基于回归模型与降维技术的矿山  
监测数据分析与优化预测研究

摘要

随着数据分析在各个领域的广泛应用，尤其在矿山监测领域，海量数据的存储、处理和建模成为亟待解决的关键问题。本研究旨在通过高效的回归模型与降维技术，深入分析矿山监测数据并进行优化预测。

在问题一中，数据变换与误差分析。我们建立了多元回归模型，并对比了不同的回归方法，包括线性回归、Lasso回归、支持向量回归（SVR）和决策树回归。通过比较不同模型的性能指标（如R^2和 MSE），我们发现决策树回归在预测精度上表现最佳，适用于处理复杂的非线性数据

在问题二中，要求我们对数据进行压缩和还原，使用降维技术并进一步探讨其对数据质量的影响。首先对数据进行KMO检验，判定出数据集各指标之间并无关联关系，本题后续建模将会采用非线性降维模型。我们使用非负矩阵分解（NMF）将标准化后的数据降到指定维度，遍历各种降维尺度，最终选择接指定维度k=15，计算数据的压缩效率，通过比较原始数据与降维后数据的存储大小，得到压缩比和存储空间节省率。然后，使用低维表示矩阵和基矩阵进行数据还原，重构出数据矩阵，并计算还原数据与原始数据之间的均方误差（MSE），评估还原的准确度。

在问题三中，我们分析了矿山监测数据中的噪声问题，并通过去噪与标准化处理，提高了模型的预测精度。采用不同的去噪方法，包括简单移动平均（SMA）、加权移动平均（WMA）、中值滤波、高斯滤波和小波去噪，对数据进行处理。接下来，针对去噪后的数据，应用三种回归模型：Lasso回归、支持向量回归（SVR）和决策树回归，计算每个模型的拟合优度R^2和均方误差（MSE）。最后，使用可视化手段绘制实际值与预测值的散点图、残差图和模型比较图，以展示各模型的性能和评估结果。

问题四,主要聚焦于优化回归模型的自适应调整。我们对输入数据进行标准化处理，将数据归一化到 [0, 1] 范围。然后，添加常数项作为截距项，为支持向量回归（SVR）模型做准备。使用网格搜索方法对SVR的超参数C和 ϵ\epsilonϵ 进行调优，遍历多个参数组合，计算每组参数下的R^2值，并选择最优的参数组合。随后，使用最佳参数训练SVR模型，计算模型的均方误差（MSE）和平均预测误差，并进行交叉验证，最终输出模型的评估结果。

在问题五中，我们探讨了降维与重构之间的平衡，使用NMF降维后进行数据重构，并分析了重构数据与目标变量之间的关系。使用非负矩阵分解（NMF）将数据降至15 维，针对降维后的数据，建立并评估四种回归模型：线性回归、Lasso 回归、支持向量回归（SVR）和决策树回归。每个模型的性能通过计算R^2 和均方误差（MSE）来评估。最后，使用条形图进行模型性能的可视化对比，展示各回归模型的拟合优度与预测误差。

**关键词**：矿山监测数据、回归模型、降维技术、支持向量回归、自适应调整

目 录

[一、 问题重述 5](#_Toc13371)

[1.1 问题背景 5](#_Toc4575)

[1.2 问题回顾 6](#_Toc840)

[1.3 研究现状 6](#_Toc9216)

[二、 问题分析 8](#_Toc1262)

[2.1 问题一分析 8](#_Toc11367)

[2.2 问题二分析 8](#_Toc21584)

[2.2 问题三分析 8](#_Toc31621)

[2.2 问题四分析 8](#_Toc31561)

[三、 模型假设 8](#_Toc25924)

[四、 符号说明 9](#_Toc20275)

[五、 模型的建立与求解 10](#_Toc11956)

[5.1 数据变换与误差分析 10](#_Toc14708)

[5.1.1 数据预处理 10](#_Toc22999)

[5.1.2 线性模型 16](#_Toc16758)

[5.1.3 其他模型对比 20](#_Toc20531)

[5.2 矿山监测数据压缩与还原 24](#_Toc31530)

[5.2.1 数据预处理 24](#_Toc12687)

[5.2.2 多模型对比 25](#_Toc32756)

[5.3 矿山监测数据噪声去除与模型拟合 27](#_Toc12094)

[5.3.1 数据预处理 27](#_Toc17506)

[5.3.2 去噪模型精度对比 29](#_Toc1183)

[5.3.3 基于多种去噪下的回归模型 31](#_Toc19166)

[5.4 矿山监测数据建模与自适应参数调整 36](#_Toc4462)

[5.5 矿山监测数据噪声去除与模型拟合 39](#_Toc10050)

[六、 模型总结 41](#_Toc7027)

[6.1 模型优点 41](#_Toc6781)

[6.2 模型缺点 42](#_Toc8688)

[6.3 模型推广 43](#_Toc1169)

[七、 参考文献 43](#_Toc7600)

[八、 附录 44](#_Toc19328)

# 问题重述

## 问题背景

随着科技的飞速进步，尤其是在遥感技术、传感器技术、物联网（IoT）技术和人工智能（AI）技术的推动下，矿山监测技术得到了显著提升。矿山监测数据的获取方式已经不仅限于传统的人工检查和手工记录，现代技术手段使得矿山的实时监测成为可能。传感器、无人机、遥感卫星等设备的应用，使得矿山的各项数据得到了高频率和高精度的采集。这些数据不仅涵盖了矿山的地质特征、矿物储量、开采过程、矿区环境等信息，还具备了高空间分辨率、多时相特性和多维属性等复杂特点。

这些海量且多样化的数据为矿山行业带来了前所未有的机会和挑战。首先，矿山监测数据的高空间分辨率和多时相特性使得可以在不同时间和空间维度上对矿山的情况进行动态监控与分析。这为地质特征提取、储量估算、矿区环境监测、矿山安全预警等提供了重要的数据支撑。通过对矿山地质数据的精准提取，能更加精确地评估矿区的资源储量，并有效预测矿山开采过程中的潜在风险，进而实现更科学的决策支持，优化矿山开采与管理。

然而，随着监测技术的进步，数据量也呈现出爆炸式增长的趋势。尤其是矿山监测数据通常来源于多种异构数据源，如遥感图像、传感器数据、气象数据、环境监测数据等，这些数据之间不仅在数据格式、采集频率、精度等方面存在显著差异，而且有时还需要结合多种算法和模型进行融合与分析。如何处理和利用这些大规模、多源异构的数据，成为当前矿山监测领域面临的一大挑战。

一方面，海量数据的存储、传输与处理都需要大量的计算资源和高效的技术支持。数据的传输延迟、存储容量的不足以及计算能力的局限，都直接影响到矿山监测的实时性和准确性。尤其是在实际应用中，矿山监测数据通常需要跨越多个设备、多个平台和多个部门进行处理，这就使得数据传输的效率和安全性问题变得尤为复杂。如何在保证数据处理效率和精度的前提下，解决海量数据的存储和传输问题，是当前矿山监测技术亟待解决的重要课题。

另一方面，矿山监测数据的多源异构性带来了融合与分析上的难题。不同来源的数据在采集方式、数据格式、时间尺度等方面存在差异，这就要求数据分析模型能够对这些异构数据进行有效的融合。传统的数据处理方法往往无法满足实时、高效和高精度的需求，而现代的数据融合与分析技术（如机器学习、深度学习、大数据分析等）虽然在一定程度上可以提升数据处理的精度和效率，但其高计算需求和复杂的算法实现，仍然是一个挑战。因此，如何在高效处理海量数据的同时，保证数据分析的精度，仍然是矿山监测技术需要克服的一个难题。

此外，矿山监测数据的多维属性特征使得数据分析更加复杂。例如，矿山环境监测数据通常涉及多个维度，如空气质量、噪音、温湿度、光照强度等，这些数据维度不仅之间存在内在联系，还与矿山开采过程、地质结构等因素有着紧密的关联。如何有效地提取和利用这些多维数据的相关性，以提升矿山监测的精准性和可预测性，是数据分析中需要解决的核心问题之一。

总的来说，矿山监测技术虽然为矿山行业的智能化管理提供了强有力的数据支持，但海量数据的存储、传输与处理，以及多源异构数据的融合与分析，仍然是当前技术发展面临的重要瓶颈。为应对这些挑战，未来的矿山监测技术需要在数据处理效率、精度和智能化水平上进一步提升，推动矿山行业朝着更加智能、高效、安全的方向发展。

## 问题回顾

问题1：

根据附件1中的数据，建立一个数学模型对数据A进行变换，使得变换后的结果与原数据尽量接近。计算变换后的结果与原数据之间的误差，并探讨误差来源（如噪声、模型误差等）对结果的影响。

问题2：

针对附件2提供的矿山监测数据，构建数据压缩模型，进行降维处理，并计算压缩效果（如压缩比和存储空间节省率等）。随后，建立还原模型对降维后的数据进行恢复，分析降维与恢复过程对数据质量的影响，并提供还原数据的准确度（MSE应低于0.005）以及误差来源的分析，确保压缩效率和还原质量的平衡。

问题3：

在矿山监测数据分析中，经常需要处理噪声问题。分析附件3中的两组数据，对数据X进行去噪和标准化处理，建立X与Y之间的数学关系模型，并计算模型拟合度。进行统计检验，确保模型具备良好的解释能力，并详细说明数据预处理、建模及误差分析的过程。

问题4：

根据附件4中的两组矿山监测数据，构建数学模型分析X与Y之间的关系，设计自适应调整算法来优化模型的拟合度。分析自适应参数与模型拟合度之间的关系，并计算模型的平均预测误差，评估模型的稳定性和适用性。

问题5：

针对矿山监测的高维数据，进行降维处理，并为降维后的数据建立重构模型，以恢复数据的主要特征并保持可解释性。探讨降维与重构之间的平衡关系，构建数学模型处理附件5中的数据X，评估所构建模型的效果，特别是其泛化能力和相关算法的复杂度分析。

## 研究现状

随着矿山监测技术的不断进步，相关领域的研究工作也日益深入。在过去几十年里，随着遥感技术、物联网技术、大数据分析技术等的迅猛发展，矿山监测数据的获取、存储、传输、处理等方面的技术不断得到提升。然而，面对海量、多源异构的数据，如何高效、精准地处理和分析这些数据，依然是当前研究的热点问题。

1. 矿山监测数据获取技术的进展

矿山监测数据获取技术在近年来有了显著的进步。遥感技术作为矿山监测的重要手段，能够通过卫星、无人机等平台对矿区进行高分辨率的监测，为地质调查、矿区环境监控、矿山安全管理等提供了大量实时、精确的数据。近年来，随着遥感卫星的发射频率不断增加，遥感影像的时效性和空间分辨率也有了显著提升。例如，Sentinel-1和Landsat系列卫星提供了高频率和高精度的影像数据，有助于对矿山资源变化和环境影响的动态监控。

同时，随着物联网技术的发展，传感器技术逐渐应用于矿山监测。矿区环境、设备状态、矿山开采过程中的各类数据，如气温、湿度、振动、气体浓度等，都可以通过传感器进行实时采集。这些数据能够反映矿山作业过程中的关键变化，为矿山安全预警、环境监控、生产调度等提供了科学依据。

此外，近年来，无人机和机器人技术也在矿山监测中得到了广泛应用。无人机能够在矿区复杂的环境中进行精确的拍摄，提供矿山地表的高清影像，尤其在一些危险区域的监测中，无人机的使用提高了数据获取的安全性和效率。

2. 数据融合与处理技术的现状

矿山监测数据的多源异构性使得数据融合与处理成为研究中的重要课题。传统的矿山数据分析方法大多依赖于单一数据源，无法有效融合多源数据的信息，这限制了数据分析的深度与精度。近年来，随着大数据技术和机器学习算法的广泛应用，矿山数据的处理技术得到了显著发展。

在数据融合方面，许多研究聚焦于如何将不同来源、不同格式、不同时间尺度的数据进行有效的融合。例如，针对遥感图像与传感器数据的融合，研究者提出了基于图像处理与信号处理相结合的融合方法，通过多层次的特征提取与数据融合模型，提高了对矿山资源和环境的监测精度。此外，机器学习与深度学习技术也被广泛应用于矿山数据的分析与处理中。例如，卷积神经网络（CNN）被用于遥感图像的特征提取与分类，支持向量机（SVM）则被用于矿区环境数据的识别与预测。

在数据存储与处理效率上，云计算与边缘计算技术的引入，也为矿山监测数据的存储和处理提供了有力的支持。云计算平台能够提供强大的计算资源和高效的数据存储能力，为海量矿山监测数据的处理提供了保障。而边缘计算则通过在数据采集源附近进行实时计算，减少了数据传输的延迟，提升了系统响应速度。

3. 数据分析与决策支持技术

在数据分析方面，近年来的研究更多关注如何通过智能算法进行精准的决策支持。矿山安全预警、地质灾害监测、资源估算等问题需要依赖于复杂的数学模型和分析算法。传统的统计方法已难以满足高精度、高效率的需求，因此，人工智能和机器学习的应用成为了研究的热点。

例如，基于深度学习的时序数据预测方法可以用于矿山开采过程中设备状态的监控与故障预警，结合传感器数据和历史记录，预测设备的运行状况，为设备维护提供支持。在矿山地质勘探中，结合遥感数据与地质数据，利用数据挖掘和模式识别技术，可以精确估算矿区的资源储量，并进行动态调整。

此外，近年来，智能决策支持系统（DSS）在矿山管理中的应用逐渐增多。这些系统能够通过对矿山监测数据的实时分析，为矿山管理者提供决策建议，帮助优化矿山开采计划、环境保护措施和安全管理策略。

4. 面临的挑战与未来研究方向

尽管矿山监测技术和数据分析方法在近年来取得了显著进展，但仍面临诸多挑战。首先，数据量庞大且多样化，如何实现高效的数据存储与实时处理仍是一个亟待解决的问题。其次，数据的高质量处理要求更为先进的算法和技术，尤其是在大数据、人工智能领域，如何提升算法的效率和精度是未来研究的重点。

此外，数据安全与隐私保护也成为矿山监测技术发展中的一个重要议题。随着数据采集与分析系统的逐渐联网，如何确保矿山监测数据的安全性，防止数据泄露和篡改，也是当前需要关注的问题。

# 问题分析

## 2.1 问题一分析

在问题一中，数据变换与误差分析。我们建立了多元回归模型，并对比了不同的回归方法，包括线性回归、Lasso回归、支持向量回归（SVR）和决策树回归。通过比较不同模型的性能指标（如R^2和 MSE），我们发现决策树回归在预测精度上表现最佳，适用于处理复杂的非线性数据

## 2.2 问题二分析

在问题二中，要求我们对数据进行压缩和还原，使用降维技术并进一步探讨其对数据质量的影响。首先对数据进行KMO检验，判定出数据集各指标之间并无关联关系，本题后续建模将会采用非线性降维模型。我们使用非负矩阵分解（NMF）将标准化后的数据降到指定维度，遍历各种降维尺度，最终选择接指定维度k=15，计算数据的压缩效率，通过比较原始数据与降维后数据的存储大小，得到压缩比和存储空间节省率。然后，使用低维表示矩阵和基矩阵进行数据还原，重构出数据矩阵，并计算还原数据与原始数据之间的均方误差（MSE），评估还原的准确度。

## 2.3 问题三分析

在问题三中，我们分析了矿山监测数据中的噪声问题，并通过去噪与标准化处理，提高了模型的预测精度。采用不同的去噪方法，包括简单移动平均（SMA）、加权移动平均（WMA）、中值滤波、高斯滤波和小波去噪，对数据进行处理。接下来，针对去噪后的数据，应用三种回归模型：Lasso回归、支持向量回归（SVR）和决策树回归，计算每个模型的拟合优度R^2和均方误差（MSE）。最后，使用可视化手段绘制实际值与预测值的散点图、残差图和模型比较图，以展示各模型的性能和评估结果。

## 2.4 问题四分析

问题四,主要聚焦于优化回归模型的自适应调整。我们对输入数据进行标准化处理，将数据归一化到 [0, 1] 范围。然后，添加常数项作为截距项，为支持向量回归（SVR）模型做准备。使用网格搜索方法对SVR的超参数C和 ϵ\epsilonϵ 进行调优，遍历多个参数组合，计算每组参数下的R^2值，并选择最优的参数组合。随后，使用最佳参数训练SVR模型，计算模型的均方误差（MSE）和平均预测误差，并进行交叉验证，最终输出模型的评估结果。

在问题五中，我们探讨了降维与重构之间的平衡，使用NMF降维后进行数据重构，并分析了重构数据与目标变量之间的关系。使用非负矩阵分解（NMF）将数据降至15 维，针对降维后的数据，建立并评估四种回归模型：线性回归、Lasso 回归、支持向量回归（SVR）和决策树回归。每个模型的性能通过计算R^2 和均方误差（MSE）来评估。最后，使用条形图进行模型性能的可视化对比，展示各回归模型的拟合优度与预测误差。

# 模型假设

为了方便模型的建立与模型的可行性，我们这里首先对模型提出一些假设，使得模型更加完备，预测的结果更加合理。

1. 假设数据中可能存在一定的噪声和异常值，但这些噪声对模型的整体影响是可控的，且不至于使得回归模型失效。
2. 题目给出的数据涉及矿山监测的不同方面，包括环境数据、地质数据等。假设这些特征变量（无论是空间分辨率、时间特征，还是维度属性）能够有效代表目标变量
3. 对于回归模型的评估，通常假设模型的残差（预测值与真实值的差异）是独立的，并且遵循正态分布。

# 符号说明

为了方便我们模型的建立与求解过程 ，我们这里对使用到的关键符号进行以下说明：

|  |  |
| --- | --- |
| 符号 | 说明 |
|  | 原始输入数据矩阵，大小为 ，表示 10000 个样本和 100 个特征。 |
|  | 目标变量的列向量，大小为 ，表示对应于每个样本的真实值。 |
|  | 数据归一化后的矩阵，大小为 ，用于标准化处理过的输入数据。 |
|  | NMF 降维后的维度，即选择保留的特征维度数量，设置为 50 。 |
| W | 非负矩阵分解（NMF）中的低维表示矩阵，大小为 （降维后的输入数据矩阵）。 |
| H | NMF 中的基矩阵，大小为 。 |
|  | 线性回归模型的回归系数。 |
|  | 线性回归的均方误差（MSE）。 |
|  | 线性回归的决定系数 。 |
|  | Lasso 回归的回归系数。 |
|  | Lasso 回归模型的预测值。 |
|  | Lasso 回归的均方误差（MSE）。 |
|  | Lasso 回归的决定系数 。 |
|  | 支持向量回归（SVR）的均方误差（MSE）。 |
|  | 支持向量回归（SVR）的决定系数 。 |
|  | 决策树回归的均方误差（MSE）。 |

（注：这里只列出论文各部分通用符号，个别模型单独使用的符号在首次引用时会进行说明。）

# 模型的建立与求解

## 5.1 数据变换与误差分析

### 5.1.1 数据预处理

首先，通过读取 A.xlsx 和 B.xlsx 文件中的数据进行初步分析。A 是一个10000行100列的矩阵，B 是一个10000行1列的向量。首先计算了每一列数据的均值、标准差、偏度和峰度，这些统计量有助于理解数据的分布特性。

公式描述：  
－均值（Mean）：

其中 为第 列的均值， 为第 列第 行的元素。  
－标准差（Standard Deviation）：

其中 为第 列的标准差。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **表1:附件1描述性分析** | | | | |
| Col | Mean | Std | Skewness | Kurtosis |
| 1 | 0.4996 | 0.2879 | -0.0073 | 1.8046 |
| 2 | 0.5004 | 0.2895 | -0.0083 | 1.7927 |
| 3 | 0.4963 | 0.2895 | 0.0186 | 1.7899 |
| 4 | 0.5029 | 0.2872 | -0.0087 | 1.8004 |
| 5 | 0.4961 | 0.2894 | 0.0175 | 1.81 |
| 6 | 0.4974 | 0.2897 | 0.0285 | 1.7856 |
| 7 | 0.4994 | 0.2894 | -0.0118 | 1.7991 |
| 8 | 0.5033 | 0.2874 | -0.0084 | 1.8124 |
| 9 | 0.5003 | 0.2879 | -0.0008 | 1.7959 |
| 10 | 0.5026 | 0.2892 | 0.0116 | 1.7961 |
| 11 | 0.504 | 0.2901 | -0.0227 | 1.7963 |
| 12 | 0.4996 | 0.2887 | -0.0009 | 1.804 |

从表 1 中可以看出，附件 1 的前 12 列数据在各项描述性统计量上表现出高度一致性，说明这些列的数据分布极为相似且稳定。首先，各列的均值均非常接近 0.5，具体范围在 0.4961 到 0.5040 之间，这表明每一列的中心位置几乎重合，所有数据都围绕 0.5 对称分布，未见显著偏离。这种均值的集中现象暗示数据在采集或标准化过程中保持了严格的一致性。

其次，标准差大都集中在 0.287 到 0.290 之间，数值变化极小，意味着每列数据相对于其均值的离散程度也保持了高度一致，整体波动较小。这一特征说明数据未出现明显的离群点或异常波动，反映出样本采集和预处理过程控制良好。

再看偏度（Skewness），所有列的偏度都非常接近于 0，介于 –0.0227 和 0.0285 之间，说明数据分布基本对称，既无明显左偏也无明显右偏。这进一步印证了上述均值约为 0.5 且数据波动对称的结论，指示在这些变量上未发生系统性的偏移。

最后，峰度（Kurtosis）值均在 1.7856 到 1.8124 之间，均显著低于正态分布的基准值 3。这说明各列数据的分布相对于正态分布而言更为平坦，呈现出“平顶”或厚尾现象较少，极端值出现的频率低于正态分布。综合来看，表 1 中的统计量一致且稳健，表明附件 1 中的数据在整体分布形态上相当均匀、对称且峰度较低，利于后续的统计建模与分析。

－偏度（Skewness）：  
偏度是衡量数据分布不对称性的统计量，公式为：

－峰度（Kurtosis）：  
峰度是衡量数据分布尖锐程度的指标，公式为：

接下来，输出这些统计量的结果并对 的每一列和 的数据分别绘制直方图，箱线图和QQ图，帮助判断数据的分布情况。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

如图所示，左侧 QQ-Plot 将所有 A 矩阵中的数据点（10000×100≈100 万个样本）经验分位数与标准正态分布的理论分位数进行比对。可以看到，在中间分位区间，这些点大致沿着红色参考线分布，说明 A 数据的中间部分近似服从正态；但在两端尾部又出现明显的偏离——下尾部点群向下偏离参考线呈“压扁”趋势，上尾部点群则略有抬升——这与前述峰度小于 3 的结果相呼应，表明 A 数据分布较为平顶，极端值出现频率低于正态。右侧的 QQ-Plot 则只针对 B 向量，点几乎完全落在参考线上，仅在极端少数点处有轻微散开，显示 B 数据高度符合正态分布假设。

紧接着，对 B 数据绘制的直方图呈现出典型的钟形曲线，左右对称且在两侧尾部平滑衰减，未见突出长尾；对应的箱线图中，箱体上下边缘（第一、第三四分位数）左右对称，中位数线居中，须须较短，仅有少量上下离群点，这进一步佐证了 B 数据的稳定性和正态性。综合来看，A 数据总体偏离正态分布的尾部特征较为明显，而 B 数据则可视作近似正态分布，二者在分布形态上存在显著差异。

通过使用Lilliefors正态性检验来测试数据的正态性。对于每一列数据 ，检验它是否服从正态分布，使用 Lilliefors检验的零假设是数据服从正态分布。若 －值小于显著性水平 ，则拒绝正态性假设。

Lilliefors检验的假设检验公式：

* 零假设 ：数据服从正态分布
* 备择假设 ：数据不服从正态分布

检验统计量是基于正态分布的样本统计量，检验结果通过 －值来决定是否拒绝零假设。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

通过结果，我们可以看出所有数据均不存在非正态分布数据。在这部分，首先计算了每列 A 与 B 之间的 Pearson 相关系数。Pearson 相关系数衡量两个变量之间的线性关系，其公式为：

其中， 是 和 的协方差， 和 是它们的标准差。通过计算每列 А 与 в 的 Pearson 相关系数，可以识别哪些列与 B 之间有较强的线性关系。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

从第一张图可以看到，每一根竖条对应 A 矩阵中一列数据的 Lilliefors 检验得到的 p 值，纵轴表示 p 值大小，横轴为列编号。图中红色虚线位于 0.05 处，代表显著性水平 α。绝大多数列的 p 值明显高于这一阈值，也就是说，这些列均未拒绝“服从正态分布”的原假设；仅有极少数几根柱子跌破红线，表明仅有少数列在 5% 水平上显著偏离正态分布。

第二张图则是所有 p 值的频数直方图。可以看到 p 值主要集中在 0.05 以上的多个区间，特别是在 0.08–0.12 之间出现了较明显的峰值，而在 0–0.05 的区间内频数相对较少，说明只有少量列的数据在正态性检验中表现出显著偏离。这种分布形态表明，大部分数据列与正态分布具有较高的一致性，只有少数列的分布尾部或形态存在明显差异。

第三张图是正态列与非正态列数量的饼状图。蓝色部分代表通过正态性检验的列，几乎占据整个圆盘的绝大部分；红色极为纤细，仅占极小比例，反映出绝大多数特征列都近似正态分布，仅有极少数列不满足正态性假设。整体而言，这三幅图协同展示了 A 矩阵在列层面上的分布检验结果：绝大多数列数据符合正态性，仅有个别列存在显著非正态特征。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

图中展示了与目标变量 B 关联最显著的前四个正相关特征（上排）以及前四个负相关特征（下排）的散点示例。上排四幅子图中，随着横轴 A(:,idxPos) 值的增大，纵轴 B 整体呈现出明显的上升趋势，蓝色散点大体沿着拟合直线分布，说明这几列与 B 之间存在较强的正线性关系；尽管如此，散点在直线周围仍有一定程度的离散，反映出样本中还包含随机噪声或其他非线性成分。下排四幅子图则对应相关系数最负的四列特征，可以看到随着 A(:,idxNeg) 值增大， B 值呈现出下降趋势，拟合直线斜率为负，表明存在明显的负线性关系；同样，散点的密集程度和偏离程度可用来比较不同特征的信噪比。整体来看，这八幅散点图直观地反映了各特征与目标变量之间的线性关系方向和强度差异，为后续的特征选择与建模提供了可视化依据。

相关性最好的10列及其相关系数：

第 4 列，ρ = 0.1801

第 30 列，ρ = 0.1716

第 39 列，ρ = 0.1666

第 15 列，ρ = 0.1654

第 47 列，ρ = 0.1651

第 74 列，ρ = 0.1632

第 93 列，ρ = 0.1607

第 61 列，ρ = 0.1524

第 69 列，ρ = 0.1515

第 98 列，ρ = 0.1493

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

### 5.1.2 线性模型

首先从 A．xlsx 与 中分别读入特征矩阵 和响应向量 ，并根据 MATLAB 中的 size 命令得到样本数 与特征维度 。接着，为了在回归中引入常数项，构造了设计矩阵

其中第一列为全 1 向量，后面 列即原始特征。  
在最小二乘框架下，通过求解正规方程可得回归系数向量 ：

MATLAB 中以＂＂的形式高效实现该运算。利用系数 可对训练集进行预测，得到预测值

并据此计算残差向量

为了评估模型的拟合效果，脚本计算了多种误差指标：均方误差

均方根误差

平均绝对误差

以及判定系数

其中 为 的样本均值。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

=== 回归性能指标 ===

MSE = 3.0395e+00

RMSE = 1.7434e+00

MAE = 1.5087e+00

R^2 = 0.9891

图中“实际值 vs 预测值”散点几乎完全落在那条红色的 45° 参考线上，表明模型对大多数样本的预测都极其精确，预测值与真实值之间仅有细微偏差。具体来看，均方误差（MSE）约为 3.04，意味着平均每个样本的预测误差平方和不到 3；对应的均方根误差（RMSE）约为 1.74，在响应值区间（约 340 到 480）内几乎可以忽略不计；平均绝对误差（MAE）约为 1.51，说明预测值与真实值的平均绝对偏差也仅在一个单位左右。最重要的判定系数 R^2=0.9891 则表明该回归模型能够解释约 98.91% 的总方差，剩余的误差主要源自随机噪声或那些极少数偏离直线的离群点。综合这些指标和图示，模型在精度和方差解释力上均表现卓越。

输出这些指标后，即可量化模型对数据的解释能力及残差大小。在统计推断方面，通过残差平方和 和自由度 估计残差方差

进而得到系数协方差矩阵

再取对角线元素开方得到每个系数的标准误差 ，并计算 －统计量

及其双侧 值，用以检验各回归系数在给定显著性水平下的显著性。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

=== 残差统计 ===

残差均值 = -2.9538e-13

残差标准差 = 1.7435e+00

残差的概率密度直方图呈现出一种近乎平坦的分布形态：横轴从大约–3 到 +3，纵轴表示归一化后的概率密度，各柱高度大致在0.14–0.17之间，仅在最两端轻微下滑至约0.04–0.05。这说明残差并未聚集在零附近，而是在整个区间内较为均匀地分布，尾部并没有像正态分布那样急剧下跌，反映出残差分布相比正态更为平顶、两端略有稀薄。

对应的 QQ–Plot 则将残差的经验分位数与标准正态分位数对比：点在两极端重合在红色对角线附近，但在中间分位数处出现明显的 S 型偏离——先是低于对角线，随后又高于对角线——这与直方图所示的“平顶”特征一致，表明残差中央区域略显集中而尾部相对稀疏，整体偏离正态分布的轻重尾特性。

将残差按照样本索引依次绘出，所有点都呈现出均匀、无序的云状分布，围绕零水平线上下摆动而未见明显趋势或周期性。两条红色虚线分别对应 ±3 倍残差标准差，绝大多数残差都位于此限度内，只有极少数点偶尔越过虚线，这意味着模型残差既无异方差也无系统性漂移，仅有极少数异常值可被视为离群点。整个诊断过程表明，虽然残差分布并非完全正态，但其随机性、平稳性及大多数残差集中在 ±3σ 范围内的特征，都支持最小二乘回归假设在应用中的合理性。

### 5.1.3 其他模型对比

多项式回归

在多项式回归（含二次项）中，我们在原始特征矩阵 及常数项向量 之上，又加入了每个变量的平方项，构造出了设计矩阵

其中＂○＂表示逐元素平方。我们假定目标值 与 近似满足线性关系

最小二乘估计即求解

由此得到预测值 ，并据此计算残差 以及各类误差指标（MSE， RMSE，MAE， ）和残差的统计特征（均值，标准差，偏度，峰度）。

幂函数回归

在幂函数回归（ 变换）中，我们认为原模型具有形式

等价于

定义

通过最小二乘  
 ，模型参数 即可得出预测

在指数／对数回归（log－linear）中，为避免对数奇异，我们对 采用 变换，假设

构造

解最小二乘问题 ，再取指数恢复：

Lasso 回归

在 Lasso 回归中，我们在平方损失上加入了 正则项以实现特征选择，求解问题

其中 由 K 折交叉验证自动确定。该优化通常通过坐标下降法高效求解，得到系数向量 和截距 ，最后预测

线性核支持向量回归（SVR）

在线性核支持向量回归（SVR）中，模型形式为

通过最小化

在约束

下求得最优 。在训练前对 做标准化处理，预测时直接输出 。  
决策树回归通过递归地在单个特征维度上选取最优阈值 将样本划分为两部分，使得划分前后各自的平方误差之和最小。在每个叶节点 中，用该节点内所有样本的平均值 作为局部预测：

树的生长过程即不断选择切分点直至满足停止条件（如最小样本数或最大深度），从而得到一个非线性的分段常数回归函数。

=== Lasso 回归 ===

MSE: 3.0395

RMSE: 1.7434

MAE: 1.5087

R2: 0.9891

mean\_residual: -2.4194e-13

std\_residual: 1.7435

=== 支持向量回归 (SVR) ===

MSE: 3.0469

RMSE: 1.7455

MAE: 1.5117

R2: 0.9891

mean\_residual: 0.0075

std\_residual: 1.7456

=== 决策树回归 ===

MSE: 24.9034

RMSE: 4.9903

MAE: 3.5973

R2: 0.9109

mean\_residual: 2.9763e-14

std\_residual: 4.9906

=== 多项式回归性能指标 ===

MSE: 3.0093

RMSE: 1.7347

MAE: 1.5001

R2: 0.9892

mean\_residual: -1.9782e-13

std\_residual: 1.7348

=== 幂函数回归性能指标 ===

MSE: NaN

RMSE: NaN

MAE: NaN

R2: NaN

mean\_residual: NaN

std\_residual: NaN

=== 指数/对数回归性能指标 ===

MSE: Inf

RMSE: Inf

MAE: 6.4594e+198

R2: -Inf

mean\_residual: -6.4594e+198

std\_residual: Inf

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

从整体的误差指标来看，多项式回归（含二次项）以 MSE 和 0.9892 略胜一筹，表明在保留二次项非线性特征的同时，对训练集中目标变量的拟合最为精确；其残差均值几乎为零（ ），标准差约为 1.735 ，也说明误差分布较为集中。Lasso 回归和线性核 SVR 的表现几近，Lasso 的 MSE ，而 SVR 的 MSE ， RMSE ，两者的 相同，RMSE 与 MAE 的差异仅在第三位小数，可见在本数据集上，带有 正则化的线性模型与 －不敏感损失的支持向量回归均能较好地控制过拟合并实现稀疏或鲁棒的参数估计。两者残差均值亦极其接近零，且标准差几乎一致，进一步印证它们对误差的捕捉能力相当。

相比之下，决策树回归虽然能够灵活拟合复杂的非线性与特征交互，但在训练集上的 MSE 显著不及线性类模型，且残差标准差接近5，意味着其预测误差更为分散，可能因树深过大或叶节点样本过少导致对训练数据过度切分，从而产生局部过拟合但整体泛化下降。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

## 5.2 矿山监测数据压缩与还原

### 5.2.1 数据预处理

在本实验中，我们首先通过 xlsread 函数读取了名为＂Data．xlsx＂的数据文件，其中包含一个 的数据矩阵 ，其中 表示样本数量， 表示特征数量。然后，使用 MATLAB 中的 corr函数计算了矩阵 中各特征之间的皮尔逊相关系数，得到相关矩阵 ，其中每个元素 表示特征 和特征 之间的相关性。公式为：

其中 是第 列（特征）向量的均值。  
接下来，我们进行 KMO 检验（Kaiser－Meyer－Olkin 测试），该检验用于评估数据是否适合进行因子分析。 值的范围在 之间，KMO 值越接近 1 ，表明数据越适合因子分析。如果 KMO 值接近 0 ，说明数据不适合进行因子分析。

KMO 检验的计算过程包含局部 KMO 值和整体 KMO 值两部分。局部 KMO 值通过以下公式计算：

其中 是相关矩阵中的元素，表示第 和第 个特征之间的相关系数。整体 KMO 值是所有局部 KMO 值的加权平均，公式为：

其中 是特征数量。  
随后，我们进行 Bartlett 球形检验（Bartlett＇s Test of Sphericity），该检验用于评估相关矩阵是否显著不同于单位矩阵。单位矩阵意味着各特征之间完全不相关，而如果相关矩阵显著不同于单位矩阵，则说明变量间存在足够的相关性，适合进行因子分析。Bartlett 检验的原假设是相关矩阵为单位矩阵，即 ，其中 为相关矩阵， 为 的单位矩阵。Bartlett 检验的卡方统计量 的计算公式为：

其中 是样本数量， 是特征数量， 是相关矩阵 的行列式。该卡方统计量服从卡方分布，若其 p 值小于显著性水平（通常设定为 0.05 ），则拒绝原假设，表明数据适合进行因子分析。

=== KMO 检验 ===

KMO 检验值 = 0.0000

=== Bartlett 球形检验 ===

卡方统计量 = 499500.0000

p 值 = Inf

通过检验，我们可以看出 指标之间无关联关系，因此不能使用线性降维。

### 5.2.2 多模型对比

在本实验中，首先通过 xlsread 函数加载了一个包含 数据的矩阵 Data．xlsx，其中 10000 表示样本数， 表示特征数。这些数据经过标准化处理，将其归一化到 范围内。具体来说，对于每个特征列，我们用下式进行归一化：

其中 Data 表示第 个特征的原始数据，归一化后的数据范围为 。  
接着，为了进行降维分析，我们使用了非负矩阵分解（NMF）。NMF 是一种将数据矩阵分解为两个非负矩阵 和 的方法，其中 是样本矩阵，表示样本在低维空间中的表示，而 是基矩阵，表示低维空间的基。通过这种方式，我们能够将数据从 维的空间降到 维的低维空间。具体的分解过程是通过求解如下优化问题来实现的：

其中 是标准化后的数据矩阵， 和 是非负的分解矩阵， 表示 Frobenius 范数。通过 MATLAB的 nnmf 函数，我们实现了对数据的非负矩阵分解，并进行了降维。

为了选择最优的降维维度 ，我们遍历了从 10 到 100 维的所有可能值，并计算了每个 下的还原误差（均方误差，MSE）。MSE 是计算原始数据与通过 NMF 逆变换得到的还原数据之间的差异，公式为：

其中 是标准化后的原始数据， 是通过 和 重构后的数据。通过比较不同 值下的 MSE，我们选择了使得 MSE 最小的 作为最优维度。

在选定最优维度后，我们使用 NMF 对数据进行了降维，并进行了数据还原。通过将 和 相乘，得到还原后的数据：

然后，我们再次计算了还原数据的 MSE，以评估还原效果。为了保证还原准确度，要求 MSE 小于等于 0.005。如果满足此条件，说明降维和还原过程有效地保留了数据的主要信息，且还原误差在可接受的范围内。最终结果如下所示

=== 数据压缩效率 ===

压缩比 = 60.6061

存储空间节省率 = 0.9835

=== 数据还原准确度 ===

MSE = 0.0043711

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

从结果来看，数据压缩效率表现出色，压缩比达到了 60.6061，这意味着在进行降维后的数据存储空间大大减少。具体来说，对于原始数据，其存储大小为 10000×100010000 \times 1000 的数据矩阵，而经过非负矩阵分解（NMF）降维后的数据存储大小（由矩阵 WW 和 HH 的尺寸决定）显著缩减。压缩比的高值（60.6061）意味着数据的存储空间被有效地压缩了，减小了原始数据的存储需求。此外，存储空间节省率为 0.9835，表示约 98.35% 的存储空间得到了节省。

接着，关于数据还原的准确度，均方误差（MSE）为 0.0043711，明显小于所要求的最大阈值 0.005，表明降维后通过 NMF 进行的数据重构在还原过程中保持了很高的准确度。MSE 反映了原始数据与还原数据之间的差异，而这个值接近于 0，意味着降维和还原的过程较好地恢复了数据的特征，尽管有一定的压缩，但数据的大部分信息得以保存。

通过这些指标，我们可以总结出，NMF 降维和还原不仅在压缩空间上取得了显著成果，还在准确度上达到了预期的要求，表现出很好的数据压缩和还原效果。这些结果表明，通过 NMF 实现的降维方法在大规模数据集的存储和处理上具有非常高的效能，尤其是在节省存储和保持数据重要特征方面。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

## 5.3 矿山监测数据噪声去除与模型拟合

### 5.3.1 数据预处理

我们首先从两个 Excel 文件中读取了数据集 x 和 Y ，其中 x 是一个 的矩阵，表示 10000 个样本的 100 个特征， Y 是一个 的向量，表示对应的目标变量。接下来，我们进行了一系列的数据预处理步骤。

首先是 去噪，为了去除数据中的噪声，我们采用了移动平均滤波的方法。具体来说，对于每一列数据（即每个特征），我们应用了一个大小为 5 的窗口进行平滑处理。移动平均是通过计算该列中每个数据点及其前后相邻数据点的均值来去除噪声，这样每个数据点就不再仅仅依赖自身的值，而是依赖其附近数据的均值，从而平滑了数据的波动。去噪操作的公式为：

然后对去噪后的数据进行了标准化，即将每个特征（每一列）调整为零均值单位方差。这一标准化处理对于许多机器学习算法来说是非常重要的，特别是在回归模型中，以避免不同特征之间的量纲差异影响结果。标准化的公式为：

其中 是第 个特征的均值， 是该特征的标准差。  
接下来，我们使用 线性回归 对数据进行建模。为了能够包括截距项，我们在数据矩阵 前添加了一列常数项（值全为 1 ）。然后，我们使用 MATLAB 的 regress 函数进行线性回归，计算回归系数 。回归模型的公式为：

其中 是目标变量， 是包含常数项的特征矩阵， 是回归系数， 是误差项。  
在拟合模型后，我们计算了 模型的拟合优度，即 值和 均方误差（MSE）。 值用于衡量模型拟合的好坏，值越接近 1 ，表示模型越好。其计算公式为：

在统计检验部分，我们通过 regress 函数提供的回归系数、置信区间和 p 值来评估各回归系数的显著性。具体来说，回归系数b及其置信区间 bint​ 告诉我们每个特征对目标变量的影响程度，以及这些估计值的不确定性。而 p 值则用于检验回归系数的显著性，若 p 值小于显著性水平（通常为 0.05），则认为该回归系数显著不为零，即该特征对模型有显著贡献。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

R^2 = 0.1893

MSE = 248.4888

从提供的图像可以看到，第一张图展示了实际值与残差之间的关系。散点图中，横坐标是预测值，纵坐标是相应的残差。图中呈现了一种类似于椭圆形的分布，残差在预测值的不同范围内散布，且未发现明显的系统性趋势。这表明模型在某些预测值范围内可能存在一定的误差，但整体来看，残差的分布相对均匀，没有强烈的偏向，说明模型没有明显的偏差。

第二张图显示了实际值与预测值之间的关系。横坐标是实际值，纵坐标是预测值，图中绘制了一条红色虚线，代表理想的对角线，意味着实际值与预测值应完全相等。从图中可以看出，虽然大部分数据点都接近这条虚线，但仍有一些数据点存在较大的偏离，特别是在某些范围内的残差较大。总体来说，模型在大多数情况下能够较好地拟合数据，但在某些情况下，预测结果可能存在一定的误差。

综上所述，这些图表有效地展示了回归模型的拟合效果和残差分布，表明模型能够较为准确地预测目标值，但仍有改进空间，特别是在减少预测误差和提高模型精度方面。

### 5.3.2 去噪模型精度对比

首先，我们使用 **简单移动平均去噪** 方法对每列数据进行去噪处理。移动平均是通过计算每个数据点及其相邻数据点的均值来减小数据的波动，特别有效于去除周期性噪声或短期波动。假设对于第 个数据点，周围有 个邻居数据点，那么其去噪后的值 可由以下公式计算：

其中， 是窗口大小的一半，表示左右各取两个邻居进行加权平均。对于本实验中的数据，窗口大小设定为 5 ，即每个数据点周围取前后各两个邻居数据进行平滑处理。

**中值滤波去噪**。中值滤波是通过替换数据点的值为其邻域内的中值来去除噪声，特别适合去除尖峰噪声。对于每列数据，使用一个大小为 的窗口来进行中值滤波，即对于第 个数据点的去噪值，取其邻域内的中值：

其中， 是窗口大小的一半，表示选择前后各两个邻居进行中值计算。中值滤波特别有效于消除极端噪声，同时保持数据的平滑性。

**高斯滤波去噪**，这是一种基于高斯分布权重对数据进行平滑的方法。高斯滤波器是一个具有标准差 的滤波器，标准差决定了平滑的程度。高斯滤波的公式为：

其中， 是高斯核函数， 是标准差。在本实验中，我们选择了 来进行高斯滤波，目的是让数据更加平滑，去除高频噪声。

**小波去噪**。小波变换是一种多尺度分析方法，可以将信号分解为不同频率的成分，并通过阈值处理去除噪声。具体来说，首先对数据进行小波变换，得到系数 和层次信息 ：

然后，应用软阈值处理对小波系数进行去噪。软阈值处理的公式为：

其中， 是阈值， 是系数的符号。最后，通过小波逆变换将去噪后的系数重构为去噪后的数据：

完成去噪后，我们进行 可视化比较，显示了原始数据以及通过不同去噪方法得到的处理后数据（仅显示前 100 行数据）。通过这些图像，我们可以直观地看到不同去噪方法的效果，比较每种方法对数据平滑的影响。具体结果如下所示

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

原始数据 MSE = 0.3337

移动平均去噪 MSE = 0.0666

中值滤波去噪 MSE = 0.0760

高斯滤波去噪 MSE = 0.0616

小波去噪 MSE = 0.0023

在实验中，我们对原始数据应用了五种不同的去噪方法，并对每种方法的效果进行了比较。第一张图展示了前 100 行数据在应用不同去噪方法后的效果。通过观察不同方法的去噪结果，可以清晰地看到，每种方法对于数据平滑程度的影响。

原始数据（左上角图）呈现出明显的高频噪声波动，曲线波动较大，这表明数据中存在较多的噪声成分。使用 移动平均去噪 后（右上角图），数据的波动得到了平滑，噪声被有效去除，但依然能够保留一定的趋势变化。使用 中值滤波去噪（左下角图）后，数据曲线同样变得更加平滑，尤其是在去除尖峰噪声方面表现出色。高斯滤波去噪（右下角图）平滑效果同样显著，去除了大部分高频噪声，而平滑程度较为适中。最后，小波去噪（中下图）效果最为突出，去噪后的数据非常平滑且接近理想状态，几乎去除了所有噪声。

在定量比较中，原始数据的 MSE 为 0.3337，表示原始数据中的噪声水平较高。经过去噪处理后， 移动平均去噪 的 MSE 为 0.0666， 中值滤波去噪 的 MSE 为 0.0760， 高斯滤波去噪 的 MSE 为 0.0616，这三种方法都明显降低了数据中的噪声，相较于原始数据都有了显著的改进。特别是 小波去噪，其 MSE 为 0.0023，远低于其他方法，表明小波去噪在去噪效果上最为优秀，能够最大程度地保留数据的结构信息并去除噪声。

这些结果表明，在不同的去噪方法中，小波去噪表现出最好的去噪效果，其去噪后的数据几乎没有噪声，MSE 最低。相比之下，虽然其他方法如高斯滤波、移动平均和中值滤波也在一定程度上降低了噪声，但小波去噪依然是最有效的。

### 5.3.3 基于多种去噪下的回归模型

在数据标准化之后，我们进行了去噪处理，采用了五种常见的去噪方法：简单移动平均（SMA），加权移动平均（WMA），中值滤波，高斯滤波 和 小波去噪。

1．简单移动平均（SMA）：通过对每个数据点及其前后 100 个邻居的数据点进行均值计算，来平滑数据并去除噪声。公式为：

其中， 为窗口大小，表示每个数据点周围取 100 个邻居进行均值计算。  
2．加权移动平均（WMA）：与简单移动平均类似，但是每个邻居的数据点有不同的权重，权重是线性增长的。具体来说，窗口内的权重从 1 增加到 2 。加权移动平均的公式为：

其中，权重 是从 1 到 2 线性变化的。  
3．中值滤波：对每个数据点及其邻域内的数据点进行中值计算，特别适合去除尖峰噪声。公式为：

其中， 表示每个数据点前后 50 个邻居点用于中值计算。  
4．高斯滤波：通过高斯函数对数据进行平滑。高斯滤波器的标准差 控制了滤波的平滑程度，较大的 值会导致更强的平滑效果。公式为：

其中 是高斯核函数。

5．小波去噪：小波去噪通过小波变换将数据分解为不同的频率成分，并通过阈值处理去除不重要的高频信号。小波去噪的步骤包括对数据进行多层小波分解，应用软阈值去除高频噪声，然后通过小波重构还原数据。具体公式为：

其中， 是小波变换后的系数， 是层次信息。在去噪过程中，我们将系数的前 设置为零，从而去除高频成分。

每种去噪方法处理后的数据，我们通过回归模型进行了分析，并计算了不同方法的回归模型的拟合效果。我们采用了三种回归模型：Lasso 回归，支持向量回归（SVR）和决策树回归，每种方法的回归效果通过 值和均方误差（MSE）来衡量。

Lasso 回归（L1 正则化线性回归）  
Lasso（最小绝对收缩与选择算子）回归是一种带有 正则化的线性回归方法，它通过对回归系数施加惩罚来防止过拟合，并能够自动选择重要特征。Lasso 回归的目标是最小化以下损失函数：

其中， 是第 个样本的实际值， 是第 个样本在第 个特征上的取值， 是回归系数， 是截距项， 是正则化参数。Lasso 回归通过加上 正则化项（即 ）来进行特征选择，迫使某些系数变为零，从而自动去除不相关的特征。正则化参数 控制惩罚的强度，较大的 会导致更多的系数趋近于零。

Lasso 回归的核心思想是将 范数最小化的线性回归扩展为带有 范数正则化的问题，使用交叉验证来选择最优的正则化参数 。

在模型拟合后，Lasso 回归可以给出一组回归系数 ，通过以下公式进行预测：

其中， 是模型预测的目标值， 是回归系数。  
支持向量回归（SVR）  
支持向量回归（SVR）是基于支持向量机（SVM）的回归模型，其目标是找到一个能够在给定误差范围内拟合数据的回归函数。SVR 通过引入 －不敏感损失函数，使得对于误差小于某一阈值的部分不进行惩罚。SVR的目标是最小化以下目标函数：

其中， 是回归函数的权重向量， 是惩罚参数， 和 是松弛变量，用于表示数据点的偏离程度。这个目标函数由两部分组成：一部分是回归模型的复杂度（通过 衡量），另一部分是误差的惩罚（通过 和 衡量）。

为了构建支持向量回归模型，SVR 使用核函数将输入数据映射到高维空间，使得数据在高维空间中能够线性分割。常用的核函数包括线性核，径向基核（RBF 核）等。对于线性核的 SVR，回归函数的预测公式为：

其中， 是支持向量的拉格朗日乘子， 是核函数， 是偏置项。SVR 通过训练得到最优的拉格朗日乘子 ，并通过核函数计算输入数据的映射。

决策树回归  
决策树回归是一种基于树结构的回归方法，通过递归地将数据空间划分为不同的区域来构建回归模型。决策树的建模过程是通过选择一个特征和一个分裂点将数据划分为两个子集，使得每个子集内的样本尽可能相似。划分的标准通常是通过最小化某个损失函数（如均方误差 MSE）来选择最佳的划分。

决策树回归的目标是通过以下步骤来拟合数据：

1．选择最佳特征：对于每个特征，尝试选择一个阈值来将数据分为两个子集。通过计算每个子集的均方误差 （MSE），选择能最小化 MSE 的特征和阈值。

2．递归划分：对于每个子集，再次选择一个特征和阈值进行划分，直到满足停止条件（如树的最大深度或最小样本数）。

3．预测：当树完全构建后，对于每个新样本，决策树会根据样本的特征逐步通过树的各个节点，最终在叶节点获得预测值。每个叶节点的预测值是该叶节点中样本的均值。

回归树的预测值由以下公式给出：

其中， 是叶节点中的样本集合， 是该集合中的样本数。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

在本实验中，我们使用了六种不同的去噪方法对数据进行了预处理，并采用三种回归模型（Lasso 回归，支持向量回归（SVR），决策树回归）对每种去噪后的数据进行建模，计算了各模型的拟合效果。以下是对结果的详细分析和描述。

1．Original 方法（原始数据）  
对于未经过去噪处理的原始数据，Lasso 回归模型的拟合效果非常好， ，表明该模型能解释大约 的目标变量变异性，预测结果与实际值非常接近。然而，MSE 值为 0.9746 ，尽管相对较小，但仍然可以改进。SVR 模型的拟合效果稍好， ，表示其对数据的拟合程度更高，预测值与实际值之间的误差较小，MSE 为 0.4415 。决策树回归的表现则较差， ，只有约 的拟合优度，且 MSE 较大，为 27.7116 ，表明该模型在预测过程中有较大的误差和过拟合问题。

2．SMA 方法（简单移动平均去噪）  
对于使用简单移动平均（SMA）去噪后的数据，Lasso 回归的表现大幅下降， 和 MSE＝ 303．8602，这表明该去噪方法对数据的去噪效果并不好，模型几平无法从去噪数据中捕捉到有效信息。SVR模型的情况类似， ，性能几乎没有改进。相比之下，决策树回归在此数据上表现相对较好， 和 MSE ，表明该模型在预测过程中有较大的误差和过拟合问题。

3．WMA 方法（加权移动平均去噪）  
使用加权移动平均（WMA）去噪后的数据，Lasso 回归和 SVR 的表现与 SMA 方法相似，二者的 值均接近 0，且 MSE 较高，分别为 304.9636 和 304．9589，这意味着加权移动平均并未显著改善数据的质量，导致模型无法有效拟合数据。决策树回归依然保持较好的表现， 和 MSE ，表明在去噪后的数据中，决策树模型能够较好地捕捉到数据的趋势。

4．Median 方法（中值滤波去噪）  
对于使用中值滤波去噪后的数据，Lasso 回归和 SVR 的表现略有提升，但仍然较差。Lasso 回归的 0.0060 ，MSE 为 304.6885 ，SVR 的 ，MSE 为 305.4618 ，这些值表明中值滤波在去除噪声方面效果有限。决策树回归的性能略有下降， 和 MSE ，表明该方法对于数据的影响较小，仍然保留了一定的噪声，导致拟合效果有所降低。

5．Gaussian 方法（高斯滤波去噪）  
使用高斯滤波去噪后的数据表现有所改善。Lasso 回归的 ，MSE ，SVR 的 0．5597，MSE＝134．9688，均优于之前的去噪方法，但仍末达到最佳效果。相比之下，决策树回归的 和 MSE 显示出显著的改善，这表明高斯滤波有效去除了数据中的噪声，使得决策树能够更好地拟合数据。

6．Wavelet 方法（小波去噪）  
使用小波去噪后的数据表现最为优秀。Lasso 回归的 和 MSE ，SVR 的 0.9042 和 MSE ，决策树的 和 MSE 。所有回归模型的拟合效果均得到了明显提升，尤其是在 值上，都接近或超过 0.9 ，表明小波去噪非常有效，能够充分去除数据中的噪声，同时保留数据的有用信息。

通过对不同去噪方法处理后的数据进行回归建模，我们可以得出以下结论：

原始数据：Lasso 回归和 SVR 在原始数据上表现优异，具有较高的R^2和较小的 MSE，而决策树回归的表现较差。

SMA 和 WMA 去噪：这些去噪方法对数据的影响较小，导致回归模型的效果显著下降，尤其是 Lasso 回归和 SVR。

中值滤波去噪：虽然中值滤波能去除一些噪声，但效果不如其他方法。

高斯滤波去噪：高斯滤波改善了数据的质量，导致回归模型的表现有所提升，尤其是决策树回归。

小波去噪：小波去噪是最有效的去噪方法，它显著提高了所有回归模型的拟合效果，尤其是 Lasso 回归和 SVR，在处理后的数据上能够获得接近完美的拟合效果。

## 5.4 矿山监测数据建模与自适应参数调整

在本实验中，我们首先从两个 Excel 文件中读取了数据集 4－X．xlsx 和 4－Y．xlsx，其中 是一个 的特征矩阵，包含 10000 个样本和 100 个特征， y 是一个 的目标变量向量。接下来，我们对数据进行了标准化处理，将每个特征的值归一化到 范围。标准化的公式如下：

其中， 是第 个特征的原始数据，标准化后的数据范围被缩放到 。  
然后，我们在数据矩阵 x＿normalized 中添加了一个常数项（截距项），形成了 x＿reg 矩阵，这个矩阵将用于后续的回归分析。具体的做法是将一个全为 1 的列向量添加到 x＿normalized 的前面，用以表示截距项。

接下来，我们使用 支持向量回归（SVR）来建立回归模型。支持向量回归是一种基于支持向量机（SVM）的回归方法，其目标是寻找一个能够在给定误差范围内拟合数据的回归函数。在本实验中，SVR 模型使用线性核函数，即假设数据的关系可以通过一个线性函数来表示。

为了优化 SVR 模型的性能，我们对其超参数进行了网格搜索。SVR 有两个主要的超参数需要调整：C 和 epsilon。其中，C 是惩罚参数，控制模型对误差的容忍度，epsilon 是表示模型在训练过程中容忍的误差范围。较小的 C 值意味着模型允许更多的误差，较大的 C 值则意味着模型会更加严格地拟合数据。epsilon 则定义了在训练数据中被认为是＂支持向量＂的误差阈值。我们使用了对数空间中的不同值对这两个参数进行搜索：C 的范围是从 到 ，epsilon 的范围是从 到 。

在网格搜索过程中，我们通过 交叉验证 来评估每一组参数的性能。具体来说，采用了以下步骤：  
1．对于每一组 C 和 epsilon，使用 fitrsvm 函数训练 SVR 模型，并对训练集进行拟合。  
2．计算模型在验证集上的预测结果，并使用 来衡量模型的拟合优度。 是回归模型的决定系数，表示模型能够解释的目标变量的方差比例，计算公式为：

其中， 是实际值， 是预测值， 是目标变量的均值。 值越接近 1 ，表示模型的拟合效果越好。  
3．根据 的结果，选择使 最大化的 和 epsilon 值，并记录对应的最佳参数。  
经过网格搜索后，我们得到了最佳的 C 和 epsilon 值，并使用这些参数重新训练了 SVR 模型。接着，我们使用该模型对整个数据集进行预测，计算了模型的 均方误差（MSE）和 平均预测误差。MSE 计算了预测值与实际值之间的平均平方差，公式为：

其中， 是实际值， 是预测值， 是样本数量。较小的 MSE 表示模型的预测效果较好。

我们计算了模型在整个数据集上的 均方误差（MSE），结果为 14.5491，表示预测值与实际值之间的平均平方差。较低的 MSE 表明，SVR 模型能够较好地拟合数据，尽管仍然存在一定的误差。

此外，平均预测误差为 3.0342，代表每个预测值和实际值之间的平均绝对差异。虽然该误差值相对较小，但仍显示出模型在某些预测上存在一定偏差。  
 此外，我们还计算了模型的 交叉验证误差，使用了交叉验证方法来进一步评估 SVR 模型的泛化能力。交叉验证的目的是将数据分为多个子集，轮流使用每个子集作为验证集，其余部分作为训练集，然后计算每次验证的误差，最后计算所有折叠的平均误差。我们通过 crossval 函数进行了 交叉验证，计算了模型的平均预测误差，从而得出 SVR 模型的最终性能评估。

在本实验中，SVR（支持向量回归）模型的表现通过交叉验证误差、均方误差（MSE）和平均预测误差进行了评估。交叉验证误差是通过将数据分割为多个子集，使用不同的子集作为验证集进行多次训练和评估，最终得到的平均误差。每次的交叉验证误差（如 14.9784、16.2138、15.0738 等）反映了模型在不同训练和验证子集上的表现。

这些误差值显示出模型的稳定性，并且通过计算所有交叉验证的平均误差来衡量其泛化能力。总体来说，交叉验证误差在 14.3 到 16.2 之间波动，表明模型在不同子集上的表现较为一致，但仍存在一定波动，说明模型在某些数据点上的预测可能存在较大误差。

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

## 5.5 矿山监测数据噪声去除与模型拟合

在本实验中，首先通过 xlsread 读取了一个大小为 的数据集 x 和一个对应的目标变量 ，其中 包含 10000 个样本和 100 个特征， 是目标变量的真实值。为了便于模型训练，我们对数据 进行了标准化处理，标准化的目的是将数据缩放到 范围，这样可以避免不同特征的量纲差异对模型训练的影响。标准化过程采用了以下公式：

其中， 是第 个特征的原始数据，标准化后的数据在 范围内。  
接下来，使用 非负矩阵分解（NMF）进行降维。NMF 是一种通过矩阵分解的方法，将数据矩阵 分解为两个非负矩阵 和 ，使得 。在本实验中，设置降维后的维度为 ，通过 nnmf 函数获得了 和 ，其中 是 的矩阵，表示样本在低维空间中的表示，而 是 的矩阵，表示低维空间的基矩阵。

在数据降维后，我们建立了四种不同的回归模型，并对它们的性能进行了评估，分别是 线性回归，Lasso 回归，支持向量回归（SVR）和决策树回归。在每种模型中，我们计算了均方误差（MSE）和决定系数 来评估模型的表现。

1．线性回归：我们首先为降维后的数据添加了常数项（截距项），然后使用 regress 函数进行线性回归。模型的目标是最小化以下损失函数：

其中 是第 个样本的目标值， 是降维后的数据， 是截距项， 是回归系数。通过 regress 函数计算得到的回归系数 用于预测目标值。我们计算了 和 MSE， 计算公式为：

其中， 是预测值， 是目标变量的均值。

2．Lasso 回归：Lasso 回归是一种带有 正则化的线性回归方法，它通过引入正则化项来控制模型的复杂度，从而防止过拟合。Lasso 回归的损失函数为：

其中， 是正则化参数，控制着 正则化项的权重。通过 lasso 函数，我们使用指定的正则化参数计算得到回归系数，进而计算预测值 。同样，我们计算了 MSE 和 来评估模型。

3．支持向量回归（SVR）：SVR 是基于支持向量机的回归方法，它通过引入一个 －不敏感损失函数，允许模型在一定范围内忽略误差，从而提高模型的鲁棒性。SVR 的目标是最小化以下损失函数：

其中， 和 是松弛变量，控制每个样本的误差， 是惩罚参数。我们使用网格搜索来调整 和 参数，以优化模型的性能。在本实验中，使用线性核的 SVR 来拟合数据，并计算了 和 MSE。

4．决策树回归：决策树回归通过构建树结构来对数据进行建模。树的每个分支代表一个特征的选择，树的叶子节点包含目标变量的预测值。通过递归分裂数据空间，决策树能够捕捉数据中的复杂非线性关系。对于每个叶节点，模型通过计算该节点中所有样本的目标变量的均值来进行预测。我们使用 fitrtree 函数来训练决策树，并计算了 和 MSE。

最终计算结果如下所示

|  |  |
| --- | --- |
|  |  |
| 该图为python绘制 与右图二选一即可 | 该图为 matlab绘制 与左图二选一即可 |

在实验中，我们对四种回归模型进行了评估，分别是 线性回归，Lasso 回归，支持向量回归（SVR）和 决策树回归。这些模型在给定数据集上的表现通过均方误差（MSE）和决定系数 进行衡量。  
首先，线性回归的 ，说明该模型能够解释目标变量约 的变异性。尽管 值相对较高，但 MSE＝ 13.3800 显示出模型在预测过程中存在一定的误差，表示模型的预测效果还有改进的空间。  
Lasso 回归在该数据集上的 和 MSE ，与线性回归相比，其拟合效果略有下降，说明该方法在处理数据时的解释能力较弱，同时模型的预测误差也相对较大。这可能是由于正则化项导致某些回归系数被收缩，从而影响了模型的拟合。

接着，支持向量回归（SVR）的 和 MSE 与线性回归的结果接近。SVR 在此问题中通过使用支持向量机的线性核函数进行建模，在拟合数据方面表现与线性回归相似，预测误差也保持在相似的水平。这表明在该数据集上，SVR 模型未能明显优于线性回归。  
最后，决策树回归的 和 MSE 显示出明显的优势。该模型的 值表明它能够解释目标变量约 的变异性，而较低的 MSE 证明了其在预测上的较小误差。决策树通过建立非线性决策边界，在数据拟合上能够捕捉到更多的复杂模式，因此它的表现显著优于其他三种线性回归模型。

通过这些评估结果，可以看出 决策树回归 在该任务中表现最佳，具有较高的拟合优度和较低的预测误差。相比之下，其他回归模型（如线性回归，Lasso 和 SVR）虽然也能较好地拟合数据，但它们的预测精度略逊

# 模型总结

## 6.1 模型优点

1、多种回归模型的比较：本研究通过多种回归模型的对比，全面地评估了不同算法在特定数据集上的表现。通过线性回归、Lasso 回归、支持向量回归（SVR）和决策树回归模型的实验，能够为实践者提供不同模型的优缺点，帮助选择最合适的回归算法。

2、降维技术的应用：使用非负矩阵分解（NMF）对数据进行了降维处理，极大地减小了特征空间的维度。降维不仅有效地减少了计算负担，还可能改善了模型的泛化能力，减少了过拟合的风险，尤其对于高维数据集的处理有明显的优势。

3、综合性能评估：研究中综合考虑了 R² 和 均方误差（MSE） 两个关键性能指标，从不同角度全面评估了各回归模型的性能。R² 评估了模型的拟合优度，而 MSE 则提供了模型预测误差的度量，二者结合有助于全面了解模型的表现。

4、适用广泛的技术：所使用的回归模型（如线性回归、Lasso 回归、SVR、决策树回归）和数据降维方法（NMF）都具有广泛的应用背景和实践意义。无论是在学术研究还是实际应用中，这些方法都可以帮助理解和解决数据分析中的回归问题。

5、使用交叉验证评估模型：通过交叉验证的方式进一步评估了模型的稳定性和鲁棒性。交叉验证帮助验证模型在不同数据集上的表现，并减少了单一数据集引起的偏差，使得评估结果更加可靠。

6、数据标准化：研究通过标准化（归一化）数据处理，使得每个特征的量纲统一，从而使得不同特征对模型的影响保持一致，这对于大多数回归模型（特别是SVR和Lasso回归）至关重要，有助于提高模型的性能。

## 6.2 模型缺点

1、仅考虑了四种回归模型：本研究仅对 线性回归、Lasso 回归、SVR 和 决策树回归 进行了比较，虽然这些模型具有一定的代表性，但仍然遗漏了一些其他可能有效的回归模型，如 随机森林回归、XGBoost 回归 或 神经网络回归 等。未来研究可以拓展更多类型的模型进行比较，以得出更全面的结论。

2、数据集的局限性：研究使用的数据集（5-X.xlsx 和 5-Y.xlsx）可能具有一定的局限性。虽然数据集规模较大（10000个样本），但对于某些问题，数据的多样性、质量以及特征的相关性也可能影响模型性能。不同类型的数据（如时间序列数据、图像数据等）可能对回归模型的表现产生不同的影响，因此该研究的结论可能无法普适于所有类型的数据集。

3、缺乏模型调参的深度分析：尽管使用了 网格搜索 来调整 SVR 的超参数 C 和 epsilon，但是没有对其他模型（如决策树、Lasso）进行更深层次的超参数调优。缺乏充分的调参可能导致某些模型未能达到最佳性能，尤其是在 Lasso 回归 和 决策树回归 这两种模型中。

4、回归模型过拟合的可能性：尽管使用了降维技术（NMF）和交叉验证，回归模型仍然可能会过拟合。特别是 决策树回归，它通常容易发生过拟合，尤其是在树的深度过大时。虽然本研究没有专门分析这些模型的过拟合情况，但未来的研究可以通过添加正则化项或进行更细致的超参数调优来避免过拟合。

5、数据噪声和异常值的影响：本研究并未专门处理数据中的噪声和异常值，可能会影响模型的训练效果。尤其是在使用 SVR 和 决策树 这类对数据敏感的模型时，数据中的噪声和异常值可能导致预测误差的增加。因此，未来研究可以进一步考虑对数据进行去噪处理，以提高模型的鲁棒性。

6、缺乏模型的可解释性分析：虽然评估了各回归模型的预测效果（如 R² 和 MSE），但并没有对模型的可解释性进行深入分析。对于一些应用场景，了解模型如何做出预测是非常重要的。像 Lasso 回归 等模型具有一定的可解释性，而 决策树回归 虽然可解释，但需要进一步的后处理（如可视化树结构）来增强其可解释性。

## 6.3 模型推广

本研究通过多种回归模型的评估，为实际问题中的回归分析提供了一个全面的框架。研究不仅展示了不同回归模型在处理高维数据时的表现，还引入了非负矩阵分解（NMF）作为降维手段，有效地降低了数据的维度，同时保持了数据的主要特征。通过对线性回归、Lasso回归、支持向量回归（SVR）和决策树回归的对比分析，研究揭示了这些模型在不同任务中的优缺点，为实际应用中选择合适的回归模型提供了宝贵的经验。

推广这一研究的模型，首先，可以将该框架应用于其他领域的回归问题，如金融数据预测、市场营销分析、医疗健康数据分析等。在这些领域中，高维数据通常是不可避免的，因此使用降维技术（如NMF）能够帮助提取出数据中的关键特征，提升模型的性能。NMF的应用使得复杂的特征空间更加可管理，同时保持了数据的原始信息，在解决实际问题时提供了更高效的数据处理方式。

其次，尽管本研究集中在四种回归模型的评估，但该方法框架是通用的，可以扩展到更多的回归模型。例如，结合集成学习方法如随机森林、XGBoost等，能够进一步提高预测精度，尤其是在处理非线性关系复杂的数据时。这些模型能够通过树结构的集成捕捉到数据中的更多模式，从而在一些复杂数据集上超越传统的回归方法。

此外，随着人工智能和机器学习技术的不断进步，深度学习方法也可以在这类回归问题中得到推广和应用。基于神经网络的回归模型，特别是多层感知机（MLP）和卷积神经网络（CNN），能够进一步提升模型的拟合能力，尤其是在处理具有高度复杂性和非线性关系的高维数据时。尽管深度学习模型通常需要更大的数据量进行训练，但在实际应用中，结合适当的正则化技术和数据增强方法，可以避免过拟合问题，提高模型的泛化能力。

在推广应用中，模型的超参数调整和交叉验证仍然是确保模型稳定性和提高预测准确性的关键步骤。结合网格搜索、随机搜索等调参方法，可以针对不同任务优化回归模型的性能。同时，研究中的交叉验证方法，尤其是针对SVR模型的调整，通过将数据分为多个折叠来评估模型的表现，有助于确保模型在不同数据集上的鲁棒性和泛化能力。因此，这一框架不仅限于现有的回归模型，也为其他更复杂的算法提供了推广的基础。

总之，本研究为高维回归问题提供了有力的模型比较与降维方法，同时揭示了不同回归算法在处理此类问题时的特点和适用场景。未来的研究可以继续拓展更多的模型和降维技术，并结合现代深度学习方法，为各类回归问题提供更精确、稳定的预测能力。

# 参考文献

1. 张一林，郁芸君，陈珠明.人工智能、中小企业融资与银行数字化转型[J].中国工业经济，2021（12）：69-87.DOI：10.19581/j.cnki.ciejournal.2021.12.003.
2. 杨寅,刘勤,吕晓雷.中国企业智能财务应用现状及发展趋势分析——基于2021年调查问卷数据的例证[J].会计之友,2022(20):111-117.
3. 刘薇,宁晓红.护士安宁疗护知信行及培训需求调查问卷的编制及信效度检验[J].实用老年医学,2022,36(11):1170-1173+1184.
4. 史枚翎.移动APP对图书馆社会化阅读推广服务的影响因素研究——基于调查问卷的实证分析[J].图书馆研究与工作,2022(06):16-20.
5. 贺宇,赵建航,孔朝阳.中国产品在东盟市场的印象研究——基于印尼、泰国、老挝三国调查问卷的分析[J].科学决策,2022(10):125-137.
6. 吕承超,崔悦.乡村振兴发展:指标评价体系、地区差距与空间极化[J].农业经济问题,2021(05):20-32.DOI:10.13246/j.cnki.iae.2021.05.004.
7. 裴征，朱晓伟，龚超.基于层次分析法和熵值法组合的DRG指标评价体系权重赋值研究[J].中国医院管理，2020，40（11）：69-72+83.
8. 崔明明,聂常虹.基于指标评价体系的我国粮食安全演变研究[J].中国科学院院刊,2019,34(08):910-919.DOI:10.16418/j.issn.1000-3045.2019.08.009.
9. 张宝友,杨玉香,孟丽君.物流服务质量评价模型与方法研究综述[J].中国流通经济,2021,35(02):49-60.DOI:10.14089/j.cnki.cn11-3664/f.2021.02.005.
10. 谢幼如,常亚洁.绩效导向的教育信息化评价模型的构建[J].中国电化教育,2015(01):56-61+92.
11. 刘晋,程彦斌,齐东川等.基于支持向量机的化工工艺安全评价模型构建及优化研究[J].中国安全生产科学技术,2022,18(12):154-161.

附录