

doi: 10.19920/j.cnki.jmsc.2023.12.003

试验设计及参数优化的 LS-SVR 显著性因子筛选^①

崔庆安¹, 崔楠^{2*}

(1. 上海海事大学经济管理学院, 上海 201306; 2. 武汉大学经济与管理学院, 武汉 430072)

摘要: 针对最小二乘支持向量回归机 (LS-SVR) 应用于试验设计建模及参数优化而产生的可解释性差、难以识别显著性影响因子等不足, 提出一种适用于 LS-SVR 的拟合不足检验及显著性因子筛选方法. 首先在重复性试验设计条件下, 将 LS-SVR 拟合模型的“残差平方和”分解为“拟合不足平方和”与“纯误差平方和”; 进而给出了“拟合不足均方”与“纯误差均方”比值的近似非中心 F -分布, 构造出拟合不足检验的方差分析表; 在此基础上, 提出一种两阶段的显著性因子筛选方法, 通过考察某个因子(组合)移除后模型拟合不足显著性的变化, 来推断该因子(组合)显著性. 仿真研究与实证表明, 所提方法不仅能够增强 LS-SVR 的统计可解释性, 有效识别出显著性因子; 而且可以得到预测性能更优的简化模型; 有助于提升试验设计建模及参数优化效率, 降低质量改进成本.

关键词: 因子筛选; 参数优化; 试验设计; 拟合不足检验; 最小二乘支持向量回归机

中图分类号: F406.2 **文献标识码:** A **文章编号:** 1007-9807(2023)12-0042-20

0 引言

高质量产品是企业赢得竞争优势的关键. 在工业 4.0 时代, 随着制造过程复杂化以及顾客需求日益苛刻, 对质量优化的要求也日渐提高. 传统方法基于经典线性模型, 难以描述复杂的质量形成规律, 优化结果仍有较大提升空间. 因此, 形式灵活的机器学习模型被引入到质量优化. 然而此类模型具有“黑箱”特性, 虽然可以实现全局建模, 但内部结构解析程度低, 难以对质量形成规律给出有说服力解释, 显著影响了使用者信心. 如何增强该类模型应用于质量优化的可解释性, 是亟待解决的科学问题. 事实上, 即使在机器学习范畴, “可解释性 (interpretability)” 也是当前研究难点^[1,2]. 对于制造过程, 质量优化主要途径之一是参数优化, 即通过选择合适的加工工艺参数值, 使产品质量特性达到最优. 而机器学习模型应用于

参数优化的瓶颈, 是如何根据模型解释各工艺参数对质量特性影响的显著程度, 进而识别关键工艺参数, 降低质量优化难度与成本.

制造业实际中, 由于过程噪声影响, 工艺参数与质量特性之间呈统计相关而非确定性函数关系, 因此很难建立精确数学模型. 目前多是通过试验设计 (design of experiments, DOE)^[3] 获取样本集, 拟合工艺参数与质量特性之间的经验回归模型, 而后寻优得到质量特性优化值及对应参数值. 在试验设计建模及参数优化中, 工艺参数被称为“因子”, 质量特性被称为“响应”, 首先需确定“因子—响应”回归模型的基本形式. 因子试验, 响应曲面分析 (response surface methodology, RSM)^[4] 等等, 采用一阶、二阶多项式模型, 这是一种局部模型. 对于较为复杂的作用关系 (例如响应存在多个极值, 因子与响应高度非线性相关) 效果有

① 收稿日期: 2020-11-16; 修订日期: 2022-01-15.

基金项目: 国家自然科学基金资助项目 (71571168; U1904211; 72172110); 国家科技部创新方法专项资助项目 (2019IM020200); 河南省高校科技创新人才人文社科类支持计划项目 (2019cx007).

通讯作者: 崔楠 (1979—), 男, 河南安阳人, 博士, 教授, 博士生导师. Email: nancui@whu.edu.cn

限^[5,6]. 为此, 研究者采用非参数模型^[7], 试图在可行域全局范围内近似“因子—响应”的作用关系. 此类模型形式灵活, 其前提假设弱化为响应的光滑变化, 而不要求与因子线性相关. Vining 和 Bohn^[8] 最早将核函数回归(kernel function regression, KFR) 应用于印刷过程质量优化; 此后采用非参数模型进行实验设计建模的研究逐渐丰富^[9,10]. 除 KFR 外, 高斯过程回归(Gaussian process regression, GPR) 或 Kriging 模型也被引入试验设计建模, 用于无重复误差的计算机试验^[11]. 也有研究者将上述模型应用于实际过程的质量预测与参数优化^[12]. 近年来, 研究者还将机器学习中的神经网络(artificial neural networks, ANN) 与试验设计结合^[13], 利用样本数据训练权值并拟合作用关系模型^[14]. KFR, GPR, ANN 优势在于其较强的非线性拟合能力, 但此类模型均是基于“经验风险最小化”, 即认为拟合性能好则预测性能也好. 这是一种大样本渐进性理论, 只有样本量足够大, 预测性能才与拟合性能一致^[15]. 较典型的是 ANN 的“过拟合”现象. 由于试验设计样本量较为有限, 上述模型效果也受到限制. 虽然可以通过改善训练过程来降低“过拟合”, 但是目前并没有非经验性方法来确保得到预测性能好的模型.

作为改进, 支持向量回归机(support vector regression, SVR)^[16] 被引入到试验设计建模及参数优化. SVR 是目前最佳的小样本机器学习工具, 基于“结构风险最小化”, 能够平衡拟合误差与模型复杂度, 在有限样本下取得较好预测性能, 更适合于试验设计建模. 崔庆安和何桢^[17] 给出了 SVR 应用于试验设计建模及参数优化的基本框架, 验证了 SVR 相对于 KFR, ANN 的优势. 孙林和杨世元^[18] 利用正交设计建立铣削表面粗糙度的 SVR 模型, 得到稳健优化结果. 向国齐和殷国富^[19] 针对两杆结构, 提出基于 SVR 和粒子群算法的稳健优化方法, 与经典 RSM, ANN 和 Kriging 对比说明了 SVR 优势. 曲兴田等^[20] 提出基于 SVR 和粒子群的焊点布置优化方法. Zhou 等^[21] 分析了 SVR, 径向基函数, GPR 在稳健性优化中的应用, 验证了此类模型对于多项式的优势.

上述研究表明, 非参数和机器学习模型形式灵活, 应用于复杂作用关系过程的试验设计建模

及参数优化优势明显; 然而这种灵活性也带来了“可解释性”差的弊端^[8], 较难提供类似于经典线性模型的推断性统计解释, 很大程度上降低了使用者对模型的信任, 主要表现之一就是难以实现显著性因子筛选(factors screening). 显著性因子是指在统计意义上对响应有影响的因子, 而非显著性因子引起的响应变化则较小, 与随机波动区别不明显. 将非显著性因子剔除可以简化作用关系模型. 在实际制造中, 只需控制显著性因子(对应为关键工艺参数), 而允许其他工艺参数在较宽范围波动, 从而降低制造难度和成本.

线性回归模型是“因子—响应”结构形式, 表现为因子及其系数加权和, 各因子(或交互作用)以分立形式存在, 可以借助系数的显著性检验来判断因子显著性. 此类研究较为成熟, 包括部分因子设计与超饱和设计^[5], 贝叶斯方法^[22], 序贯分支方法^[23]等等. 然而对于非参数和机器学习模型而言, 其结构的“可解释性”差是本质问题; 表现为所有因子在模型中以整体样本点而非分立形式存在. 这是一种“样本点—响应”结构, 难以直接考察因子影响. 例如 KFR, GPR, SVR 模型结构为核函数及其系数加权和, 所有因子作为样本点, 整体由核函数进行映射变换, 很难转化成“因子—响应”结构. 而对于 ANN, 其复杂的网络联结以及非线性激励函数映射, 使因子与响应之间的影响变得更加难以探查和解释.

目前基于非参数和机器学习模型进行显著性因子筛选的成果还很少. 而与之相关的, 对于复杂作用关系过程的显著性因子筛选研究可分三类: 其一是直接应用线性模型. 刘丽君和马义中等^[24] 将序贯分支方法应用于供应链仿真, 考虑位置与散度效应筛选显著性因子; 施文等^[25] 研究了基于动态效应的序贯分支方法. 但是序贯分支方法以多项式模型为基础, 且假设各因子对于响应的影响方向已知, 并不完全适用于非参数或机器学习模型. 此外, Xue 和 Wang 等^[26] 利用偏相关分析识别舰载机陆下沉速度的关键影响参数, 并利用 Kriging 建模, 但识别依据是线性显著性, 而且需在模型拟合之前进行. 其二是根据非参数和机器学习模型特点, 直接从样本点角度考察显著性. 崔庆安^[27] 采用 SVR 拟合过程作用关系模型, 利用因子可行域内的支持向量比率, 来识别显著性样

本点. 该研究者^[28]进一步基于最小二乘 SVR 进行序贯设计及建模, 构造了样本点显著性检验的正态分布统计量. 此类方法虽然可以识别出显著性样本点, 但是并未单独识别出各因子显著性. 其三, 在计算机试验中, 可采用灵敏度分析识别关键因子. 即改变各因子水平, 根据响应协方差计算 Sobol 指数, 较大者即对应关键因子. 该方法只能给出各因子的相对重要度, 难以判断其统计显著性. Kleijnen^[29]采用序贯分支方法对 Kriging 模型进行变量筛选. Denimal 等^[30]将 GPR 或 Kriging 模型拓展到随机误差情形. 对于此类过程, Sobol 指数较难区分因子或噪声影响. 如何利用非参数和机器学习模型, 从统计显著角度检验因子显著性, 仍然值得深入研究.

本研究将开展试验设计条件下的机器学习模型的显著性因子筛选研究. 考虑到 SVR 模型在样本量有限时的优良性能, 将其选作基础回归模型, 利用推断性统计检验, 从该模型的“样本点一响应”结构中直接筛选显著性因子. 进一步地, 采用最小二乘 SVR (least squares support vector regression, LS-SVR) 代替标准 SVR 进行研究. 与标准 SVR 相比, LS-SVR 的拟合值是观测值线性组合, 更有利于构造推断性统计量^[31]. 本研究首先介绍 LS-SVR 基本原理; 而后提出基于重复观测样本集的 LS-SVR 模型拟合不足的近似 F -检验方法; 进而提出两阶段的显著性因子筛选方法; 最后通过仿真与实证说明方法的有效性.

1 理论简介

LS-SVR 是一种回归估计方法. 设 p 维变量 $\mathbf{x} = [x_1, \dots, x_p]^T \in X$ 与一维随机变量 $y \in Y$ 之间存在未知依赖关系 $y = f(\mathbf{x}) + \varepsilon$, 其中 $\varepsilon \sim N(0, \sigma^2)$ 为随机误差. 回归估计就是利用来自于总体 (X, Y) 的训练样本集 $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \dots, n\}$ 去拟合 $f(\mathbf{x})$. 根据 $f(\mathbf{x})$ 形式及拟合方法不同, 形成各类回归模型. 例如多项式形式的线性模型

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^p \beta_i x_i + \sum_{i=1}^p \sum_{j=i+1}^p \beta_{ij} x_i x_j \quad (1)$$

LS-SVR 将 $f(\mathbf{x})$ 的形式设定为

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b \quad (2)$$

其中 $\boldsymbol{\phi}(\mathbf{x}) : \mathbb{R}^p \rightarrow \mathcal{H}$ 是非线性映射函数, \mathbf{w} 是权值向量, 改变 $\boldsymbol{\phi}(\mathbf{x})$ 可得不同 $f(\mathbf{x})$ 形式. 利用样本集进行估计, 建立 LS-SVR 模型的优化函数为

$$\min_{\mathbf{w}, b} J(\mathbf{w}, \xi_i) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \gamma \sum_{i=1}^n \xi_i^2$$

$$\text{s. t. } y_i = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b + \xi_i, \quad i = 1, 2, \dots, n \quad (3)$$

其中 γ 是惩罚系数, $\frac{1}{2} \mathbf{w}^T \mathbf{w}$ 和 $\sum_{i=1}^n \xi_i^2$ 分别代表模型复杂度和经验风险; 引入 Lagrange 乘子 $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]^T$; 令 $\mathbf{I}_n = [1, \dots, 1]^T$, \mathbf{I} 是 n 阶单位阵, $\mathbf{y} = [y_1, \dots, y_n]^T$, 求解上式可得

$$\begin{pmatrix} \mathbf{0} & \mathbf{I}_n^T \\ \mathbf{I}_n & \boldsymbol{\Omega} + \frac{1}{\gamma} \mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ b \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{y} \end{pmatrix}$$

其中 $\boldsymbol{\Omega} = [\omega_{ij}]_{n \times n}$, $\omega_{ij} = \varphi(\mathbf{x}_i) \varphi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$ 是核函数矩阵. 适用于强非线性过程的是 Gaussian 核函数^[32] $K(\mathbf{x}, \mathbf{x}_j) = \exp(-\|\mathbf{x} - \mathbf{x}_j\|^2/h^2)$, 其中 h^2 为窗宽参数. 令 $\mathbf{Z} = \boldsymbol{\Omega} + \frac{1}{\gamma} \mathbf{I}$, $\boldsymbol{\ell} = \mathbf{I}_n^T \mathbf{Z}^{-1} \mathbf{I}_n$,

则有估计值 $\hat{\boldsymbol{\alpha}} = (\mathbf{Z}^{-1} - \frac{1}{c} \mathbf{Z}^{-1} \mathbf{I}_n \mathbf{I}_n^T \mathbf{Z}^{-1}) \mathbf{y}$, $\hat{b} = \frac{1}{c} \mathbf{I}_n^T \mathbf{Z}^{-1} \mathbf{y}$, 进而得到与式 (2) 相应的经验回归方程

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}, \mathbf{x}_i) + \hat{b} = \boldsymbol{\Omega}^T(\mathbf{x}) \hat{\boldsymbol{\alpha}} + \hat{b} \quad (4)$$

以 $\hat{\mathbf{y}} = [\hat{y}_1, \dots, \hat{y}_n]^T$ 表示 \mathbf{y} 估计值, 即可得样本集的拟合模型 $\hat{\mathbf{y}} = \boldsymbol{\Omega} \hat{\boldsymbol{\alpha}} + \mathbf{I}_n \hat{b} = \mathbf{L} \mathbf{y}$. 其中

$$\mathbf{L} = \boldsymbol{\Omega} (\mathbf{Z}^{-1} - \frac{1}{c} \mathbf{Z}^{-1} \mathbf{I}_n \mathbf{I}_n^T \mathbf{Z}^{-1}) + \frac{1}{c} \mathbf{I}_n \mathbf{I}_n^T \mathbf{Z}^{-1} \quad (5)$$

由于 $\boldsymbol{\Omega}$, \mathbf{Z}^{-1} , c 仅与 \mathbf{x}_i 有关, 于是 \mathbf{L} 与 \mathbf{y} 无关. 则 $\mathbf{L} \mathbf{y}$ 是 \mathbf{y} 的线性组合或称线性光滑 (linear smoother), 有利于推断性统计. 设 $E(\mathbf{y}) = \boldsymbol{\mu}$, $\text{Var}(\mathbf{y}) = \sigma^2 \mathbf{I}$, 则 $E(\mathbf{L} \mathbf{y}) = \mathbf{L} \boldsymbol{\mu}$, $\text{Var}(\mathbf{L} \mathbf{y}) = \sigma^2 \mathbf{L} \mathbf{L}^T$, $E(\mathbf{y}^T \mathbf{L} \mathbf{y}) = \sigma^2 \text{trace}(\mathbf{L}) + \boldsymbol{\mu}^T \mathbf{L} \boldsymbol{\mu}$.

2 方法研究

本节将进行基于 LS-SVR 拟合模型的显著性因子筛选研究. 由式 (4) 的拟合模型结构可知, 样本点 \mathbf{x}_j 首先经过核函数 $K(\mathbf{x}_j, \mathbf{x}_i)$ 映射, 再与 $\hat{\alpha}_i$ 加

权求和形成输出 \hat{y}_j ; 这种结构很难单独分离出因子 x_1, \dots, x_p 并转化成式(1)所示的因子和系数乘积之和,也就无法直接利用经典线性模型中的显著性检验方法.但是从另外角度考虑,对于回归模型,如果某些重要因子没有参与建模,那么有可能引起“拟合不足(lack of fit)”.此时拟合值 \hat{y} 的变化,将显著小于噪声波动,拟合模型已失去意义.从因子角度思考,如果将某个因子从模型中移除之后,模型开始出现拟合不足,那么可以认为该因子对响应的影响是重要的,因而可以通过拟合不足检验来考察因子显著性.此外,“拟合不足”关注的是模型拟合值与观测值之间差异,对模型形式没有严格要求,因此 LS-SVR 模型也有可能发展出“拟合不足”的统计检验,进而实现显著性因子筛选.具体涉及到两个关键问题:其一是试验设计条件下适用于 LS-SVR 模型的“拟合不足”检验;其二是基于拟合不足检验的显著性因子筛选.

2.1 LS-SVR 模型拟合不足的近似 F-检验方法

方差分析是进行拟合不足检验的成熟和规范性方法^[33],通过对 \hat{y} 的方差分量分解,考察由模型带来的波动占比.但目前均是针对线性模型,需要开发适用于 LS-SVR 模型的方差分析方法.

1) 重复观测的样本集与 LS-SVR 拟合模型.

设 x 共有 m 个不同水平,有 $x_i = [x_{i,1}, \dots, x_{i,p}]^T$, $i = 1, \dots, m$. 在每个 x_i 处各有 n_i 次重复试验,其观测值为 y_{ij} , $j = 1, \dots, n_i$; 假设 n_i 个 y_{ij} 独立同分布即 $y_{ij} \sim N(f(x_i), \sigma^2)$, 则建模样本集可表示为

$$^{(0)} = \{ (x_1, y_{1,1}), \dots, (x_1, y_{1,n_1}), \dots, \dots, (x_m, y_{m,1}), \dots, (x_m, y_{m,n_m}) \} \quad (6)$$

其样本量大小为 $n = \sum_{i=1}^m n_i$. 设

$$Y = [Y_1^T, \dots, Y_m^T]^T \quad (7)$$

其中 $Y_i = [y_{i,1}, \dots, y_{i,n_i}]^T$, 设 $\bar{y}_i = \sum_{j=1}^{n_i} y_{i,j}$ 代表 x_i 处 n_i 次观测均值, $\bar{Y}_i = [\bar{y}_i, \dots, \bar{y}_i]^T$ 表示与 Y_i 对应的 n_i 维均值列向量, 则有 $\bar{Y}_i = \frac{1}{n_i} \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T Y_i$. 根据样本集 $^{(0)}$ 及式(5)可得 LS-SVR 拟合模型

$$\hat{Y} = LY \quad (8)$$

其中估计值 $\hat{Y} = [\hat{y}_{1,1}, \dots, \hat{y}_{1,n_1}, \dots, \dots, \hat{y}_{m,1}, \dots,$

$\hat{y}_{m,n_m}]^T$. 根据 LS-SVR 性质, 因为 y_{ij} , $j = 1, \dots, n_i$ 对应的输入均为 x_i , 则有 $\hat{y}_{i,1} = \dots = \hat{y}_{i,n_i}$, 简记为 $\hat{Y}_i = [\hat{y}_{i,1}, \dots, \hat{y}_{i,n_i}]^T = [\hat{y}_i, \dots, \hat{y}_i]^T$, $i = 1, \dots, m$. 此处的 L 为 n 阶方阵, $n = \sum_{i=1}^m n_i$.

2) LS-SVR 模型残差的平方和分解. 拟合不足检验主要通过对残差的方差分析来实现, 而残差的平方和分解是基础, 即将残差总的平方和分解为来源于纯误差的、来源于拟合不足的平方和. 对于观测值 y_{ij} , 其拟合残差可以表示为 $y_{ij} - \hat{y}_{ij}$, 并进一步可以分解为

$$y_{i,j} - \hat{y}_i = (y_{i,j} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i)$$

其中 $y_{i,j} - \bar{y}_i$ 代表在 x_i 处 j 个重复观测值与其平均值之差, 在回归估计中称为“纯误差”. 而 $\bar{y}_i - \hat{y}_i$ 则表示在 x_i 处的拟合值与观测值均值之差, 可以反映模型的拟合性能. 例如, 如果 \hat{y}_i 与 \bar{y}_i 相差较多, 则说明模型没有较好地反映 y_{ij} 的平均变化趋势, 就有可能存在“拟合不足”. 而要在样本集的范围考察纯误差与拟合不足, 就需要借助“平方和”来进行. 定义 SS_{Res} 为“残差平方和”, 表示拟合值与实际观测值之差的平方和

$$SS_{Res} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{i,j} - \hat{y}_i)^2 = (Y - \hat{Y})^T (Y - \hat{Y}) = Y^T (I - 2L + L^2) Y$$

定义 SS_{Pe} 为“纯误差平方和”, 表示重复观测值与其平均值之差的平方和

$$SS_{Pe} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = (Y - MY)^T (Y - MY) = Y^T (I - M) Y$$

$$\text{其中 } M = \begin{pmatrix} \frac{1}{n_1} \mathbf{1}_{n_1} \mathbf{1}_{n_1}^T & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \frac{1}{n_m} \mathbf{1}_{n_m} \mathbf{1}_{n_m}^T \end{pmatrix} \text{ 为 } n \text{ 阶分块方}$$

阵, 且有 $[Y_1, Y_2, \dots, Y_m]^T = M [Y_1, Y_2, \dots, Y_m]^T$, 以及如下性质: (a) M 为幂等矩阵, $M^2 = M$; (b) M 的迹, $\text{tr}(M) = m$; (c) $ML = LM = L$. 可以看出, SS_{Pe} 与模型无关. 定义 SS_{Lot} 为拟合模型的“拟合不足”平方和, 有

$$SS_{Lof} = \sum_{i=1}^m \sum_{j=1}^n (\hat{y}_i - \hat{y}_i)$$

$= (MY - LY)^T (MY - LY) = Y^T (M - 2L + L^2) Y$
 “较小”的 SS_{Lof} 说明 \hat{y}_i 与 \bar{y}_i 的差别不大, 意味着模型较好地拟合了样本集的变化; 反之, “较大”的 SS_{Lof} 说明 \hat{y}_i 与 \bar{y}_i 相差较大, 该拟合模型与样本集基本无关. 何为“较小”或“较大”, 则需进行统计检验. 根据各平方和的定义可得

$$SS_{Res} = SS_{Pe} + SS_{Lof}$$

上式即为残差平方和的分解等式, 在此基础上, 可以发展出相应的显著性检验方法.

3) 拟合不足均方与纯误差均方比值的近似 F -分布. 由定义, SS_{Res} 、 SS_{Pe} 和 SS_{Lof} 均为半正定二次型, 其对应矩阵为 $A_{Res} = I - 2L + L^2$ 、 $A_{Pe} = I - M$ 和 $A_{Lof} = M - 2L + L^2$. 根据二次型理论, SS_{Pe} 和 SS_{Lof} 的自由度分别为 $df_{Pe} = \text{tr}(A_{Pe}) = n - m$ 和 $df_{Lof} = \text{tr}(A_{Lof}) = m - \text{tr}(2L - L^2)$. 将平方和除以自由度, 即可得纯误差均方 MS_{Pe} 、拟合不足均方 MS_{Lof}

$$MS_{Pe} = SS_{Pe} / df_{Pe}; \quad MS_{Lof} = SS_{Lof} / df_{Lof} \quad (9)$$

若满足 (a) A_{Res} 、 A_{Pe} 、 A_{Lof} 幂等; (b) $A_{Pe} A_{Lof} = \mathbf{0}$, 则均方比值 $F_0 = MS_{Lof} / MS_{Pe}$ 严格服从 F -分布. 计算可知, 上述条件除 A_{Lof} 非幂等之外均成立. 对于 A_{Lof} , 作为近似, 可以证明当惩罚系数 γ 足够大时, 有 $A_{Lof}^2 \approx A_{Lof}$, 因而可得近似非中心 F -分布^②

$$F_0 = MS_{Lof} / MS_{Pe} \sim F_{df_{Lof}, df_{Pe}, \lambda_{Lof}, \lambda_{Pe}} \quad (10)$$

其中自由度 $\lambda_{Pe} = \frac{1}{\sigma^2} E^T(Y) A_{Pe} E(Y)$, $\lambda_{Lof} = \frac{1}{\sigma^2} E^T(Y) A_{Lof} E(Y)$. 此外, 期望均方

$$E(MS_{Pe}) = \sigma^2 + E^T(Y) A_{Pe} E(Y) / \text{tr}(A_{Pe}) \quad (11)$$

$$E(MS_{Lof}) = \sigma^2 + E^T(Y) A_{Lof} E(Y) / \text{tr}(A_{Lof}) \quad (12)$$

可以证明 $ME(Y) = E(Y)$ ^②, 则易证 $E^T(Y) A_{Pe} E(Y) = 0$ 及 $\lambda_{Pe} = 0$ 、 $E(MS_{Pe}) = \sigma^2$, 于是 MS_{Pe} 可以作为 σ^2 的无偏估计.

4) LS-SVR 拟合不足的显著性检验及方差分析表. 该检验的原假设 H_0 和备择假设 H_1 可设为

$$H_0: E(Y) = \hat{Y}; \quad H_1: E(Y) \neq \hat{Y}.$$

即如果 \hat{Y} 与 $E(Y)$ 不一致, 则存在拟合不足. 当 H_0 成立时, $E(Y) = \hat{Y} = LY$, $E[E(Y)] = E(LY)$

或 $E(Y) = LE(Y)$, 又 $ME(Y) = E(Y)$, 于是易证 $E^T(Y) A_{Lof} E(Y) = 0$ 及 $\lambda_{Lof} = 0$ 、 $E(MS_{Lof}) = \sigma^2$, 则 MS_{Lof} 为 σ^2 无偏估计, MS_{Lof} / MS_{Pe} 服从近似中心 F -分布

$$F_0 = MS_{Lof} / MS_{Pe} \sim F_{df_{Lof}, df_{Pe}} \quad (13)$$

而当模型拟合不足时, 根据半正定二次型性质, 有 $E^T(Y) A_{Lof} E(Y) > 0$ 因而 $E(MS_{Lof}) > \sigma^2$, $E(MS_{Lof}) > E(MS_{Pe})$. 式(14)中, 将有可能出现较大的 F_0 , 其显著性 P -value 为

$$P\text{-value} = \text{Prob.} \{ F_0 > F_{\alpha, df_{Lof}, df_{Pe}} \} \quad (14)$$

其中 α 是某个显著性水平值(例如 0.05). 由此可得拟合不足检验的方差分析表:

表 1 LS-SVR 模型拟合不足检验的方差分析表

Table 1 ANOVA table of lack-of-fit test for LS-SVR model

来源	平方和 (SS.)	自由度 (df.)	均方 (MS.)	F 值
拟合不足 (Lof)	$M - 2L + L^2$	$\text{tr}(M - 2L + L^2)$	$\frac{SS_{Lof}}{df_{Lof}}$	$\frac{MS_{Lof}}{MS_{Pe}}$
纯误差 (Pe)	$I - M$	$\text{tr}(I - M)$	$\frac{SS_{Pe}}{df_{Pe}}$	
残差 (Pes)	$I - 2L + L^2$	$\text{tr}(I - 2L + L^2)$		

式(13)中, 当 n_i 增大时, m 将增大, $df_{Pe} = n - m$ 也将增大; 但是 $df_{Lof} = m - \text{tr}(2L - L^2)$ 的变化与 n_i 的相关性不高: 因为 m 与 n_i 无关, 而 $\text{tr}(2L - L^2)$ 与 $2L$ 和 L^2 的接近程度有关, 受不同样本点数目 m 影响较大, 与单个样本点重复次数 n_i 无关. 对于 F -分布而言, 当第二自由度(即 df_{Pe})增大时, 反映其不对称性的指标“偏斜”将减少, 显著性检验的稳健性将增加, 因此增加重复试验次数将有利于显著性检验. 但是 n_i 也不能无限制增加, 还需受到实验成本与时间约束.

2.2 基于拟合不足检验的 LS-SVR 模型显著性因子筛选

1) 利用两阶段的 LS-SVR 模型考察因子的显著性. 其一是对初始 LS-SVR 模型的拟合不足检验, 此时所有因子 x_1, \dots, x_p 均参与建模. 根据样本集 \mathcal{D} 拟合出初始的 LS-SVR 模型 $\mathcal{M}^{(0)}$, 其中 $x_i = [x_{i1}, \dots, x_{ip}]^T, i = 1, \dots, m$. 如果由式(13)

② 证明过程见附录 A 和附录 B.

和式 (14) 来计算出的 $P^{(0)}\text{-value} < 0.05$, 则有较强理由认为 $\mathcal{M}^{(0)}$ 存在拟合不足. 此时应该重新选择和优化模型的超参数, 或者增加新的建模样本点, 重新拟合模型, 直至拟合不足不显著为止.

其二是改变参数建模的因子组合, 拟合新的 LS-SVR 拟合模型, 再进行拟合不足检验考察各因子(组合)的显著性. 对于 $x = [x_1, \dots, x_p]^T$ 若需考察其中某个因子 x_k 的显著性, 一个比较直观的思路, 是将 x_k 从 x 中移除, 形成新的因子组合 $x^{(-k)} = [x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_p]^T$ 及建模样本集 $\mathcal{M}^{(-k)} = \{ (x_1^{(-k)}, y_{1,1}), \dots, (x_1^{(-k)}, y_{1,n_1}), \dots, (x_m^{(-k)}, y_{m,1}), \dots, (x_m^{(-k)}, y_{m,n_m}) \}$ (15)

而后拟合新的 LS-SVR 模型 $\mathcal{M}^{(-k)}$. 如果 $\mathcal{M}^{(-k)}$ 拟合不足, 则说明不能将 x_k 移除, 其对 y 有显著影响. 而这种显著性的大小, 可以用 $P^{(-k)}\text{-value}$ 来表示. 该值越小, 说明因子 x_k 的移除对模型影响越大, 因而 x_k 也就越显著. 若需考察因子组合 $x_c = [x_{c_1}, \dots, x_{c_k}]^T, \{c_1, \dots, c_k\} \subset \{1, \dots, p\}$ 的显著性, 可将 x_c 从 x 中移除, 而后拟合新的 LS-SVR 模型进行拟合不足检验. 还可根据各因子(组合) $P\text{-value}$ 值的相对大小进一步比较其显著性.

2) 两阶段的 LS-SVR 模型超参数选择. 其一, 选择 $K(x, x_j) = \exp\left(\frac{1}{h^2} \|x - x_j\|^2\right)$ 作为 $\mathcal{M}^{(0)}$ 和 $\mathcal{M}^{(-k)}$ 的核函数. 则其超参数为窗宽参数 h^2 与惩罚系数 γ . 对于 $\mathcal{M}^{(0)}$ h^2 和 γ 的选择应使模型拟合不足不显著, 可对常见的“留一法(Leave-one-out)”超参数寻优算法进行改进, 增加关于 $P^{(0)}\text{-value}$ 的约束. 此外, 应满足考虑式(10)近似 F -分布成立的条件 $A_{\text{Lof}}^2 \approx A_{\text{Lof}}$. 可增加 $\|A_{\text{Lof}}^2 - A_{\text{Lof}}\|_1 < T_{\text{Lof}}$ 的约束^③ 其中 $T_{\text{Lof}} > 0$ 为一个较小阈值. 而当 γ 过大时, 有 $L \approx I$, 将会导致“过拟合”^[28]. 为此需增加 $\|L - I\|_1 > T_{\text{Overfitting}}$ 的约束, 其中 $T_{\text{Overfitting}} > 0$ 为一个较小阈值. T_{Lof} 和 $T_{\text{Overfitting}}$ 的引入, 可以在足够大的 γ 和较小的 LS-SVR 模型复杂度之间取得折中. 由此形成基于

“留一法”的 LS-SVR 超参数约束优化方程

$$\begin{aligned} \min_{h, \gamma} C(h, \gamma) &= \frac{1}{n} \sum_i \sum_j [y_{ij} - \hat{f}_{(-ij)}(x_i)]^2 \\ \text{s. t. } &\begin{cases} \|L - I\|_1 > T_{\text{Overfitting}} \\ \|A_{\text{Lof}}^2 - A_{\text{Lof}}\|_1 < T_{\text{Lof}} \\ P^{(0)}\text{-value} > \alpha \end{cases} \end{aligned} \quad (16)$$

其中 $\hat{f}_{(-ij)}(x_i)$ 表示将 (x_i, y_{ij}) 从样本集 $\mathcal{M}^{(0)}$ 移出后的模型在 x_i 处估计值; α 可设为 0.05; 当 x_i 规范化至 $[0, 1]$ 时, 可经验性设定 $T_{\text{Lof}} = T_{\text{Overfitting}} = 0.5$.

其二, 关于 $\mathcal{M}^{(-k)}$ 超参数的选择. LS-SVR 模型结构主要由 h^2, γ 以及样本集来确定, 可以依样本集分布而光滑变化. 仅仅根据样本集 $\mathcal{M}^{(-k)}$, 改变 h^2, γ 也可得到某个拟合模型 $\mathcal{M}^{(-k)}$. 但是与 $\mathcal{M}^{(0)}$ 相比, $\mathcal{M}^{(-k)}$ 的 h^2, γ 和 x_i 均发生了变化, 不论其拟合不足的显著性如何, 均不能看作是仅仅由于将 x_k 移除而导致的. 为反映 x_k 移除对于 $\mathcal{M}^{(0)}$ 的影响, 恰当做法是将 $\mathcal{M}^{(0)}$ 的 h^2, γ 直接应用于 $\mathcal{M}^{(-k)}$. 由于样本观测值 y_{ij} 在 $\mathcal{M}^{(-k)}$ 与 $\mathcal{M}^{(0)}$ 中保持一致, 此时变化的只有因子组合, 由含有 x_k 的 x_i 变为移除 x_k 的 $x_i^{(-k)}$. 如果 $\mathcal{M}^{(-k)}$ 出现拟合不足, 则说明不能移除 x_k , 其对于 LS-SVR 拟合模型而言是显著的.

2.3 显著性因子筛选的实现步骤

根据上述分析, 形成试验设计条件下基于 LS-SVR 拟合不足检验的显著性因子筛选步骤. 总体上可以分为初始模型拟合, 显著性因子筛选两个阶段. 基于研究的聚焦性, 假设已经确定了合适的试验设计方式.

阶段 I 初始模型拟合检验.

步骤 1 设共有 p 个因子 $x = [x_1, \dots, x_p]^T$, 采用空间填充设计, 例如拉丁超立方(Latin hypercube sampling, LHS)设计, 共 m 个试验点(即试验水平) $x_i, i = 1, \dots, m$; 每个 x_i 处重复试验次数 $n_i > 1$, 观测值为 $y_{ij}, j = 1, \dots, n_i$; 总试验次数 $n = \sum_{i=1}^m n_i$, 形成式(6)所示样本集 $\mathcal{M}^{(0)}$.

步骤 2 选择 $K(x, x_j) = \exp\left(\frac{1}{h^2} \|x - x_j\|^2\right)$ 为

③ 证明过程见附录 B.

核函数 利用式(16)对 h^2 和 γ 寻优. 利用式(4)和样本集 $\mathcal{M}^{(0)}$ 形成初始拟合模型 $\mathcal{M}^{(0)}$.

步骤3 利用表1和式(13)、式(14)对 $\mathcal{M}^{(0)}$ 进行拟合不足检验. 如果 $P^{(0)}\text{-value} < 0.05$, 则说明 $\mathcal{M}^{(0)}$ 存在拟合不足, 此时需返回步骤3, 重新选择 h^2 和 γ , 直至拟合不足不显著.

阶段II 显著性因子筛选阶段.

步骤4 令 $k = 1, \dots, p$, 依次将 x_k 从 x 中移出, 得到 $x^{(-k)} = [x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_p]^T$, 以及如式(15)所示的新建模样本集 $\mathcal{M}^{(-k)}$, 采用和 $\mathcal{M}^{(0)}$ 相同的 h^2 和 γ , 形成简化后的拟合模型 $\mathcal{M}^{(-k)}$.

步骤5 对 $\mathcal{M}^{(-k)}$ 进行拟合不足检验, 计算 $P^{(-k)}\text{-value}$, 其值越小, 说明 x_k 越重要.

步骤6 $k = 1, \dots, p$, 重复步骤4、步骤5, 将得到的各 $P^{(-k)}\text{-value}$ 由小到大排序 $P^{(-1)}\text{-value} \leq P^{(-2)}\text{-value} \leq \dots \leq P^{(-p)}\text{-value}$, 该序列即是各因子显著性程度由高至低排序. 在 0.05 的显著性水平下, $P^{(-k)}\text{-value} < 0.05$ 对应的 x_k 均可看作显著性因子, 其余因子均不显著.

步骤7 若不显著因子多于 1 个, 还需考察这些因子组合的显著性. 设不显著因子集合为 $\{\tilde{x}_1, \dots, \tilde{x}_r\}$, $r < p$. 该集合所有元素个数 ≥ 2 的子集, 即为待考察显著性的因子组合, 将其依次从 x 中移出, 再重复步骤4和步骤5进行显著性检验.

步骤8 对比分析步骤6与步骤7的结果, 确定可以从 LS-SVR 模型中移除的不显著因子, 得到显著性因子组合以及对应的简化模型 $\mathcal{M}^{(\text{Red.})}$.

需要说明的是, 表1的计算与核函数形式无关, 可以适用于各种形式的核函数(例如多项式核、B-样条核等等); 但是需要根据其不同的超参数类型, 参照式(16)进行约束最优化求解, 以使 A_{Lof} 满足近似幂等的条件.

3 仿真研究

利用两个典型仿真函数, 考察 MS_{Lof} 与 MS_{Pe} 比值的近似 F -分布的经验概率密度曲线与理论概率密度曲线的符合性; 对基于拟合不足检验的

因子显著性筛选方法进行仿真验证.

3.1 仿真函数

仿真函数 I 是带有白噪声的 Camel 函数

$$f_{\text{Camel}} = 2x_1^2 - 1.05x_1^4 + \frac{1}{6}x_1^6 + x_1x_2 + x_2^2 + 0 \sum_{t=3}^6 x_t + \varepsilon \tag{17}$$

其中 $\varepsilon \sim N(0, \sigma^2)$; $x_1, \dots, x_6 \in [-2, 2]$. 上式中仅含有 x_1 或 x_2 的部分是 Camel 函数, 已被应用于各类算法仿真^[34]. f_{Camel} 在 x_1, x_2 维度上较为显著, 如图1所示. 而对于因子 $x_t, t = 3, 4, 5, 6$, 其对应系数均为 0, 本质上不显著性. 对于 f_{Camel} , 一阶、二阶多项式难以实现全局性建模. 但若采用高阶多项式, 将会导致“维数灾难”. 例如 6 因子六阶多项式, 需估计 529 个系数; 若只考虑不超过三阶的交互作用, 则需估计 101 个系数. 因此采用 LS-SVR 进行全局性建模.

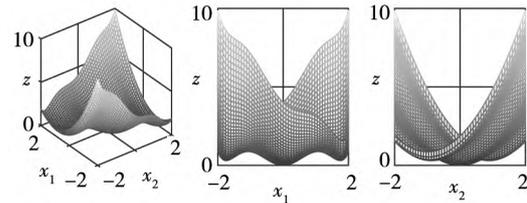


图1 Camel 函数的 3 维和 2 维投影图

Fig. 1 3D and 2D projection of function camel

仿真函数 II 是有工程背景的 Piston 函数, 由一系列非线性函数组成链式结构, 用于对活塞在气缸内的圆周运动建模^[35], 其形式为

$$f_{\text{Piston}} = 2\pi \sqrt{\frac{M}{k + S^2 P_0 (V_0/T_0) (T_a/V^2)}} + \varepsilon \tag{18}$$

其中 $V = \frac{1}{2k} S \sqrt{A^2 + \frac{1}{T_0} 4kP_0V_0T_a - A}$; $A = P_0S + 19.62M - \frac{kV_0}{S}$; $M, S, V_0, k, P_0, T_a, T_0$ 分别表示活塞重量/(kg), 活塞表面积/(m^2), 气体初始体积/(m^3), 弹性系数/(N/m), 气压/(N/m^2), 介质温度/(K), 充气平衡温度/(K). 响应则是活塞运行周期/(s), 此外, $\varepsilon \sim N(0, \sigma^2)$.

3.2 仿真过程

根据仿真函数范围 $f_{\text{Camel}} \in [0, 9.8667] + \varepsilon$, $f_{\text{Piston}} \in [0.1642, 1.1991] + \varepsilon$, 各选取两个噪声

水平 σ 以及因子水平组合数 m . f_{Camel} 仿真: $\sigma = 0.5, 1.0$; $m = 50, 100$; f_{Piston} 仿真: $\sigma = 0.05, 0.1$; $m = 50, 100$. 而后在每一 (σ, m) 组合处各进行 10 次仿真. 每次仿真的因子水平组合由 LHS 设计生成, 表示为 x_1, \dots, x_m , 且规范化至 $[0, 1]$ 内. 对于每个 x_i , 随机生成重复试验次数 $n_i \leq 3$. 根据利用式 (17)、式 (18) 生成初始样本集, 利用 3.3 节步骤识别出显著性因子 (或因子组合). 作为对比, 利用 LHS 设计生成大小为 1 000 的预测样本集, 分别计算初始和简化 LS-SVR 模型的预

测均方误差 $MSPE$. 对于相同的 x_i, n_i 和 σ , 固定 LS-SVR 超参数不变, 重复 5 000 次试验, 得到含有随机噪声的响应观测值. 根据式 (9), 计算出 5 000 个 $F_0 = MS_{\text{Lof}}/MS_{\text{Pe}}$ 值, 由此估计出对应的经验概率密度曲线 (采用核估计方法); 再根据式 (10) 计算理论概率密度曲线.

3.3 仿真结果

1) 关于近似 F -分布的经验与理论概率密度曲线. 由上述过程得到 f_{Camel} 和 f_{Piston} 各 40 组经验与理论概率密度曲线, 如图 2 (a)、图 2 (b) 所示.

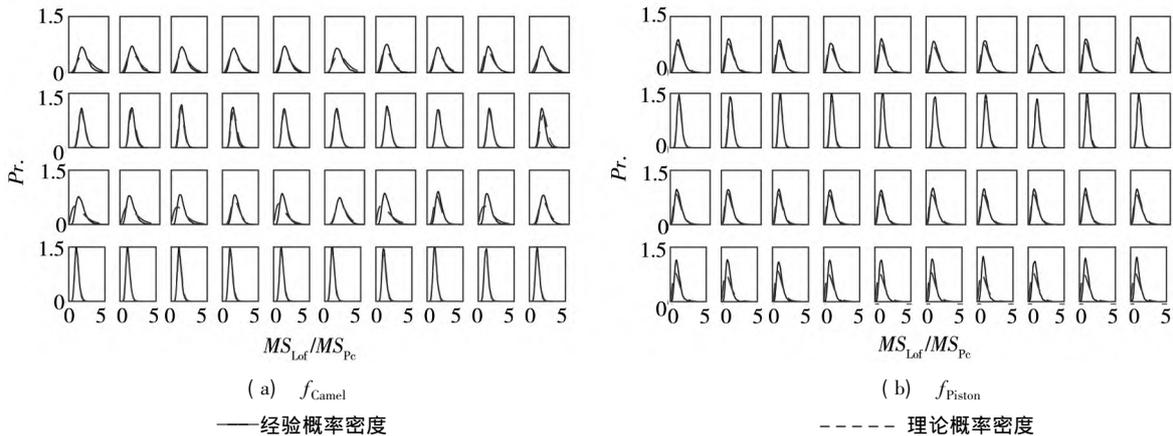


图 2 $MS_{\text{Lof}}/MS_{\text{Pe}}$ 的经验与理论概率密度曲线对比

Fig. 2 Comparison of empirical and theoretical probability density curves of $MS_{\text{Lof}}/MS_{\text{Pe}}$

2) 关于显著性因子的筛选. 表 2 和表 3 为 f_{Camel} 和 f_{Piston} 仿真的显著性因子筛选, 给出了移除某个因子后的 LS-SVR 模型拟合不足 P -value. 该值越小, 被移除因子对模型影响越为显著. 表中还给出了能够被移除的最大因子组合.

以表 2 的 $(\sigma, m) = (0.5, 50)$ 为例, 设定显著性水平为 0.05, 第 1 次仿真, 单独移除 x_1 或 x_2 , 模型均出现拟合不足, P -value 分别为 $4e - 11$ 和 $9e - 17$ (科学计数法); 而单独移除 x_3, x_4, x_5, x_6 中任一个, P -value 均大于 0.05, 不会导致拟合不足. 如果将因子组合 $[x_3, x_4, x_5, x_6]$ 移除, 拟合不足也不显著 (P -value = 0.294 9). 包含 x_1 或 x_2 的因子组合共 45 组, 将其分别移除后所得 P -value 最大值为 $1e - 12$, 拟合不足均为高度显著; 而对于所有不包含 x_1 或 x_2 的因子组合 (共 11 组), 分别将其移除后 P -value 均大于 0.05, 拟合不足均不显著. 此时, 最大因子组合是 $[x_3, x_4, x_5, x_6]$, 而 $[x_1, x_2]$ 即为显著性因子.

为进一步说明, 表 4 给出了 $f_{\text{Piston}}, (\sigma, m) = (0.05, 50)$ 第 1 次仿真确定可移除的最大因子组合的过程: 首先发现因子 x_1, x_5, x_6, x_7 对应的 P -value 均大于 0.05, 于是考虑将该因子组合移除; 但发现移除后 P -value = 0.032 6 < 0.05, 因而不能整体移除; 而后考虑移除其中 3 个因子, 计算得知因子组合 $[x_5, x_6, x_7]$ 对应 P -value = 0.414 6, 而其他 3 因子组合 P -value 均小于 0.05, 不能被移除. 由此可移除的最大因子组合为 $[x_5, x_6, x_7]$ 对应的显著性因子为 $[x_1, x_2, x_3, x_4]$.

3) 关于简化后的 LS-SVR 模型的预测性能. 为考察移除不显著因子对 LS-SVR 模型影响, 利用表 2 和表 3 的样本集, 采用拟合均方误差 MSE 和预测均方误差 $MSPE$ 来对比初始模型 $\mathcal{M}^{(0)}$ 和简化模型 $\mathcal{M}^{(\text{Red.})}$ 性能, 每次仿真均由 LHS 设计生成一个大小为 1 000 的测试样本集. 而后相同 (σ, m) 的 10 次仿真形成一组 Box 图, f_{Camel} 和 f_{Piston} 各形成 4 组对比 Box 图, 如图 3 和图 4 所示.

表 2 f_{Camel} 仿真的拟合不足 P -value
Table 2 P -value of lack-of-fit for f_{Camel} simulations

(σ, m)	被移除的单个因子 或最大因子组合	仿真次数									
		第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次
(0.5, 50)	x_1	4e-11	3e-14	1e-15	3e-10	1e-12	1e-12	4e-11	8e-16	4e-09	3e-09
	x_2	9e-17	7e-17	1e-27	7e-18	7e-17	3e-15	2e-15	2e-19	3e-16	5e-18
	x_3	0.156 1	0.396 1	0.023 6	0.422 7	0.545 4	0.122 1	0.179 2	0.161 9	0.152 3	0.156 1
	x_4	0.161 9	0.521 6	0.010 8	0.403 0	0.211 5	0.391 3	0.150 0	0.142 1	0.249 6	0.161 9
	x_5	0.136 5	0.520 3	0.226 7	0.169 3	0.204 7	0.014 7	0.411 9	0.027 6	0.174 0	0.349 4
	x_6	0.332 4	0.314 6	0.103 6	0.275 2	0.417 7	0.176 3	0.628 8	0.133 7	0.199 1	0.374 3
	$[x_5, x_6]$			0.129 0							
	$[x_3, x_5, x_6]$										0.142 5
	$[x_3, x_4, x_5, x_6]$	0.294 9	0.342 0		0.150 7	0.185 9	0.102 8	0.391 2	0.050 3	0.202 2	
(0.5, 100)	x_1	4e-27	1e-23	1e-28	4e-25	1e-23	1e-28	3e-24	5e-22	2e-27	9e-26
	x_2	7e-28	4e-35	1e-33	2e-35	3e-26	1e-38	5e-38	5e-34	2e-35	5e-34
	x_3	0.533 3	0.442 8	0.491 6	0.283 5	0.108 9	0.156 9	0.109 9	0.262 3	0.033 1	0.737 6
	x_4	0.506 4	0.144 3	0.462 7	0.215 6	0.225 3	0.165 4	0.311 0	0.455 9	0.096 4	0.717 0
	x_5	0.290 9	0.214 5	0.387 6	0.175 2	0.091 8	0.122 6	0.358 5	0.452 3	0.026 6	0.605 4
	x_6	0.501 5	0.374 5	0.524 8	0.061 8	0.212 5	0.121 1	0.202 3	0.429 2	0.060 8	0.661 0
	$[x_4, x_6]$									0.058 0	
	$[x_3, x_4, x_5, x_6]$	0.491 2	0.113 7	0.476 0	0.106 3	0.064 1	0.092 2	0.108 6	0.459 6		0.471 6
(1.0, 50)	x_1	1e-05	7e-2	2e-4	8e-06	9e-05	4e-5	2e-4	5e-06	8e-06	3e-07
	x_2	6e-11	3e-06	1e-07	9e-07	4e-05	1e-4	2e-07	4e-06	9e-07	2e-09
	x_3	0.120 8	0.457 4	0.574 3	0.376 4	0.478 3	0.129 1	0.268 5	0.238 4	0.329 7	0.083 6
	x_4	0.114 7	0.467 3	0.641 1	0.436 9	0.143 3	0.365 3	0.252 9	0.283 8	0.321 5	0.252 7
	x_5	0.116 1	0.477 3	0.604 4	0.344 6	0.223 9	0.280 8	0.163 4	0.121 1	0.358 4	0.236 4
	x_6	0.184 9	0.470 0	0.384 7	0.482 1	0.079 2	0.180 0	0.490 9	0.252 7	0.164 8	0.060 9
	$[x_3, x_4, x_5, x_6]$	0.050 1	0.311 3	0.751 6	0.713 6	0.116 7	0.179 8	0.173 2	0.308 0	0.142 2	0.252 8
(1.0, 100)	x_1	4e-17	2e-14	2e-13	4e-11	1e-12	1e-07	3e-07	3e-09	7e-12	3e-09
	x_2	3e-21	4e-17	3e-18	9e-18	5e-17	6e-15	3e-15	4e-17	4e-21	2e-15
	x_3	0.320 4	0.360 7	0.217 0	0.309 1	0.382 7	0.751 2	0.446 6	0.240 6	0.132 4	0.422 7
	x_4	0.185 5	0.610 4	0.060 0	0.287 2	0.378 2	0.468 3	0.539 1	0.239 5	0.121 3	0.373 3
	x_5	0.270 4	0.530 8	0.183 7	0.553 7	0.527 0	0.590 7	0.599 2	0.129 2	0.077 3	0.420 9
	x_6	0.099 7	0.663 1	0.130 7	0.376 5	0.396 3	0.506 4	0.468 4	0.102 7	0.104 1	0.315 1
	$[x_3, x_4, x_5, x_6]$	0.064 4	0.419 9	0.102 6	0.288 8	0.630 4	0.747 3	0.529 4	0.112 5	0.098 1	0.552 3

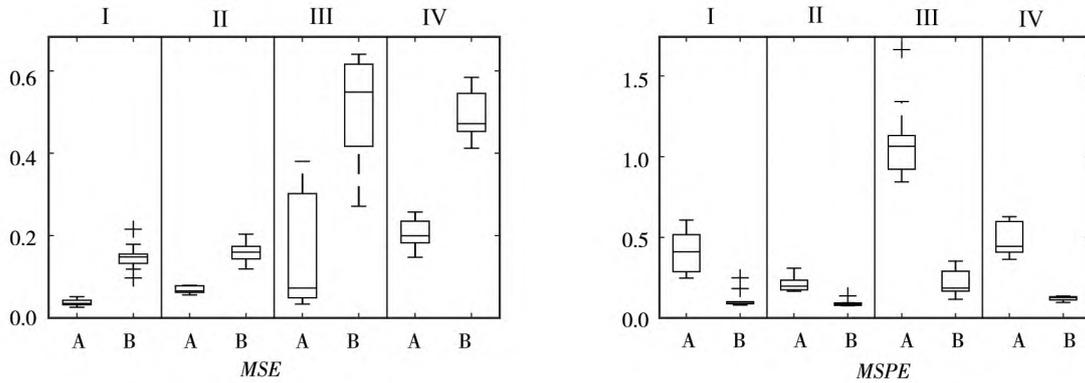
表 3 f_{Piston} 仿真的拟合不足 P -value
Table 3 P -value of lack-of-fit for f_{Piston} simulations

(σ, m)	被移除的单个因子 或最大因子组合	仿真次数									
		第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次
(0.05, 50)	x_1	0.152 6	0.447 8	0.042 6	0.010 5	0.045 2	0.305 1	0.132 4	0.475 6	0.154 7	0.279 8
	x_2	2e-07	4e-09	7e-09	1e-09	4e-09	3e-13	1e-09	1e-06	6e-11	2e-08
	x_3	5e-09	4e-08	5e-05	2e-06	5e-11	2e-09	2e-04	2e-07	5e-07	1e-04
	x_4	0.006 8	0.159 4	0.545 8	0.013 3	0.033 8	0.037 6	0.016 1	0.349 1	0.037 1	0.081 8
	x_5	0.590 5	0.657 3	0.408 0	0.318 9	0.174 0	0.079 3	0.207 7	0.652 6	0.342 0	0.244 9
	x_6	0.543 0	0.727 4	0.407 4	0.322 4	0.506 8	0.233 2	0.605 4	0.640 8	0.313 3	0.370 4
	x_7	0.767 0	0.782 9	0.585 1	0.260 5	0.369 6	0.106 1	0.481 2	0.462 7	0.304 4	0.436 9
	$[x_5, x_6, x_7]$	0.414 6		0.361 3	0.178 6	0.341 2	0.129 8				
	$[x_1, x_5, x_6, x_7]$							0.190 0			0.171 2
(0.05, 100)	x_1	0.003 5	0.112 8	0.025 4	0.000 6	0.000 3	0.003 9	0.003 4	0.001 2	0.003 3	0.048 1
	x_2	2e-21	3e-16	3e-19	7e-24	1e-29	2e-22	1e-22	9e-24	2e-20	4e-21
	x_3	2e-14	6e-10	8e-10	6e-17	2e-21	6e-12	5e-16	6e-21	4e-13	2e-12
	x_4	0.009 4	0.107 7	0.001 4	0.015 5	8e-07	0.005 3	0.000 9	0.002 4	0.041 8	0.066 2
	x_5	0.117 9	0.593 2	0.544 0	0.105 5	0.186 6	0.348 7	0.163 7	0.109 9	0.237 2	0.453 0
	x_6	0.082 3	0.577 2	0.514 4	0.225 1	0.206 0	0.371 1	0.294 2	0.161 0	0.354 4	0.335 5
	x_7	0.109 7	0.498 9	0.527 9	0.184 6	0.131 8	0.259 4	0.268 4	0.109 4	0.285 2	0.328 0
	$[x_5, x_6, x_7]$	0.101 4		0.382 3	0.117 8	0.229 1	0.451 7	0.207 8	0.159 5	0.435 9	0.209 9
	$[x_1, x_5, x_6, x_7]$		0.099 1								
(0.1, 50)	x_1	0.220 6	0.447 3	0.490 3	0.828 6	0.164 4	0.303 3	0.221 4	0.349 2	0.758 7	0.344 7
	x_2	0.008 8	0.007 1	0.044 6	0.001 2	1e-05	0.000 7	0.001 3	0.002 8	0.007 0	0.058 0
	x_3	0.001 1	0.163 9	0.004 5	0.042 1	0.001 9	0.016 7	0.048 9	0.006 5	0.027 1	0.003 7
	x_4	0.061 6	0.403 6	0.473 1	0.729 9	0.106 1	0.265 3	0.388 6	0.267 5	0.399 2	0.718 4
	x_5	0.197 3	0.567 7	0.530 0	0.507 3	0.081 7	0.339 1	0.218 8	0.200 2	0.736 4	0.634 8
	x_6	0.263 2	0.614 4	0.748 0	0.784 0	0.347 7	0.672 2	0.542 4	0.289 0	0.811 1	0.697 8
	x_7	0.315 9	0.421 3	0.538 3	0.854 1	0.275 5	0.406 5	0.201 2	0.439 6	0.788 3	0.598 8
	$[x_1, x_5, x_6, x_7]$					0.067 2					
	$[x_1, x_4, x_5, x_6, x_7]$	0.129 9	0.255 4	0.441 2	0.388 6		0.449 3	0.063 7	0.216 9	0.414 9	0.227 9
(0.1, 100)	x_1	0.258 4	0.312 8	0.612 5	0.540 4	0.170 9	0.239 3	0.415 8	0.459 7	0.840 1	0.187 2
	x_2	8e-8	3e-5	2e-2	3e-3	4e-9	6e-3	4e-4	1e-2	2e-5	4e-6
	x_3	0.101 7	0.005 5	0.046 4	0.082 3	0.000 0	0.455 6	0.002 0	0.000 8	0.066 9	0.001 3
	x_4	0.339 6	0.253 3	0.332 3	0.545 7	0.008 5	0.844 5	0.313 4	0.661 4	0.146 6	0.136 2
	x_5	0.159 0	0.649 1	0.592 7	0.716 1	0.108 4	0.957 5	0.691 0	0.753 8	0.668 0	0.232 1
	x_6	0.371 8	0.650 1	0.576 0	0.847 0	0.467 8	0.851 8	0.659 0	0.673 3	0.891 9	0.281 5
	x_7	0.161 4	0.489 5	0.346 6	0.966 9	0.187 6	0.849 7	0.542 3	0.370 4	0.671 7	0.074 7
	$[x_5, x_6, x_7]$										0.152 2
	$[x_4, x_5, x_6, x_7]$		0.125 6			0.096 1					
$[x_1, x_4, x_5, x_6, x_7]$	0.069 4		0.738 3	0.144 4		0.609 7	0.103 0	0.527 3	0.305 3		

表 4 LS-SVR 模型的不显著因子移除过程示意
Table 4 Steps of nonsignificant factors removing of LS-SVR model

被移除因子(组合)	(1)							(2)	(3)			
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_1, x_5, x_6, x_7	x_1, x_5, x_6	x_1, x_5, x_7	x_1, x_6, x_7	x_5, x_6, x_7
P -value	0.152 6	2e-07	5e-09	0.006 8	0.590 5	0.543 0	0.767 0	0.032 6	0.036 1	0.041 4	0.040 9	0.414 6

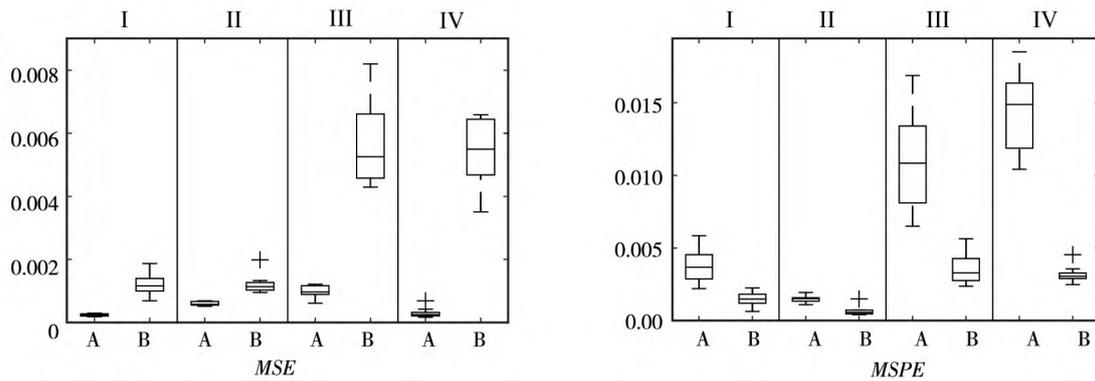
注: 此为 $f_{\text{Piston}}, \sigma = 0.05, m = 50$ 表 3 的第 1 次仿真.



“I”, “II”, “III”, “IV”对应 (σ, m) 为 $(0.5, 50)$ $(0.5, 100)$ $(1.0, 50)$ $(1.0, 100)$; “A”初始模型, “B”简化模型

图3 初始模型与简化模型的 MSE 及 MSPE 对比(f_{Camel})

Fig. 3 Comparison of MSE and MSPE for the initial and reduced models (f_{Camel})



“I”, “II”, “III”, “IV”对应 (σ, m) 为 $(0.5, 50)$ $(0.5, 100)$ $(1.0, 50)$ $(1.0, 100)$; “A”初始模型, “B”简化模型

图4 初始模型与简化模型的 MSE 及 MSPE 对比(f_{Piston})

Fig. 4 Comparison of MSE and MSPE for the initial and reduced models (f_{Piston})

4) 关于与逐步回归方法的对比. 在经典理论中, 与显著性因子筛选相关的交叉验证法主要有逐步回归法 (step wise regression, SWR)^[36], 序贯分支法等等. 后者无法包含二阶主效应, 有可能会引起残差的异方差. 因此采用含一阶、二阶主效应以及二阶交互作用效应的线性回归模型, 运行 SWR 方法. 其模型为

$$y = \beta_0 + \sum_{i=1}^6 \beta_i x_i + \sum_{i=1}^5 \sum_{j=i+1}^6 \beta_{ij} x_i x_j + \sum_{i=1}^6 \beta_{ii} x_i^2 + \varepsilon$$

将 $x_i x_j$ 以及 x_i^2 定义为新变量 x_i ($i = 7, \dots$,

27) 形成模型 $y = \beta_0 + \sum_{i=1}^{27} \beta_i x_i + \varepsilon$.

令 p_1 表示显著但未被识别为显著的因子总个数, p_{11} 表示不显著性但被识别为显著的因

子总个数; p_1 、 p_{11} 以每次仿真计, 多次仿真需重复计. $p_{sig} = 2$ 表示仿真函数中显著性因子个数, $n_{sim} = 10$ 表示仿真次数. 引入“漏判率 (miss acceptance rate, MAR)”与“误判率 (false acceptance rate, FAR)”评价方法性能 $MAR =$

$$\frac{p_1}{p_{sig} \times n_{sim}}, FAR = \frac{p_{11}}{(p - p_{sig}) \times n_{sim}}. \text{ 以 } f_{Camel}$$

仿真函数为例, 表 5 给出了 SWR 方法和所提方法的性能对比, 包括识别出的显著性因子、MAR 和 FAR. 工程实际中“漏判”影响比“误判”大. 前者会遗漏某些显著性因子, 从而无法实现对响应的有效控制; 后者则会保留某些不显著因子, 复杂度增加, 但不影响有效控制.

表 5 各类方法识别出的显著性因子及漏判率、误判率对比(f_{Camel})

Table 5 Comparison of significant factors ,MAR and FAR of different approaches(f_{Camel})

(σ, m)	(0.5, 50)		(0.5, 100)		(1.0, 50)		(1.0, 100)	
方法	SWR	本文	SWR	本文	SWR	本文	SWR	本文
第 1 次	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2, x_5, x_6]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$
第 2 次	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2, x_4, x_5]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$
第 3 次	$[x_1, x_2]$	$[x_1, x_2, x_3, x_4]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2, x_4]$	$[x_1, x_2]$
第 4 次	$[x_4]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_3, x_6]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$
第 5 次	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$				
第 6 次	$[x_1, x_2, x_4, x_5]$	$[x_1, x_2]$	$[x_1, x_2, x_3, x_4]$	$[x_1, x_2]$	$[x_1, x_2, x_3]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$
第 7 次	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_3, x_5, x_6]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$
第 8 次	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$				
第 9 次	$[x_1, x_2, x_4]$	$[x_1, x_2]$	$[x_1, x_2, x_3]$	$[x_1, x_2, x_3, x_5]$	$[x_1, x_2, x_3, x_4, x_6]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$
第 10 次	$[x_1, x_2, x_3]$	$[x_1, x_2, x_4]$	$[x_1, x_2, x_3, x_5]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$	$[x_1, x_2]$
MAR	0.1	0	0	0	0.15	0	0	0
FAR	0.125	0.075	0.175	0.050	0.275	0	0.025	0

4 实证研究

4.1 案例背景

3D 打印可以应用于汽车、航空航天、医用材料等领域,是智能制造的代表技术之一^[37]。熔融沉积成型是使用最为广泛的 3D 打印技术,其基本原理是将高分子等热塑性材料从喷嘴处挤出,分层堆积凝固形成实体。此过程的工艺参数主要包括层厚、打印速度、喷嘴温度、热床温度、填充率、最短冷却时间等等;质量特性则包括零件的尺寸精度及形状符合度等等。成型过程需经历从固态到熔融,再到冷却成型的复杂变化,使得工艺参数与制成品质量特性之间呈高度非线性相关。对于大多数 3D 打印制成品而言,翘曲(warping)和飞边(flash)是主要质量缺陷,而通过基于试验设计建模的工艺参数优化,可以较为显著性地降低翘曲程度和飞边长度。然而上述工艺参数的影响大小不一致,在质量优化之前,需要先识别出显著性因子,以降低优化控制的难度。考虑到 3D 打印过程的强非线性,这里采用所提出的 LS-SVR 建模和显著性因子筛选方法,再通过遗传算法寻优,得到优化的工艺参数值及质量特性值,并通过试验验证有效性。

试验设备为“Z-603S”型“极光尔沃”牌 3D

打印机。采用直径 1.75 mm 的聚乳酸材料进行打印。试验零件为 15 mm × 15 mm × 3 mm 的立方体。工艺参数分别为层厚/(mm): $x_1 \in [0.06, 0.20]$, 打印速度/(mm/s): $x_2 \in [20, 70]$, 喷嘴温度/(°C): $x_3 \in [190, 220]$, 热床温度/(°C): $x_4 \in [50, 70]$, 填充率/(%): $x_5 \in [20, 50]$, 最短冷却时间/(s): $x_6 \in [3, 6]$ 。质量特性为翘曲/(mm): y_{warp} 和飞边/(mm): y_{flash} 均为望小型。采用游标卡尺进行测量,以制成品水平放置时,其表面距离水平面的最大高度为 y_{warp} ,以各条边的最大飞边长度为 y_{flash} 。

4.2 试验设计样本集获取及显著性因子筛选

采用样本量 $m = 50$ 的 LHS 设计,各试验点处随机重复 $n_i \leq 3$ 次,总计 106 次,得到表 6 所示因子组合及响应值。每个响应值为翘曲和飞边长度的平均,即 $y = \frac{1}{2}(y_{\text{warp}} + y_{\text{flash}})$ 。采用所提方法,得到将各因子或因子组合移除后 LS-SVR 模型拟合不足的显著性 P-value,如图 5 所示。能够被移除的最大因子组合为 $[x_2, x_5, x_6]$ (P-value 为 0.1623) 相应的显著性因子组合为 $[x_1, x_3, x_4]$,即层厚、喷嘴温度、热床温度。据此建立简化后的 LS-SVR 模型。图 6 给出了该模型在 $x_1 - x_3$ 、 $x_1 - x_4$ 、 $x_3 - x_4$ 视角的曲面及等高线图。

表6 3D打印试验LHS设计的因子组合及响应值

Table 6 Factor combinations and responses of LHS runs for 3D printing experiments

i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	y_{i1}	y_{i2}	y_{i3}	i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}	y_{i1}	y_{i2}	y_{i3}
1	0.64	0.28	0.95	0.27	0	0.28	0.72	0.71	0.81	26	0.93	0.88	0.45	0.8	0.87	0.74	0.43	0.52	0.63
2	0.86	0.52	0.50	0.67	0.80	0.62	0.45	0.51		27	0	0.86	0	0.50	0.73	0.04	0.67	0.60	0.73
3	0.21	0.74	0.90	0.27	0.47	0.16	1.10	0.67	1.05	28	0.57	0.36	0.65	0.20	0.23	0.52	0.61	0.77	0.79
4	0.07	0.06	0.70	0.57	0.63	0.81	1.00	1.10		29	0.71	0.72	0.45	0.37	0.20	0.06	0.41	0.60	0.77
5	0.21	0.78	0.40	0.63	0.20	0.74	0.64	0.75	1.01	30	0.21	0.06	0.05	0.17	0.33	0.84	0.57		
6	0.36	0.80	0.60	0.60	0.50	0.92	1.20	0.67		31	0.93	0.56	0.85	0.80	0.83	0.44	0.53	0.51	
7	0.29	0.24	0.80	0.40	0.67	0.50	0.96			32	0.71	0.16	0.30	0.33	0.83	0.86	0.40	0.82	
8	0.86	0.36	0.15	0.07	0.97	0.69	0.68	0.64		33	0.79	0.38	0.20	0.97	0.50	0.34	0.28	0.31	0.34
9	0.07	0.48	0.35	0.50	0.37	0.86	0.78			34	0.50	0.80	0.15	0.03	0.53	0.58	0.49	0.69	0.51
10	0.57	0.62	0.95	0.33	0.07	0.55	0.82	1.01		35	0.71	0.12	0.30	0.87	0.30	0.95	0.15	0.14	0.28
11	0.07	0.68	0.35	0.17	0.03	0.36	0.78	0.85	0.97	36	0.36	0.18	0.20	0.67	0.67	0.09	0.52		
12	0.79	0.70	0.60	0.47	0.53	0.99	0.71	0.97		37	0.14	0.40	0.60	0.53	0.80	0.11	0.83	0.71	
13	0.07	0.28	0.85	0.40	0.77	0.08	1.34			38	0.36	0.22	0.10	0.70	0.37	0.01	0.41		
14	0.50	0.34	1.00	0.53	0.63	0.41	0.79	0.52		39	0.57	0.72	0.25	0.77	0.40	0.71	0.59		
15	1.00	0.92	0.60	0.87	0.37	0.98	0.44	0.66	0.44	40	0.79	1.00	0.25	0.13	0.43	0.63	0.74		
16	0.43	0.48	0.10	0.77	0.97	0.47	0.39	0.38	0.30	41	0.29	0.10	0.80	0.93	0.27	0.22	0.51	0.58	0.73
17	0	0.84	0	0.90	0.47	0.39	0.49			42	0.43	0.14	0.35	0	0.57	0.78	0.73	0.81	
18	0.14	0.18	0.55	0.07	0.97	0.23	0.96	0.96	1.13	43	0.86	0.64	0.75	0.27	0.90	0.68	0.73	0.81	1.16
19	0.43	0.58	0.50	0.23	0.07	0.91	0.75	0.57	0.84	44	0.64	0.26	0.50	0.20	0.90	0.15	0.86	0.76	
20	0.36	0.52	0.90	0.43	0.17	0.25	0.84	1.32		45	1	0.98	0.75	0.93	0.70	0.12	0.50		
21	0.21	0.92	0.70	1.00	0.70	0.44	0.78	0.75	0.66	46	0.14	0.94	0.15	0.10	0.23	0.28	0.61	0.61	
22	0.93	0	0.65	0.83	0.93	0.89	0.33			47	0.79	0.62	0.95	0.60	0.60	0.76	0.61	0.66	
23	0.64	0.44	0.45	0.43	0.73	0.52	0.43	0.58		48	0.64	0.84	0.75	0.70	0.10	0.20	0.53	0.71	
24	0.93	0.44	0.25	0.33	0.30	0.65	0.41	0.46		49	0.29	0.30	0.05	0.73	0.10	0.30	0.29	0.43	0.53
25	0.50	0.04	0.90	0.07	0.13	0.58	0.99			50	0.50	0.58	0.40	0.90	0.17	0.32	0.43		

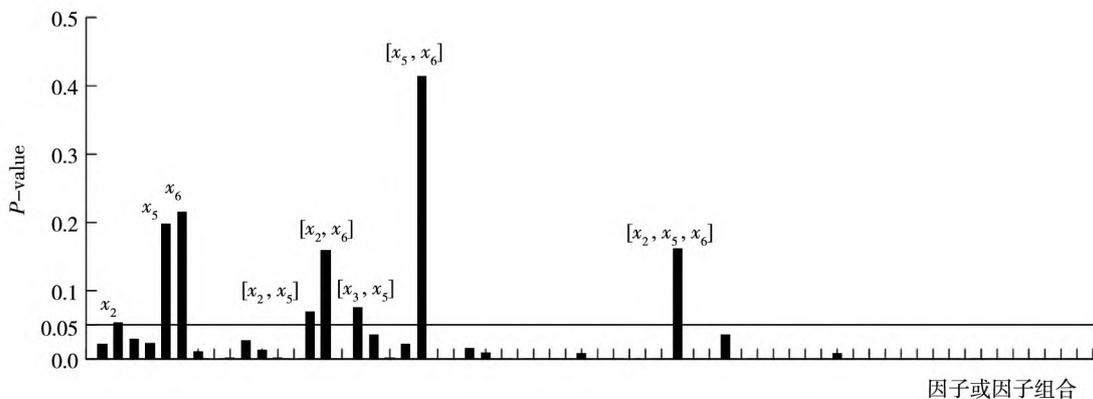


图5 3D打印过程因子(或因子组合)移除后的拟合不足显著性

Fig. 5 Significance of lack-of-fit after removing factor or factor combinations for 3D printing process

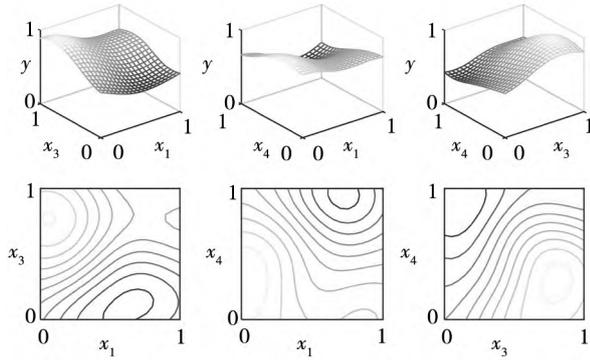


图 6 3D 打印的简化 LS-SVR 模型曲面及等高线

Fig. 6 Surface and contour plots of reduced LS-SVR model for 3D printing process

4.3 质量特性寻优

对简化模型进行遗传算法寻优, 得到参数优化值 $[x_1^* \ x_3^* \ x_4^*] = [0.691\ 9, 0.201\ 3, 0.903\ 4]$ (规范化后) 和对应的质量特性预测值 $y^* = 0.287\ 5$. 对寻优结果进行 15 次验证试验, 如表 7 所示. 为体现“显著性因子筛选”的意义, 将 $[x_1 \ x_3 \ x_4]$ 固定为优化后水平, 仅将不显著性因子 $[x_2 \ x_5 \ x_6]$ 安排为随机水平. 15 次试验质量特性的均值为 0.276 7, 方差为 0.001 4, 与预测值 y^* 接近.

表 7 3D 打印的简化 LS-SVR 模型寻优结果验证试验

Table 7 Verifying experiments of reduced LS-SVR model for 3D printing process

x_2	x_5	x_6	y^{ver}	x_2	x_5	x_6	y^{ver}
0.972 3	0.630 4	0.548 5	0.34	0.649 0	0.392 7	0.431 3	0.28
0.010 5	0.156 8	0.737 8	0.24	0.112 1	0.062 4	0.275 9	0.22
0.544 5	0.830 4	0.468 5	0.31	0.258 7	0.573 0	0.367 7	0.29
0.449 6	0.218 6	0.987 2	0.22	0.737 3	0.770 1	0.640 4	0.30
0.492 4	0.965 2	0.124 1	0.32	0.690 3	0.508 9	0.888 6	0.28
0.827 9	0.267 1	0.820 2	0.30	0.168 0	0.415 8	0.217 2	0.27
0.297 3	0.918 8	0.193 7	0.23	0.357 4	0.691 5	0.698 7	0.26
0.875 3	0.126 3	0.031 6	0.31				

注: $[x_1 \ x_3 \ x_4] = [0.691\ 9, 0.201\ 3, 0.903\ 4]$ 在试验中保持不变.

5 结果讨论

5.1 所提方法的有效性

1) 非中心 F -分布的符合性. 从图 7 (a) 和图 7

(b) 可以看出, 非中心 F -分布的经验与理论概率密度曲线的符合程度较好.

其一, 各条经验概率密度曲线均表现出典型的非中心 F -分布形态, 且与理论曲线符合性较好. 说明 $A_{Lof}^2 \approx A_{Lof}$ 时, MS_{Lof}/MS_{Pe} 的实际分布非常接近其理论分布. 也说明在超参数选择中, 经验性设定阈值 $T_{Lof} = 0.5$ 与 $T_{Overfitting} = 0.5$ 具有合理性.

其二, 经验与理论概率密度曲线的尾部符合性更高, 对于较为显著的拟合不足判断影响较小. 拟合不足意味着存在一个较大 MS_{Lof}/MS_{Pe} , 对应于曲线右偏尾部位置. 图 7 表明, 两类曲线的差异主要集中在最高点附近, 而尾部差异不大. 作为归一化, 采用各概率密度相对于近似 F -分布 (理论分布) 分位数 F_α 的曲线作为参照, 定义

$$D^2(F_\alpha) = \frac{1}{1 - F_\alpha} \int_{F_\alpha}^1 [g_e(F_\alpha) - g_t(F_\alpha)]^2 dF_\alpha$$

其中 $g_e(F_\alpha)$ 和 $g_t(F_\alpha)$ 表示 F_α 处的经验和理论概率密度, $D^2(F_\alpha)$ 表示累积到 F_α 处的两类概率密度曲线的偏差均方. 图 7 (a) 和图 7 (b) 给出了 f_{Camel} 与 f_{Piston} 仿真的 $D^2(F_\alpha)$ 曲线. 可以看出, 当 F_α 较大 (例如 >0.5) 时, $D^2(F_\alpha)$ 呈快速减小趋势.

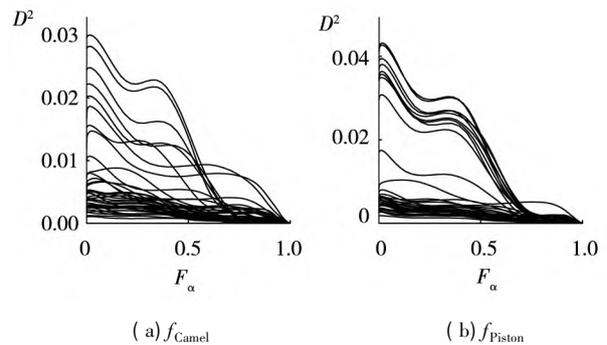


图 7 经验与理论概率密度关于 F -分布分位数的偏差变化
Fig. 7 Deviations between empirical and theoretical probability density of F -distribution quantiles

2) 显著性因子识别及简化模型结构的一致性. 表 2 和表 3 说明, 所提方法能够有效识别显著性因子, 给出可移除的不显著因子最大组合.

其一, 所提方法在识别显著性因子方面的一致性较高. f_{Camel} 的 40 次仿真, x_1, x_2 均被识别为显著, 其对应的拟合不足 P -value 均较小; 与作为白

噪声添加进模型的不显著因子 x_3, x_4, x_5, x_6 对应的 P -value 有明显差异; 40 次仿真中, 37 次(比例为 92.5%) 识别出了所有不显著因子. f_{Piston} 的 40 次仿真, x_2 有 39 次(比例为 97.5%) 被识别为显著, x_3 有 35 次(比例为 87.5%) 被识别为显著; 40 次仿真中, 能够被移除的最大因子组合均包含 $[x_5, x_6, x_7]$.

其二, 简化模型的基本结构具有一致性. 表 2 和表 3 说明, σ 和 m 不同, 被移除最大因子组合不同, 对应的简化模型也不同. f_{Camel} 的 $(\sigma, m) = (0.5, 50)$ 第 1 次仿真, 得到了含有 x_1, x_2 的简化模型; 而 $(\sigma, m) = (0.5, 50)$ 第 10 次仿真, 得到了含有 x_1, x_2, x_4 的简化模型, 但是所有 40 次仿真, 真实的显著性因子 x_1, x_2 均被无差别识别并包含于简化模型. 也就是说, 简化模型基本一致, 且描述了初始模型主要特征. 对于 f_{Piston} , 其仿真函数没有特别加入权重系数为 0 的白噪声因子, 但是 40 次仿真中, x_2, x_3 均包含于简化模型, 说明了简化模型结构的基本一致性.

3) 简化模型复杂度降低和预测性能改善. 直观理解, 如果将某些因子移除, 模型表现能力将降低. 但图 3 和图 4 说明, 与初始模型相比, 简化模型拟合误差增加, 但预测误差却明显降低.

其一, 简化模型的预测均方误差 $MSPE$ 降低且波动减小. 对于 f_{Camel} 和 f_{Piston} , 简化模型的拟合性能降低, 表现为拟合均方误差 MSE 的均值和波动性均有所增加, 但是预测均方误差 $MSPE$ 却优于初始模型, 并且其 IQR(四分位距) 也有明显减小, 说明方法预测性能也较为稳定.

其二, 对机器学习模型而言, 小的拟合误差并不意味着小的预测误差, 却有可能意味着“过拟合”, 即模型复杂度过高, 使得 \hat{y} 在相邻样本点之间产生较大“起伏”, 从而影响预测性能. 而所提方法由于移除了不显著因子, 所得简化模型的复杂性较初始模型有实质降低, 这就是其预测性能改善的根本原因. 对于使用者而言, 预测性能比拟合性能更重要. 由此不仅说明了对机器学习模型进行简化的必要性, 也验证了所提方法

的优势.

5.2 与经典线性模型显著性因子筛选方法的比较

1) 均是以 F -分布和方差分析为基础. 所提方法基于式(10)的近似 F -分布和方差分析, 与经典线性模型相比, 这也是一种统计检验方法. 因此在单次检验中, 会由于过程噪声 σ 影响, 得到不同的显著性结果, 表 2 和表 3 也说明, 最终被移除的最大因子组合不完全相同. 但在多次重复检验下, 某个因子组合将会以较大概率出现, 例如表 2 中的 $[x_3, x_4, x_5, x_6]$ 和表 3 的 $[x_5, x_6, x_7]$, 这也是统计检验方法的特点. 类似地, 经典线性模型中, 对于同一个“因子—响应”过程, 不同观测值也可能导致不同的显著性因子筛选结果.

2) 显著性因子筛选的实现方式不同. 经典线性模型直接通过式(1)系数 $\hat{\beta}_i$ 或 $\hat{\beta}_{ij}$ 的显著性检验判断 x_i 或 x_{ij} 重要程度. 这是一种“正向”方式, 可以通过精确 F -分布实现, 但是对模型形式要求严格, 难以适用于非“因子—系数”加权的形式. 此外, 如果样本量较小, 所拟合线性模型未包含某些高阶交互作用, 则无法检验其显著性. 而所提方法通过拟合不足来考察因子(组合)显著性, 这是一种“逆向”方式, 对模型结构无前提假设, 但是需要通过近似 F -分布来实现. 而这种近似对 P -value 的影响是可以接受的: 一方面, 仿真与实证表明, 由此导致的理论与实际概率密度曲线偏差很小; 另一方面, 判断显著性是依据区间(例如 P -value < 0.05) 而非单值, 而且 0.05 并非严格界限, 当 P -value 接近 0.05 时, 各类检验方法均需小心判断或做补充试验.

3) 对于样本量大小的要求不同. 经典线性模型的 F -分布要求样本量 $n \rightarrow +\infty$. 然而试验设计建模样本量一般有限且较小. 而所提方法的近似 F -分布不以 $n \rightarrow +\infty$ 为基础, 试图在有限样本条件下通过调整 LS-SVR 超参数 h^2 与 γ 来获得, 具有较强的工程实践现实意义. 再者, 近似 F -分布对于样本量的要求与 LS-SVR 模型本身的小样本

特性是一致的,更有利于体现模型性能。图 3 和图 4 也说明,通过近似 F -分布剔除不显著因子得到的简化模型,其预测性能得到了较好改善。

4) 关于因子效应的“稀疏性”、“交互作用”与“因子组合”的显著性。因子效应的“稀疏性”指试验设计中的重要因子(即显著性因子)是少数的^[38]。例如析因试验中,只有一部分因子效应呈现为统计学意义上的显著,而其他(尤其是高阶交互作用的效应)不显著。本研究也体现了这一原则。对于 3D 打印试验,总计 6 个工艺参数,能够被同时移除的有 3 个,其余为显著性因子。由于 LS-SVR 模型的特点,所提方法未涉及各因子间的交互作用。“交互作用”项指线性模型的因子乘积项(例如 x_1x_2)。LS-SVR 模型中不存在因子乘积项,而实际制造过程也不存在能够单独控制的“交互作用”项。所提方法可以从因子组合整体(而非多个因子简单加和)角度判断显著性,可以较好地适应实际过程参数优化和控制的需要。3D 试验也说明了这一点。

5) 与逐步回归方法的“漏判率”与“误判率”对比。从表 5 可以看出,所提方法的 MAR 与 FAR 均优于 SWR 方法。所提方法的 MAR 在不同 m 与 σ 组合下均为 0,而 SWR 方法最高达 0.15,意味着在 10 次仿真中有 3 个(次)显著性因子未被有效识别出。所提方法的 FAR 最高为 0.075,而 SWR 方法最高达 0.275,是本文方法的 3.67 倍。此外,当 σ 增大时,所提方法的 FAR 均变为 0,更

适合于工程实践噪声较大的情形。

6 结束语

本研究根据 LS-SVR 特点,提出了基于拟合不足检验的显著性因子筛选方法,用以在试验设计建模及参数优化中识别关键工艺参数、降低质量改进成本。利用 LS-SVR 线性光滑估计的特点,将拟合模型残差平方和分解为拟合不足与纯误差的平方和;利用重复性试验,给出了拟合不足与纯误差均方比值的近似 F -分布,构造了拟合不足检验的方差分析表;利用移除某个因子(组合)导致的拟合不足显著性 P -value 的变化,提出了显著性因子的筛选方法;还通过增加满足近似 F -分布的约束,给出了 LS-SVR 模型的超参数选择方法。仿真与实证研究表明,本研究较好地实现了具有“样本点—响应”及核函数结构的机器学习模型的显著性因子筛选,简化了模型形式,改善了预测性能。本研究有助于增强在试验设计建模及参数优化中应用 LS-SVR 模型的“可解释性”,有利于从因子(而非样本点)角度探究其影响,拓展机器学习模型在质量改进的应用。所提方法也适用于大数据集的因子显著性筛选。一般而言,大数据集缺乏严格的重复性样本集,可以考虑将某一样本点附近较小邻域内的样本点近似作为重复测量。而邻域半径的确定、近似重复性对于显著性检验的影响等等,是需要进一步研究的问题。

参考文献:

- [1] Miller T. Explanation in artificial intelligence: Insights from the social sciences [J]. *Artificial Intelligence*, 2019, 267(1): 1–38.
- [2] 曾大军, 李一军, 唐立新, 等. 决策智能理论与方法研究 [J]. *管理科学学报*, 2021, 24(8): 18–25.
Zeng Dajun, Li Yijun, Tang Lixin, et al. Decision intelligence research: Theory and methods [J]. *Journal of Management Sciences in China*, 2021, 24(8): 18–25. (in Chinese)
- [3] Wang G, Shao M, Cui Q, et al. Dual-response optimization for lifetime improvement experimental designs [J]. *Computers & Industrial Engineering*, 2021, 158(7): 107437.
- [4] Samson J, Murugan K, Abhra P, et al. Optimization of process parameters using response surface methodology: A review [J]. *Materials Today: Proceedings*, 2021, 37(2): 1301–1304.

- [5]Montgomery D C. Design and Analysis of Experiments[M]. 10th Edition. New York: John Wiley & Sons ,2019.
- [6]王 宁,张 帅,刘玉敏,等. 多工序串并联制造过程关键质量特性识别[J]. 系统工程学报,2019,34(6): 855 - 866.
- Wang Ning ,Zhang Shuai ,Liu Yumin ,et al. Identification the key quality characteristics in multistage series-parallel manufacturing process [J]. Journal of Systems Engineering ,2019 ,34(6) : 855 - 866. (in Chinese)
- [7]Myers R H ,Montgomery D C ,Vining G G ,et al. Response surface methodology: A retrospective and literature survey [J]. Journal of Quality Technology ,2004 ,36(1) : 53 - 78.
- [8]Vining G G ,Bohn L L. Response surfaces for the mean and variance using a nonparametric approach [J]. Journal of Quality Technology ,1998 ,30(3) : 282 - 291.
- [9]Wan W ,Birch J B. A semiparametric technique for the multi-response optimization problem [J]. Quality & Reliability Engineering International ,2011 ,27(1) : 47 - 59.
- [10]Osocha P ,Podgórski J. The response surface methodology revisited-comparison of analytical and non-parametric approaches [J]. Production Engineering Archives ,2018 ,20(1) : 49 - 53.
- [11]Fang K T ,Lin Y ,Peng H. A new type of robust designs for chemometrics and computer experiments [J]. Chemometrics and Intelligent Laboratory Systems ,2022 ,221(2) : 104474.
- [12]Lee S H. Optimization of cold metal transfer-based wire arc additive manufacturing processes using gaussian process regression [J]. Metals ,2020 ,10(4) : 461 - 473.
- [13]Heinisch J ,Lockner Y ,Hopmann C. Comparison of design of experiment methods for modeling injection molding experiments using artificial neural networks [J]. Journal of Manufacturing Processes ,2021 ,61(1 - 4) : 357 - 368.
- [14]Bensingh R J ,Machavaram R ,Boopathy R S ,et al. Injection molding process optimization of a bi-aspheric lens using hybrid artificial neural networks (ANNs) and particle swarm optimization (PSO) [J]. Measurement: Journal of the International Measurement Confederation ,2019 ,134(2) : 359 - 374.
- [15]Vapnik V N . Complete statistical theory of learning [J]. Automation and Remote Control ,2019 ,80(11) : 1949 - 1975.
- [16]Vapnik V ,Izmailov R. Reinforced SVM method and memorization mechanisms [J]. Pattern Recognition ,2021 ,119(11) : 108018.
- [17]崔庆安,何 楨,崔 楠. 基于 SVM 的 RSM 模型拟合方法研究 [J]. 管理科学学报,2008 ,11(1) : 31 - 41.
- Cui Qingan ,He Zhen ,Cui Nan. SVM-based RSM model fitting approach [J]. Journal of Management Sciences in China ,2008 ,11(1) : 31 - 41. (in Chinese)
- [18]孙 林,杨世元. 基于正交试验和支持向量机的参数设计方法研究 [J]. 中国机械工程,2011 ,22(8) : 971 - 975.
- Sun Lin ,Yang Shiyuan. Study on parameter design based on orthogonal test and support vector machine [J]. China Mechanical Engineering ,2011 ,22(8) : 971 - 975. (in Chinese)
- [19]向国齐,殷国富. 基于支持向量机和粒子群算法的稳健优化 [J]. 机械设计与研究,2016 ,32(6) : 6 - 9.
- Xiang Guoqi ,Yin Guofu. Robust optimization based on support vector machine and particle swarm algorithm [J]. Machine Design & Research ,2016 ,32(6) : 6 - 9. (in Chinese)
- [20]曲兴田,马 强,李笑男,等. 柔性薄板装配不等间距定形焊点布置优化 [J]. 计算机集成制造系统,2016 ,22(8) : 1857 - 1866.
- Qu Xingtian ,Ma Qiang ,Li Xiaonan ,et al. Layout optimization of shaping welding spot unequal pitch for sheet metal assembly [J]. Computer Integrated Manufacturing Systems ,2016 ,22(8) : 1857 - 1866. (in Chinese)
- [21]Zhou X J ,Ma Y Z ,Tu Y L ,et al. Ensemble of surrogates for dual response surface modeling in robust parameter design [J]. Quality and Reliability Engineering International ,2013 ,29(2) : 173 - 197.
- [22]汪建均,马义中,欧阳林寒,等. 多响应稳健参数设计的贝叶斯建模与优化 [J]. 管理科学学报,2016 ,19(2) : 85 - 94.

- Wang Jianjun, Ma Yizhong, Ouyang Linhan, et al. Bayesian modeling and optimization of multi-response robust parameter design [J]. *Journal of Management Sciences in China*, 2016, 19(2): 85–94. (in Chinese)
- [23] Liu L, Ma Y, Park C, et al. Robust sequential bifurcation for simulation factor screening under data contamination [J]. *Computers & Industrial Engineering*, 2019, 129(3): 102–112.
- [24] 刘丽君, 马义中, 欧阳林寒, 等. 基于位置和散度效应因子分类筛选及仿真试验 [J]. *系统工程理论与实践*, 2017, (37): 2854–2862.
- Liu Lijun, Ma Yizhong, Ouyang Linhan, et al. Screening for location and dispersion effects simultaneously by sequential bifurcation [J]. *Systems Engineering: Theory & Practice*, 2017, (37): 2854–2862. (in Chinese)
- [25] 施文, 周程, 刘慧. 基于动态效应的序贯分支验证仿真实验及应用 [J]. *系统工程理论与实践*, 2016, 36(8): 2003–2012.
- Shi Wen, Zhou Cheng, Liu Hui. Simulation experiments for validating sequential bifurcation: A dynamic-effect method [J]. *Systems Engineering: Theory & Practice*, 2016, 36(8): 2003–2012. (in Chinese)
- [26] Xue X F, Wang Y Z, Lu C, et al. Sinking velocity impact-analysis for the carrier-based aircraft using the response surface method-based improved kriging algorithm [J]. *Advances in Materials Science & Engineering*, 2020, 6(1): 1–13.
- [27] 崔庆安. 复杂作用关系过程的区域显著性实验设计及全局建模方法 [J]. *系统工程理论与实践*, 2013, 33(9): 2249–2262.
- Cui Qingan. Design of experiments and global modeling approach based on sub-domains significance for complicated relationship process [J]. *Systems Engineering: Theory & Practice*, 2013, 33(9): 2249–2262. (in Chinese)
- [28] 崔庆安, 季泽, 段焕姣. 基于样本点显著性及嵌套正交设计的序贯设计及建模方法 [J]. *系统工程理论与实践*, 2019, 39(9): 2398–2411.
- Cui Qingan, Ji Ze, Duan Huanjiao. Sequential design and modeling based on significance of samples and nested orthogonal design [J]. *Systems Engineering: Theory & Practice*, 2019, 39(9): 2398–2411. (in Chinese)
- [29] Kleijnen J P C. Regression and kriging metamodels with their experimental designs in simulation: A review [J]. *European Journal of Operational Research*, 2017, 256(1): 1–16.
- [30] Denimal E, Sinou J J. Advanced kriging-based surrogate modelling and sensitivity analysis for rotor dynamics with uncertainties [J]. *European Journal of Mechanics-A/Solids*, 2021, 90(6): 1–20.
- [31] Hong X, Mitchell R, Fatta G D. Simplex basis function based sparse least squares support vector regression [J]. *Neuro-computing*, 2019, 330(2): 394–402.
- [32] Bishwas A K, Mani A, Palade V. Gaussian kernel in quantum learning [J]. *International Journal of Quantum Information*, 2020, 18(15): 2050006.
- [33] Montgomery D C, Peck E A, Vining G G. *Introduction to Linear Regression Analysis* [M]. 5th ed., New Jersey: John Wiley & Sons, 2012.
- [34] Campana E F, Diez M, Iemma U, et al. Derivative-free global ship design optimization using global/local hybridization of the direct algorithm [J]. *Optimization & Engineering*, 2016, 17(1): 127–156.
- [35] Moon H, Dean A M, Santner T J. Two-stage sensitivity-based group screening in computer experiments [J]. *Technometrics*, 2012, 54(4): 376–387.
- [36] Barbara Z, Szymon J. Fast stepwise regression based on multidimensional indexes [J]. *Information Sciences*, 2021, 549(3): 288–309.
- [37] Browne M P, Redondo E, Pumera M. 3D printing for electrochemical energy applications [J]. *Chemical Reviews*, 2020, 120(5): 2783–2810.
- [38] 汪建均, 屠雅楠, 马义中. 结合 SUR 与因子效应原则的多响应质量设计 [J]. *管理科学学报*, 2020, 23(12): 12–29.

Wang Jianjun , Tu Yanan , Ma Yizhong. Multi-response quality design integration SUR models with factorial effect principles [J]. Journal of Management Sciences in China , 2020 , 23(12) : 12 - 29. (in Chinese)

Significant factors screening of LS-SVR for experimental design and parameters optimization

CUI Qing-an¹ , CUI Nan^{2*}

1. School of Economics and Management , Shanghai Maritime University , Shanghai 201306 , China;
2. Economics and Management School , Wuhan University , Wuhan 430072 , China

Abstract: The least squares support vector regression (LS-SVR) in experimental design modeling and parameters optimization is accompanied by poor interpretability and difficulties in significant factors screening. This paper proposes an under-fitting-test-based significant factors screening approach for the LS-SVR model. Firstly , the “residual sum of squares” is decomposed into the sum squares of “lack-of-fit” and “pure error”. Then the nearly non-central F -distribution of the ratio between the mean square of “lack-of-fit” and “pure error” is given. Consequently , a two-stage factor screening method is developed , which infers the significance of a certain factor (or combination) by investigating the change of the significance of model’s under-fitting after factor removing. The simulation and case studies show that the proposed approach can enhance the statistical inference interpretability of LS-SVR , screen the significant factors effectively and obtain a reduced LS-SVR model with lower complexity and higher predictive performance. Furthermore , the proposed approach is conducive to increase the efficiency of experimental design and parameters optimization , and to reduce the cost of quality improvement.

Key words: factors screening; parameters optimization; experimental design; lack-of-fit test; least squares support vector regression

附录 A 关于矩阵 M 的性质

1) M 的幂等性. 因为 $\mathbf{I}_{n_i}^T \mathbf{I}_{n_i} = n_i$, 所以 $\left(\frac{\mathbf{I}_{n_i}^T \mathbf{I}_{n_i}}{n_i} \right)^2 = \frac{\mathbf{I}_{n_i}^T \mathbf{I}_{n_i}}{n_i}$. 于是对于 n 阶分块矩阵 M , 有

$$M^2 = \begin{pmatrix} \frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} \end{pmatrix}^2 = \begin{pmatrix} \left(\frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} \right)^2 & & 0 \\ & \ddots & \\ 0 & & \left(\frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} \right)^2 \end{pmatrix} = \begin{pmatrix} \frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} \end{pmatrix} = M$$

2) M 的迹 $\text{tr}(M) = m$. 因为

$$\text{tr}\left(\frac{\mathbf{I}_{n_i}^T \mathbf{I}_{n_i}}{n_i}\right) = \frac{1}{n_i} \text{tr}(\mathbf{I}_{n_i}^T \mathbf{I}_{n_i}) = \frac{n_i}{n_i} = 1 , i = 1 , \dots , m$$

$$\text{所以 } \text{tr}(M) = \text{tr}\left(\frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1}\right) + \dots + \text{tr}\left(\frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m}\right) = 1 + \dots + 1 = m$$

3) $ML = L$. 根据 LS-SVR 理论, 式(8)中 L 的结构为 $L = \begin{pmatrix} L_{1,1} & & L_{1,m} \\ & \ddots & \\ L_{m,1} & & L_{m,m} \end{pmatrix}$, 其中 $L_{i,j} = \begin{pmatrix} l_{i,j} & & l_{i,j} \\ & \ddots & \\ l_{i,j} & & l_{i,j} \end{pmatrix} = l_{i,j} \mathbf{I}_{n_i} \mathbf{I}_{n_j}^T$,
 $i, j = 1, \dots, m$; 则 $\frac{\mathbf{I}_{n_i} \mathbf{I}_{n_i}^T}{n_i} L_{i,j} = \frac{\mathbf{I}_{n_i} \mathbf{I}_{n_i}^T l_{i,j} \mathbf{I}_{n_i} \mathbf{I}_{n_j}^T}{n_i} = l_{i,j} \frac{1}{n_i} (\mathbf{I}_{n_i}^T \mathbf{I}_{n_i}) \mathbf{I}_{n_j}^T = l_{i,j} \mathbf{I}_{n_i} \mathbf{I}_{n_j}^T = L_{i,j}$,

$$ML = \begin{pmatrix} \frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} \end{pmatrix} \begin{pmatrix} L_{1,1} & & L_{1,m} \\ & \ddots & \\ L_{m,1} & & L_{m,m} \end{pmatrix} = \begin{pmatrix} \frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} L_{1,1} & & \frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} L_{1,m} \\ & \ddots & \\ \frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} L_{m,1} & & \frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} L_{m,m} \end{pmatrix} = L$$

4) $ME(Y) = E(Y)$. 因为 $Y = [Y_1^T, \dots, Y_m^T]^T$, 于是 $E(Y) = [E(Y_1^T), \dots, E(Y_m^T)]^T$. 又因为 $y_{i,j} \sim N(f(x_i), \sigma^2)$, 所以 $E(y_{i,j}) = f(x_i) = E(y_i)$, 且 $E(Y_i) = [E(y_{i,1}), \dots, E(y_{i,n_i})]^T = E(y_i) \mathbf{I}_{n_i}$. 于是 $\frac{1}{n_i} \mathbf{I}_{n_i} \mathbf{I}_{n_i}^T E(Y_i) = \frac{1}{n_i} \mathbf{I}_{n_i} \mathbf{I}_{n_i}^T E(y_i) = \mathbf{I}_{n_i} E(y_i) = E(Y_i)$. 则

$$ME(Y) = \begin{pmatrix} \frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} \end{pmatrix} \begin{pmatrix} E(Y_1) \\ \vdots \\ E(Y_m) \end{pmatrix} = \begin{pmatrix} \frac{\mathbf{I}_{n_1} \mathbf{I}_{n_1}^T}{n_1} E(Y_1) \\ \vdots \\ \frac{\mathbf{I}_{n_m} \mathbf{I}_{n_m}^T}{n_m} E(Y_m) \end{pmatrix} = [E(Y_1^T), \dots, E(Y_m^T)]^T = E(Y)$$

附录 B 关于 $A_{\text{Lof}}^2 \approx A_{\text{Lof}}$ 的条件

因为 $A_{\text{Lof}} = M - 2L + L^2$, 由式(5)及 $Z = \Omega + \frac{1}{\gamma} I$ 可得 $L = I - \frac{1}{\gamma} (Z^{-1} - Z^{-1} \frac{1}{c} \mathbf{I}_n \mathbf{I}_n^T Z^{-1})$. 则

$$A_{\text{Lof}}^2 - A_{\text{Lof}} = -\frac{1}{\gamma^2} (Z^{-1} - Z^{-1} \frac{1}{c} \mathbf{I}_n \mathbf{I}_n^T Z^{-1})^2 - \frac{1}{\gamma^4} (Z^{-1} - Z^{-1} \frac{1}{c} \mathbf{I}_n \mathbf{I}_n^T Z^{-1})^4$$

因此当 $\gamma^2 \rightarrow +\infty$ 有 $\frac{1}{\gamma^2} \rightarrow 0$ 和 $\frac{1}{\gamma^4} \rightarrow 0$, $A_{\text{Lof}}^2 - A_{\text{Lof}} \rightarrow 0$. 于是 $\forall \xi > 0, \exists \delta_\gamma > 0$, 使得 $\gamma^2 > \delta_\gamma$ 时, $A_{\text{Lof}}^2 - A_{\text{Lof}} \rightarrow 0$.

因此, 由 1-范数的定义, 当 γ 足够大时, 有 $\|A_{\text{Lof}}^2 - A_{\text{Lof}}\|_1 \rightarrow 0$, 且 $A_{\text{Lof}}^2 \approx A_{\text{Lof}}$.