



# 大规模数据分析中基于外推的调节参数选取

任好洁<sup>1</sup>, 邹长亮<sup>2\*</sup>, 李润泽<sup>3</sup>

1. 上海交通大学数学科学学院, 上海 200240;

2. 南开大学统计与数据科学学院, 天津 300071;

3. Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

E-mail: haojieren@gmail.com, nk.chlzou@gmail.com, rzli@psu.edu

收稿日期: 2020-07-20; 接受日期: 2020-08-24; 网络出版日期: 2021-03-26; \* 通信作者

国家自然科学基金 (批准号: 11931001, 11690014, 11690015 和 11925106) 和天津市自然科学基金 (批准号: 18JCJQJC46000) 资助项目

**摘要** 许多统计建模方法需要通过一个或多个调节参数来控制模型的复杂性. 这些调节参数可以是非参数回归或密度估计中核光滑方法的带宽, 也可以是在高维建模中利用正则化方法进行特征选择的相关正则化参数. 调节参数选择在统计建模和机器学习中起着关键作用. 对于大规模数据分析, 诸如基于信息准则的网格点搜索等常用的调节参数选择方法往往需要巨大的计算成本. 即使利用并行计算平台, 常用方法的可行性仍值得怀疑. 本文旨在开发一种快速算法来有效地近似最优调节参数. 该算法需要 (a) 假设一个参数模型来描述最优调节参数与样本大小之间的趋势, (b) 利用子采样数据拟合模型并建立趋势, 进而 (c) 将这种趋势外推至大样本的情形. 为了确定子采样样本大小, 在给定总计算成本预算的约束下, 本文推导子采样的理论最优设计, 并进一步证明估计的设计具有渐近最优性. 本文的数值研究表明, 在多个统计应用中, 即使利用一个简单的两参数幂函数模型, 所提出的算法所挑选的调节参数与基于全数据集所得到的调节参数几乎一致, 同时显著地减少了计算时间和存储空间, 表现出明显的优势.

**关键词** 渐近性 外推法 非线性模型 最优设计 预测误差 正则化方法

**MSC (2020) 主题分类** 62J02, 62J07

## 1 引言

大多数统计或机器学习方法都需要通过调节一个或多个参数来达到所期望的性能. 这样的调节参数通常控制着模型的复杂性, 因此, 调节参数的选择在理论分析和实际应用中都具有重要意义. 统计建模中有许多涉及调节参数选择的例子, 如非参数回归或密度估计中核光滑方法的带宽选取<sup>[1]</sup>、通过惩罚最小二乘方法进行变量选择时的正则化参数选择<sup>[2,3]</sup>、近邻分类中邻位顺序的确定<sup>[4]</sup>及基于阈值的方法中阈值的选择<sup>[5]</sup>等. 最优调节参数往往依赖于未知的总体参数, 所以不能直接应用于数据分

英文引用格式: Ren H J, Zou C L, Li R Z. Extrapolation-based tuning parameters selection in massive data analysis (in Chinese). Sci Sin Math, 2022, 52: 689–708, doi: 10.1360/SCM-2020-0622

析. 因此, 在实践中, 研究人员通常利用数据驱动的方法来选择调节参数. 但是, 对于具有超大样本量的数据, 传统数据驱动的方式常常难以适用. 这需要我们针对大数据分析开发新工具, 关于这方面的研究也引起了越来越多的关注.

现今, 尽管计算能力已经得到了稳步提升, 但大规模数据的激增仍然给有效选择调节参数带来了挑战. 在统计学习问题中, 通常利用交叉验证、广义交叉验证、Mallow  $C_p$  和 Bayes 信息准则等网格搜索准则进行调节参数的挑选, 以便于统计模型达到较好的预测能力. 这种网格点搜索的计算复杂度为  $O(d^q T)$ , 其中  $T$  为给定调节参数下整个过程所需计算时间,  $d$  为搜索每个调节参数所需的网格点数,  $q$  为模型中调节参数的个数. 计算复杂度以  $q$  的指数倍速度增加, 并且在实际中  $d$  通常选取在 20 到 100 之间, 因此, 调节参数选取可能非常耗时. 尽管一些可拓展的计算方法已经被开发用于减少计算或者降低存储, 如分治策略<sup>[6]</sup>, 但这些方法仅能用于减少计算时间  $T$ . 实际中, 即使  $O(T)$  是可以承受的, 实际应用者也可能不愿使用计算时间  $O(d^q T)$  仅用于选择调节参数. 因此, 需要开发在计算上可拓展的、同时可以保持良好统计特性的新方法.

当一种已建立的统计方法由于计算资源有限而不再适用时, 常用的解决方案就是从数据中提取有用信息的子抽样法<sup>[7,8]</sup>. 然而, 子抽样方法并不能直接应用于调节参数选取的问题, 因为最优调节参数往往依赖于样本大小. 本文提出一种全新的算法, 称为“基于子采样的外推法” (subsampling-based extrapolation, 以下记为 SUBEX), 来解决这样一个问题: “在大数据分析中, 当给定计算时间预算时, 如何最佳地利用该预算来找到有效可靠的调节参数组合?”

本文提出的算法需要假设一个参数模型来描述最优调节参数与样本量之间的趋势, 通过子采样的数据拟合该模型并建立趋势, 进而将这种趋势外推至样本量较大的情形. 在一些较弱条件假设下, 我们研究了所挑选的调节参数的渐近相合性和其收敛速度. 实现该算法的一个关键问题是如何设计子抽样样本大小. 我们得到一个最优设计, 该设计在总计算时间给定的约束下使得外推估计的渐近均方误差最小, 并提出一种求解约束优化问题的有效算法. 由于所建议的最优设计涉及一些未知参数, 我们建议首先获得相关参数的初步估计, 进而将这些估计代入最优设计中. 我们从渐近理论的角度证明了利用插值所得到的最优设计与其总体版本具有相同的性能. 数值研究表明, 在广泛的实际应用中, 即使对最优调节参数与样本量之间建立一个简单的两参数幂函数模型, 基于所提出算法得到的调节参数与使用全样本数据集所获得的调节参数具有几乎等效的表现, 我们的算法可以同时显著降低计算时间和存储空间.

本文余下内容的安排如下: 第 2 节提出一种全新算法, 并介绍子抽样最优设计的推导和估计及其在理论上的论证. 第 3 节讨论在实际应用中所建议的算法的一些实现问题. 第 4 节主要进行数值研究和展示一些实际数据示例. 第 5 节给出本文的讨论和总结. 附录 A 详述了相关理论证明.

## 2 调节参数选择的新方法

本节介绍调节参数选择基本算法、子抽样最优设计的推导和估计及一些渐近理论结果.

### 2.1 外推估计

假设全样本集合为  $\mathcal{X} = \{\mathcal{X}_1, \dots, \mathcal{X}_N\}$ ,  $\mathcal{A}(\mathcal{X}; \boldsymbol{\theta})$  表示当给定调节参数  $\boldsymbol{\theta} \in \boldsymbol{\Lambda} \subset \mathbb{R}^q$  时所得到的估计 (预测) 模型. 记挑选过程为  $f$ , 该函数将数据集映射到调节参数空间  $\boldsymbol{\Lambda}$ , 如交叉验证或 Bayes 信息准则. 假设  $\{\mathcal{X}_i\}_{i=1}^n$  是一个样本大小为  $n \leq N$  的子集,  $\boldsymbol{\lambda}(n) \in \boldsymbol{\Lambda}$  表示基于该子集利用  $f$  所确定的最

优调节参数, 即

$$\lambda(n) = f(\{\mathcal{X}_i\}_{i=1}^n). \quad (2.1)$$

本文所提出的外推步骤是将  $\lambda(n)$  建模为一个关于  $n$  的函数, 这里要求  $n \geq n_0$ ,  $n_0$  为预先指定的正整数, 进而将该模型外推至  $n = N$  的情形, 得到调节参数的外推估计量, 记为  $\hat{\lambda}(N)$ . 假设  $\lambda(n) = \{\lambda_1(n), \dots, \lambda_q(n)\}^\top$  沿着  $n$  的趋势可以用以下模型描述:

$$\lambda(n) = g(n; \theta) + \{D(n)\}^{1/2} \varepsilon, \quad (2.2)$$

其中  $n \geq n_0$ ,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_q)^\top$  满足  $E(\varepsilon) = 0$  和  $\text{cov}(\varepsilon) = I_q$ ,  $D(n) = \text{diag}\{\sigma_1^2(n), \dots, \sigma_q^2(n)\}$ . 令  $g(n; \theta) = \{g_1(n; \theta_1), \dots, g_q(n; \theta_q)\}^\top$ , 且  $g_j(n; \theta_j)$  ( $j = 1, \dots, q$ ) 为关于  $\theta = (\theta_1^\top, \dots, \theta_q^\top)^\top$  的给定 (非线性) 函数, 其中  $\theta_j \in \Theta_j \subset \mathbb{R}^{p_j}$  为  $p_j$  维参数向量. 例如,  $g_j(n; \theta_j) = \alpha_j n^{\beta_j}$  和  $\theta_j = (\alpha_j, \beta_j)^\top$ . 在模型 (2.2) 中, 允许  $\lambda_j(n)$  的方差可随着样本大小而变化, 这样的模型假设在实际中是合理的, 这是由于我们通常认为  $\lambda_j(n)$  相对于均值函数  $g_j(n; \theta_j)$  的变化将随着  $n$  的增长而单调递减. 图 1 展示了关于  $\lambda_j(n)$  变化的一个示例.

给定一组样本  $\mathcal{V}_m = \{n_1, \dots, n_m\}$ , 其中  $m \geq \max_{1 \leq j \leq q} p_j \equiv p$ , 可以利用加权最小二乘方法估计模型 (2.2), 即

$$\hat{\theta}_j = \arg \min_{\theta_j \in \Theta_j} \sum_{i=1}^m w_{ji} \{\lambda_j(n_i) - g_j(n_i; \theta_j)\}^2, \quad j = 1, \dots, q, \quad (2.3)$$

其中  $\lambda(n_1), \dots, \lambda(n_m)$  通过独立随机子采样利用 (2.1) 获得. 考虑到方差异质性,  $w_{ji}$  的标准选择是  $1/\sigma_j^2(n_i)$  的某个原始估计. 由此可以得到  $\hat{\lambda}(N) = g(N; \hat{\theta})$  和  $\hat{\theta} = (\hat{\theta}_1^\top, \dots, \hat{\theta}_q^\top)^\top$ .

**注 2.1** 模型 (2.2) 是一般化多响应非线性模型的一种特殊情形 (参见文献 [9, 第 11 章]), 其中简单地假设不同调节参数对应的  $\theta_j$  都是不同的, 这使得我们可以分别逐个估计它们. 否则,  $\theta$  是由  $\theta_j$  中所有不同元素构成的向量, 然后需要利用联合最小二乘进行估计. 模型 (2.2) 和相关的估计过程 (2.3) 极大地简化了计算, 并且足以满足我们的需求.

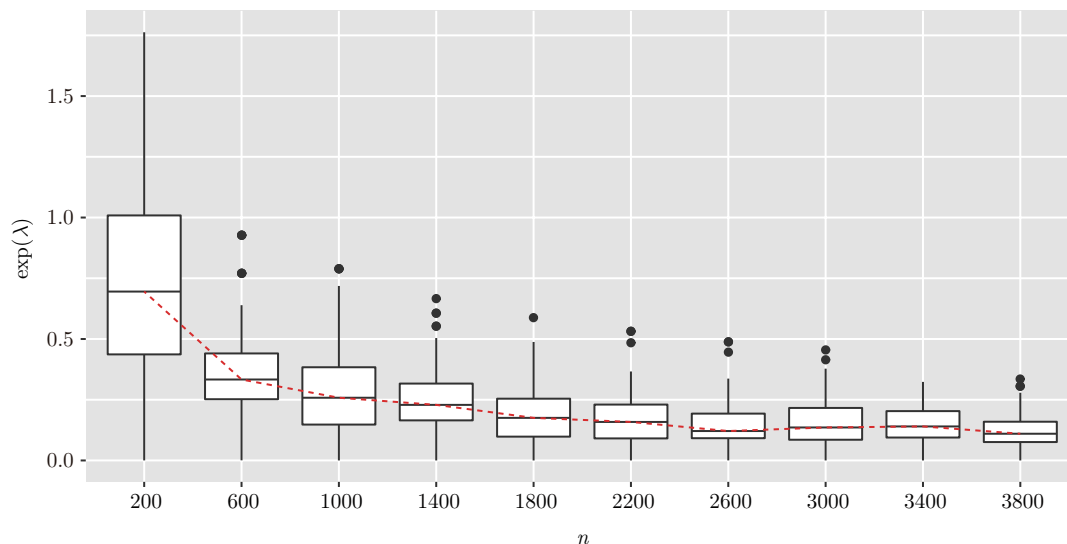


图 1 (网络版彩图) 例 4.1(i) 所对应的  $\lambda(n)$  关于  $n$  变化的盒子图, 虚线表示每个  $n$  下对应的平均  $\lambda$

我们已经在许多应用中证实,  $\lambda_j(\cdot)$  是一个单调函数, 特别是在  $n$  较大时. 例如, 在核光滑中, 渐近最优带宽通常是  $N$  的多项式形式<sup>[10]</sup>. 同样, 图 1 中的示例也提供了一些见解. 因此, 在调节参数选取问题中找到一个合适的外推法并不难. 第 3.3 小节详细地讨论了一个两参数幂函数模型, 并将该模型广泛地应用于第 4 节中. 数值模拟结果表明, 即使使用这种简单的外推法也可以获得相当优良的预测  $\hat{\lambda}(N)$ , 而且该过程可以广泛应用在大量的统计学习问题中. 相比于外推估计的准确性, 更重要的是, 我们发现对于任何特定的学习方法, 基于该外推估计可以获得一个足够好的  $\mathcal{A}(\mathcal{X}; \lambda(N))$  模型近似. 关于此问题的更多讨论可参见第 2.4 小节.

## 2.2 子抽样样本大小的最优设计

要实现外推法, 我们必须指定子抽样样本  $\mathcal{V}_m = \{n_1, \dots, n_m\}$  的大小, 其中  $n_1 \leq \dots \leq n_m$ . 一个简单的选择是从  $n_0$  开始等间距地选取样本大小. 但是, 基于等间距设计的算法可能不是最优的选择, 同时  $m$  的选择也仍然是一个问题. 我们提出了一个更有效的过程, 通过选择  $\mathcal{V}_m$  来获得最佳的预测  $\hat{\lambda}(N)$ .

在确定  $\mathcal{V}_m$  时, 需要有效地平衡统计估计准确性和计算成本, 因为较大的  $n_i$  和  $m$  通常会得到  $\lambda(N)$  的更精确估计, 但同时计算更加烦琐. 给定计算时间预算  $\mathcal{T}$ , 我们的目标是最小化外推预测  $\hat{\lambda}(N)$  的均方误差, 其中设计点  $\mathcal{V}_m$  可通过  $\lambda(n_1), \dots, \lambda(n_m)$  发挥其影响, 即最小化

$$\mathbb{E}\{[\hat{\lambda}(N) - \lambda(N)]^\top \{\mathbf{D}(N)\}^{-1} [\hat{\lambda}(N) - \lambda(N)]\} \quad (2.4)$$

使得

$$\sum_{i=1}^m t(n_i) \leq \mathcal{T}, \quad (2.5)$$

其中  $\mathbf{D}(N) = \text{diag}\{\sigma_1^2(N), \dots, \sigma_q^2(N)\}$ ,  $t(n)$  表示  $f(\{\mathcal{X}_i\}_{i=1}^n)$  “平均” 的计算时间. 实际上,  $t(n)$  可以近似为  $(n/n_0)^\zeta t(n_0)$ , 其中  $\zeta \in \mathbb{N}^+$  是在给定  $\boldsymbol{\theta}$  时  $\mathcal{A}(\{\mathcal{X}_i\}_{i=1}^n; \boldsymbol{\theta})$  的计算复杂度 (相对于样本大小), 通常是已知的先验信息. 例如, 在线性回归估计中,  $\zeta = 1$ ; 在支持向量机分类问题中,  $\zeta = 2$ . 在目标函数 (2.4) 中, 我们简单地考虑具有对角矩阵  $\mathbf{D}(N)$  的标准 Euclid 距离, 而不是 Markov 距离, 因为后者需要对协方差函数  $\text{cov}\{\lambda_{j_1}(n), \lambda_{j_2}(n)\}$  建模, 这在实际应用中可能难以处理, 而且可能不会带来太多的收益.

如果  $n_m \ll N$ , 则可以合理地假设  $\{\boldsymbol{\varepsilon}(n_1), \dots, \boldsymbol{\varepsilon}(n_m)\}$  近似独立于  $\boldsymbol{\varepsilon}(N)$ . 利用在  $n = N$  时的 (2.2) 替换  $\lambda(N)$ , 进而, 最小化 (2.4) 近似等价于最小化

$$\mathbb{E}\{g(N; \boldsymbol{\theta}^*) - g(N; \hat{\boldsymbol{\theta}}(\mathcal{V}_m))\}^\top \{\mathbf{D}(N)\}^{-1} \{g(N; \boldsymbol{\theta}^*) - g(N; \hat{\boldsymbol{\theta}}(\mathcal{V}_m))\}, \quad (2.6)$$

其中  $\boldsymbol{\theta}^*$  表示模型 (2.2) 中的真实调节参数,  $\hat{\boldsymbol{\theta}}(\mathcal{V}_m)$  是利用  $\mathcal{V}_m$  得到的 (2.3) 估计量.

在一定的条件下,  $g_j(n_i; \hat{\boldsymbol{\theta}}_j(\mathcal{V}_m))$  的 Taylor 展开为

$$g_j(n_i; \hat{\boldsymbol{\theta}}_j(\mathcal{V}_m)) \approx g_j(n_i; \boldsymbol{\theta}_j^*) + \{g_j'(n_i; \boldsymbol{\theta}_j^*)\}^\top \{\hat{\boldsymbol{\theta}}_j(\mathcal{V}_m) - \boldsymbol{\theta}_j^*\}.$$

因此, 可以得到

$$\hat{\boldsymbol{\theta}}_j(\mathcal{V}_m) - \boldsymbol{\theta}_j^* \approx (\mathbf{Z}_j^\top \mathbf{W}_j \mathbf{Z}_j \mathbf{V}_m)^{-1} \mathbf{Z}_j^\top \mathbf{W}_j \mathbf{Y}_j, \quad j = 1, \dots, q, \quad (2.7)$$

其中

$$\begin{aligned} \mathbf{W}_j &= \text{diag}\{w_{j1}, \dots, w_{jm}\}, \\ \mathbf{Z}_{j\mathcal{V}_m} &= \{g'_j(n_1; \boldsymbol{\theta}_j^*), \dots, g'_j(n_m; \boldsymbol{\theta}_j^*)\}^\top, \\ \mathbf{Y}_j &= \{\lambda_j(n_1) - g(n_1; \boldsymbol{\theta}_j^*), \dots, \lambda_j(n_m) - g(n_m; \boldsymbol{\theta}_j^*)\}^\top. \end{aligned}$$

类似地, 对  $\mathbf{g}(N; \hat{\boldsymbol{\theta}}(\mathcal{V}_m))$  进行 Taylor 展开, (2.6) 可以近似为

$$\sum_{j=1}^q \frac{\mathbb{E}[\mathbf{Z}_{jN}^\top \{\hat{\boldsymbol{\theta}}_j(\mathcal{V}_m) - \boldsymbol{\theta}_j^*\}]^2}{\sigma_j^2(N)}, \quad (2.8)$$

其中  $\mathbf{Z}_{jN} = g'_j(N; \boldsymbol{\theta}_j^*)$ . 然后, 将 (2.7) 代入 (2.8), 目标函数 (2.6) 则变成

$$\text{AMSE}(\mathcal{V}_m) = \sum_{j=1}^q \sigma_j^{-2}(N) \mathbf{Z}_{jN}^\top (\mathbf{Z}_{j\mathcal{V}_m}^\top \mathbf{W}_j \mathbf{Z}_{j\mathcal{V}_m})^{-1} (\mathbf{Z}_{j\mathcal{V}_m}^\top \mathbf{W}_j \boldsymbol{\Sigma}_j \mathbf{W}_j \mathbf{Z}_{j\mathcal{V}_m}) (\mathbf{Z}_{j\mathcal{V}_m}^\top \mathbf{W}_j \mathbf{Z}_{j\mathcal{V}_m})^{-1} \mathbf{Z}_{jN}, \quad (2.9)$$

其中

$$\boldsymbol{\Sigma}_j = \text{diag}\{\sigma_j^2(n_1), \dots, \sigma_j^2(n_m)\}.$$

在 (2.5) 的约束下,  $m \leq \lfloor \mathcal{T}/t(n_0) \rfloor \equiv m_{\max}$ . 对于每个  $m$ ,  $\mathcal{V}_m$  中的最大值应该小于

$$n_{\max}(m) = \lfloor t^{-1}(\{\mathcal{T} - (m-1)t(n_0)\}) \rfloor.$$

因此, 为我们外推预测找最优设计的问题转变为求解

$$\min_{m \in [p, m_{\max}]} \min_{(n_1, \dots, n_m) \in [n_0, n_{\max}(m)]} \text{AMSE}(\mathcal{V}_m) \quad \text{s.t.} \quad \sum_{i=1}^m t(n_i) \leq \mathcal{T}, \quad (2.10)$$

使得上式达到最小值所对应的设计记为  $(m^*, \mathcal{V}_m^*)$ . 称此调节参数选取过程为基于子采样的外推法 (SUBEX).

自文献 [11] 的开创性工作以来, 非线性模型的最优设计在理论方面得到了广泛的发展, 但是据我们所知, 设计准则 (2.10) 是全新的, 详见文献 [12]. 我们所关心的问题与  $D$  最优设计相关, 也就是最小化  $\hat{\boldsymbol{\theta}}$  的渐近协方差矩阵的行列式, 但是目标函数 (2.10) 是针对我们问题 (即, 预测  $\boldsymbol{\lambda}(N)$ ) 量身定制的.

### 2.3 最优设计的估计

如 (2.9) 所示的  $\text{AMSE}(\mathcal{V}_m)$  依赖于未知变量  $\boldsymbol{\theta}_j^*$  和  $\sigma_j^2(\cdot)$ , 因此, 不可以直接应用精确的 SUBEX 方法. 一个解决方案就是通过使用一组可行的样本量进行初步研究以获得未知变量原始且富含信息的估计, 然后将所得到的估计值代入 (2.10) 中以逼近最优设计, 以便在之后的阶段中可以实现更精确的模型估计.

具体而言, 从全数据集中随机采样  $m_0 = q$  个具有不同大小的子样本  $\tilde{\mathcal{V}}_{m_0} = \{\tilde{n}_1, \dots, \tilde{n}_{m_0}\}$  两次, 并通过 (2.1) 获得相应的调节参数, 记为  $\tilde{\boldsymbol{\lambda}}^{(l)}(\tilde{n}_1), \dots, \tilde{\boldsymbol{\lambda}}^{(l)}(\tilde{n}_{m_0})$ ,  $l = 1, 2$ . 为了估计变量  $\sigma_j^2(\cdot)$ , 我们再次假设一个参数模型, 即  $\sigma_j^2(n) = h_j(n; \boldsymbol{\gamma}_j)$ , 其中  $\boldsymbol{\gamma}_j \in \mathbb{R}^{r_j}$ ,  $1 \leq r_j \leq m_0$ . 然后, 利用最小二乘方法 [13] 估计  $\boldsymbol{\gamma}_j$ , 即

$$\tilde{\boldsymbol{\gamma}}_j = \arg \min_{\boldsymbol{\gamma}} \sum_{i=1}^{m_0} \{\tilde{\sigma}_j^2(n_i) - h_j(n_i; \boldsymbol{\gamma}_j)\}^2, \quad j = 1, \dots, q, \quad (2.11)$$

其中  $\tilde{\sigma}_j^2(n_i) = \{\tilde{\lambda}^{(1)}(\tilde{n}_i) - \tilde{\lambda}^{(2)}(\tilde{n}_i)\}^2/2$  是  $\sigma_j^2(n_i)$  的一个粗略的估计.

通过将 (2.3) 中的  $\lambda(n_i)$  和  $w_{ji}$  分别替换为  $\{\tilde{\lambda}^{(1)}(\tilde{n}_i) + \tilde{\lambda}^{(2)}(\tilde{n}_i)\}/2$  和  $1/\tilde{\sigma}_j^2(\tilde{n}_i)$ , 可以得到  $\theta$  的初步估计, 记为  $\tilde{\theta}$ . 将研究限制在设计点的个数 ( $m_0$ ) 等于调节参数的个数 ( $q$ ) 的情形, 并且对于每个  $\tilde{n}_i$  仅进行两次随机采样, 这是获得一个初步估计的最低计算需求, 而且我们发现这样的估计在大部分情形下都表现较好.

因此,  $\text{AMSE}(\mathcal{V}_m)$  的估计版本为

$$\widehat{\text{AMSE}}(\mathcal{V}_m) = \sum_{j=1}^q \frac{\tilde{\mathbf{Z}}_{jN}^\top (\tilde{\mathbf{Z}}_{j\mathcal{V}_m}^\top \mathbf{W}_j \tilde{\mathbf{Z}}_{j\mathcal{V}_m})^{-1} \tilde{\mathbf{Z}}_{jN}}{h_j(N; \tilde{\gamma}_j)}, \quad (2.12)$$

其中  $\text{AMSE}(\mathcal{V}_m)$  中的  $\theta^*$ 、 $\sigma_j^2(\cdot)$  和  $\Sigma_j$  被分别替换为  $\tilde{\theta}$ 、 $h_j(\cdot; \tilde{\gamma}_j)$  和  $\mathbf{W}_j^{-1}$ . 进而, 估计的最优设计为

$$(\hat{m}, \hat{\mathcal{V}}_{\hat{m}}) = \arg \min_{m, \mathcal{V}_m} \widehat{\text{AMSE}}(\mathcal{V}_m) \quad \text{s.t.} \quad \sum_{i=1}^m t(n_i) \leq \mathcal{T}. \quad (2.13)$$

在上面的讨论中, 一个隐含的假设是, 初步研究中使用的计算时间相较于总时间约束  $\mathcal{T}$  可忽略不计, 否则, 需要修改 (2.13) 中的约束, 即减去初步研究中所利用的计算时间. 此外, 我们主要关注计算时间  $t(n_i)$ , 忽略了其他影响计算时间的因素, 如拟合非线性模型和寻找最优设计, 因为这两者在大多数情形下都可以通过非常快速有效的计算得到, 具体参见第 3 节.

## 2.4 渐近理论性质

本小节研究预测变量  $\hat{\lambda}(N)$  的渐近相合性及其收敛速度, 以及最优设计估计的相合性. 我们在此主要讨论实现 SUBEX 过程所需的理论条件, 相关结果的理论证明将在附录中阐述.

在渐近理论框架中, 我们处理以下情形:  $\mathcal{V}_m$  中的样本大小  $n_i$ 、 $\mathcal{V}_{m_0}$  中的  $\tilde{n}_i$  以及  $N$  都是随着  $n_0$  而递增, 也就是当  $n_0 \rightarrow \infty$  时  $n_i, \tilde{n}_i, N \rightarrow \infty$ . 关于最小二乘估计的相合性和渐近正态性在非线性模型中已经被广泛研究, 参见文献 [14, 15]. 然而, 现有的大多数理论都是建立在  $m \rightarrow \infty$  之下的, 也就是要求“观测”的数量趋于无穷大, 这显然不适用于我们考虑的情形. 因此, 考虑一种不同类型的渐近理论, 即固定  $m$ , 但让  $n_0 \rightarrow \infty$ . 这通常被称为“小误差”渐近性<sup>[16]</sup>, 即假设误差项  $\sigma_j(n_i)\varepsilon_{ji}$  足够小且  $\sigma_j^2(\cdot)$  的幂可展开, 相关讨论可参见文献 [17]. 值得注意的是, 模型 (2.2) 具有特殊的结构, 也就是说“协变量”依赖于  $n_0$ , 并且随着  $n_0 \rightarrow \infty$  均值函数本身可能会发散到无穷大或趋近于 0, 这就使得我们的理论无法从现有的工作中直接得到.

为了便于理论结果证明, 进行如下假设 ( $j = 1, \dots, q$ ).

**假设 2.1** 对于一个固定的  $C < \infty$  和  $i = 1, \dots, m_0$ ,  $E(\varepsilon_{ji}^4) \leq C < \infty$ .

**假设 2.2**  $\Theta_j$  是  $\mathbb{R}^{p_j}$  的一个紧子集, 且  $\theta_j^*$  是  $\Theta_j$  的一个内点.

**假设 2.3**  $g_j(\cdot; \theta)$  是  $\theta_j \in \Theta_j$  的连续函数, 且对所有的  $\theta_j \in \Theta_j^*$  具有连续一阶导数, 其中  $\Theta_j^* \in \Theta_j$  是  $\theta_j^*$  的一个开邻域.

**假设 2.4** 存在序列  $a_{jn_0} \rightarrow \infty$  (该序列可能依赖于  $\theta_j^*$ ) 使得

$$a_{jn_0}^{-1} \sum_{i=1}^m \frac{g'_j(n_i; \theta_j^*) \{g'_j(n_i; \theta_j^*)\}^\top}{\sigma_j^2(n_i)} \rightarrow \Omega_j(\theta_j^*), \quad \text{当 } n_0 \rightarrow \infty \text{ 时,}$$

其中  $\Omega_j = \Omega_j(\theta_j^*)$  是非退化的.

**假设 2.5** 对于任意  $t > 0$  和  $n \geq n_0$ , 函数  $g'_j(n; x)$  满足

$$\frac{\|g'_j(n; x+t) - g'_j(n; x)\|}{\|g'_j(n; x)\|} \leq M_{jn}t,$$

其中  $M_{jn}$  是某个满足当  $n_0 \rightarrow \infty$  时  $M_{jn}/a_{jn_0} \rightarrow 0$  的正序列.

**假设 2.6**  $h_j(\cdot; \gamma_j)$  是关于  $\gamma_j$  的连续函数;  $\gamma_j$  的参数空间  $\Gamma_j$  是  $\mathbb{R}^{r_j}$  的一个紧子集, 并且在异方差的模型中, 真实参数  $\gamma_j^*$  是  $\Gamma_j$  的一个内点.

**假设 2.7** (i) 存在序列  $b_{jn_0} \rightarrow \infty$  (可能依赖于  $\gamma_j$  和  $\tilde{\gamma}_j$ ) 使得  $b_{jn_0}^{-1} \sum_{i=1}^{m_0} h_j(\tilde{n}_i; \gamma_j) h_j(\tilde{n}_i; \tilde{\gamma}_j)$  对于所有  $\Gamma_j$  中的  $\gamma_j$  和  $\tilde{\gamma}_j$  一致收敛至函数  $B(\gamma_j, \tilde{\gamma}_j)$ ;

(ii) 假设  $D(\gamma_j, \gamma_j^*) = 0$  当且仅当  $\gamma_j = \gamma_j^*$ , 其中

$$D(\gamma_j, \tilde{\gamma}_j) = B(\gamma_j, \gamma_j) + B(\tilde{\gamma}_j, \tilde{\gamma}_j) - 2B(\gamma_j, \tilde{\gamma}_j);$$

(iii) 对于  $i = 1, \dots, m_0$ , 随着  $n_0 \rightarrow \infty$ ,  $\sigma_j^2(\tilde{n}_i)/b_{jn_0} \rightarrow 0$ .

**注 2.2** 在研究  $\hat{\lambda}(N)$  的性质时, 假设 2.2 中  $\Theta_j$  的有界性并不是一个严格的限制, 因为大多数参数都受到所建模型本身约束的限制. 假设 2.3 是建立非线性最小二乘估计存在性的一个标准假设. 假设 2.4 类似于文献 [9] 的第 12 章中的条件 (A4), 也就是推导收敛速率的必要条件. 由于均值和方差函数都依赖于  $n_0$ , 可将  $a_{jn_0}$  看作是一个“新的”收敛速率. 假设 2.5 是一个技术要求, 在此条件下  $\hat{\theta}$  的渐近展开成立 (即等式 (2.7)). 大多数常用的函数都满足这个假设, 如不同  $M_{jn}$  下的函数  $x^a$ 、 $\exp(x)$ 、 $\log(x)$  和  $x/(1+x)$ . 假设 2.6 和 2.7 给出的是建立  $\tilde{\gamma}_j$  相合性所需的条件, 它们与假设 2.2–2.5 是对应的, 除了我们使用假设 2.7 而不是  $h(n; \gamma_j)$  的一阶导数条件, 因为我们的主要结果不需要  $\tilde{\gamma}_j$  的收敛速度. 一个类似于假设 2.7 的条件也常常在文献中被用于研究传统渐近设置, 即 “ $m \rightarrow \infty$ ” 的情形, 具体可参见文献 [9, 第 564 页].

我们以一个例子来验证这些假设的有效性, 令  $g_j(n; \theta_j) = \alpha_{gj} n^{\beta_{gj}}$ ,  $h_j(n; \theta_j) = \alpha_{hj} n^{\beta_{hj}}$ ,  $\mathcal{V}_m = \{C_1 n_0, \dots, C_m n_0\}$ , 其中  $1 \leq C_1 < \dots < C_m < \infty$  为一些常数. 在这种情形下, 当  $M_{jn} \sim \log n$  和  $a_{jn_0} \sim (\log n_0)^2 n_0^{2\beta_{gj} - \beta_{hj}}$  时, 假设 2.4 和 2.5 成立. 假设  $\beta_{hj} < 2\beta_{gj}$  是合理的, 因为  $\lambda_j(n)$  的变化通常比其均值函数有更小的阶数. 同样, 我们可以很容易验证假设 2.6 和 2.7 也是成立的.

我们在第一个定理中证明了所提出的预测变量  $\hat{\lambda}(N)$  的相合性及其收敛至  $\mathbf{g}(N; \theta^*)$  的速度. 需要特别指出的是,  $\hat{\theta}$  是基于变量的初步估计所得到的, 也就是 (2.3) 中  $w_{ji}$  为  $1/h_j(n_i; \tilde{\gamma}_j)$ ,  $\tilde{\gamma}_j$  是由 (2.11) 在初步研究中得到的.

**定理 2.1** 若假设 2.1–2.7 成立, 且  $n_0 \rightarrow \infty$ , 则

$$\frac{\|\hat{\lambda}(N) - \mathbf{g}(N; \theta^*)\|}{\|\mathbf{g}'(N; \theta^*)\|} = O_p\left\{\left(\min_j a_{jn_0}\right)^{-1/2}\right\}.$$

这一结果为我们的方法提供了理论依据. 定理中的收敛速度由假设 2.4 中指出的  $\hat{\theta}$  的协方差收敛速率所决定.

**定理 2.2** 若假设 2.1–2.7 成立, 则当  $n_0 \rightarrow \infty$  时,

$$\frac{\text{AMSE}(\hat{V}_{\hat{m}})}{\text{AMSE}(\mathcal{V}_{m^*}^*)} \xrightarrow{p} 1,$$

即说明估计的设计与最优设计非常接近.

该定理的证明是通过建立带有插值估计的优化准则 (2.12) 的一致收敛性得到的. 尽管这个定理并不能保证  $\hat{\mathbf{V}}_{\hat{m}}$  收敛至  $\mathbf{V}_{m^*}^*$  (这需要关于  $\mathbf{g}$  更严格的条件), 但这个定理仍然表明至少从渐近的角度来看估计的设计与理论上的最优设计一样有效, 因为它们所预测的调节参数可以导致渐近等效的 AMSE (asymptotic mean square error).

### 3 新方法的实际应用指南

本节将为所提出的方法 SUBEX 在实际应用中提供一些指南, 并为 (2.13) 提供一个数值优化算法.

#### 3.1 参数及模型的选取

首先, 讨论 SUBEX 在实际应用中如何选取非线性模型和  $\tilde{\mathbf{V}}_{m_0}$ , 以及如何确定  $n_0$ .

**模型选取** 是否可以利用外推法预测调节参数是所提出方法面对的一个挑战. 如果不能令人信服且有效地实现外推, 本文方法的效用将非常有限. 但是, 无论外推是否充分, 我们关于调节参数估计的方法仍有较大的价值, 因为  $\boldsymbol{\lambda}(n)$  与  $n$  之间的函数趋势始终是有帮助的, 尤其对于复杂数据或计算密集型统计学习方法而言, 它们中最优调节参数的理论表达经常是很难得到的. 预先利用可视化展示  $\boldsymbol{\lambda}(n)$  与  $n$  之间的关系往往能够为选取适当形式的外推模型提供一些见解.

如前所述,  $g_j(n; \boldsymbol{\theta})$  和  $h_j(n; \boldsymbol{\gamma})$  通常都是关于  $n$  的单调函数, 因此, 我们发现一些典型的增长模型 (参见文献 [9, 第 7 章]) 不仅足够灵活, 而且也可以很好地用来拟合数据, 例如, 指数模型  $g(x) = a + b \exp(-cx)$ , 或者不像指数模型增长那么快的幂函数模型  $g(x) = ax^b$ .

**$n_0$  的确定** 显然, 如果不考虑计算成本的话, 较大的  $n_0$  将会导致更好的结果. 在给定  $\mathcal{T}$  下,  $n_0$  应该为使得  $\mathcal{A}(\{\mathcal{X}_i\}_{i=1}^{n_0}; \boldsymbol{\lambda}(n_0))$  得到一定程度上稳定结果的样本大小, 参见文献 [18]. 因此, 选取多大的  $n_0$  是相当主观的, 不仅依赖于计算限制, 也取决于应用者在实现统计学习算法上的经验和先验知识.

**初步研究中  $\tilde{\mathbf{V}}_{m_0}$  的选取** 我们在第 2.4 小节中的理论结果表明, 只要从初步研究中可以得到  $\boldsymbol{\theta}_j^*$  和  $\boldsymbol{\gamma}_j^*$  的相合估计, 无论其收敛速度有多快, 估计的设计都非常接近最优设计. 因此, 从经验上讲, 我们可以从  $n_0$  开始等间距选取样本大小作为  $\tilde{\mathbf{V}}_{m_0}$ , 即  $\{(1 + (i-1)\rho)n_0\}_{i=1}^q$  足以获得有效的估计, 其中  $\rho > 0$  为一个预先指定的常数. 数值研究中, 在没有偏好情形下, 建议选择  $\rho = 0.5$ .

#### 3.2 最小化 (2.13) 的算法

目标函数 (2.13) 是一个  $m$  维的整数约束优化问题. 对所有网格点的穷尽搜索非常耗时, 需要一个更快的替代方案. 实际上, 通常只需从一个“小”的集合中挑选可能的  $n_i$ , 如集合

$$\mathcal{S} = (s_1, \dots, s_d) \subset [n_0, n_{\max}(m_{\max})].$$

由此, (2.13) 等价于

$$\begin{aligned} \min_{k_1, \dots, k_d} \quad & \sum_{j=1}^q \tilde{\mathbf{Z}}_{jN}^\top \left( \sum_{i=1}^d k_i w_{js_i} \tilde{\mathbf{Z}}_{js_i} \tilde{\mathbf{Z}}_{js_i}^\top \right)^{-1} \frac{\tilde{\mathbf{Z}}_{jN}}{h_j(N; \tilde{\boldsymbol{\gamma}}_j)} \\ \text{s.t.} \quad & k_i \in \mathbb{N}, \quad k_i \geq 0, \quad \sum_{i=1}^d k_i t(s_i) \leq \mathcal{T}. \end{aligned}$$

$k_i = 0$  表示  $s_i$  不会在  $\mathcal{V}_m$  中使用, 而  $k_i > 1$  意味着设计点  $s_i$  重复了  $k_i$  次. 通过简单的运算, 最小化函数可以重写为

$$\begin{aligned} \min_{k_1, \dots, k_d} \sum_{j=1}^q & \left[ -\log \left\{ \det \left( \sum_{i=1}^d k_i w_{js_i} \tilde{\mathbf{Z}}_{js_i} \tilde{\mathbf{Z}}_{js_i}^\top \right) \right\} \right. \\ & \left. + \log \left\{ \sum_{i=1}^d k_i^{p_j-1} \tilde{\mathbf{Z}}_{jN}^\top (w_{js_i} \tilde{\mathbf{Z}}_{js_i} \tilde{\mathbf{Z}}_{js_i}^\top)^* \frac{\tilde{\mathbf{Z}}_{jN}}{h_j(N; \tilde{\gamma}_j)} \right\} \right] \\ \text{s.t. } & k_i \in \mathbb{N}, \quad k_i \geq 0, \quad \sum_{i=1}^d k_i t(s_i) \leq \mathcal{T}. \end{aligned} \quad (3.1)$$

可以通过忽略或放松  $k_i$  为整数的约束来找到 (3.1) 的近似解. 放松后的 (3.1) 是一个凸差优化问题, 因为目标函数由两个关于  $(k_1, \dots, k_d)^\top$  的凸函数组成. 可以验证, 最优结果在  $\sum_{i=1}^d k_i t(s_i) = \mathcal{T}$  处达到. 凸差问题在非凸函数优化中占有重要地位, 并且在近三十年得到了广泛的研究, 一些算法可用于解决约束最小化问题 (3.1), 如文献 [19] 中提出的近点算法. 更多关于凸差问题的研究可参见文献 [20]. 约束条件  $\sum_{i=1}^d k_i t(s_i) \leq \mathcal{T}$  ( $k_i \geq 0$ ) 本质上与加权的  $L_1$  惩罚<sup>[21]</sup> 相关, 且会有一个稀疏的解, 即最小化的估计中一部分元素精确为零<sup>[2]</sup>.

显然, 放松约束后得到的最优解为优化问题 (3.1) 的最优解提供了一个下界, 因为问题 (3.1) 有一个附加的约束. 可以通过四舍五入放松约束后的最优解, 构造问题 (3.1) 的次最优解; 对于较大的  $\mathcal{T}$ , 约束条件  $\sum_{i=1}^d k_i t(s_i) = \mathcal{T}$  可以几乎得到满足. 此方法与穷举搜索在模拟研究中表现差异不显著.

### 3.3 一个具体模型例子

本小节考虑使用幂函数模型  $g(x) = ax^b$  作为例证. 我们已知一些问题是精确符合该模型假设的, 例如, 核光滑方法中带宽的选取, 以及在估计一个具有  $J$  维的大协方差矩阵时, 协方差正则化方法中的阈值设置  $c\sqrt{\log J/n}$  (参见文献 [5]). 幂指数模型也满足许多其他问题的需求. 我们对每个调节参数进行  $\log$  转换,

$$\log\{\lambda_j(n)\} = a_j + b_j \log n + \sigma_j(n) \varepsilon_j(n), \quad j = 1, \dots, q, \quad (3.2)$$

其中  $\sigma_j(n)$  和  $\varepsilon_j(n)$  与模型 (2.2) 中的假设一致.

在给定一组样本大小  $\mathcal{V}_m$  时, 转换后的模型可以通过线性回归方法轻松地被估计. 也就是说,

$$\mathbf{Z}_{jN} = (1, \log N)^\top, \quad \mathbf{Z}_{j\mathcal{V}_m} = \{1_m, (\log n_1, \dots, \log n_m)^\top\},$$

$\mathbf{Z}_{j\mathcal{V}_m}^\top \mathbf{W}_j \mathbf{Z}_{j\mathcal{V}_m}$  对应的形式可简化为

$$\mathbf{H} = \begin{pmatrix} \sum_{i=1}^m \sigma_j^{-2}(n_i) & \sum_{i=1}^m \sigma_j^{-2}(n_i) \log n_i \\ \sum_{i=1}^m \sigma_j^{-2}(n_i) \log n_i & \sum_{i=1}^m \sigma_j^{-2}(n_i) (\log n_i)^2 \end{pmatrix}.$$

进而, 目标函数  $\text{AMSE}(\mathcal{V}_m)$  变为

$$\text{AMSE}(\mathcal{V}_m) = \sum_{j=1}^q \frac{\sum_{i=1}^m \sigma_j^{-2}(n_i) (\log n_i - \log N)^2}{\sigma_j^{-2}(N) \det(\mathbf{H})}.$$

为了找到该函数的最优设计, 只需要在初步研究中获得  $\sigma_j(n)$  的估计, 这也可以类似地拟合  $\log$  变换后的

$$\sigma^2(n) = cn^d.$$

这样做使得当  $m$  不是很大时, 即使利用穷尽搜索也可以很快地获得最优设计. 在下一节中, 我们的模拟结果表明该模型在诸多问题中均可很好地发挥作用.

## 4 数值分析

本节将针对 5 个典型的调节参数挑选问题, 利用模拟数据和实际数据来评估所提出 SUBEX 过程的性能.  $g(n, \theta)$  和  $h(n, \gamma)$  均考虑幂指数模型, 初步研究中设置  $\tilde{\mathcal{V}}_{m_0} = \{(1 + 0.5(i-1))n_0\}_{i=1}^3$ . 在所有的例子中, 全样本数据只生成一次, 然后保持不变. 每一次通过子采样和拟合外推模型实现模拟重复, 所有的模拟结果基于 100 次重复所得. 本文利用 R-3.4.3 实现数值模拟.

我们将与两种相关方法进行比较. 第一种方法是将本文所提出的外推算法与等距设计一起使用, 即在给定计算预算要求下, 设置  $\mathcal{V}_m^E$  为从  $n_0$  起等距选择样本大小的集合, 其中  $m = 5$ , 将该方法称为等距外推过程 (equally-spaced extrapolation, EQEX). 第二种是一个简单的子采样方法 (naive subsampling method, NASUB), 即在给定的时间预算内直接随机抽取一个子采样集合, 并基于该子集进行调节参数的挑选. 在总计算时间可承受的例子中, 我们还将本文的方法与基于全样本数据得到的“所谓”最优参数  $f(\mathcal{X})$  进行比较. 除了比较不同方法下估计的  $\lambda$  的均方误差和计算时间外, 我们还比较了基于不同方法所得到模型拟合的精度, 如线性回归的预测误差等.

**例 4.1** (基于阈值的协方差正则化) 文献 [5] 提出了基于阈值进行协方差矩阵估计的方法, 即

$$T_\lambda(\mathbf{S}) = [s_{ij}I(|s_{ij}| \geq \lambda)],$$

其中  $s_{ij}$  为样本协方差矩阵  $\mathbf{S}$  第  $(i, j)$  个元素. 文献 [5] 证明了当  $\lambda = c\sqrt{\log J/n}$  时,  $T_\lambda(\mathbf{S})$  在算子范数下是协方差矩阵  $\Sigma$  的相合估计, 其中  $c$  是一个充分大的常数,  $J$  是协方差矩阵的维数. 显然, 这个  $\lambda$  完美地满足我们的模型假设  $g(n, \theta) = an^b$ , 其中  $a = c\sqrt{\log J}$  和  $b = -0.5$ . 因此, 将这个例子作为一个完美情形来说明所提出的外推法如何实施. 此外, SUBEX 中的  $f(\cdot)$  采用文献 [5] 推荐的两折交叉验证.

我们将研究两类总体协方差矩阵:

- (i) 自回归协方差矩阵, 即  $\Sigma = (\rho^{|j_1-j_2|})_{J \times J}$ , 其中  $\rho = 0.8$ ;
- (ii) 带状协方差矩阵, 即  $\Sigma = (\rho^{|j_1-j_2|}I(|j_1-j_2| < \lfloor J/3 \rfloor))_{J \times J}$ , 其中  $\rho = 0.5$ .

模拟数据由一个多元正态分布  $N(0, \Sigma)$  生成, 全样本大小固定为  $N = 10^6$ , 维数为  $J = 200$ , 时间预算设置为不同  $n_0$  下的  $T = 20t(n_0)$ .

表 1 给出了例 4.1(i) 中方法 SUBEX、EQEX 和 NASUB 得到的估计  $\hat{\lambda}$  的偏差和方差、 $(T_\lambda(\mathbf{S}) - \Sigma)$  的  $L_1$  和 Frobenius 范数以及相应的计算时间. 通过比较  $\hat{\lambda}$  的偏差和方差, 我们发现基于 SUBEX 和 EQEX 所得到的  $\hat{\lambda}$  非常接近基于全样本挑选的结果, 这意味着  $\lambda(n)$  的趋势可以通过外推算法近似地拟合. 如我们所料, 直接的子抽样方法 NASUB 效果不佳.

本文提出的方法 SUBEX 在矩阵范数方面表现与基于全样本数据得到的几乎相同, 但是本文的方法所需的计算时间要少得多. 同时, SUBEX 的表现优于 EQEX, 因为前者通常产生较小的方差. 我们可以从图 2 更好地理解这一点, 该图展示了在矩阵范数的意义下, SUBEX 和 EQEX 两个方法与基于

表 1 例 4.1(i) 下的模拟结果. 表中比较了  $\hat{\lambda}$  的标准差 (sd) 和估计偏差 (bias),  $(T_{\lambda}(S) - \Sigma)$  的  $L_1$  范数和 Frobenius 范数, 以及计算时间; 括号内为相应的标准误. 表中第 3 和 4 列放大了 1,000 倍, 第 5 和 6 列放大了 100 倍

| $n_0$ | 方法    | sd( $\hat{\lambda}$ ) | bias( $\hat{\lambda}$ ) | $L_1$                   | Frobenius                | 计算时间 (分)                |
|-------|-------|-----------------------|-------------------------|-------------------------|--------------------------|-------------------------|
| 200   | SUBEX | 1.09                  | 0.39                    | 12.17 <sub>(1.72)</sub> | 15.32 <sub>(1.67)</sub>  | 11.19 <sub>(0.42)</sub> |
|       | EQEX  | 1.29                  | 0.42                    | 12.73 <sub>(2.25)</sub> | 15.80 <sub>(2.07)</sub>  | 9.69 <sub>(0.35)</sub>  |
|       | NASUB | 3.52                  | 45.53                   | 50.01 <sub>(3.63)</sub> | 140.13 <sub>(9.67)</sub> | 11.84 <sub>(0.96)</sub> |
| 600   | SUBEX | 1.08                  | 0.48                    | 12.36 <sub>(1.73)</sub> | 15.45 <sub>(1.60)</sub>  | 33.37 <sub>(1.72)</sub> |
|       | EQEX  | 1.23                  | 0.51                    | 12.76 <sub>(2.25)</sub> | 15.78 <sub>(1.95)</sub>  | 30.80 <sub>(1.05)</sub> |
|       | NASUB | 2.52                  | 24.23                   | 31.69 <sub>(2.23)</sub> | 78.74 <sub>(8.01)</sub>  | 47.98 <sub>(2.42)</sub> |
| 1,000 | SUBEX | 1.01                  | 0.38                    | 11.84 <sub>(1.47)</sub> | 15.12 <sub>(1.41)</sub>  | 68.12 <sub>(2.57)</sub> |
|       | EQEX  | 1.17                  | 0.26                    | 12.19 <sub>(1.57)</sub> | 15.47 <sub>(1.58)</sub>  | 64.38 <sub>(1.19)</sub> |
|       | NASUB | 2.05                  | 18.03                   | 26.85 <sub>(2.07)</sub> | 61.11 <sub>(6.05)</sub>  | 85.03 <sub>(4.32)</sub> |
|       | 全样本   |                       |                         | 11.17                   | 14.05                    | 469.01                  |

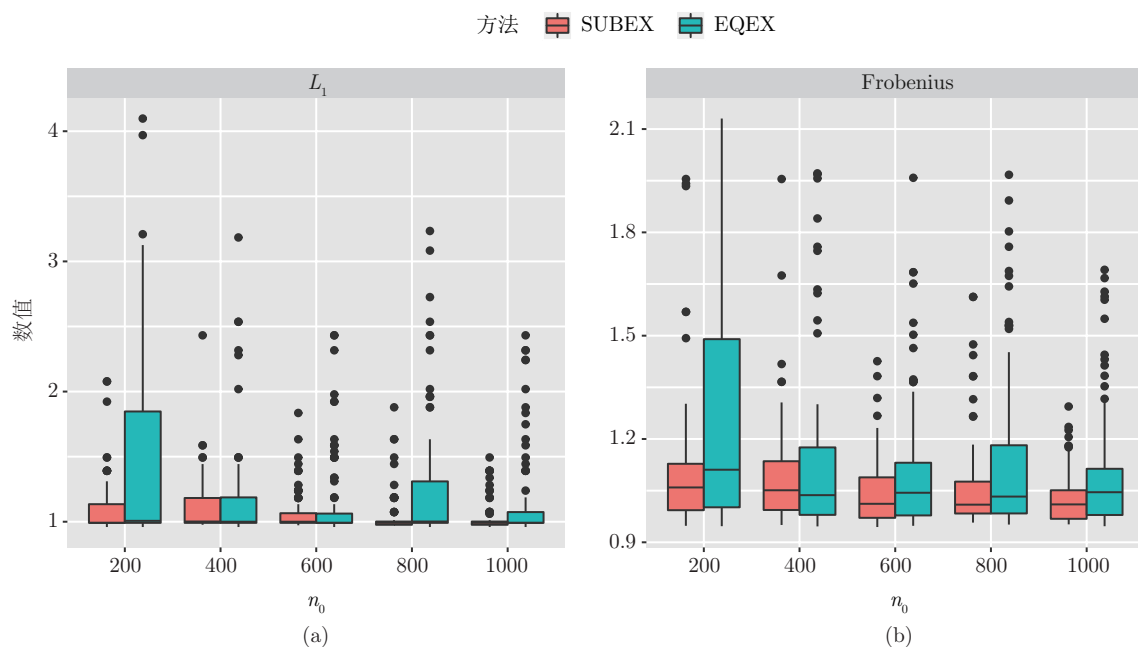


图 2 (网络版彩图) 例 4.1(ii) 下相对  $L_1$  和 Frobenius 范数的盒子图

全样本数据的方法的比值. 尽管两者的平均值都接近于 1, 但是 SUBEX 可以比 EQEX 更快地收敛. SUBEX 的优势在  $n_0$  较小时更加明显, 这与本文的定理 2.2 所示结果相吻合, 即 SUBEX 得到的估计设计  $\hat{v}_m$  可以获得比 EQEX 方法下更小的 AMSE.

**例 4.2** (基于 LASSO 的线性回归) 考虑 LASSO (least absolute shrinkage and selection operator), 它是线性模型中用于高维变量选择和估计的最常用的方法<sup>[3]</sup>. 该方法可以通过 R 包 “glmnet” 有效地实现, 并且调节参数  $\lambda$  可利用基于交叉核实准则的函数 “cv.glmnet()” 得到.

在这个例子中, 我们采用 BlogFeedback 数据集来评估所提出方法的性能, 该数据集可以从 UCI 机器学习数据库通过以下链接下载: <https://archive.ics.uci.edu/ml/datasets/BlogFeedback>. 该数据集

共包含 2010–2012 年 52,397 个观测, 目标是利用 281 个变量预测未来 24 小时内的评论数量. 我们将 2010–2011 年的数据作为训练集建立模型, 预测 2012 年 2 月 1 日 151 个测试集中的评论数.

图 3 展示 SUBEX、EQEX 和 NASUB 关于估计  $\hat{\lambda}$ 、预测误差及模型大小的比较. 我们发现 SUBEX 表现最好, 其次是 EQEX, 最后是 NASUB. 与例 4.1 所示结果类似, NASUB 会挑选一个较大的调节参数  $\lambda$ , 从而导致模型拟合不足. 随着  $n_0$  的增加, SUBEX 的表现趋向于最优的, 并且其变化一般比 EQEX 小很多.

**例 4.3 (基于 SVM 的分类)** 考虑用支持向量机 (support vector machine, SVM) 进行两类分类的问题, SVM 是机器学习中一个常用的分类方法, 参见文献 [22]. 给定  $N$  对训练集  $(\mathbf{x}_i, y_i)$ ,  $i = 1, \dots, N$ , 其中  $\mathbf{x}_i \in \mathbb{R}^p$  及  $y_i \in \{-1, 1\}$ , 计算 SVM 分类器等价于优化

$$\begin{aligned} \min_{\mathbf{w}, b, \mathbf{x}_i} \quad & \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i (\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \end{aligned}$$

其中  $C > 0$  为惩罚参数;  $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$  为核函数, 此处考虑常用的径向基函数  $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ ;  $\gamma > 0$  为核参数. 在此问题中, 我们的目标是识别两个可靠的调节参数  $(C, \gamma)$ . 最常用的策略是利用交叉核实的方法对调节参数  $C$  和  $\gamma$  进行“网格搜索”. 我们利用 R 包“e1071”实现 SVM 过程, 并使用函数“tune.svm()”的五折交叉核实挑选参数. 我们主要比较两个指标: AUC (area under the curve) 和分类正确率. 由于 SVM 算法的计算复杂度为  $O(N^2)$ , 因此考虑略大的时间预算  $T = 50t(n_0)$ , 其中  $n_0$  为 100、200 或 400.

考虑实际数据集 SUSY, 具体可参见文献 [23]. 该数据集可从 UCI 机器学习数据库下载, 链接为 <https://archive.ics.uci.edu/ml/datasets/SUSY>. SUSY 数据集的目标是利用数据集中 18 个运动学特征

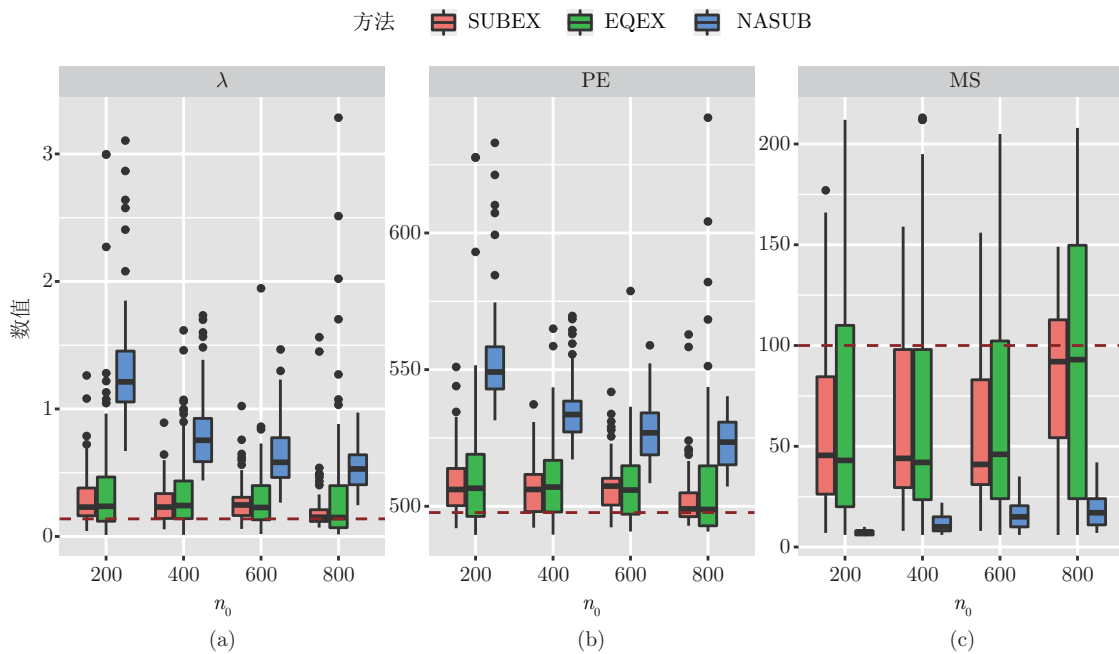


图 3 (网络版彩图) 例 4.2 中  $\hat{\lambda}$  的估计值、预测误差 (MS) 及模型大小 (MS) 的盒子图, 其中虚线为基于全样本数据搜索所得结果

来区分产生的超对称粒子和背景粒子. 我们将  $N = 40,000$  作为训练集, 其他的 5,000 个观测作为测试集.

如前两个例子所示, NASUB 方法并不具有竞争力, 因此, 我们在表 2 中仅比较 SUBEX 和 EQEX 的结果. 就 AUC 和分类正确率而言, SUBEX 的表现要优于 EQEX, 因为前者获得了更精确的分类, 同时在相似的计算时间下结果更加稳定. 我们发现即使在  $n_0$  较小的情形下, 基于 SUBEX 挑选的调节参数  $(C, \gamma)$  也会获得与所谓的最优调节参数 (基于全样本数据搜索所得) 相似的分类结果. 此外, 在表 2 的最后两列中, 我们列举了基于外推算法搜索与基于全样本数据搜索的相对时间, 可以看到与全样本数据搜索相比, SUBEX 在计算时间方面的优势非常明显.

**例 4.4 (联合图 LASSO)** 近年来, 研究者非常关注如何估计多类不同但相关的高维 Gauss 图模型. 众所周知的方法是文献 [24] 提出的联合图 LASSO (joint graphical LASSO, JGL) 的方法. 令  $K$  为类的数量, 其中  $K \geq 2$ ,  $\Theta_k$  是我们寻求的精度矩阵, 而  $S_k$  表示对应的经验协方差矩阵. 文献 [24] 建议通过最大化惩罚对数似然

$$L(\{\Theta\}) = \sum_{k=1}^K m_k [\log\{\det(\Theta_k)\} - \text{tr}(S_k \Theta_k)] - P(\{\Theta\})$$

去寻找  $\{\Theta\} = \{\Theta_1, \dots, \Theta_K\}$ . 特别地, 考虑损失函数

$$P(\{\Theta\}) = \lambda_1 \sum_{k=1}^K \sum_{i \neq j} |\theta_{ij}^k| + \lambda_2 \sum_{k < k'} \sum_{i, j} |\theta_{ij}^k - \theta_{ij}^{k'}|,$$

其中  $m_k$  为每类的样本大小,  $\theta_{ij}^k$  是精度矩阵  $\Theta_k$  第  $(i, j)$  个元素,  $\lambda_1$  和  $\lambda_2$  是两个非负的调节参数. 如果  $\lambda_1$  很大, 则 JGL 会导致稀疏估计  $\hat{\Theta}_1, \dots, \hat{\Theta}_K$ ; 如果  $\lambda_2$  很大, 则  $\hat{\Theta}_1, \dots, \hat{\Theta}_K$  中的很多元素都将在不同类中是相同的. 文献 [24] 建议最小化 Akaike 信息准则 (Akaike information criterion, AIC) 来挑选调节参数. R 包 “JGL” 可用来实现该方法.

此例采用文献 [24] 中的模拟设置, 即生成一个具有  $K = 3$  类和  $p = 200$  个特征的网络结构, 其中该网络是由幂律度分布生成的 10 个相等大小的子网络形成. 全样本大小固定为  $N = 50,000$ , 时间预算设置为  $T = 20t(n_0)$ . 多个指标用于评估方法的性能, 包括平均 Frobenius 损失和误报率 [25]. 表 3 总结了 SUBEX 基于 100 次模拟的结果. 从标准差和偏差列中可以看出, 随着  $n_0$  的增加, 基于 SUBEX 方法估计的  $\lambda_1$  和  $\lambda_2$  趋近于所谓最优值, 这是由于随着  $n_0$  的增加可得到更可靠的  $\lambda(n)$  的趋势. 从误报率的比较也可以得出类似的结论. 此外, SUBEX 得到的 Frobenius 损失与基于全样本搜索得到的损失相当.

**例 4.5 (基于核逻辑回归的分类)** 核逻辑回归 (kernel logistic regression, KLR) [26] 是如 SVM 一

**表 2** 例 4.3 中 SUBEX 和 EQEX 关于平均 AUC、分类正确率、相对计算时间及相应的标准误比较; 基于全样本搜索得到的调节参数对应结果列在最后一行

| $n_0$ | AUC (%)                 |                         | 正确率 (%)                 |                         | 相对时间 ( $10^{-4}$ )      |                         |
|-------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
|       | SUBEX                   | EQEX                    | SUBEX                   | EQEX                    | SUBEX                   | EQEX                    |
| 100   | 86.12 <sub>(0.22)</sub> | 78.78 <sub>(1.17)</sub> | 78.89 <sub>(0.38)</sub> | 69.64 <sub>(1.15)</sub> | 5.33 <sub>(0.09)</sub>  | 6.97 <sub>(0.18)</sub>  |
| 200   | 86.77 <sub>(0.09)</sub> | 85.21 <sub>(0.47)</sub> | 79.77 <sub>(0.07)</sub> | 77.69 <sub>(0.61)</sub> | 17.93 <sub>(0.36)</sub> | 18.07 <sub>(0.46)</sub> |
| 400   | 86.95 <sub>(0.06)</sub> | 86.06 <sub>(0.21)</sub> | 79.91 <sub>(0.04)</sub> | 78.62 <sub>(0.44)</sub> | 71.31 <sub>(2.20)</sub> | 68.22 <sub>(2.26)</sub> |
| 全样本   | 87.32                   |                         | 80.06                   |                         | 42 (小时)                 |                         |

表 3 例 4.4 中 SUBEX 的模拟结果. 表中列了  $\hat{\lambda}_1$  和  $\hat{\lambda}_2$  估计的标准差和偏差、平均的 Frobenius 损失 (Frobenius loss, FL) 以及平均误报率 (false positive, FP). 表中第 2 和 3 列放大了 1,000 倍, 第 4 和 5 列放大了 100 倍

| $n_0$ | $\text{sd}(\hat{\lambda}_1)$ | $ \text{bias}(\hat{\lambda}_1) $ | $\text{sd}(\hat{\lambda}_2)$ | $ \text{bias}(\hat{\lambda}_2) $ | FL   | FP(%) |
|-------|------------------------------|----------------------------------|------------------------------|----------------------------------|------|-------|
| 200   | 0.74                         | 3.53                             | 0.44                         | 4.80                             | 0.67 | 3.50  |
| 400   | 1.28                         | 2.97                             | 0.30                         | 2.71                             | 0.67 | 2.69  |
| 600   | 0.04                         | 2.75                             | 0.17                         | 1.76                             | 0.67 | 1.23  |
| 800   | 0.87                         | 1.72                             | 0.13                         | 1.21                             | 0.67 | 0.63  |
| 1,000 | 0.51                         | 1.79                             | 0.12                         | 0.85                             | 0.67 | 0.55  |
| 全样本   |                              |                                  |                              |                                  | 0.68 | 0.13  |

表 4 例 4.5 中 SUBEX 与基于全样本搜索关于分类正确率、相对计算时间及对应的标准误的比较

| 数据集                 | $p$ | $N$    | 正确率<br>(全样本) | 正确率<br>(SUBEX)         | 相对时间                   |
|---------------------|-----|--------|--------------|------------------------|------------------------|
| Spambase            | 57  | 3,680  | 0.94         | 0.93 <sub>(0.02)</sub> | 0.38 <sub>(0.02)</sub> |
| Wilt                | 5   | 4,339  | 0.73         | 0.74 <sub>(0.03)</sub> | 0.34 <sub>(0.02)</sub> |
| Occupancy detection | 5   | 8,143  | 0.94         | 0.95 <sub>(0.01)</sub> | 0.23 <sub>(0.06)</sub> |
| Banking marketing   | 16  | 36,168 | 0.94         | 0.94 <sub>(0.01)</sub> | 0.13 <sub>(0.07)</sub> |

样强有力的分类方法. 利用与例 4.3 中相同的定义, 拟合 KLR 分类问题等价于解决如下的优化问题:

$$\min_{\mathbf{w}, b, \mathbf{x}_i} \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_{i=1}^N \log[1 + \exp\{-y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b)\}],$$

其中  $C$  是惩罚参数,  $\phi(\mathbf{x}_i)$  是包含参数  $\gamma$  的核函数. R 包 “CVST” 可利用交叉核实方法识别调节参数  $(C, \gamma)$ . 由于 KLR 的计算复杂度通常为  $O(N^3)$ , 以下实验中的时间预算均设置为  $T = 100t(n_0)$ , 其中  $n_0 = 50$ .

为了评估 SUBEX 在实际应用中的性能, 将其应用于 UCI 机器学习数据库中的 4 个数据集 (即 Spambase、Wilt、Occupancy detection 和 Banking marketing). 表 4 介绍了每个数据集的样本大小、维度以及基于 SUBEX 和全样本搜索得到的分类正确率. 相比于全样本搜索, SUBEX 可以获得相似的分类正确率, 但是花费更少的时间. 我们也发现随着样本数量的增加, SUBEX 的优势变得更加明显.

本节中所有结果均表明, 只要最优的调节参数依赖于样本量大小, 本文提出的 SUBEX 始终能够为调节参数提供可靠的选取. 在大部分情形下, 本文的方法可以很好地近似基于全样本数据选择的所谓最优调节参数, 但是本文的方法花费的计算时间却少得多. 即使没有最优设计, 我们仍可以使用基于等距设计的外推模型来获取关于  $\lambda(n)$  趋势的某些模式.

## 5 结论

基于数据驱动进行调节参数选取在理论分析和实际应用中都非常重要. 本文在统一框架下开发了 SUBEX 过程用于海量数据分析中调节参数的选取. 通过优化插值相关估计后的 AMSE 准则来解决子样本设计问题. 理论分析和数值论证均显示了所提出方法的有效性, 表明 SUBEX 可以作为实践中的一个有用工具. 在最优调节参数与样本量之间的函数形式可以通过一个参数模型来近似的假设下, 所

提出的方法充分满足了关于调节参数选取的需求. 我们认为这种假设并不苛刻, 因为许多应用已经证实最优调节参数与样本量之间是一个单调且平滑的函数.

我们以两个讨论结束本文. 首先, 我们进行模型拟合时没有考虑调节参数之间的相关性. 因此, 如何将相关结构融合到估计过程中将是一个有意义的问题. 其次, 假设  $q$  是固定的. 但是对于某些学习方法 (如基于神经网络的方法),  $q$  可能会相当大, 在高维 ( $q \rightarrow \infty$ , 随着  $N \rightarrow \infty$ ) 情形下所提出方法的理论和数值研究是另一个值得研究的问题.

**致谢** 作者感谢主编、副主编及审稿人的建设性意见. 该工作为任好洁在宾夕法尼亚州立大学做博士后期间完成的.

## 参考文献

- 1 Hardle W, Marron J S. Optimal bandwidth selection in nonparametric regression function estimation. *Ann Statist*, 1985, 13: 1465–1481
- 2 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc*, 2001, 96: 1348–1360
- 3 Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Stat Methodol*, 1996, 58: 267–288
- 4 Hall P, Park B U, Samworth R J. Choice of neighbor order in nearest-neighbor classification. *Ann Statist*, 2008, 36: 2135–2152
- 5 Bickel P J, Levina E. Covariance regularization by thresholding. *Ann Statist*, 2008, 36: 2577–2604
- 6 Battey H, Fan J, Liu H, et al. Distributed testing and estimation under sparse high dimensional models. *Ann Statist*, 2018, 46: 1352–1382
- 7 Kleiner A, Talwalkar A, Sarkar P, et al. A scalable bootstrap for massive data. *J R Stat Soc Ser B Stat Methodol*, 2014, 76: 795–816
- 8 Ma P, Mahoney M W, Yu B. A statistical perspective on algorithmic leveraging. *J Mach Learn Res*, 2015, 16: 861–911
- 9 Seber G A F, Wild C J. *Nonlinear Regression*. New York: John Wiley & Sons, 1989
- 10 Fan J, Gijbels I. *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability, vol. 66. London: Chapman Hall, 1996
- 11 Box G E P, Lucas H L. Design of experiments in non-linear situations. *Biometrika*, 1959, 46: 77–90
- 12 Ford I, Titterton D M, Kitsos C P. Recent advances in nonlinear experimental design. *Technometrics*, 1989, 31: 49–60
- 13 Carroll R J, Ruppert D. Robust estimation in heteroscedastic linear models. *Ann Statist*, 1982, 10: 429–441
- 14 Jennrich R I. Asymptotic properties of non-linear least squares estimators. *Ann Math Statist*, 1969, 40: 633–643
- 15 Wu C F. Asymptotic theory of nonlinear least squares estimation. *Ann Statist*, 1981, 9: 501–513
- 16 Beale E M L. Confidence regions in non-linear estimation. *J R Stat Soc Ser B Stat Methodol*, 1960, 22: 41–76
- 17 Johansen S. Some topics in regression. *Scand J Statist*, 1983, 10: 161–194
- 18 Shah R D, Samworth R J. Variable selection with error control: Another look at stability selection. *J R Stat Soc Ser B Stat Methodol*, 2013, 75: 55–80
- 19 Sun W, Sampaio R J B, Candido M A B. Proximal point algorithm for minimization of DC function. *J Comput Math*, 2003, 21: 451–462
- 20 Le Thi H A, Pham Dinh T. DC programming and DCA: Thirty years of developments. *Math Program*, 2018, 169: 5–68
- 21 Zou H. The adaptive Lasso and its oracle properties. *J Amer Statist Assoc*, 2006, 101: 1418–1429
- 22 Cortes C, Vapnik V. Support-vector networks. *Mach Learn*, 1995, 20: 273–297
- 23 Baldi P, Sadowski P, Whiteson D. Searching for exotic particles in high-energy physics with deep learning. *Nat Commun*, 2014, 5: 4308
- 24 Danaher P, Wang P, Witten D M. The joint graphical lasso for inverse covariance estimation across multiple classes. *J R Stat Soc Ser B Stat Methodol*, 2014, 76: 373–397
- 25 Guo J, Levina E, Michailidis G, et al. Joint estimation of multiple graphical models. *Biometrika*, 2011, 98: 1–15
- 26 Zhu J, Hastie T. Kernel logistic regression and the import vector machine. *J Comput Graph Statist*, 2005, 14: 185–205
- 27 Amemiya T. Non-linear regression models. In: Griliches Z, Intriligator M D, eds. *Handbook of Econometrics*, vol. 1. Amsterdam: Elsevier, 1981, 333–389
- 28 van der Vaart A W. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000

## 附录 A

尽管本文考虑了多响应模型 (2.2), 但实际上我们对  $g_j(\cdot; \cdot)$  ( $j = 1, \dots, q$ ) 逐个进行估计. 因此, 为了便于阐述, 除非另有说明, 否则在以下引理证明中, 我们舍去了  $g_j(\cdot; \cdot)$  和相关符号对  $j$  的依赖性. 对于  $p \times p$  矩阵  $\mathbf{A}$ , 令  $\|\mathbf{A}\|$  为其频谱范数 (绝对值中的最大特征值). 令  $A_p \approx B_p$  表示  $A_p$  与  $B_p$  渐近等价, 即在某种意义上存在常数  $C > 1$  使得  $B_p/C \leq A_p \leq B_p C$  依概率趋向于 1. 为方便起见, 在附录中记  $n_0$  为  $n$ , 这应不会引起混淆.

在证明定理之前, 首先阐明两个引理.

**引理 A.1** 设假设 2.2-2.5 成立, 对  $i = 1, \dots, m$ , 如果随着  $n \rightarrow \infty$ ,  $w_i \sigma^2(n_i)$  收敛至某个常数  $c_i \in (0, \infty)$ , 则给定  $w_i$ , 有

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + O_p(a_n^{-1/2}).$$

**证明** 因为证明过程对  $q > 1$  的情形类似, 我们将主要集中在  $q = 1$  的情形. 给定假设 2.2 和 2.3, 文献 [14] 证明了最小二乘估计存在, 即对于给定常数  $w_i > 0$ ,  $\hat{\boldsymbol{\theta}}$  可通过最小化下式得到:

$$S_n(\boldsymbol{\theta}) = \sum_{i=1}^m w_i \{\lambda(n_i) - g(n_i; \boldsymbol{\theta})\}^2. \quad (\text{A.1})$$

接下来, 先证明  $\hat{\boldsymbol{\theta}}$  是  $\boldsymbol{\theta}^*$  的一个弱相合估计. 对于一个充分大的  $n$ , 在假设 2.4 下, (A.1) 中给出的目标函数  $S_n(\boldsymbol{\theta})$  是一个关于  $\boldsymbol{\theta}$  的严格凸函数. 因此, 只要可以证明  $S_n(\boldsymbol{\theta})$  有一个  $\sqrt{a_n}$  相合的局部最小值, 它就一定是  $\sqrt{a_n}$  相合的全局最小. 进而可以直接得到  $\hat{\boldsymbol{\theta}}$  是  $\sqrt{a_n}$  相合的.

为此, 证明存在一个  $\sqrt{a_n}$  相合的局部最小值, 也就是证明对于任意小的  $\epsilon > 0$ , 存在一个充分大的常数  $C$  使得

$$\liminf_n \Pr \left\{ \inf_{\mathbf{u} \in \mathbb{R}^p: \|\mathbf{u}\|=C} S_n(\boldsymbol{\theta}^* + a_n^{-1/2} \mathbf{u}) > S_n(\boldsymbol{\theta}^*) \right\} > 1 - \epsilon. \quad (\text{A.2})$$

这说明以至少  $1 - \epsilon$  的概率在球  $\{\boldsymbol{\theta}^* + a_n^{-1/2} \mathbf{u} : \|\mathbf{u}\| = C\}$  中存在一个局部最小值, 因此可以得到  $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = O_p(a_n^{-1/2})$ . 根据  $S_n(\boldsymbol{\theta})$  的定义, 通过简单的代数运算可以得到

$$\begin{aligned} D_n &= S_n(\boldsymbol{\theta}^* + a_n^{-1/2} \mathbf{u}) - S_n(\boldsymbol{\theta}^*) \\ &= 2 \sum_{i=1}^m w_i \varepsilon_i \{g(n_i; \boldsymbol{\theta}^*) - g(n_i; \boldsymbol{\theta}^* + a_n^{-1/2} \mathbf{u})\} + \sum_{i=1}^m w_i \{g(n_i; \boldsymbol{\theta}^* + a_n^{-1/2} \mathbf{u}) - g(n_i; \boldsymbol{\theta}^*)\}^2. \end{aligned}$$

对  $g(n_i; \boldsymbol{\theta}^* + a_n^{-1/2} \mathbf{u})$  作 Taylor 展开, 可以得到

$$\begin{aligned} D_n &= -2 \sum_{i=1}^m w_i \varepsilon_i a_n^{-1/2} \mathbf{u}^\top g'(n_i; \boldsymbol{\theta}^*) - 2 \sum_{i=1}^m w_i \varepsilon_i a_n^{-1/2} \mathbf{u}^\top \Delta(n_i; \boldsymbol{\xi}) \\ &\quad + \sum_{i=1}^m w_i a_n^{-1} \{\mathbf{u}^\top g'(n_i; \boldsymbol{\theta}^*)\}^2 + \sum_{i=1}^m w_i a_n^{-1} \{\mathbf{u}^\top \Delta(n_i; \boldsymbol{\xi})\}^2 \\ &\quad + 2 \sum_{i=1}^m w_i a_n^{-1} \{\mathbf{u}^\top g'(n_i; \boldsymbol{\theta}^*)\} \{\mathbf{u}^\top \Delta(n_i; \boldsymbol{\xi})\} \\ &=: R_1 + R_2 + R_3 + R_4 + R_5, \end{aligned}$$

其中  $\boldsymbol{\xi}$  在  $\boldsymbol{\theta}^*$  与  $\boldsymbol{\theta}^* + a_n^{-1/2} \mathbf{u}$  之间,  $\Delta(n_i; \boldsymbol{\xi}) = g'(n_i; \boldsymbol{\xi}) - g'(n_i; \boldsymbol{\theta}^*)$ .

记  $0 < \underline{c} < c_1 < \cdots < c_m < \bar{c}$ . 注意到在给定  $\mathcal{X}$  的条件测度下,  $\varepsilon_i$  相互独立且均值为 0. 因此, 对于充分大的  $n$  及一致在  $\mathcal{X}$  上, 有

$$\mathbb{E}(R_1 | \mathcal{X}) = 0, \quad \text{var}(R_1 | \mathcal{X}) = \sum_{i=1}^m w_i^2 \sigma^2(n_i) a_n^{-1} \{\mathbf{u}^\top g'(n_i; \boldsymbol{\theta}^*)\}^2 \leq 2\bar{c}C^2 \text{eig}_p(\boldsymbol{\Omega}),$$

其中  $\text{eig}_1(\boldsymbol{\Omega}) < \cdots < \text{eig}_p(\boldsymbol{\Omega})$  为  $\boldsymbol{\Omega}$  的特征值. 进而可以得到  $R_1 = O_p(C)$ . 利用假设 2.5 和类似的讨论, 可以得到  $R_2 = O_p(C^2 a_n^{-1} M_{n_m})$ , 其中不失一般性地假设  $M_{n_m} = \max_{1 \leq i \leq m} M_{n_i}$ .

类似地, 对充分大的  $n$ , 也可以证明

$$R_3 \geq 0.5\underline{c}C^2 \text{eig}_1^2(\boldsymbol{\Omega})$$

及  $R_4 = O_p(M_{n_m}^2 a_n^{-2})$ . 通过选择一个充分大的  $C$ , 在  $\|\mathbf{u}\| = C$  下,  $R_3$  一致地大于  $R_1$ . 此外, 利用条件  $M_{n_m}/a_n \rightarrow 0$ , 可以验证  $R_2 = o_p(R_3)$  和  $R_4 = o_p(R_3)$ . 进一步, 利用 Cauchy 不等式, 可得

$$R_5 = O_p(a_n^{-1} M_{n_m}) = o_p(R_3).$$

因此, 只要常数  $C$  足够大,  $R_3$  项就可以以任意大的概率主导  $D_n$  中的其他四项, 这意味着不等式 (A.2) 成立, 即完成证明.  $\square$

**引理 A.2** 如果假设 2.1、2.6 和 2.7 成立, 则

$$\tilde{\gamma} = \gamma^* + o_p(1).$$

**证明** 定义  $Q_n(\gamma) = b_n^{-1} \sum_{i=1}^{m_0} \{\tilde{\sigma}^2(n_i) - h(n_i; \gamma)\}^2$ . 文献 [27] 给出了  $\tilde{\gamma}$  相合性证明必要部分的一个有用框架. 核心就是证明  $\gamma^*$  可唯一地最小化  $Q_n(\gamma)$  的极限 (依概率)  $\text{plim}_n Q_n(\gamma)$ . 在这种情形下, 如果  $n$  充分大, 使得  $Q_n(\gamma)$  与  $\text{plim}_n Q_n(\gamma)$  很接近, 那么使得  $Q_n(\gamma)$  达到最小的  $\tilde{\gamma}$  将会非常接近使得  $\text{plim}_n Q_n(\gamma)$  达到最小的  $\gamma^*$ , 参见文献 [28, 定理 5.9] 及其注释. 现在

$$\begin{aligned} Q_n(\gamma) &= b_n^{-1} \sum_{i=1}^{m_0} \eta_i^2 + b_n^{-1} \sum_{i=1}^{m_0} \eta_i \{h(\tilde{n}_i; \gamma) - h(\tilde{n}_i; \gamma^*)\} + b_n^{-1} \sum_{i=1}^{m_0} \{h(\tilde{n}_i; \gamma) - h(\tilde{n}_i; \gamma^*)\}^2 \\ &=: R_1 + R_2 + R_3, \end{aligned}$$

其中  $\eta_i = \tilde{\sigma}^2(n_i) - h(n_i; \gamma^*)$ . 假设 2.7 使得  $R_3 \rightarrow D(\gamma, \gamma^*)$ . 在给定  $\mathcal{X}$  的条件测度下,  $\eta_i$  相互独立, 且  $\mathbb{E}(\eta_i | \mathcal{X}) = 0$  及  $\text{var}(\eta_i | \mathcal{X}) = \sigma^4(n_i) \{\mathbb{E}(\varepsilon_i^4) + 1\}/2$ . 根据假设 2.1, 利用 Cauchy 不等式可以得到  $R_1 = o_p(1)$  和  $R_2 = o_p(1)$ . 利用假设 2.7(ii) 可得到  $D(\gamma, \gamma^*)$  在  $\gamma^*$  处有唯一的最小值, 进而, 可以得到  $\text{plim}_n Q_n(\gamma)$  在  $\gamma^*$  处有唯一的最小值. 证毕.  $\square$

**定理 2.1 的证明** 利用引理 A.1 和 A.2, 得到  $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + O_p(a_n^{-1/2})$ , 其中估计值  $\hat{\boldsymbol{\theta}}$  通过利用  $1/h(n_i; \tilde{\gamma})$  作为  $w_{n_i}$  得到. 在假设 2.5 和条件  $M_N/a_n \rightarrow 0$  下,  $g(N; \hat{\boldsymbol{\theta}})$  在  $\boldsymbol{\theta}^*$  作 Taylor 展开使得

$$|g(N; \hat{\boldsymbol{\theta}}) - g(N; \boldsymbol{\theta}^*)| = O_p(|g'(N; \boldsymbol{\theta}^*)| a_n^{-1/2}).$$

证毕.  $\square$

在证明定理 2.2 之前, 首先证明一个引理. 记

$$\mathbf{H}(\mathbf{Z}_{V_m}, \mathbf{W}) = (\mathbf{Z}_{V_m}^\top \mathbf{W} \mathbf{Z}_{V_m})^{-1}, \quad \mathbf{J}(\mathbf{W}, \boldsymbol{\Sigma}) = \mathbf{H}(\mathbf{Z}_{V_m}, \mathbf{W}) \{\mathbf{H}(\mathbf{Z}_{V_m}, \mathbf{W} \boldsymbol{\Sigma} \mathbf{W})\}^{-1} \mathbf{H}(\mathbf{Z}_{V_m}, \mathbf{W}).$$

**引理 A.3** 假设定理 2.1 的条件都成立, 则有如下结果:

$$\frac{\tilde{\mathbf{Z}}_N^\top \mathbf{H}(\tilde{\mathbf{Z}}_N, \mathbf{W}) \tilde{\mathbf{Z}}_N - \mathbf{Z}_N^\top \mathbf{J}(\mathbf{W}, \Sigma) \mathbf{Z}_N}{\mathbf{Z}_N^\top \mathbf{J}(\mathbf{W}, \Sigma) \mathbf{Z}_N} = o_p(1), \quad \text{在 } \mathcal{V}_m \text{ 上一致成立.}$$

**证明** 这个表达式小于

$$\begin{aligned} & \frac{|(\tilde{\mathbf{Z}}_N - \mathbf{Z}_N)^\top \mathbf{J}(\mathbf{W}, \Sigma)(\tilde{\mathbf{Z}}_N - \mathbf{Z}_N)| + |\tilde{\mathbf{Z}}_N^\top \{\mathbf{J}(\mathbf{W}, \Sigma) - \mathbf{H}(\tilde{\mathbf{Z}}_N, \mathbf{W})\} \tilde{\mathbf{Z}}_N|}{\mathbf{Z}_N^\top \mathbf{J}(\mathbf{W}, \Sigma) \mathbf{Z}_N} \\ & \leq \frac{\|\mathbf{Z}_N - \tilde{\mathbf{Z}}_N\|^2 \text{eig}_p(\mathbf{J}(\mathbf{W}, \Sigma))}{\mathbf{Z}_N^\top \mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma) \mathbf{Z}_N} R_3 + \frac{\|\tilde{\mathbf{Z}}_N\|^2 \|\mathbf{J}(\mathbf{W}, \Sigma) - \mathbf{H}(\tilde{\mathbf{Z}}_N, \mathbf{W})\|}{\mathbf{Z}_N^\top \mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma) \mathbf{Z}_N} R_3 \\ & = R_1 R_3 + R_2 R_3, \end{aligned}$$

其中

$$R_3 = \frac{\mathbf{Z}_N^\top \mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma) \mathbf{Z}_N}{\mathbf{Z}_N^\top \mathbf{J}(\mathbf{W}, \Sigma) \mathbf{Z}_N}.$$

首先处理  $R_1 R_3$  部分. 根据假设 2.4, 可得  $a_n^{-1} \mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma) \rightarrow \Omega$ , 进而有

$$\text{eig}_1(\mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma)) \approx a_n, \quad \text{eig}_p(\mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma)) \approx a_n.$$

利用假设 2.4 和引理 A.2, 有

$$\begin{aligned} & \|\mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma) - \mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \mathbf{W} \Sigma \mathbf{W})\| \\ & = \left\| \sum_i w_{n_i} \mathbf{Z}_{n_i} \mathbf{Z}_{n_i}^\top - \sum_i w_{n_i}^2 \sigma^2(n_i) \mathbf{Z}_{n_i} \mathbf{Z}_{n_i}^\top \right\| \\ & = o_p(a_n), \end{aligned}$$

以及相应地,

$$\|\mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma) - \mathbf{J}(\mathbf{W}, \Sigma)\| = o_p(a_n). \quad (\text{A.3})$$

因此,

$$|R_3 - 1| = \frac{o_p(a_n) \|\mathbf{Z}_N\|^2}{O_p\{\|\mathbf{Z}_N\|^2 \text{eig}_1(\mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma))\}} = o_p(1).$$

利用假设 2.5 和引理 A.1, 得到

$$\frac{\|\mathbf{Z}_N - \tilde{\mathbf{Z}}_N\|}{\|\mathbf{Z}_N\|} = O_p(M_N a_n^{-1}) = o_p(1).$$

注意  $R_1$  的分母的下界为  $\|\mathbf{Z}_N\|^2 \text{eig}_1(\mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma))$ . 综合上述式子, 得到  $R_1 R_3 = o_p(1)$  在  $\mathcal{V}_m$  上一致成立.

现在处理  $R_2$ . 根据上述的讨论, 有

$$\|\tilde{\mathbf{Z}}_N\| = \|\mathbf{Z}_N\| \{1 + o_p(1)\}.$$

同时, 利用  $\|\tilde{\mathbf{Z}}_{\mathcal{V}_m} - \mathbf{Z}_{\mathcal{V}_m}\| = O_p(a_n^{-1} M_{n_m})$  和  $\|\mathbf{W} \Sigma\| = o_p(1)$  这两个事实, 可以得到

$$\|\mathbf{H}(\mathbf{Z}_{\mathcal{V}_m}, \Sigma) - \mathbf{H}(\tilde{\mathbf{Z}}_{\mathcal{V}_m}, \mathbf{W})\| = o_p(a_n),$$

进而利用 (A.3) 得到结论

$$\|\mathbf{H}(\tilde{\mathbf{Z}}_{\mathcal{V}_m}, \mathbf{W}) - \mathbf{J}(\mathbf{W}, \boldsymbol{\Sigma})\| = o_p(a_n).$$

因此,  $R_2 = o_p(1)$  在  $\mathcal{V}_m$  上一致成立, 由此引理结论成立.  $\square$

**定理 2.2 的证明** 对任意给定的  $m$ , 令

$$\mathcal{V}_m^* = \arg \min_{\mathcal{V}_m \in \mathbb{F}} \text{AMSE}(\mathcal{V}_m), \quad \widehat{\mathcal{V}}_m = \arg \min_{\mathcal{V}_m \in \mathbb{F}} \widetilde{\text{AMSE}}(\mathcal{V}_m),$$

其中  $\mathbb{F}$  表示  $\mathcal{V}_m$  的可行域. 我们将证明

$$\frac{\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m)}{\text{AMSE}(\mathcal{V}_m^*)} \xrightarrow{p} 1. \quad (\text{A.4})$$

根据引理 A.3, 有

$$\sup_{\mathcal{V}_m \in \mathbb{F}} \left| \frac{\widetilde{\text{AMSE}}(\mathcal{V}_m)}{\text{AMSE}(\mathcal{V}_m)} - 1 \right| \xrightarrow{p} 0. \quad (\text{A.5})$$

因为当  $\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m) \geq \text{AMSE}(\mathcal{V}_m^*)$  时,

$$0 \leq \frac{\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m) - \text{AMSE}(\mathcal{V}_m^*)}{\text{AMSE}(\mathcal{V}_m^*)} \leq \frac{\widetilde{\text{AMSE}}(\mathcal{V}_m^*) - \text{AMSE}(\mathcal{V}_m^*)}{\text{AMSE}(\mathcal{V}_m^*)} = o_p(1);$$

当  $\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m) \leq \text{AMSE}(\mathcal{V}_m^*)$  时,

$$\begin{aligned} 0 &\leq \frac{\text{AMSE}(\mathcal{V}_m^*) - \widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m)}{\text{AMSE}(\mathcal{V}_m^*)} \leq \frac{\text{AMSE}(\widehat{\mathcal{V}}_m) - \widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m)}{\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m)} \\ &\leq \frac{\text{AMSE}(\widehat{\mathcal{V}}_m) - \widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m)}{\text{AMSE}(\widehat{\mathcal{V}}_m)} \frac{\text{AMSE}(\widehat{\mathcal{V}}_m)}{\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_m)} = o_p(1), \end{aligned}$$

所以, 我们可以说明 (A.4) 成立.

最后, 我们得到

$$\begin{aligned} 0 &\leq \frac{\text{AMSE}(\widehat{\mathcal{V}}_{\widehat{m}})}{\text{AMSE}(\mathcal{V}_{m^*}^*)} - 1 \\ &= \frac{\text{AMSE}(\widehat{\mathcal{V}}_{\widehat{m}}) - \widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_{\widehat{m}})}{\text{AMSE}(\mathcal{V}_{m^*}^*)} + \frac{\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_{\widehat{m}}) - \widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_{m^*}^*)}{\text{AMSE}(\mathcal{V}_{m^*}^*)} + \frac{\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_{m^*}^*) - \text{AMSE}(\mathcal{V}_{m^*}^*)}{\text{AMSE}(\mathcal{V}_{m^*}^*)} \\ &=: C_1 + C_2 + C_3, \end{aligned}$$

其中根据定义  $C_2 \leq 0$  和 (A.4) 获得  $C_3 \xrightarrow{p} 0$ . 同时, 再次利用 (A.4) 和 (A.5), 有

$$\begin{aligned} C_1 &= \frac{\text{AMSE}(\widehat{\mathcal{V}}_{\widehat{m}}) - \widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_{\widehat{m}})}{\widetilde{\text{AMSE}}(\mathcal{V}_{m^*}^*)} \frac{\widetilde{\text{AMSE}}(\mathcal{V}_{m^*}^*)}{\text{AMSE}(\mathcal{V}_{m^*}^*)} \\ &\leq \frac{\text{AMSE}(\widehat{\mathcal{V}}_{\widehat{m}}) - \widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_{\widehat{m}})}{\widetilde{\text{AMSE}}(\widehat{\mathcal{V}}_{\widehat{m}})} \frac{\widetilde{\text{AMSE}}(\mathcal{V}_{m^*}^*)}{\text{AMSE}(\mathcal{V}_{m^*}^*)} \\ &\xrightarrow{p} 0, \end{aligned}$$

由此可以得出定理成立.  $\square$

## Extrapolation-based tuning parameters selection in massive data analysis

Haojie Ren, Changliang Zou & Runze Li

**Abstract** Many statistical modeling procedures involve one or more tuning parameters to control the model complexity. These tuning parameters can be the bandwidth in the kernel smoothing method in the nonparametric regression and density estimation or be the regularization parameter in the regularization method for feature selection in the high dimensional modeling. Tuning parameter selection plays critical roles in the statistical modeling and machine learning. For the massive data analysis, commonly-used methods such as grid-point search with information criteria become prohibitively costly in computation. Their feasibility is questionable even with modern parallel computing platforms. This paper aims to develop a fast algorithm to efficiently approximate the best tuning parameters. The algorithm entails (a) assuming a parametric model to describe the trend between the best tuning parameters and sample sizes, (b) establishing the trend via fitting the model with subsampling data, and (c) extrapolating this trend to the case of huge sample size. To determine the subsampling sample sizes to be taken, we derive optimal designs for settings that allow a constraint on the budget of total computational cost. We show that the proposed designs possess an asymptotic optimality property. Our numerical studies demonstrate that with a simple two-parameter polynomial model, the proposed algorithm performs almost equivalently to the procedure using the full data set in several different statistical settings, while it has a significant reduction in computing time and storage.

**Keywords** asymptotic, extrapolation, nonlinear models, optimal design, prediction error, regularized methods

MSC(2020) 62J02, 62J07

doi: 10.1360/SCM-2020-0622