DOI:10.13546/j.cnki.tjyjc.2021.08.011

分布式计算中统计方法的拓展

任图南[1,2]

（1.中国银联博士后科研工作站，上海 201201；2.复旦大学网络空间安全博士后科研流动站，上海 200433）

摘 要：在数据体量逐渐增大的时代，处理大体量数据已经成为科学研究必需的途径。分布式计算为处理这样的大体量数据提供了方案，但站在统计学的角度，分布式计算所带来的便捷性也会造成统计学性质的损失。文章针对分布式计算与统计理论结合问题进行综述，并分析了这些方法的优势和不足，指出了在这一领域进一步研究的方向。

关键词：分布式计算；One-shot方法；高维稀疏回归

中图分类号：F222.1 文献标识码：A 文章编号：1002-6487(2021)08-0054-04

0 引言

在实际工作中所产生的数据集可能太大，以至于无法存储在一台计算机的硬盘上，只能存储在分布式系统中。数据集一旦被存储在分布式系统中，会给传统的统计推断带来相应的挑战。因为数据不被存储在同一台计算机上，传统的统计学方法将不再适用，需要对其进行相应的改造[1-3]，才能在分布式计算中得到统计量的估计，而尽可能少地损失统计有效性。不仅是统计有效性的损失，在分布式系统中的“主仆”模式中，“仆”计算机之间是无法进行数据传输的，这限制了某些方法的应用，如Newton-Raphson迭代。如何减少通讯成本也是统计学在分布式系统研究的一个热点。

众多的统计学家在近些年来对这一领域做出了贡献。本文将介绍分布式计算中的最新统计研究成果，包括高维稀疏回归问题在分布式计算中的拓展，One-shot方法的应用及其理论性质，以及一些统计方法在分布式计算的拓展。

1 分布式计算中的统计理论

1.1 高维稀疏回归问题

Chen和Xie(2014)[3]针对高维稀疏回归问题，在分布式计算中提出了One-shot(Split and Conquer)方法。考虑广义线性模型 $E(y_i=g(x_i\beta),i=1,\cdots,n$ ，y_i 是响应变量，x_i 是 $p\times1$ 的解释变量，β 是 $p\times1$ 的未知参数，$g(\bullet)$ 是连接函数。高维稀疏数据假设 p 会随着 n 的增长而增长，所以 β 是稀疏的。带惩罚项的极大似然估计通常用来解决这样的问题：

$$\hat{\beta}=\arg\max_{\beta}\{\frac{l(\beta;y,X)}{n}-\rho(\beta;\lambda)\} \tag{1}$$

其中，$\rho(\bullet)$ 是惩罚函数，λ 是超参数，不同的 $\rho(\bullet)$ 会有不同的参数估计，如LASSO估计量[2]、LARS估计量[4]、SCAD估计量[5]和MCP估计量[6]等。作者在分布式计算中考虑这一问题，假设数据过大只能存储在 K 台计算机上，在第 k 台计算机上分别得到局部估计如下：

$$\hat{\beta}_k=\arg\max_{\beta}\{\frac{l(\beta;y_k,X_k)}{n_k}-\rho(\beta;\lambda)\} \tag{2}$$

在得到局部估计之后，本文采用多数表决法得到新的组合估计 $\hat{\beta}^{(c)}$ ，并证明 $\hat{\beta}^{(c)}$ 的符号一致性。

在Zhang等(2013)[7]工作的基础上，Lee等(2017)[8]认为通讯成本是分布式计算的瓶颈。同样是针对高维稀疏回归问题，相较于Chen和Xie(2014)[3]、Lee等(2017)[8]提出的方法可以保证符号一致性，统计量与全局估计量的收敛性质是一样的，并且设计了使通讯成本有效的算法。考虑线性回归 $y=X\beta^{*}+\varepsilon$ ，y 是响应变量，X 是 $n\times p$ 的解释变量矩阵，β 是 $p\times1$ 的未知参数，且稀疏性为 s 。当 $p\gg n$ 的时候，正则化是必须的，本文考虑Lasoo正则化，全局估计量为：

$$\hat{\beta}=\arg\min_{\beta}\frac{1}{2n}\|y-X\beta\|_2^2+\lambda\|\beta\|_1 \tag{3}$$

然而Lasso估计量是有偏的，在分布式计算中对统计量平均可以减少方差，但不能减少偏差，所以应该在平均之前，对每台计算机上得到的局部估计量进行纠偏，参考Javanmard和Montanari(2014)[9]中的纠偏Lasso估计量：

$$\hat{\beta}^{d}=\hat{\beta}+\frac{1}{n}\hat{\Theta}X^{T}(y-X\hat{\beta}) \tag{4}$$

其中，$\hat{\beta}$ 是式(3)中的Laaso估计量，$\hat{\Theta}\in R^{p\times p}$ 是协方差矩阵的逆矩阵估计。$\hat{\Theta}$ 的估计需要全部数据都被传输到主计算机上，但是这样的通讯成本非常高，而如果各台计算机分别估计各自的 $\hat{\Theta}_k$，由于它的奇异性，计算消耗代价很高。所以本文在不损失估计精度的前提下，设计了一个通讯有效的算法，只需要两轮通讯传输，且传输的都是

作者简介：任图南(1992—)，男，山西阳泉人，博士，研究方向：应用经济学。

向量，传输成本小，且最终的估计量与全局Lasso统计量的收敛速度一致，只要“仆”计算机的个数不是特别的多。

步骤1：各台计算机依据局部数据计算局部统计量 $\hat{\beta}_k$ 和 $\frac{1}{n}X_k^T\left(y_k-X_k\hat{\beta}_k\right)$，并将其传输到“主”计算机上。

步骤2：“主”计算机对收到的统计量进行处理后，将 $\frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_k$ 和 $\frac{1}{N}\sum_{k=1}^{m}X_k^T\left(y_k-X_k\hat{\beta}_k\right)$ 再传输回各“仆”计算机上。

步骤3：第 j 台计算机，只需要自己存储的数据，就可以对局部估计量做出纠偏，$\tilde{\beta}_j=\frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_k+\hat{\Theta}_{j,\bullet}\left\{\frac{1}{N}\sum_{k=1}^{m}X_k^T\left(y_k-X_k\hat{\beta}_k\right)\right\}$，其中，$\hat{\Theta}_{j,\bullet}$ 是p维向量。将 $\tilde{\beta}_j$ 传输到“主”计算机。

步骤4：“主”计算机得到最终估计量 $\bar{\beta}=\frac{1}{m}\sum_{k=1}^{m}\tilde{\beta}_k$。

针对高维稀疏回归问题，不同于之前的学者，Yang等(2016)[10]假设数据在分布式系统中的存储方式是按照特征分割的，即不同的“仆”计算机上存有全部的样本，但只有部分的特征，这在超高维的情况下是合理的假设。本文首先利用Fan和Lv(2008)[11]的方法，对特征进行降维，然后通过sketch的转换，将所需数据传输到“主”计算机上进行估计，得到最后的统计量。该方法把按特征分布存储的超高维回归问题的计算复杂度从 $O(mN^2)$ 降低到了 $O(Nms)$，并且可以达到全局Lasso估计量的收敛速度。

1.2 One-shot统计方法

Zhang等(2013)[7]提出了两种“通讯有效”的算法，来解决分布式计算中通过最小化损失函数来得到估计量的问题。假设数据有 N 条观测数据，存储在 m 台计算机上，每台机器上独立同分布地存储 n 条数据。若可以得到全局的估计量，则估计量的MSE的收敛速度应该是 $O\left(N^{-1}\right)$。让 $f(\theta,x)$ 表示实值的损失函数，θ 为未知参数，x 表示数据，S 表示全数据集，$i=1,\cdots,m$，$S_{1,i}$ 表示存储在第 i 台计算机上的数据。全局最优估计量由式(5)得到：

$$\hat{\theta}=\arg\min_\theta\frac{1}{|S|}\sum_{x\in S}f(\theta,x) \tag{5}$$

本文提出的第一种方法为Average Mixture Algorithm，首先在第 i 台计算机上得到局部的无偏估计量 $\hat{\theta}_{1,i}=\arg\min_\theta\frac{1}{|S_{1,i}|}\sum_{x\in S_{1,i}}f(\theta,x)$，之后对 m 个局部估计量进行平均，得到估计量：

$$\bar{\theta}_1=\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{1,i} \tag{6}$$

在适当的假设下，$\bar{\theta}_1$ 的MSE的收敛速度是 $O(N^{-1}+n^{-2})$，$m\backsim\sqrt{N}$ 时，$\bar{\theta}_1$ 与全局统计量有一样的收敛速度。方法一的优点是计算简单，不需要额外的通讯成本。文中第二种方法为Subsampled Average Mixture Algorithm。首先在第 i 台计算机存储的数据 $S_{1,i}$ 中无放回地选出 $|rn|$ 大小的 $S_{2,i}$ 子样本，其中 $r\in[0,1]$ 是固定的采样率。之后第 i 台计算机计算 $\hat{\theta}_{2,i}=\arg\min_\theta\frac{1}{|S_{2,i}|}\sum_{x\in S_{2,i}}f(\theta,x)$，$\bar{\theta}_2=\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{2,i}$。最后的估计量为：

$$\bar{\theta}_{SAVGM}=\frac{\bar{\theta}_1-\bar{\theta}_2}{1-r} \tag{7}$$

$\bar{\theta}_{SAVGM}$ 的MSE的收敛速度是 $O\left(N^{-1}+n^{-3}\right)$，只要 $m\backsim n^2$，$\bar{\theta}_{SAVGM}$ 就可以达到全局估计量的收敛速度，且偏差的二阶项比 $\bar{\theta}_1$ 更小。第二种方法的优点是对 m 的假设更放松，且估计量的效果更好，但需要付出更多的计算消耗。

Rosenblatt和Nadler(2014)[12]指出在分布式计算中不同的极限性质，当 $N\to\infty$，m、p 固定的时候，分布式计算中，One-shot及平均各个“仆”计算机的局部变量得到最终估计量的方法与全局估计量是一阶项一致的，然而当 $p,n\to\infty,\frac{p}{n}\to\mu_l\in(0,1)$ 时，平均估计量则是次最优的。

Battey和Fanhan(2015)[1]进一步讨论了分布式计算中的One-shot(Divideand Conquer)方法的理论性质。其针对众多的假设检验和参数估计方法，提出了针对不断变大的数据量 n 的分布式计算机台数 k 的上界，使得可以保持估计量与全局统计量有同样的有效性，但是全局统计量在巨大的数据集前无法得到。One-shot方法是指在分布式系统中的各个“仆”计算机上得到统计量，然后在“主”计算机上按照适合的方法进行整合，以得到最终统计量的方法。本文首先针对高维Wald检验和Rao score检验提出了通讯成本有效的算法，并且证明在线性模型中，$k=O((s\times\log d)^{-1}\sqrt{n})$ 是保证统计推断有效性与全局统计量一致的上界；广义线性模型中的上界 $k=O(((s\vee s_1)\times\log d)^{-1}\sqrt{n})$。$s$ 指的是参数向量的稀疏性，d 代表参数向量的维度，n 是全样本的数据量，s_1 是信息矩阵逆的稀疏性。之后针对高维数据的估计准确性问题，参照假设检验的思路，在线性模型中给出 k 的上界为 $k=O\left(\sqrt{(s^2\times\log d)^{-1}n}\right)$，在广义线性模型中 $k=O\left(\sqrt{((s\vee s_1)^2\times\log d)^{-1}n}\right)$。因此，要保证Divideand Conquer方法的统计有效性与全局统计量一致，k 的增长速度不能太快，若是将模型限制在线性回归中，可以适当放松对 k 的限制。

Jordan等(2019)[13]指出了One-shot方法的三个缺点：(1)该方法只能得到待估计参数的点估计量，无法得到置信区间，不能统计推断和假设检验。(2)该方法对“仆”计算机的个数有较强的限制，不能太大，即要求每台计算机上存储的样本必须足够多，这是一个很强的假设，实际情况不一定满足。(3)该方法针对非线性的估计量表现极差。本文认为，通讯成本有效的替代极大似然估计的框架(CSL)，可以用于低维度的参数估计、高维的正则化估计以及贝叶斯估计。$Z_1^N=\left\{Z_{ij}:i=1,\cdots,n;j=1,\cdots,k\right\}$ 表示 N 条观测数据，存储在 k 台计算机上，数据服从分布 P_{θ^*}，$Z_j=\left\{Z_{ij}:i=1,\cdots,n\right\}$ 表示存储在第 j 台计算机上的 n 条数据。$L_j(\theta)=\frac{1}{n}\sum_{i=1}^{n}L(\theta,Z_{ij})$ 为局部的损失函数，整体的损

失函数如式(8)所示：

$$L_N(\theta)=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{k}L(\theta,Z_{ij})=\frac{1}{k}\sum_{j=1}^{k}L_j(\theta) \tag{8}$$

本文通过全局损失函数的Taylor展开以及用局部估计代替全局估计，提出了一个代替全局损失函数的 $\tilde{L}(\theta)=L_1(\theta)-\left\langle\theta,\nabla L_1(\bar{\theta})-\nabla L_N(\bar{\theta})\right\rangle$。$\bar{\theta}$ 是在第一台计算机上得到的局部估计，∇ 代表一阶求导。令 $\hat{\theta}=\arg\min_{\theta}L_N(\theta)$ 表示全局最优估计量，$\tilde{\theta}=\arg\min_{\theta}\tilde{L}(\theta)$ 表示通讯成本有效的估计量。本文针对低维数据的情况证明了 $\tilde{\theta}$ 与 $\hat{\theta}$ 的高阶一致性，且可以利用第一台计算机上的局部数据得到置信区间；在高维正则化问题的情况下，$\tilde{\theta}=\arg\min_{\theta}\tilde{L}(\theta)+\lambda\|\theta\|_1$ 和全样本数据下的Lasso估计量有同样的收敛速度；在贝叶斯框架中，CSL方法可以得到和全局一样好的后验估计，并依据服务器的个数 k 减少运算复杂度。

1.3 统计方法在分布式计算中的拓展

1.3.1 主成分分析(PCA)

Fan等(2017)[14]将主成分分析(PCA)拓展到分布式计算中，提出了一种新的算法，证明了在“仆”计算机的数量不是非常大的情况下，得到的PCA和全样本PCA表现一样好。本文将数据集按照样本分割开存储在分布式系统中。假设有 N 条观测数据、d 个维度的特征分布式地存储在 m 台机器上，每台机器存储 n 条数据，Σ 是样本方差矩阵，$V_K=(v_1,\cdots,v_K)$ 是样本方差矩阵的前 K 个特征向量。首先在各个机器上计算方差矩阵的 K 个特征向量，记为 $\left\{\hat{V}_K^{(l)}=(\hat{v}_1^{(l)},\cdots,\hat{v}_K^{(l)})\right\}_{l=1}^{m}$。然后在“主”计算机上进行整合，得到局部特征向量和整体方差矩阵的估计，$\tilde{\Sigma}=\frac{1}{m}\sum_{l=1}^{m}\hat{V}_K^{(l)}\hat{V}_K^{(l)T}$。最后得到 $\tilde{\Sigma}$ 的 K 个特征向量，$\tilde{V}_K=(\tilde{v}_1,\cdots,\tilde{v}_K)$。算法需要的通讯成本是 $O(mKd)$。把统计误差 $\|\tilde{V}_K\tilde{V}_K^T-V_KV_K^T\|_F$ 作为评价标准，只要每台机器上的样本数 n 足够大，分布式算法得到的PCA与全局PCA的统计误差具有同样的收敛速率。

1.3.2 特征筛选

Li等(2019)[15]研究了在分布式计算中实现特征筛选的方法。特征筛选是针对高维数据的一种降维方法，依据的标准是某种特定的相关性度量，Fan和Lv(2008)[11]中SIS特征筛选方法依据的相关性度量是Pearson相关性。因为特征筛选不用对数据做模型假设，所以适用于数据结构复杂的情况。现有的方法大多假定样本的特征个数 $p\to\infty$，而样本个数 N 是有限的。但在实际的数据中，p 和 N 都很大，考虑到 $p,N\to\infty$ 的情况，且存储在 m 台机器上，提出ACS变量筛选框架，在适当的条件下可证明ACS方法和传统的全局筛选方法一样有效，而且也不需要做模型假定。

$D=\{(Y_i,X_i)\}_{i=1}^{N}$ 代表全数据集，是独立同分布的，$X_i=(X_{i1},\cdots,X_{ip})^T$ 是 p 维解释变量。面对高维数据集时，通常只有部分解释变量与因变量 Y_i 相关，使用 M 指代与因变量相关的解释变量的指标集合，$M^c=\{1,\cdots,p\}\backslash M$ 指代无关变量的指标集合。变量筛选的目标就是移除 M^c 指代的无关变量，基本方法是依据某种相关性度量给解释变量排序，根据先验给定的阈值进行筛选，移除相关性低于阈值的变量。在分布式系统中，假设全数据集 D 被平均分割存储在 m 台机器上，每台机器存储 n 条数据。SAS表示简单的平均方法，即在第 l 台机器上得到 X_j 与 Y 的相关性度量 $\hat{\omega}_{l,j}$，然后在主机器上对其进行平均，得到整合的相关性度量 $\bar{\omega}_j=\frac{1}{m}\sum_{l=1}^{m}\hat{\omega}_{l,j}$，然后进行变量筛选。SAS方法计算简单，但缺点是可能会因为 $\hat{\omega}_{l,j}$ 的偏差导致 $\bar{\omega}_j$ 的偏差，在实际中的变量筛选是有偏差的。本文提出ACS方法，首先把 ω_j 表示成一些参数的函数：

$$\omega_j=g(\theta_{j,1},\cdots,\theta_{j,s}) \tag{9}$$

其中，g 是给定的函数，$\theta_{j,1},\cdots,\theta_{j,s}$ 是需要估计的参数，文中通过局部U-统计量给出 $\theta_{j,1},\cdots,\theta_{j,s}$ 的无偏估计。将各个“仆”计算机上得到的 $\theta_{j,1},\cdots,\theta_{j,s}$ 传输到“主”计算机上，对其进行平均之后，通过函数 g 得到估计 $\tilde{\omega}_j$，然后进行变量筛选。相较于SAS方法，本文提出的ACS方法更加准确和稳健。

1.3.3 分布式计算中的一步迭代估计量

Huang和Huo(2015)[16]提出了针对分布式计算的一步迭代估计量。不同于One-shot(Divideand Conquer)方法，用局部统计量的平均来获得最终估计量，本文针对M-估计量提出了一步迭代的估计量，其和全局估计量享有共同的收敛性质。数据模拟的结果显示一步迭代估计量比One-shot估计量有更好的表现。$m(x,\theta)$ 表示关心的标准函数，且二阶连续可导；数据集 S 包含 N 条观测数据，平均分布式存储在 k 台计算机上，每台计算机存储 n 条数据，记为 $S_i,i=1,\cdots,k$。$\dot{m}(x,\theta)$ 表示对 θ 的梯度向量，$\ddot{m}(x,\theta)$ 表示对 θ 的Hessian矩阵。全局标准函数 $M(\theta)=\frac{1}{k}\sum_{i=1}^{k}M_i(\theta)$，$M_i(\theta)=\frac{1}{|S_i|}\sum_{x\in S_i}m(x,\theta)$ 指的是第 i 台计算机上的局部标准函数。全局估计量 $\hat{\theta}=\arg\max_{\theta}M(\theta)$，第 i 台计算机的局部估计量为 $\hat{\theta}_i=\arg\max_{\theta}M_i(\theta)$。One-shot估计量为所有局部估计量的平均，记为 $\hat{\theta}^{(0)}=\frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i$。一步迭代估计量是在 $\hat{\theta}^{(0)}$ 的基础上，按照下列步骤得到一步迭代估计量 $\hat{\theta}^{(1)}$：

步骤1：“主”计算机将 $\hat{\theta}^{(0)}$ 传输给各个“仆”计算机，然后第 i 台计算机可以计算得到 $\dot{M}_i(\hat{\theta}^{(0)})$ 与 $\ddot{M}_i(\hat{\theta}^{(0)})$ 的值。

步骤2：各“仆”计算机把 $\dot{M}_i(\hat{\theta}^{(0)})$ 与 $\ddot{M}_i(\hat{\theta}^{(0)})$ 传输到“主”

计算机后通过平均的方法得到 $\dot{M}\left(\hat{\theta}^{(0)}\right)=\frac{1}{k}\sum_{i=1}^{k}\dot{M}_i\left(\hat{\theta}^{(0)}\right)$ 与 $\ddot{M}\left(\hat{\theta}^{(0)}\right)=\frac{1}{k}\sum_{i=1}^{k}\ddot{M}_i\left(\hat{\theta}^{(0)}\right)$ 的估计值。

步骤3：最后利用迭代公式 $\hat{\theta}^{(1)}=\hat{\theta}^{(0)}-\left[\ddot{M}\left(\hat{\theta}^{(0)}\right)\right]^{-1}\dot{M}\left(\hat{\theta}^{(0)}\right)$ 得到估计量。

证明 $\hat{\theta}^{(1)}$ 与全局估计量 $\hat{\theta}$ 有着相同的极限分布，而One-shot统计量只能达到同样的收敛速度。可以看到，虽然提出的是一步估计量，但是在实现的过程中多次用到了One-shot估计量的平均思想，初始估计是One-shot估计量，迭代步骤中的梯度与Hessian矩阵也是通过平均的方法得到的，而且需要多一轮的通讯，消耗更多的通讯成本，以换来更好的统计性质。

2 进一步研究方向

针对目前分布式计算中的统计学理论，有三个方面值得更进一步的研究：

（1）One-shot统计量是目前公认的比较成熟的分布式计算方法，计算简单，通讯成本低是其优点，但是也有其相应的缺点，如无法得到统计量的置信区间以及对分布式系统中的计算机个数有较严格的假设。已经有学者在这一方面提出新的研究方法，如何进一步地改善One-shot统计量是未来研究的一大方向。

（2）目前的研究都假设在不同的计算机上存储的数据样本是独立同分布的，这样可以保证每台计算机上得到的局部统计量是一致无偏的。但是在实际应用中，这样的假设是很难满足的。比如中国移动的通信数据就是按照地域不同来分割存储的，从而导致各计算机中的数据不服从独立同分布，得到的局部估计量则是有偏的。在这种情况下，One-shot统计量完全失效，需要新的方法来保证统计有效性。

（3）可以看到有的学者将主成分分析（PCA）与变量筛选拓展到分布式计算领域，这是很好的尝试。传统统计学中有很多经典的方法，如因子分析、逻辑回归、分位数回归、Bootstrap等。在数据体量日渐增大的今天，分布式计算是未来的主流发展趋势，如果能将这些传统经典的统计学模型拓展到分布式计算领域，将会是非常重要的工作。

3 结束语

分布式计算虽然目前还处于起步阶段，但众多统计学家已经在这一领域做出了贡献，本文重点介绍了一些统计学者在分布式计算当中做出的工作，展示了这一领域的广阔前景。本文只是对现有文献中关于分布式计算的一个基本介绍，另外还有诸如模型估计具体步骤、模型的统计特征、估计结果的渐近性质等问题尚未涉及，还需进一步进行研究。

参考文献：

[1]Battey H, Fanhan J. Distributed Estimation and Inference With Statistical Guarantees [J]. Statistics Theory,2015.

[2]Chen S S, Donoho D L, Saunders M A. Atomic Decomposition by Basis Pursuit [J].Siam Review,2001,43(1).

[3]Chen X, Xie M. A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data [J]. Statistical Sinica,2014,24(4).

[4]Efron B, Hastie T, Johnstone I, et al. Least Angle Regression [J]. Annals of Statistics,2004,(32).

[5]Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties [J].Journal of the American Statistical Association,2001,(96).

[6]Zhang C H. Nearly Unbiased Variable Selection Under Minimax Concave Penalty [J].Annals of Statistics,2010,(38).

[7]Zhang Y, Duchi J C, Wainwright M J. Communication-efficient Algorithms for Statistical Optimization [J].Journal of Machine Learning Research,2013,14(1).

[8]Lee J D, Liu Q, Sun Y, et al. Communication-efficient Sparse Regression [J].Journal of Machine Learning Research,2017,18(5).

[9]Javanmard A, Montanari A. Condence Intervals and Hypothesis Testing for High- dimensional Regression [J].The Journal of Machine Learning Research, 2014, 15(1).

[10]Yang J, Mahoney M W, Saunders M A, et al. Feature-distributed Sparse Regression: A Screen-and-clean Approach [J].Neural Information Processing Systems,2016.

[11]Fan J, Lv J. Sure Independence Screening for Ultrahigh Dimensional Feature Space [J]. Journal of the Royal Statistical Society Series B-Statistical Methodology,2008,70(5).

[12]Rosenblatt J, Nadler B. On the Optimality of Averaging in Distributed Statistical Learning [J].Information and Inference,2014,(1407).

[13]Jordan M I, Lee J D, Yang Y. Communication-efficient Distributed Statistical Inference [J].Journal of the American Statistical Association,2019,114(526).

[14]Fan J, Wang D,Wang K, et al. Distributed Estimation of Principal Eigenspaces [J]. Annals of Statistics, 2017,47(6).

[15]Li X, Li R, Xia Z, et al. Distributed Feature Screening via Componentwise Debiasing [J].Journal of Machine Learning and Research, 2019,21(24).

[16]Huang C, Huo X. A Distributed One-step Estimator [J]. Computer Science,2015.

（责任编辑/亦　民）

DOI:10.13546/j.cnki.tjyjc.2021.08.012

稳健MEWMA控制图的构建与应用

李雄英[1a],黄时文[1b],王斌会[2]

(1.广东财经大学a.经济学院;b.金融学院,广州 510320;2.暨南大学 管理学院,广州 510632)

摘 要:针对传统MEWMA控制图对离群值比较敏感,导致监控效果与实际情况不符这一现象,文章引入稳健统计的思想,将稳健MM估计与传统MEWMA控制图相结合,构造出稳健MEWMA控制图以达到抵御离群值影响的目的,同时进行了模拟和实证分析。模拟和实证分析的结果均表明:当数据中不存在离群值时,传统MEWMA控制图方法与稳健MEWMA控制图方法得到的结果基本保持一致;当数据中存在离群值时,传统MEWMA控制图不能很好地监测出过程的不受控状态,容易发生漏报的现象,而稳健MEWMA控制图可以很好地监测出过程的不受控状态,并发出出界报警信号。相对于传统MEWMA控制图,稳健MEWMA控制图能更有效地抵抗离群值的影响,具有良好的抗干扰性和抗差性,同时能够更好地监测到过程的失控状态。

关键词:MEWMA控制图;稳健统计;离群点;均值

中图分类号:O212.1 文献标识码:A 文章编号:1002-6487(2021)08-0058-05

0 引言

产品的质量具有多个方面的特征,在生产加工过程中,为了保证产品的质量需要同时对每个方面的质量特征都加以监控。多变量指数加权滑动平均(MEWMA:Multivariate Exponentially Weighted Moving Average)控制图技术将传统的单变量质量控制图技术拓展到了多变量质量监控,用一个控制图实现了对生产过程中产品多种质量特征的同时监控,从而提高了监控效率。

多变量指数加权滑动平均控制图技术一经提出,很快得到了较为广泛的应用[1—8],然而,随着计算机技术和物联网技术的发展,对企业生产过程中产品质量特征的监测已由过去间断的定期抽样监测转变为不间断的实时监测,大数据时代的到来使得传统的统计控制技术遇到了一些新问题,其中一个重要的问题是大数据中通常都含有较多的离群值,而离群值的存在会使控制图的监控效果与实际不相符。针对此问题,本文拟采用稳健统计的思想对传统MEWMA控制图进行改进,从而提高对离群值的识别和处理能力,并构建出稳健MEWMA控制图方法,以便能更好地应用于大数据时代的统计质量检测控制中。

1 传统MEWMA控制图的计算模型及其不稳健性

假设在生产或运营管理过程中需要被监测和控制的质量特征有 n 个,这些质量特性组成的随机向量为 $X=(x_1^T,x_2^T,\cdots,x_m^T)^T$,且服从均值向量为 μ 、协方差矩阵为 $\sum$ 的

基金项目:广东省教育厅特色人才类项目(人文社科)(2019KTSCX043);广州市哲学社会科学“十三五”规划一般课题(2019GZYB48);广东省哲学社会科学规划共建项目(GD17XGL08);广州市社会科学规划青年项目(2018GZQN36)

作者简介:李雄英(1987—),女,广东梅州人,博士,副教授,研究方向:大数据分析方法。

王斌会(1965—),男,陕西陇县人,教授,博士生导师,研究方向:管理统计方法。

Expansion of Statistical Methods in Distributed Computing

Ren Tunan[1,2]

(1. China UnionPay Post-doctoral Research Station, Shanghai 201201, China; 2. Post-doctoral Research Station of Computer Science and Technology, Fudan University, Shanghai 200433, China)

Abstract: In the era of increasing data volume, processing large volume of data has become a necessary approach for scientific researches. Distributed computing provides a solution for handling such a large volume of data, but from the statistical perspective, the convenience brought by distributed computing can also lead to loss of statistical properties. This paper reviews the integration of distributed computing and statistical theory, and analyzes both the advantages and disadvantages of these methods, pointing out the direction of further researches in this field.

Key words: distributed computing; one-shot method; high-dimensional sparse regression