的响应变量， X_i =(X_{i1}，……， X_{ip})$^\top \in \mathbb{R}^p$ 是相应的 p维协变量向量。我们假设 Z_i ($1 \le i \le N$)相互独立且同分布。令 $\ell_i(\theta) = \ell(\theta;$ Z_i) 为一个合理的平滑损失函数。于是，全局损失函数定义为 $\mathcal{L}(\theta) = \sum_{i=1}^{N} \ell_i(\theta)$, 其极小值点为 $\widehat{\theta} = \arg\min_\theta \mathcal{L}(\theta)$。在全文中

$\widehat{}$，我们将 θ 称为全局估计量，以强调它是基于完整样本得到的。令 θ_0 为未知参数的真实值 θ，$\widehat{}$其定义为 $\theta_0 = \text{argmin}_\theta E\{\ell_i(\theta)\}$。在许多情况下，全局估计量 θ 满足以下渐近规律（Shao，2003；Van der Vaart，2000）：

$$\sqrt{N}(\widehat{\theta} - \theta_0) \to^d \mathcal{N}(0, \Sigma)$$

对于某个正定矩阵 $\Sigma \in \mathbb{R}^{p\times p}$，当 $N \to \infty$ 时。然而，当完整样本的样本量 N 极大时，计算 θ 可能会很困难，尤其是在从业者的计算资源非常有限的情况下。因此，我们需要一种计算上更可行的方法。为此，我们认为装袋法是适用于大规模统计估计与推断的实用解决方案。

具体而言，令 $\mathbb{S} = \{1, \ldots, N\}$ 表示完整样本的索引集。令 n 表示装袋子样本的大小，K 表示子样本的数量。将 $\mathcal{S}_k = \{i_1^{(k)}, \ldots, i_n^{(k)}\} \subset \mathbb{S}$ 记为第 k 个装袋子样本，其中 $i_m^{(k)}$（对于任意 $1 \leq m \leq n$，$1 \leq k \leq K$）通过有放回简单随机抽样的方法从 $\mathbb{S}$ 中独立抽取。换言之，给定 $\mathbb{S}$，$i_m^{(k)}$ 相互独立且同分布，其概率为 $P(i_m^{(k)} = i) = 1/N$，对任意 $i \in \mathbb{S}$ 均成立。因此，对于每个 $1 \leq k \leq K$，我们可以基于装袋子样本 $\mathcal{S}_k$ 得到一个估计量：$\widehat{\theta}^{(k)} = \arg\min_{\theta\in\mathbb{R}^p} \mathcal{L}_k(\theta)$，其中 $\mathcal{L}_k(\theta) = \sum_{i\in\mathcal{S}_k} \ell_i(\theta)$。随后，我们可以将这些子样本估计量聚合成一个更稳定的估计量：$\widehat{\theta}_{\text{BAG}} = K^{-1}\sum_{k=1}^K \widehat{\theta}^{(k)}$。$\widehat{}$下文将 θ_{BAG} 称为装袋估计量。

2.2 装袋估计量的统计性质

为研究装袋估计量的理论性质，我们假设以下技术条件。

(C1) **(参数空间)** 参数空间 Θ 是一个紧致凸集 i n

$\mathbb{R}^p$。此外，真值 θ_0 位于 Θ的内部。

(C2) (凸性) 损失函数 $\ell(\theta; Z)$ 关于 θ 是凸函数，对几乎所有 $Z \in \mathbb{R}^{p+1}$ 均成立。

(C3) (梯度与海森矩阵) 假设损失函数的总体梯度在 θ_0 处为零，即 $E\{\dot{\ell}_i(\theta_0)\} = 0$。令 $V(\theta) = \text{var}\{\dot{\ell}_i(\theta)\}$，为简便起见，记 $V = V(\theta_0)$。令 $H(\theta) = E\{\ddot{\ell}_i(\theta)\}$ 表示损失函数在 θ 处的总体海森矩阵，为简便起见，记为 $H = H(\theta_0)$。假设 V 和 H 均为正定矩阵。此外，定义 $\Sigma = V^{-1}HV^{-1}$。

(C4)（光滑性）定义 $B(\delta) = \{\theta \in \Theta : \|\theta - \theta_0\| < \delta\}$ 为以 θ_0 的真实值为中心、半径为 $\delta > 0$ 的球。接下来作出以下两个假设。**(C4.1)** 存在某些常数 G_1 和 G_2，使得对于任意 $\theta \in B(\delta)$，$E\{\|\dot{\ell}_i(\theta_0)\|^8\} \leq G_1^8$ 和$E\{\|\ddot{\ell}_i(\theta_0) - H(\theta)\|^8\} \leq G_2^8$。**(C4.2)** 对几乎所有 $Z \in \mathbb{R}^{p+1}$，$\ddot{\ell}(\theta; Z)$ 海森矩阵 $\ell(\theta; Z)$ 均为 $L(Z)$-利普希茨连续，即对于任意 θ、$\theta' \in B(\delta)$，$\|\ddot{\ell}(\theta; Z) - \ddot{\ell}(\theta'; Z)\| \leq L(Z)\|\theta - \theta'\|$，其中 $L(Z)$ 满足 $E\{L^8(Z)\} \leq L^8$ 和$E\left[\{L(Z) - E(L(Z))\}^8\right] \leq L^8$，且某个常数 L 存在。

条件 (C1) 假设参数空间是紧致且凸的，这一假设已广泛用于以往研究中（Zhang 等人，2013；Huang and Huo，2019年；Zhu 等人，2021a）。条件 (C2)-(C3) 是标准的正则性条件。通常假设这些条件，以建立 M-估计的一致性和渐近正态性（Van derVaart，2000；Lehmann and Casella，2006）。最后，条件 (C4) 假设损失函数在真值 θ_0附近足够光滑。否则，无法应用标准的泰勒展开方法。这也是过去文献中广泛使用的一组标准条件（Zhang 等人，2013；Huang and Huo，2019年；Jordan 等人，2019年）。接下来给出以下引理。

引理 1. 假设条件（*C1*）–（*C4*）。则有 $\sqrt{n}(\widehat{\theta}^{(k)} - \theta_0) = O_p(1)$.

该引理的证明可见附录 A。根据引理 1，我们知道$\widehat{\theta}^{(k)}$ 是 θ_0 对于每个 $1 \leq k \leq K$的相合估计量。此外，回想一下

$\dot{\mathcal{L}}_k(\widehat{\theta}^{(k)}) = 0$。根据引理 1 和中值定理（Shao，2003），我们可以得到

$$
\begin{aligned}
0 =& n^{-1}\dot{\mathcal{L}}_k(\widehat{\theta}^{(k)}) = n^{-1}\dot{\mathcal{L}}_k(\theta_0) + n^{-1}\int_0^1 \ddot{\mathcal{L}}_k\Big((1-t)\theta_0 + t\widehat{\theta}^{(k)}\Big)dt\,(\widehat{\theta}^{(k)} - \theta_0) \\
=& n^{-1}\dot{\mathcal{L}}_k(\theta_0) + H(\widehat{\theta}^{(k)} - \theta_0) + \Delta^{(k)},
\end{aligned}
$$

其中 $\Delta^{(k)} = \Big\{n^{-1}\ddot{\mathcal{L}}_k(\theta_0) - H\Big\}(\widehat{\theta}^{(k)} - \theta_0) + n^{-1}\Big\{\int_0^1 \ddot{\mathcal{L}}_k\Big((1-t)\theta_0 + t\widehat{\theta}^{(k)}\Big)dt - \ddot{\mathcal{L}}_k(\theta_0)\Big\}(\widehat{\theta}^{(k)} - \theta_0)$。于是有 $\widehat{\theta}^{(k)} = \theta_0 - H^{-1}n^{-1}\dot{\mathcal{L}}_k(\theta_0) - H^{-1}\Delta^{(k)} = \theta_0 + Q_1^{(k)} + Q_2^{(k)}$。因此，装袋估计量可表示为 $\theta_{\mathrm{BAG}} = \theta_0 + Q_1 + Q_2$，其中$Q_1 = K^{-1}\sum_{k=1}^K Q_1^{(k)}$ 且 $Q_2 = K^{-1}\sum_{k=1}^K Q_2^{(k)}$。$Q_1$ 和 Q_2 的性质见以下定理。

定理1。假设满足条件（*C1*）–（*C4*）。那么有 $\theta_{\mathrm{BAG}} - \theta_0 = Q_1 + Q_2$，其中$E(Q_1) = 0$，方差$(Q_1) = \{(nK)^{-1} + N^{-1}\}\Sigma + o\{(nK)^{-1} + N^{-1}\}$，且 $Q_2 = O_p(n^{-1})$。

该定理的详细证明见附录 B。如定理 1所示，装袋估计量的变异性主要由项 Q_1 刻画，该项进一步由两部分决定。第一部分 $N^{-1}\Sigma$ 与完整样本量 N有关，装袋法无法减少完整样本量。第二部分（nK）$^{-1}\Sigma$ 与装袋子样本量 n 及子样本数 K有关。该项反映了抽样成本与估计变异性之间的权衡。另一方面，偏差项 Q_2 主要由装袋子样本量 n决定。这是因为装袋法中的平均操作无法减少估计偏差。因此，在实践中，我们可能会期望装袋子样本量 n 应当足够大（例如，$n > \sqrt{N}$），以获得令人满意的估计精度。我们进一步建立了装袋估计量的渐近正态性，具体如下述定理所示。

—

定理 2。假设条件（*C1*）–（*C4*）且 $n/\sqrt{N} \to \infty$ 当 $N \to \infty$。那么我们有 $\{$（nK）$^{-1} + N^{-1}\}^{-1/2}(\widehat{\theta}_{\mathrm{BAG}} - \theta_0) \to_d \mathcal{N}(0, \Sigma)$ 当 $N \to \infty$。

参见 附录 C 给出了该定理的详细证明。根据定理 2 我们可以看出 t,
A

只要满足 $n \gg \sqrt{N}$ 和 $nK \gg N$，装袋估计量在渐近意义上就应与全局估计量同样有效。

对于推断，我们需要估计装袋估计量的方差。请注意，我们已经获得了基于不同子样本的 K 个 θ_0 估计值。利用这些估计值，我们可以构造方差估计量，

$$\widehat{\mathrm{SE}}^2(\widehat{\theta}_{\mathrm{BAG}}) = \left(\frac{1}{nK} + \frac{1}{N}\right) \frac{n}{K} \sum_{k=1}^{K} \left(\widehat{\theta}^{(k)} - \widehat{\theta}_{\mathrm{BAG}}\right)\left(\widehat{\theta}^{(k)} - \widehat{\theta}_{\mathrm{BAG}}\right)^{\top}. \tag{2.1}$$

其性质由下述定理给出。定理的详细证明见附录 D。由定理 $\widehat{\text{标准误}}_2(\widehat{\theta}_{\mathrm{BAG}})$ 可知，$\widehat{\text{它}}$应是 bagging 估计量渐近方差的合理估计量 θ_{BAG}。因此，我们可以用它进行进一步的推断，例如构造置信区间。我们通过大量模拟实验展示 $\widehat{\text{标准误}}_2(\widehat{\theta}_{\mathrm{BAG}})$ 的表现。

定理 3. 假设满足条件（*C1*）–（*C4*）。进一步假设当 $N \to \infty$ 时，$n \to \infty$、$n/N \to 0$ 和 $K \to \infty$。那么我们有 $E\left\{\widehat{\text{标准误}}_2(\widehat{\theta}_{\mathrm{BAG}})\right\} = \{(nK)^{-1} + N^{-1}\}\Sigma + o\{(nK)^{-1} + N^{-1}\}$.

3 数值研究

3.1 模拟研究

为验证装袋估计量在有限样本下的表现，我们开展了若干模拟研究。假设完整样本共包含 $N = 200,000$ 个观测值。对于每个 $1 \le i \le N$，我们在三种不同模型下生成 (X_i, Y_i)：线性回归模型、逻辑回归模型和泊松回归模型。对于每个模型，我们将真实参数设为 $\theta_0 = (-0.2, -0.1, 0, 0.1, 0.2)^{\top}$。更详细的模型设定如下。

示例 1.（线性回归）在此示例中，我们考虑 $p = 5$ 协变量$X_i = (X_{i1}, X_{i2}, \ldots, X_{i5})^{\top}$，其中每个协变量均独立地生成自

标准正态分布。然后，响应变量 Y_i 由与协变量 X_i 的线性关系生成，具体如下：

$$Y_i = X_i^\top \theta_0 + \varepsilon_i,$$

其中，噪声项 ε_i 独立地服从标准正态分布。

示例 2.（逻辑回归）在本例中，我们同样考虑 $p = 5$个协变量。与上文相同，每个协变量均独立地服从标准正态分布。随后，响应变量 Y_i 服从伯努利分布，其响应概率为

$$P(Y_i = 1|X_i, \theta_0) = \frac{\exp(X_i^\top \theta_0)}{1 + \exp(X_i^\top \theta_0)}.$$

示例 3.（泊松回归）在本例中，我们同样考虑 $p = 5$个协变量。它们服从多元正态分布 $\mathcal{N}(0, \Sigma_X)$，其中 $\Sigma_X = (\Sigma_{X,\,ij})_{1\le i,\,j\le p}$ 且 $\Sigma_{X,\,ij} = 0.5^{|i-j|}$。随后，响应变量 Y_i 服从泊松分布，其分布为

$$P(Y_i = k|X_i, \theta_0) = \frac{\lambda_i^k}{k!} \exp(-\lambda_i), \text{ where } \lambda_i = \exp(X_i^\top \theta_0).$$

一旦模拟出完整数据集 $\{(X_i, Y_i) : 1 \le i \le N\}$，我们便使用上述方法计算装袋估计量。为进行全面评估，我们考虑了不同的装袋子样本量（$n = 500$、750、1000）和子样本数（K =50、100、150、200、250）。设 $\widehat{\theta}^{(b)}$ BAG $= \big(\widehat{\theta}^{(b)}$ BAG $_j\big)_{j=1}^{p}$ 为在第 b 次（$1 \le b \le B$）模拟中，使用 $B = 1,000$ 得到的装袋估计量。我们将偏差计算为 Bias$_j = |\theta_{\text{BAG},j} - \theta_{0j}|$，其中 θ_{0j} 是真实值 θ_0 的第 j 个分量，且 $\theta_{\text{BAG},j} = B^{-1} \sum_{b=1}^{B} \widehat{\theta}^{(b)}$ BAG $_j$。标准误 SE$_j^{(b)}$ 可利用（2.1）估计，即 $\widehat{\text{SE}}_j^{(b)} = \sqrt{\widehat{S}_{jj}^{(b)}}$，其中 $\widehat{S}_{jj}^{(b)}$ 是 $\widehat{\text{SE}}^2(\widehat{\theta}^{(b)}$ BAG) 的第 j 个对角元素。我们报告了 –

平均值 $\widehat{\text{标准误}}_j = B^{-1}\sum_{b=1}^{B}\text{标准误}_j^{(b)}$。我们还将 $\widehat{\text{标准误}}_j$ 与 $\widehat{\theta}^{(b)}\,\text{BAG}_j$ 的蒙特卡洛标准差进行比较。该值通过 $\text{SE}_j = \sqrt{B^{-1}\sum_{b=1}^{B}\left(\widehat{\theta}^{(b)}\,\text{BAG}_j - \theta_{\text{BAG},j}\right)^2}$ 计算得出。最后，我们为 θ_j 构造一个 95% 置信区间：$\text{CI}_j^{(b)} = \left(\widehat{\theta}^{(b)}\,\text{BAG}_j - z_{0.975}\,\widehat{\text{标准误}}_j^{(b)}, \widehat{\theta}^{(b)}\,\text{BAG}_j + z_{0.975}\,\widehat{\text{标准误}}_j^{(b)}\right)$，其中 z_α 是标准正态分布的第 α 个下分位数。随后，覆盖概率计算为 $\text{ECP}_j = B^{-1}\sum_{b=1}^{B} I\{\theta_{0j} \in \text{CI}_j^{(b)}\}$，其中 $I\{\cdot\}$ 是示性函数。线性回归、逻辑回归和泊松回归的详细结果分别列于表 1–3。

一般而言，表1–3 中的模拟结果非常相似。我们可以得出以下结论。首先，随着子样本数量 K 增加，标准误值逐渐减小。这是预期的，因为根据定理1，较大的 K 会带来较小的方差。请注意，在线性回归模型中，普通最小二乘估计量是无偏的。因此，我们关注逻辑回归模型和泊松回归模型的估计偏差。从表2 和3可以看出，偏差值并未随着 K 增加而呈现出明显的下降趋势。另一方面，偏差值会随着子样本量 n 的增加而减小。这与定理1中的发现一致，因为装袋法中的平均操作无助于减小 $\widehat{\text{偏差}}$。最后，我们发现估计的标准误值与蒙特卡洛标准误值非常接近，这表明式（2.1）中构造的渐近方差公式应当是正确的。此外，我们可以看到，经验覆盖概率均在名义水平95%左右。这证实了所提出估计量的渐近正态性，也表明所提出的方差估计量具有实用价值。

接下来，我们比较装袋估计量与全局估计量的表现。具体而言，在计算装袋估计量时，我们将装袋子样本量固定为 $n = 2,000$，并令子样本数量 K 在 50 到 2，000 之间变化。令 $\widehat{\theta}^{(b)}$ 为在第 b<style id='19'}次（$1 \le b \le B$）模拟中得到的一个特定估计量（例如全局估计量），其中 $B = 500$。我们将均方误差（MSE）计算为 $\text{MSE}(\widehat{\theta}) = B^{-1}\sum_{b=1}^{B}\|\widehat{\theta}^{(b)} - \theta_0\|^2$。随后，将不同估计量的均方误差对数变换值绘于图1中。从图1可以看出，随着 K 增大，装袋估计量的均方误差值稳步下降。此外，装袋

当 $K = 2,000$ 时，装袋估计量的均方误差表现几乎与全局估计量一样好。这是预期之中的，因为根据定理2，只要 $n \gg \sqrt{N}$且 $nK \gg N$，装袋估计量渐近地应与全局估计量同样有效。

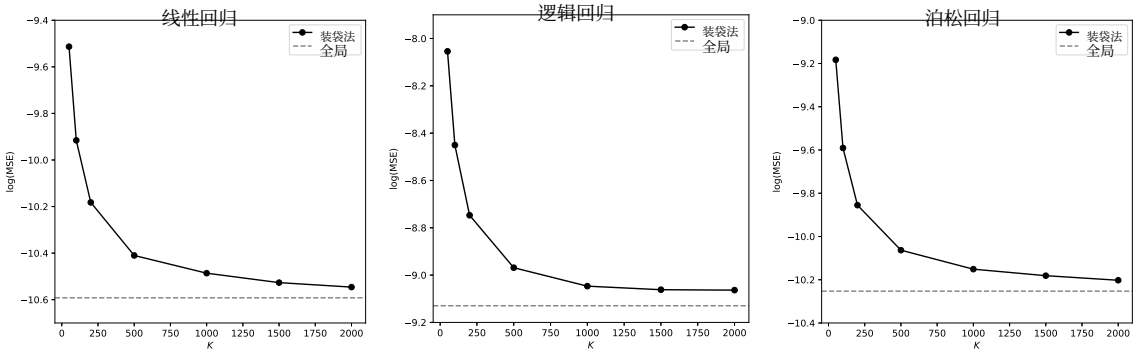

图1：三种不同模型下装袋估计量与全局估计量的对数变换均方误差值。我们将子样本量 n 固定为 2，000，并令子样本数量 K 从 50 变化至 2，000。实线表示装袋估计量，虚线表示全局估计量。

3.2 一个真实数据示例

为了说明问题，我们将装袋法应用于一个真实数据集——美国航空公司数据集（http://stat-computing.org/dataexpo/2009）。该数据集包含美国境内所有商业航班从1987年10月到2008年4月的到达和起飞详情。数据集约含1.23亿条记录，占用硬盘空间12 GB。经过数据预处理后，共保留 $N = 120{,}748{,}239$条记录。我们的任务是预测航班是否延误。我们使用二元变量 `Delayed` 表示航班的延误状态（延误为1，未延误为0）。根据美国联邦航空管理局的规定，如果航班到达时间比计划时间晚15分钟以上，则视为延误。为预测延误状态，我们考虑四个协变量。第一个是 `Distance`，表示出发地与目的地之间的距离（英里）。该数值变量经过标准化，均值为0，方差为1。第二个是计划起飞时间（`DepTime`）。它被划分为四个时段：午夜（0:00-

7:00）、早晨（7:00–12:00）、下午（12:00–18:00）和傍晚（18:00–0:00）。第三个是 `DayOfWeek`，它是一个具有七个水平的分类变量（星期一–星期日）。最后一个是 `Month`，它是一个具有十二个水平的分类变量（一月–十二月）。接下来，所有分类变量都转换为虚拟变量。请注意，为避免共线性，每个分类变量的第一个水平都会被移除。最后，模型中总共使用 $p = 22$ 个变量作为预测变量（包括截距）。

随后，我们使用逻辑回归模型研究航班的延误状态。为计算装袋估计量，我们将子样本量设为 $n = [\sqrt{N}\log\log(N)] =$32，126，并抽取总计 $K =$ ~~1,~~000 个子样本。详细的估计结果列于表 4。根据表4，我们可以得出以下有趣的发现。首先，所有协变量对应的 p-值均小于 0.001。这表明每个协变量都会对响应变量产生显著影响。其次， `Distance` 的估计系数为正，说明长距离航班更容易延误。就起飞时间而言，下午和傍晚的延误较午夜和早晨更多（系数分别为 1.057 和 0.857）。其次，星期五的航班（系数为 0.249）更常延误，而星期六的航班（系数为 -0.167）则较不容易延误。最后，我们可以看到，二月至十一月对应的估计系数均为负值。这意味着一月和十二月是航班延误较多的月份。所有这些发现都与我们的日常经验相符。

4 总结

在本文中，我们研究大规模 M估计问题的装袋估计量。我们首先从理论上考察装袋估计量的渐近偏差和方差。研究表明，装袋子样本量应足够大，以使偏差项可以忽略不计。接下来，我们建立了装袋估计量的渐近正态性。我们给出了装袋估计量达到

最优渐近效率。为了支持进一步的推断，我们还构造了一个方差估计量。最后，我们通过大量模拟实验验证了理论结果。此外，我们还提供了一个真实数据示例，以说明装袋估计量的实用性。最后，我们讨论了若干值得进一步研究的有趣课题。首先，本文假设损失函数足够光滑。然而，分位数回归和支持向量机等一些重要问题涉及不可微损失函数。研究装袋估计量在这些问题中的理论性质同样很有意义，也颇具挑战性。其次，我们发现装袋法中简单的平均操作无法降低最终估计量的偏差。因此，探索其他聚合策略十分有趣。最后，变量选择方法广泛应用于统计分析，并在过去十年中受到广泛关注。将装袋法应用于大规模变量选择问题也值得研究。

附录 A：引理 1 的证明

注意，目标函数 $\mathcal{L}_k(\theta)$ 关于 θ严格凸。因此，为证明 $\widehat{\theta}^{(k)}$ 是 $\sqrt{n}$-相合的，只需采用 Fan 和 Li (2001) 中的技巧，验证对于任意 $\epsilon > 0$，存在一个有限常数 $C > 0$ 使得，

$$\liminf_n P\Big\{ \sup_{\|u\|=1} \mathcal{L}_k(\theta_0 + Cn^{-1/2}u) > \mathcal{L}_k(\theta_0) \Big\} \geq 1 - \epsilon.$$

写作 $\theta_u = \theta_0 + Cn^{-1/2}u$，其中 $C > 0$ 是一个固定常数，$u \in \mathbb{R}^p$ 是一个单位长度向量（即 $\|u\| = 1$）。根据泰勒展开，我们有

$$\begin{aligned}\sup_{\|u\|=1} \Big\{ \mathcal{L}_k(\theta_0 + Cn^{-1/2}u) - \mathcal{L}_k(\theta_0) \Big\} =& n^{-1/2}Cu^\top \dot{\mathcal{L}}_k(\theta_0) + (2n)^{-1}C^2 u^\top \ddot{\mathcal{L}}_k(\theta_0)u + o_p(1) \\ =& Cu^\top J_1 + C^2 u^\top J_2 u + o_p(1), \end{aligned} \tag{A.1}$$

其中 $J_1 = n^{-1/2}\dot{\mathcal{L}}_k(\theta_0)$ 和 $J_2 = n^{-1}\ddot{\mathcal{L}}_k(\theta_0)$。

然后，我们按如下方式计算 $E(J_1)$ 和 $\mathrm{var}(J_1)$。事实上，

$$E(J_1) = E\{E(J_1|\mathbb{S})\} = \sqrt{n}E\Big\{E\Big(n^{-1}\sum_{i\in\mathcal{S}_k}\ell_i(\theta_0)\Big|\mathbb{S}\Big)\Big\} = \sqrt{n}E\Big\{N^{-1}\dot{\mathcal{L}}(\theta_0)\Big\} = 0,$$

其中， $\mathbb{S}$ 表示完整样本中包含的信息。就方差而言，我们有 $\mathrm{var}(J_1) =$

$$E\big\{\mathrm{var}(J_1|\mathbb{S})\big\} + \mathrm{var}\big\{E(J_1|\mathbb{S})\big\} = E\Big\{n^{-1}\,\mathrm{var}\Big(\sum_{i\in\mathcal{S}_k}\ell_i(\theta_0)\Big|\mathbb{S}\Big)\Big\} + \mathrm{var}\Big\{\sqrt{n}N^{-1}\dot{\mathcal{L}}(\theta_0)\Big\}$$

$$= E\left[N^{-1}\sum_{i=1}^{N}\Big\{\dot{\ell}_i(\theta_0) - N^{-1}\dot{\mathcal{L}}(\theta)\Big\}\Big\{\dot{\ell}_i(\theta_0) - N^{-1}\dot{\mathcal{L}}(\theta)\Big\}^{\top}\right] + \mathrm{var}\Big\{\sqrt{n}N^{-1}\dot{\mathcal{L}}(\theta_0)\Big\}$$

$$= \frac{N-1}{N}V + \frac{n}{N}V = O(1).$$

那么我们应有 $J_1 = O_p$（1）。通过类似的论证，我们可以证明 $J_2 \to_p H$。回顾一下， H 是一个正定矩阵。因此，只要 C 足够大，（A.1）中的二次项就支配其线性项。这意味着$\sup_{\|u\|=1}\Big\{\mathcal{L}_k\left(\theta_0 + Cn^{-1/2}u\right) - \mathcal{L}_k\left(\theta_0\right)\Big\} > 0$ 以趋近于 1 的概率成立，当$n \to \infty$时。这进一步表明，以趋近于 1 的概率，存在一个局部极小值点（即 $\widehat{\theta}^{(k)}$），使得 $\widehat{\theta}^{(k)} - \theta_0 = O_p\left(n^{-1/2}\right)$。引理的证明至此完成。

附录 B：定理1的证明

我们首先研究项 Q_1。可以计算得到 $E(Q_1) = E\big\{E(Q_1|\mathbb{S})\big\} = E\big\{E(Q_1^{(k)}|\mathbb{S})\big\} = -H^{-1}E\big\{N^{-1}\dot{\mathcal{L}}(\theta_0)\big\} = 0$，其中 $\mathbb{S}$ 表示完整样本中包含的信息。接下来我们研究 $\mathrm{var}(Q_1)$。注意，$\mathrm{var}(Q_1)$ =

$E\{\text{var}(Q_1|\mathbb{S})\} + \text{var}\{E(Q_1|\mathbb{S})\}$。我们可以计算得到 $E\Big\{\text{var}(Q_1|\mathbb{S})\Big\} =$

$$
\begin{aligned}
&K^{-1}H^{-1}E\Big\{\text{var}\left(n^{-1}\dot{\mathcal{L}}_k(\theta_0)|\mathbb{S}\right)\Big\}H^{-1} = K^{-1}H^{-1}E\Big\{n^{-1}\,\text{var}\left(\dot{\ell}_i(\theta_0)|\mathbb{S}\right)\Big\}H^{-1} \\
=&(nK)^{-1}H^{-1}E\left[E\Big\{\dot{\ell}_i(\theta_0)\dot{\ell}_i^{\top}(\theta_0)\Big|\mathbb{S}\Big\} - E\Big\{\dot{\ell}_i(\theta_0)\Big|\mathbb{S}\Big\}E\Big\{\dot{\ell}_i(\theta_0)\Big|\mathbb{S}\Big\}^{\top}\right]H^{-1} \\
=&(nK)^{-1}H^{-1}\left[E\Big\{\dot{\ell}_i(\theta_0)\dot{\ell}_i^{\top}(\theta_0)\Big\} - E\Big\{\Big(N^{-1}\mathcal{L}(\theta_0)\Big)\Big(N^{-1}\mathcal{L}(\theta_0)\Big)^{\top}\Big\}\right]H^{-1} \\
=&(nK)^{-1}(1-N^{-1})\Sigma.
\end{aligned}
$$

此外，我们有 $\text{var}\left\{E(Q_1|\mathbb{S})\right\} = \text{var}\ \left\{H^{-1}N^{-1}\dot{\mathcal{L}}(\theta_0)\right\} = H^{-1}\,\text{var}\{N^{-1}\dot{\mathcal{L}}(\theta_0)\}H^{-1} = N^{-1}\Sigma$。结合这些结果，我们得出

$$
\text{var}(Q_1) = (nK)^{-1}(1-N^{-1})\Sigma + N^{-1}\Sigma = \left(\frac{1}{nK}+\frac{1}{N}\right)\Sigma + o\left(\frac{1}{nK}+\frac{1}{N}\right).
$$

接下来我们研究项 Q_2。根据引理18和22以及 Huang and Huo （2019年）中的上界（B.10），我们有 $E\|\Delta^{(k)}\|^2 = O（n^{-2}）$。于是我们得到

$$
E\|Q_2\| \le K^{-1}\sum_{k=1}^{K} E\|Q_2^{(k)}\| = \|H^{-1}\| \cdot E\|\Delta^{(k)}\| = O(n^{-1}).
$$

因此 因此，我们应有 $Q_2 = O(n^{-1})$。这就完成了该定理的证明 .

附录 C：定理 2 的证明

令 $\tau = \{(nK)^{-1} + N^{-1}\}^{1/2}$。由定理1可知，在假设的条件下， $\tau^{-1}Q_2 \to_p 0$。因此，只需证明 $\tau^{-1}Q_1 \to_d \mathcal{N}(0,\Sigma)$。回顾一下，$Q_1 = H^{-1}K^{-1}\sum_{k=1}^{K}\dot{\mathcal{L}}_k(\theta_0)$，且 $\Sigma = H^{-1}VH^{-1}$。因此，我们接下来证明 $U = (\tau K)^{-1}\sum_{k=1}^{K}\dot{\mathcal{L}}_k(\theta_0) \to_d \mathcal{N}(0,V)$。若能证明特征函数 $f(t) = E\{\exp(it^{\top}U)\}\} \to \exp(-t^{\top}Vt/2)$，则结论成立。为此，我们首先定义 $\widehat{V} = \text{var}(\dot{\ell}_i(\theta_0)|\mathbb{S}) = N^{-1}\sum_{i=1}^{N}\dot{\ell}_i(\theta_0)\,\dot{\ell}_i^{\top}(\theta_0) - \left\{N^{-1}\dot{\mathcal{L}}(\theta_0)\right\}\left\{N^{-1}\dot{\mathcal{L}}(\theta_0)\right\}^{\top}$。然后

我们可以计算得出

$$
\begin{aligned}
f(t) =& E\left[\exp\left\{\frac{it^\top}{\tau K}\sum_{k=1}^{K}\Big(\dot{\mathcal{L}}_k(\theta_0)-N^{-1}\mathcal{L}(\theta_0)\Big)\right\}\exp\left\{\frac{it^\top}{\tau}N^{-1}\dot{\mathcal{L}}(\theta_0)\right\}\right]\\
=& E\left(E\left[\exp\left\{\frac{it^\top}{\tau K}\sum_{k=1}^{K}\Big(\dot{\mathcal{L}}_k(\theta_0)-N^{-1}\mathcal{L}(\theta_0)\Big)\right\}\middle|\mathbb{S}\right]\exp\left\{\frac{it^\top}{\tau\sqrt{N}}N^{-1/2}\dot{\mathcal{L}}(\theta_0)\right\}\right)\\
=& E\left(E\left[\exp\left\{\frac{it^\top\widehat{V}^{1/2}}{\tau nK}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k}\widehat{V}^{-1/2}\Big(\dot{\ell}_i(\theta_0)-N^{-1}\mathcal{L}(\theta_0)\Big)\right\}\middle|\mathbb{S}\right]\exp\left\{\frac{it^\top}{\tau\sqrt{N}}N^{-1/2}\dot{\mathcal{L}}(\theta_0)\right\}\right)\\
=& E\left[E\left\{\exp\left(\frac{it^\top\widehat{V}^{1/2}}{\tau\sqrt{nK}}Z_1\right)\middle|\mathbb{S}\right\}\exp\left(\frac{it^\top}{\tau\sqrt{N}}Z_2\right)\right],
\end{aligned}
$$

其中 $Z_1=(nK)^{-1/2}\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k}\widehat{V}^{-1/2}\Big(\dot{\ell}_i(\theta_0)-N^{-1}\mathcal{L}(\theta_0)\Big)$，且 $Z_2=N^{-1/2}\dot{\mathcal{L}}(\theta_0)$。我们通过考虑以下几种情形来研究 $f(t)$ 的极限。

情形1 ($nK/N\to\infty$)。在这种情况下，应有 $\tau\sqrt{nK}\to\infty$。然后可以验证 $\|t^\top\widehat{V}^{1/2}\|/(\tau\sqrt{nK})\to_p 0$。注意，$E(Z_1|\mathbb{S})=0$，且 $\mathrm{var}(Z_1|\mathbb{S})=I_p$，其中 $I_p\in\mathbb{R}^{p\times p}$ 是单位矩阵。由中心极限定理可知，给定 $\mathbb{S}$，$Z_1\to_d\mathcal{N}(0,I_p)$。因此，我们有 $E\Big[\exp\Big\{it^\top\widehat{V}^{1/2}Z_1/\big(\tau\sqrt{nK}\big)\Big\}\Big|\mathbb{S}\Big]\to_p 1$。进一步注意到 (1) $\tau\sqrt{N}\to_p 1$ 和 (2) $Z_2\to_d\mathcal{N}(0,V)$。由此可得，$E\Big[\exp\big\{it^\top Z_2/(\tau\sqrt{N})\big\}\Big]\to\exp(-t^\top Vt/2)$。然后，根据控制收敛定理，我们得出 $f(t)\to\exp(-t^\top Vt)$。情形1的证明至此完成。

情形2 ($nK/N\to 0$)。在这种情况下，我们应有 $\tau\sqrt{N}\to\infty$。由于 $Z_2\to_d\mathcal{N}(0,V)$，因此 $\exp\big\{it^\top Z_2/(\tau\sqrt{N})\big\}\to_p 1$。然后，根据支配收敛定理，只需证明 $E\Big[\exp\Big\{\big(it^\top\widehat{V}^{1/2}Z_1\big)/\big(\tau\sqrt{nK}\big)\Big\}\Big]\to\exp(-t^\top Vt)$。理由如下：（1）$\tau\sqrt{nK}\to 1$ 且 $\widehat{V}\to_p V$，这意味着$\widehat{V}^{1/2}/\big(\tau\sqrt{nK}\big)\to_p V^{1/2}$。（2）根据中心极限定理，$Z_1\to_d\mathcal{N}(0,I_p)$。至此，情形2的证明完成。

情形 3 ($nK/N\to C$ 对于某个常数 $C>0$)。我们首先将 $f(t)$ 分解为 $f(t)=f_1(t)+f(t)-f_1(t)$，其中 $f_1(t)=E\Big[\Gamma_1\exp\big\{it^\top V^{1/2}Z_2/(\tau\sqrt{N})\big\}\Big]$，且 $f(t)-f_1(t)=E\Big[(\Gamma_1-\Gamma_2)\exp\big\{it^\top V^{1/2}Z_2/(\tau\sqrt{N})\big\}\Big]$，其中 $\Gamma_1=\exp\big\{-t^\top\widehat{V}t/(2\tau^2nK)\big\}$，且 $\Gamma_2=E\Big[\exp\big\{it^\top\widehat{V}^{1/2}Z_1/\big(\tau\sqrt{nK}\big)\big\}\Big|\mathbb{S}\Big]$。

由于 $\tau\sqrt{nK} \to \sqrt{C+1}$ 且 $Z_1 \to_d \mathcal{N}(0, I_p)$，我们有在给定 $\mathbb{S}$ 的条件下 $\Gamma_1 - \Gamma_2 \to_p 0$。由控制收敛定理可得 $f(t) - f_1(t) \to 0$。因此，尚需考察 $f_1(t)$ 的极限。注意到 $\widehat{V} \to_p V$ 且 $\tau^2 nK \to C+1$。于是我们有 $\Gamma_1 \to_p \exp\big[-t^\top Vt/\{2(C+1)\}\big]$。由于 $Z_2 \to_d \mathcal{N}(0, V)$ 且 $\tau\sqrt{N} \to \sqrt{C^{-1}+1}$，我们应有 $E\Big[\exp\big\{it^\top V^{1/2} Z_2/(\tau\sqrt{N})\big\}\Big] \to \exp\big\{-t^\top Vt/(C^{-1}+1)\big\}$。结合上述结果和控制收敛定理，我们得出 $f_1(t) \to \exp\big[-t^\top Vt/\{2(C+1)\} - t^\top Vt/\{2(C^{-1}+1)\}\big] = \exp(-t^\top Vt/2)$。这就完成了情形 3 的证明，并完成了定理的证明。

附录 D：定理 3 的证明

设 $c_0 = n\{(nK)^{-1} + N^{-1}\}$。回顾 $\widehat{\theta}_{\mathrm{BAG}} = K^{-1}\sum_{k=1}^K \widehat{\theta}^{(k)}$。于是我们有

$$\begin{aligned}
\widehat{\mathrm{SE}}^2(\widehat{\theta}_{\mathrm{BAG}}) =& \frac{c_0}{K}\sum_{k=1}^K \Big(\widehat{\theta}^{(k)} - \widehat{\theta}_{\mathrm{BAG}}\Big)\Big(\widehat{\theta}^{(k)} - \widehat{\theta}_{\mathrm{BAG}}\Big)^\top \\
=& \frac{c_0}{K}\sum_{k=1}^K \Big(\widehat{\theta}^{(k)} - \theta_0 + \theta_0 - \widehat{\theta}_{\mathrm{BAG}}\Big)\Big(\widehat{\theta}^{(k)} - \theta_0 + \theta_0 - \widehat{\theta}_{\mathrm{BAG}}\Big)^\top \\
=& c_0\big(A_1 - A_2\big),
\end{aligned}$$

其中 $A_1 = K^{-1}\sum_{k=1}^K \big(\widehat{\theta}^{(k)} - \theta_0\big)\big(\widehat{\theta}^{(k)} - \theta_0\big)^\top$ 和 $A_2 = \big(\widehat{\theta}_{\mathrm{BAG}} - \theta_0\big)\big(\widehat{\theta}_{\mathrm{BAG}} - \theta_0\big)^\top$。接下来我们计算 $E(A_1)$ 和 $E(A_2)$。

我们首先计算 $E(A_1)$。回想一下，$\widehat{\theta}^{(k)} = \theta_0 - H^{-1}n^{-1}\dot{\mathcal{L}}_k(\theta_0) - H^{-1}\Delta^{(k)}$，其中 $\Delta^{(k)} = \Big\{n^{-1}\ddot{\mathcal{L}}_k(\theta_0) - H\Big\}(\widehat{\theta}^{(k)} - \theta_0) + n^{-1}\Big\{\int_0^1 \ddot{\mathcal{L}}_k\Big((1-t)\theta_0 + t\widehat{\theta}^{(k)}\Big)dt - \ddot{\mathcal{L}}_k(\theta_0)\Big\}(\widehat{\theta}^{(k)} - \theta_0)$。根据附录 B 中的证明，我们可以推导出 $E\|n^{-1}\dot{\mathcal{L}}_k(\theta_0)\|^2 = O(n^{-1})$，并且 $E\|\Delta^{(k)}\|^2 = O(n^{-2})$。然后我们可以计算得出 $E(A_1) =$

$$\begin{aligned}
&E\Big[E\Big\{K^{-1}\sum_{k=1}^K \big(\widehat{\theta}^{(k)} - \theta_0\big)\big(\widehat{\theta}^{(k)} - \theta_0\big)^\top \Big| \mathbb{S}\Big\}\Big] = E\Big\{\big(\widehat{\theta}^{(k)} - \theta_0\big)\big(\widehat{\theta}^{(k)} - \theta_0\big)^\top\Big\} \\
=& H^{-1}E\Big[\big\{n^{-1}\dot{\mathcal{L}}_k(\theta_0)\big\}\big\{n^{-1}\dot{\mathcal{L}}_k(\theta_0)\big\}^\top - n^{-1}\dot{\mathcal{L}}_k(\theta_0)\Delta^{(k)\top} - \Delta^{(k)}\big\{n^{-1}\dot{\mathcal{L}}_k(\theta_0)\big\}^\top + \Delta^{(k)}\Delta^{(k)\top}\Big]H^{-1} \\
=& n^{-1}\Sigma + O(n^{-3/2}).
\end{aligned}$$

同理，我们可以计算得出 $E(A_2) = E\left\{\left(\widehat{\theta}_{\mathrm{BAG}} - \theta_0\right)\left(\widehat{\theta}_{\mathrm{BAG}} - \theta_0\right)^\top\right\} = \left\{(nK)^{-1} + N^{-1}\right\}\Sigma + O(n^{-3/2})$。注意，在所假设的条件下， $c_0 = o(1)$。于是我们有

$$E\left\{\widehat{\mathrm{SE}}^2(\widehat{\theta}_{\mathrm{BAG}})\right\} = \left(\frac{1}{nK} + \frac{1}{N}\right)\Sigma + o\left(\frac{1}{nK} + \frac{1}{N}\right).$$

证明到此完成。

参考文献

Breiman, L.（1996）。Bagging 预测器。机器学习，24(2)：123–140。

B¨uhlmann, P. 和 Yu, B. (2002). Bagging 分析。《统计年刊》, 30(4):927–961.

Chang, X., Lin, S.-B., and Wang, Y. (2017). 分而治之的局部平均回归。《电子统计学期刊》，11(1):1326–1350。

Chen, X., Lee, J.D.,Li,H.和Yang,Y. (2021)。主成分分析的分布式估计：扩展特征空间分析。《美国统计学会会刊》，即将出版。

Chen,X.,Liu,W.,Mao,X. 和Yang, Z.（2020）。分布式高维回归在分位数损失函数下。《机器学习研究期刊》，21(182):1–43。

Drineas, P., Magdon-Ismail, M., Mahoney, M. W. 和 Woodruff, D. P.（2012）。矩阵相干性和统计杠杆值的快速近似。《机器学习研究期刊》，13(1):3475–3506。

Fan, J. 和 Li, R.（2001）。通过非凹惩罚似然进行变量选择及其预言机性质。《美国统计学会会刊》，96(456):1348– 1360。

Fan, J., Li, R., Zhang, C.-H., 和 Zou, H. (2020). 数据科学的统计基础. Chapman and Hall/CRC.

Fan, J., Wang, D., Wang, K., and Zhu, Z.（2019年）。主特征子空间的分布式估计。《统计年刊》，47(6):3009。

Huang, C. and Huo, X.（2019年）。分布式一步估计量。*Mathematical Pro-gramming*，174(1):41–76。

Jordan, M. I., Lee, J. D., and Yang, Y.（2019年）。通信高效的分布式统计推断。《美国统计学会会刊》，114(526):668– 681。

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I.（2014）。海量数据的可扩展自助法。皇家统计学会会刊：*B*辑（统计方法学），76(4):795–816。

Lehmann, E. L. and Casella, G.（2006）。点估计理论。Springer Science & Business Media。

Li, X., Li, R., Xia, Z., 和 Xu, C. (2020)。通过逐分量去偏差进行分布式特征筛选。《机器学习研究期刊》，21(24):1–32。

Lian, H.、Zhao, K. 和 Lv,S.（2019年）。海量数据的高维部分线性模型中非参数函数的投影样条估计。《统计年刊》，47(5)：2922–2949。

Ma, P.、Mahoney,M. W. 和 Yu, B.（2015）。算法杠杆化的统计学视角。《机器学习研究期刊》，16(27)：861–911。

Ma, P., Zhang, X., Xing, X., Ma, J. 和 Mahoney, M.（2020）。随机数值线性代数算法的抽样估计量的渐近分析。载于 Chiappa, S. 和 Calandra, R. 编，第二十三届人工智能与统计国际会议论文集，第 108 卷，机器学习研究论文集，第 1026–1035 页。PMLR。

Pan, R., Ren, T., Guo, B., Li, F., Li, G., and Wang, H. (2021). 关于通过先导抽样和单步更新进行分布式分位数回归的说明。《商业与经济统计学杂志》，即将刊出。

Shao, J.（2003）。《数理统计》。Springer 统计学丛书。Springer。Van der Vaart, A. W.（2000）。《渐近统计学》。剑桥大学出版社。

Volgushev, S., Chao, S.-K. 和 Cheng, G.（2019年）。分位数回归过程的分布式推断。《统计年刊》，47(3):1634–1662。

Wager, S.（2014）。随机森林的渐近理论。*arXiv* 预印本 *arXiv:1405.0352*。

Wang, H. 和 Ma, Y.（2021）。大数据中分位数回归的最优子抽样。《*Biometrika*》，108(1):99–112。

Wang, H., Yang, M. 和 Stufken, J.（2019年）。基于信息的用于大数据线性回归的最优子数据选择。《美国统计学会会刊》，114(525):393–405。

Wang, H.、Zhu, R. 和 Ma, P.（2018）。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。

Wu, S.、Zhu, X. 和 Wang, H.（2021）。子抽样与刀切法：一种实用便捷的解决方案，适用于计算资源有限的大数据分析。《统计辛尼卡》，预印本。

Zhang, Y.、Duchi, J. C. 和 Wainwright, M. J.（2013）。用于统计优化的通信高效算法。《机器学习研究期刊》，14(1):3321–3363。

Zhu, X.、Li, F. 和 Wang, H.（2021a）。分布式系统的最小二乘近似。《计算与图形统计学杂志》，30(4)：1004–1018。

Zhu, X.、Pan, R.、Wu, S. 和 Wang, H.（2021b）。通过子抽样进行海量数据分析的特征筛选。《商业与经济统计学杂志》，已录用待刊。

Zou, T., Li, X., Liang, X., and Wang, H.（2021）。关于海量数据的子装袋估计。*arXiv* 预印本 *arXiv:2103.00631*。

Table 1: The simulation results for the bagging estimator under the linear regression model with $B = 1,000$ replications. The Bias_j ($\times 10^2$), SE_j ($\times 10^2$), $\widehat{\text{SE}}_j$ ($\times 10^2$) and ECP_j are reported under different (n, K)-specifications.

	$n = 500$					$n = 750$					$n = 1000$				
K	50	100	150	200	250	50	100	150	200	250	50	100	150	200	250
Bias_1	0.017	0.010	0.006	0.005	0.003	0.016	0.006	0.003	0.003	0.005	0.009	0.005	0.002	0.008	0.009
SE_1	0.697	0.511	0.433	0.392	0.359	0.581	0.435	0.374	0.337	0.319	0.511	0.393	0.337	0.314	0.295
$\widehat{\text{SE}}_1$	0.663	0.500	0.429	0.388	0.361	0.557	0.427	0.372	0.341	0.321	0.495	0.385	0.340	0.316	0.300
ECP_1	0.927	0.943	0.943	0.946	0.951	0.932	0.941	0.951	0.954	0.956	0.936	0.945	0.954	0.957	0.967
Bias_2	0.026	0.040	0.029	0.028	0.031	0.021	0.030	0.030	0.029	0.025	0.039	0.028	0.029	0.025	0.022
SE_2	0.681	0.515	0.439	0.391	0.355	0.579	0.439	0.366	0.338	0.313	0.515	0.390	0.338	0.310	0.294
$\widehat{\text{SE}}_2$	0.664	0.501	0.429	0.388	0.362	0.554	0.427	0.373	0.342	0.322	0.496	0.386	0.341	0.316	0.300
ECP_2	0.939	0.939	0.935	0.941	0.960	0.930	0.935	0.958	0.953	0.957	0.935	0.938	0.953	0.960	0.958
Bias_3	0.026	0.022	0.014	0.015	0.007	0.025	0.015	0.013	0.007	0.010	0.023	0.015	0.007	0.007	0.008
SE_3	0.688	0.516	0.442	0.393	0.368	0.581	0.443	0.381	0.348	0.330	0.516	0.393	0.347	0.322	0.305
$\widehat{\text{SE}}_3$	0.668	0.501	0.430	0.389	0.362	0.557	0.427	0.373	0.342	0.323	0.496	0.386	0.341	0.317	0.300
ECP_3	0.939	0.945	0.937	0.939	0.954	0.929	0.936	0.949	0.951	0.950	0.940	0.940	0.949	0.950	0.953
Bias_4	0.006	0.011	0.006	0.011	0.015	0.003	0.006	0.011	0.017	0.016	0.013	0.010	0.017	0.014	0.011
SE_4	0.649	0.486	0.416	0.374	0.353	0.532	0.416	0.362	0.326	0.311	0.485	0.374	0.325	0.308	0.292
$\widehat{\text{SE}}_4$	0.663	0.499	0.428	0.388	0.361	0.554	0.426	0.372	0.341	0.322	0.493	0.385	0.340	0.316	0.300
ECP_4	0.958	0.953	0.958	0.961	0.951	0.957	0.959	0.955	0.960	0.950	0.947	0.954	0.956	0.953	0.957
Bias_5	0.001	0.003	0.004	0.003	0.007	0.006	0.005	0.008	0.007	0.010	0.002	0.003	0.006	0.009	0.008
SE_5	0.679	0.518	0.451	0.401	0.366	0.575	0.451	0.383	0.343	0.325	0.518	0.401	0.343	0.322	0.305
$\widehat{\text{SE}}_5$	0.665	0.499	0.429	0.389	0.362	0.555	0.427	0.373	0.342	0.323	0.495	0.387	0.342	0.317	0.301
ECP_5	0.946	0.944	0.938	0.939	0.946	0.941	0.936	0.937	0.950	0.948	0.932	0.941	0.947	0.949	0.943

Table 2: The simulation results for the bagging estimator under the logistic regression model with $B = 1,000$ replications. The Bias_j ($\times 10^2$), SE_j ($\times 10^2$), $\widehat{\text{SE}}_j$ ($\times 10^2$) and ECP_j are reported under different (n, K)-specifications.

	$n = 500$					$n = 750$					$n = 1000$				
K	50	100	150	200	250	50	100	150	200	250	50	100	150	200	250
Bias_1	0.315	0.280	0.275	0.278	0.302	0.209	0.181	0.199	0.210	0.208	0.138	0.137	0.162	0.160	0.156
SE_1	1.391	1.056	0.921	0.806	0.755	1.172	0.915	0.775	0.720	0.681	1.046	0.799	0.717	0.667	0.637
$\widehat{\text{SE}}_1$	1.379	1.035	0.890	0.806	0.751	1.146	0.880	0.769	0.706	0.665	1.017	0.796	0.702	0.651	0.618
ECP_1	0.940	0.933	0.935	0.943	0.933	0.941	0.938	0.937	0.934	0.935	0.936	0.948	0.935	0.931	0.936
Bias_2	0.206	0.159	0.154	0.129	0.142	0.153	0.104	0.090	0.090	0.086	0.089	0.056	0.065	0.069	0.067
SE_2	1.369	1.011	0.886	0.808	0.749	1.140	0.879	0.772	0.697	0.657	1.003	0.803	0.697	0.644	0.620
$\widehat{\text{SE}}_2$	1.372	1.027	0.882	0.798	0.744	1.143	0.875	0.764	0.701	0.660	1.007	0.788	0.697	0.646	0.613
ECP_2	0.944	0.951	0.933	0.942	0.938	0.942	0.931	0.942	0.946	0.945	0.945	0.941	0.950	0.944	0.951
Bias_3	0.006	0.021	0.011	0.004	0.019	0.020	0.013	0.024	0.018	0.018	0.023	0.004	0.018	0.013	0.016
SE_3	1.421	1.034	0.898	0.790	0.753	1.145	0.893	0.766	0.714	0.662	1.025	0.781	0.710	0.646	0.614
$\widehat{\text{SE}}_3$	1.357	1.022	0.878	0.795	0.741	1.132	0.872	0.761	0.698	0.657	1.005	0.784	0.694	0.643	0.611
ECP_3	0.933	0.938	0.945	0.957	0.943	0.946	0.942	0.955	0.945	0.942	0.939	0.956	0.946	0.946	0.954
Bias_4	0.153	0.126	0.143	0.152	0.147	0.095	0.094	0.101	0.096	0.084	0.057	0.081	0.074	0.060	0.060
SE_4	1.405	1.068	0.917	0.825	0.754	1.190	0.911	0.780	0.710	0.671	1.057	0.815	0.708	0.651	0.607
$\widehat{\text{SE}}_4$	1.375	1.031	0.883	0.801	0.745	1.143	0.875	0.765	0.701	0.660	1.013	0.790	0.697	0.646	0.613
ECP_4	0.939	0.931	0.935	0.926	0.942	0.928	0.941	0.933	0.939	0.943	0.933	0.931	0.942	0.950	0.946
Bias_5	0.337	0.345	0.336	0.325	0.325	0.238	0.241	0.225	0.220	0.218	0.204	0.185	0.172	0.164	0.151
SE_5	1.387	1.024	0.886	0.792	0.736	1.131	0.880	0.759	0.695	0.652	1.013	0.786	0.694	0.638	0.603
$\widehat{\text{SE}}_5$	1.383	1.038	0.891	0.807	0.751	1.147	0.880	0.769	0.706	0.664	1.024	0.798	0.705	0.652	0.619
ECP_5	0.936	0.944	0.922	0.937	0.920	0.942	0.932	0.947	0.940	0.938	0.944	0.949	0.947	0.944	0.946

Table 3: The simulation results for the bagging estimator under the Poisson regression model with $B = 1,000$ replications. The Bias_j ($\times 10^2$), SE_j ($\times 10^2$), $\widehat{\text{SE}}_j$ ($\times 10^2$) and ECP_j are reported under different (n, K)-specifications.

	$n = 500$					$n = 750$					$n = 1000$				
K	50	100	150	200	250	50	100	150	200	250	50	100	150	200	250
Bias_1	0.143	0.135	0.120	0.115	0.109	0.120	0.086	0.072	0.076	0.082	0.082	0.061	0.058	0.062	0.060
SE_1	0.753	0.583	0.486	0.446	0.405	0.655	0.486	0.425	0.379	0.353	0.583	0.446	0.378	0.345	0.327
$\widehat{\text{SE}}_1$	0.736	0.551	0.473	0.428	0.399	0.614	0.470	0.410	0.377	0.355	0.547	0.427	0.377	0.349	0.332
ECP_1	0.938	0.924	0.934	0.936	0.943	0.910	0.937	0.937	0.943	0.952	0.931	0.943	0.945	0.954	0.949
Bias_2	0.018	0.028	0.037	0.045	0.051	0.000	0.018	0.030	0.032	0.032	0.001	0.017	0.023	0.025	0.030
SE_2	0.834	0.630	0.547	0.491	0.456	0.727	0.548	0.472	0.429	0.404	0.632	0.492	0.428	0.399	0.375
$\widehat{\text{SE}}_2$	0.827	0.621	0.533	0.483	0.451	0.692	0.532	0.464	0.426	0.401	0.617	0.481	0.426	0.395	0.375
ECP_2	0.940	0.944	0.937	0.945	0.947	0.939	0.943	0.944	0.946	0.947	0.945	0.943	0.943	0.942	0.951
Bias_3	0.014	0.011	0.003	0.004	0.008	0.022	0.003	0.008	0.009	0.008	0.010	0.003	0.009	0.008	0.011
SE_3	0.860	0.631	0.526	0.476	0.440	0.712	0.528	0.458	0.416	0.394	0.630	0.475	0.413	0.386	0.368
$\widehat{\text{SE}}_3$	0.830	0.622	0.535	0.484	0.451	0.695	0.533	0.464	0.427	0.402	0.616	0.482	0.427	0.395	0.375
ECP_3	0.934	0.946	0.955	0.948	0.943	0.939	0.949	0.943	0.950	0.957	0.938	0.944	0.953	0.952	0.949
Bias_4	0.045	0.026	0.044	0.051	0.050	0.002	0.028	0.029	0.031	0.030	0.000	0.025	0.021	0.025	0.026
SE_4	0.822	0.618	0.528	0.485	0.450	0.688	0.529	0.468	0.428	0.399	0.616	0.484	0.429	0.397	0.381
$\widehat{\text{SE}}_4$	0.830	0.623	0.535	0.484	0.451	0.698	0.534	0.465	0.427	0.402	0.618	0.482	0.426	0.395	0.375
ECP_4	0.949	0.953	0.958	0.958	0.953	0.945	0.958	0.956	0.953	0.957	0.941	0.954	0.952	0.950	0.946
Bias_5	0.146	0.148	0.126	0.120	0.115	0.116	0.091	0.085	0.077	0.077	0.095	0.067	0.060	0.057	0.056
SE_5	0.761	0.551	0.473	0.422	0.391	0.622	0.474	0.409	0.369	0.356	0.553	0.423	0.370	0.350	0.328
$\widehat{\text{SE}}_5$	0.731	0.550	0.472	0.428	0.399	0.612	0.471	0.411	0.377	0.355	0.544	0.424	0.375	0.348	0.331
ECP_5	0.930	0.937	0.940	0.933	0.941	0.938	0.937	0.936	0.945	0.947	0.933	0.940	0.947	0.947	0.946

表4：逻辑回归$\widehat{\text{模}}$型的装袋估计量估计结果。我们报告了所有协变量的估计系数（估计值）、标准误（标准误）和 p- 值。

协变量	水平	估计值	$\widehat{\text{标准误}}(\times 10^2)$	p-值	协变量	水平	估计量	$\widehat{\text{标准误}}(\times 10^2)$	p-值
`Intercept`		-2.025	0.329	< 0.001	`Month`	二月	-0.064	0.243	< 0.001
`Distance`		0.122	0.046	< 0.001		三月	-0.131	0.244	< 0.001
`DepTime`	早晨	0.396	0.261	< 0.001		四月	-0.337	0.246	< 0.001
	下午	0.857	0.253	< 0.001		May	-0.341	0.249	< 0.001
	傍晚	1.057	0.260	< 0.001		June	-0.025	0.240	< 0.001
`DayOfWeek`	星期二	-0.048	0.195	< 0.001		July	-0.098	0.229	< 0.001
	星期三	0.046	0.194	< 0.001		八月	-0.152	0.241	< 0.001
	星期四	0.198	0.194	< 0.001		九月 -0.534		0.263	< 0.001
	星期五	0.249	0.186	< 0.001		十月	-0.384	0.245	< 0.001
	星期六	-0.167	0.205	< 0.001		十一月	-0.275	0.256	< 0.001
	星期日	-0.015	0.189	< 0.001		十二月	0.154	0.236	< 0.001