

# 基于随机投影的压缩离群点检测算法

*Random-Projection-Based Compressed Outlier Detection*

---

刘坤瓒

lkz18@mails.tsinghua.edu.cn

liukunzan.github.io

2020年8月13日



# 总览

## 基于随机投影的压缩离群点检测算法

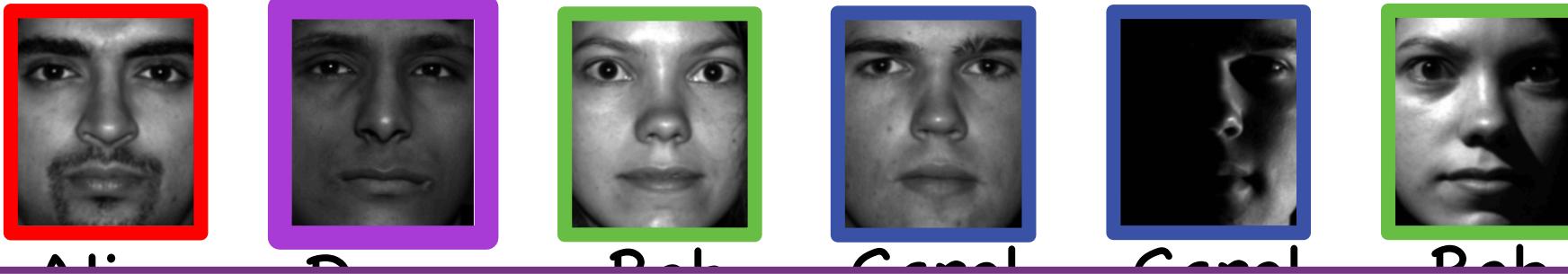
*Random-Projection-Based Compressed Outlier Detection*

本工作是一个机器学习算法设计，偏重算法理论分析，应用于**大规模高维数据处理**问题。

- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结



# 什么是离群点检测？



**离群点**：数据中的无关信息，严重干扰了数据处理的准确率和效率。



➤ 项目背景

➤ 算法提出

➤ 理论分析

➤ 实验结果

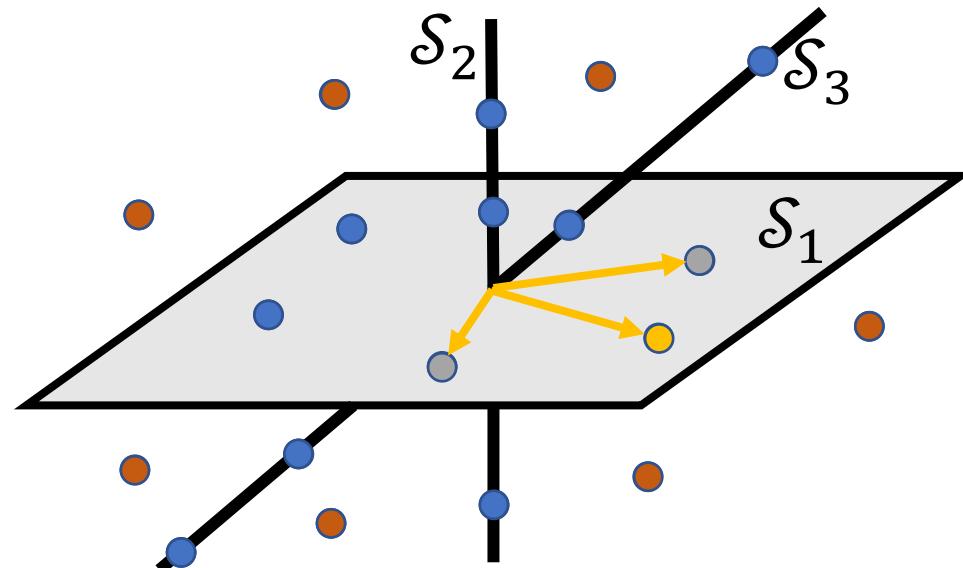
➤ 整理总结



# 目前离群点检测算法是什么？

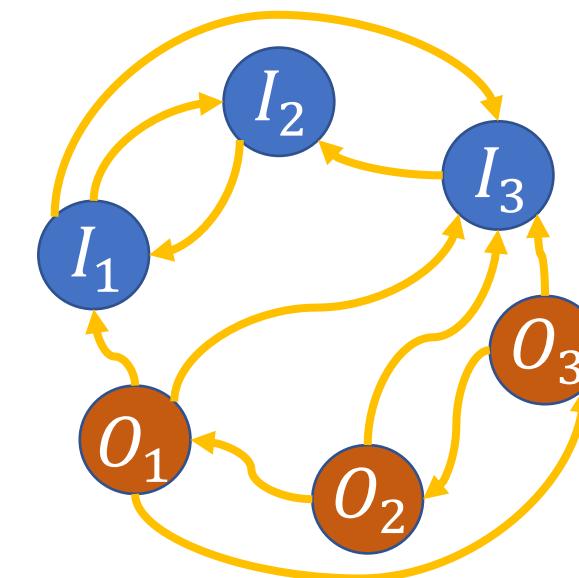
自表示向量

$$r_j = \arg \min_c \lambda \|c\|_1 + \frac{1-\lambda}{2} \|c\|_2^2 + \frac{\gamma}{2} \|x_j - Xc\|_2^2$$



步骤1：自表示，提取数据低维特征

数据集



步骤2：随机游走，区分  
离群点与非离群点

- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结



# 现存算法存在哪些问题？

甘工白圭二的离群点检测

## 回答：压缩离群点检测

集上获得较高的准确率。

- 缺点：效率随着数据维数的增长而大大降低

提问：是否能在保证较高准确率的前提下  
大幅降低计算成本？

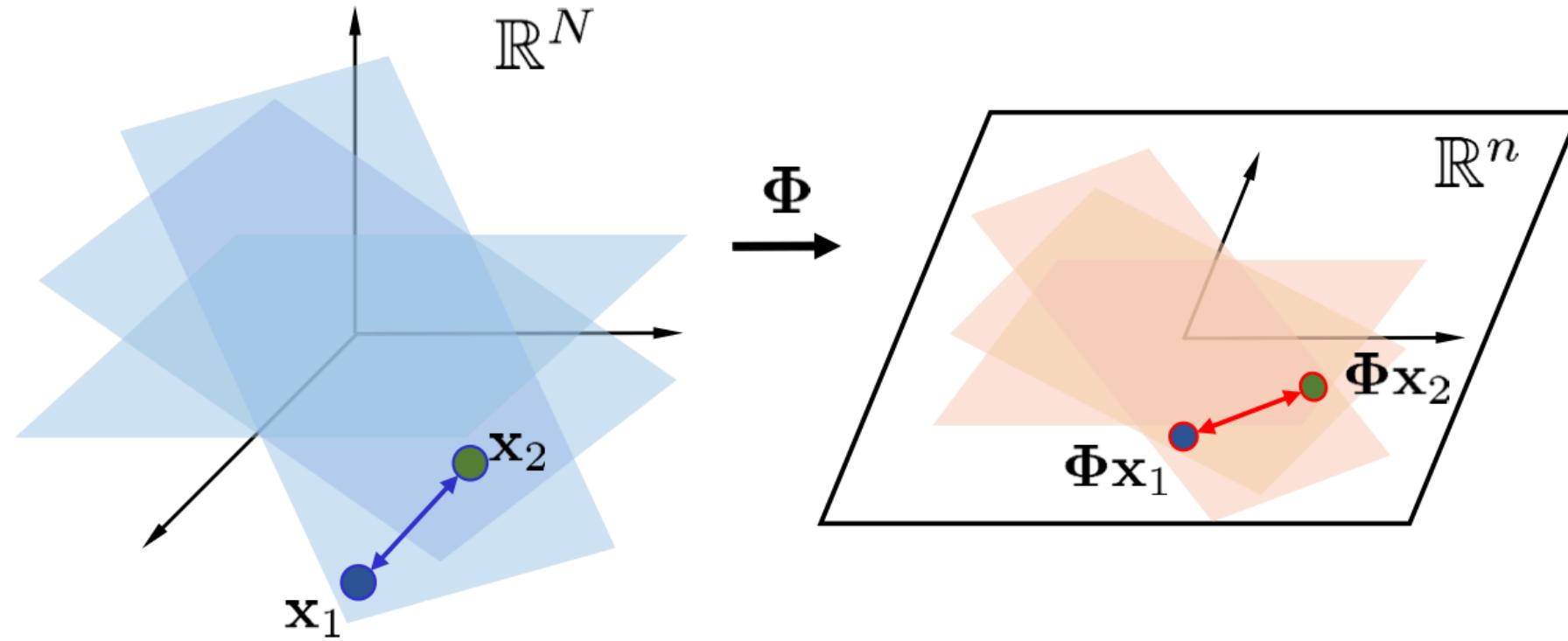
- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结

---

<sup>1</sup>C. You, D. P. Robinson, and R. Vidal, “Provable self-representation based outlier detection in a union of subspaces,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 4323–4332, 2017.



# 如何降低运算成本？



图：采用随机投影实现降维，将现有的高维空间 $\mathbb{R}^N$ 投影到低维空间 $\mathbb{R}^n$ ，适用于大规模高维数据处理，它可以保证投影前后数据集空间结构以大概率不被破坏。

- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结

<sup>2</sup> G. Li, Q. Liu and Y. Gu, "Rigorous restricted isometry property of low-dimensional subspaces," Applied and Computational Harmonic Analysis, 2019.



# 算法提出

## Algorithm 1 压缩离群点检测

输入：数据集、压缩后维数、其他参数

步骤 I：随机投影

步骤 II：自表示 **【时间瓶颈】**

步骤 III：随机游走

输出：离群点

➤ 项目背景

➤ 算法提出

➤ 理论分析

➤ 实验结果

➤ 整理总结

自表示向量

数据集

$$\mathbf{r}_j = \arg \min_{\mathbf{c}} \lambda \|\mathbf{c}\|_1 + \frac{1-\lambda}{2} \|\mathbf{c}\|_2^2 + \frac{\gamma}{2} \|\mathbf{x}_j - \mathbf{X}\mathbf{c}\|_2^2$$



# 如何证明算法优势？

提问：是否能在保证较高准确率的前提下  
大幅降低计算成本？

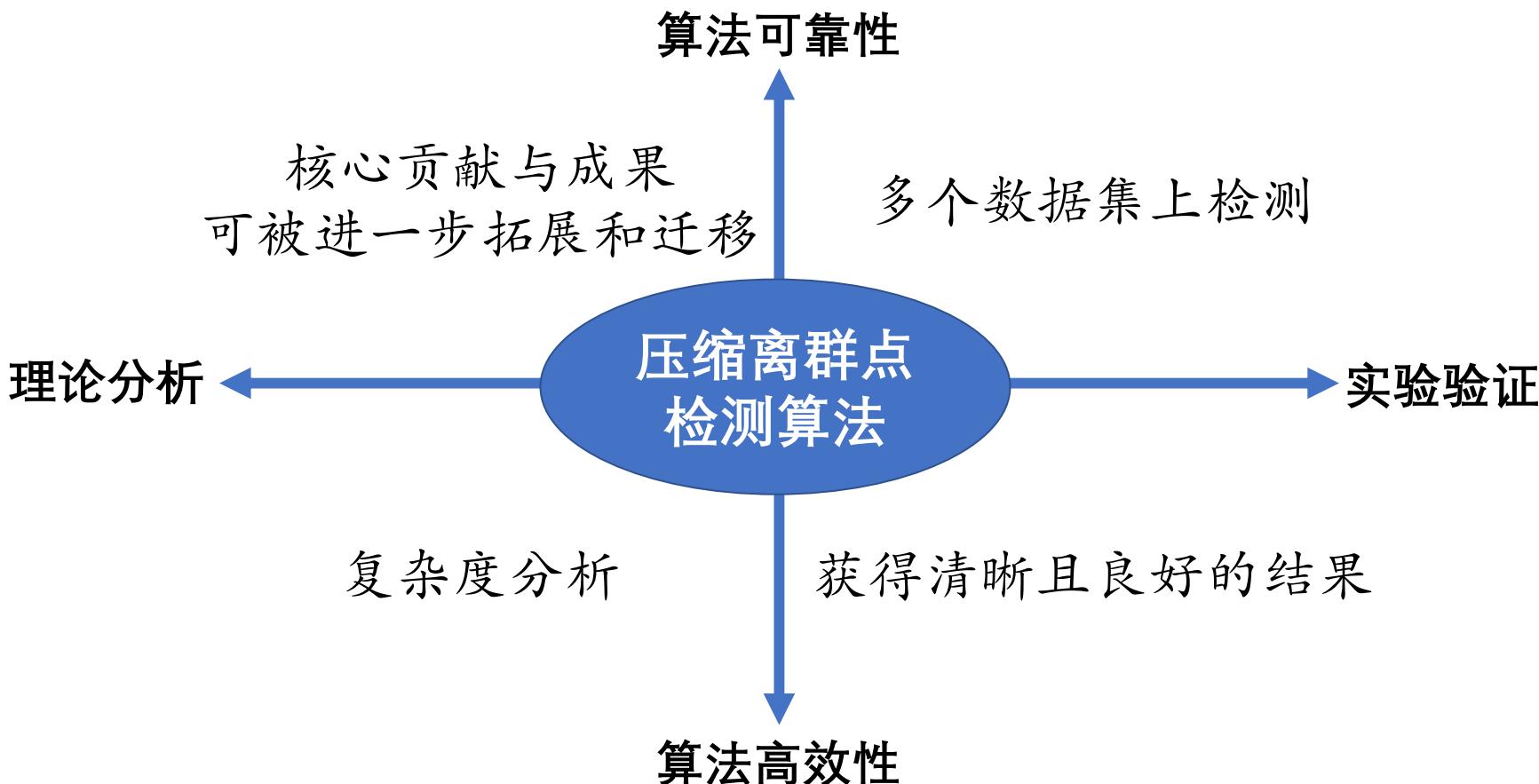
➤ 项目背景

➤ 算法提出

➤ 理论分析

➤ 实验结果

➤ 整理总结





# 理论分析

**记号**  $N$  – 原始数据维数,  $n$  – 压缩后数据维数,  $M$  – 数据量  
**定义**  $\varepsilon \rightarrow 0$  表示小量,  $c_1$  和  $c_2$  是大于0的常量

- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结

	原算法	本算法
算法		
假设	结论：准确率以近乎100%的概率保持不变，	
条件	运行时间在压缩后大幅下降。	
附加条件		$\varepsilon > \max(\varepsilon_1, \ln M)$
成功率	1	$\approx 1 - \exp(-c_2 \varepsilon^2 n) \rightarrow 1$
复杂度	$\mathcal{O}(N^2 M^3)$	$\gg \mathcal{O}(n^2 M^3)$

$$\tau := \max_j \left\{ \varepsilon \left( 1 + \|\delta_j\|_2^2 \right) + \sqrt{6\varepsilon} (1 + \varepsilon) \gamma \|\delta_j\|_2 \right\} \rightarrow 0$$

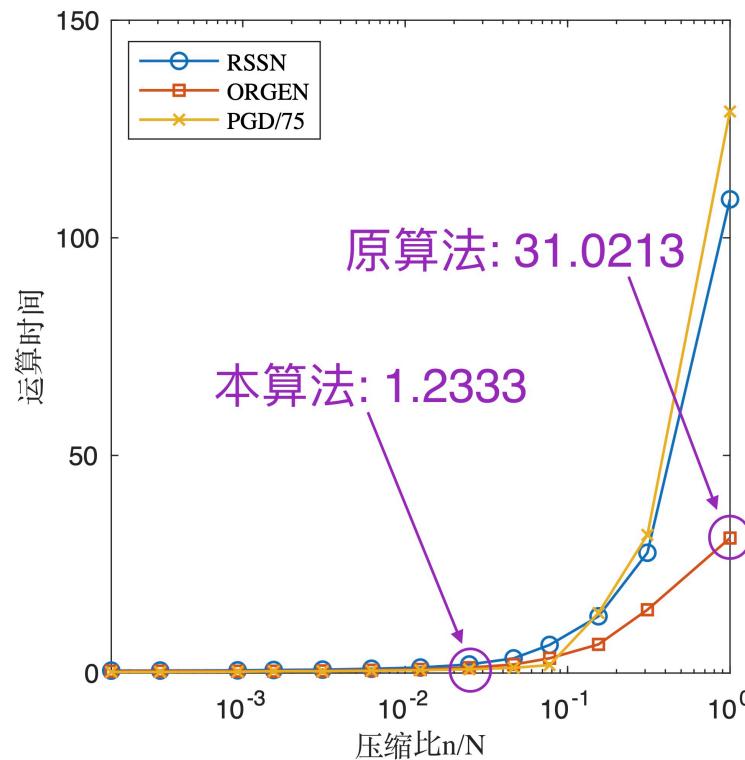


# 实验结果

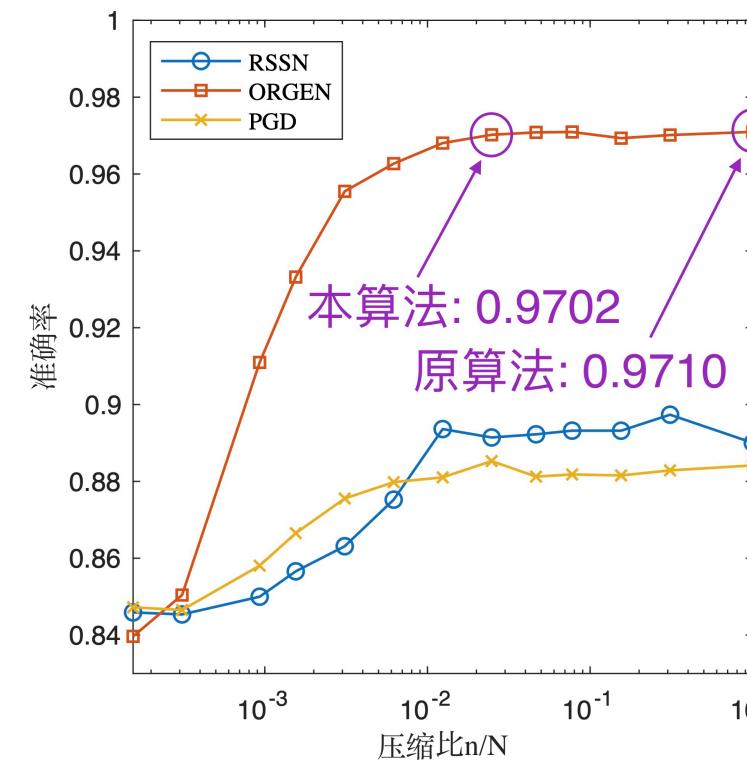
实验数据来源 Extended Yale Face B 人脸数据集 32256×262

$$\mathbf{r}_j = \arg \min_{\mathbf{c}} \lambda \|\mathbf{c}\|_1 + \frac{1-\lambda}{2} \|\mathbf{c}\|_2^2 + \frac{\gamma}{2} \|\mathbf{x}_j - \mathbf{X}\mathbf{c}\|_2^2$$

- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结



运算时间 ↓ 96.03%



准确率 ↓ 0.08%



# 创新点

## 理论创新

- 随机投影在自表示离群点检测中**首次引入**
- 随机投影下子空间保持性质**首次证明**

## 应用创新

- 大规模人脸聚类
- “**又好又快又省**”  
准确率高+速度快+省存储

## 潜在价值

网络压缩 / 受损图像处理 / 多视图子空间学习

- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结

# 基于随机投影的压缩离群点检测算法

感谢倾听，恳请指正！

---

刘坤瓒

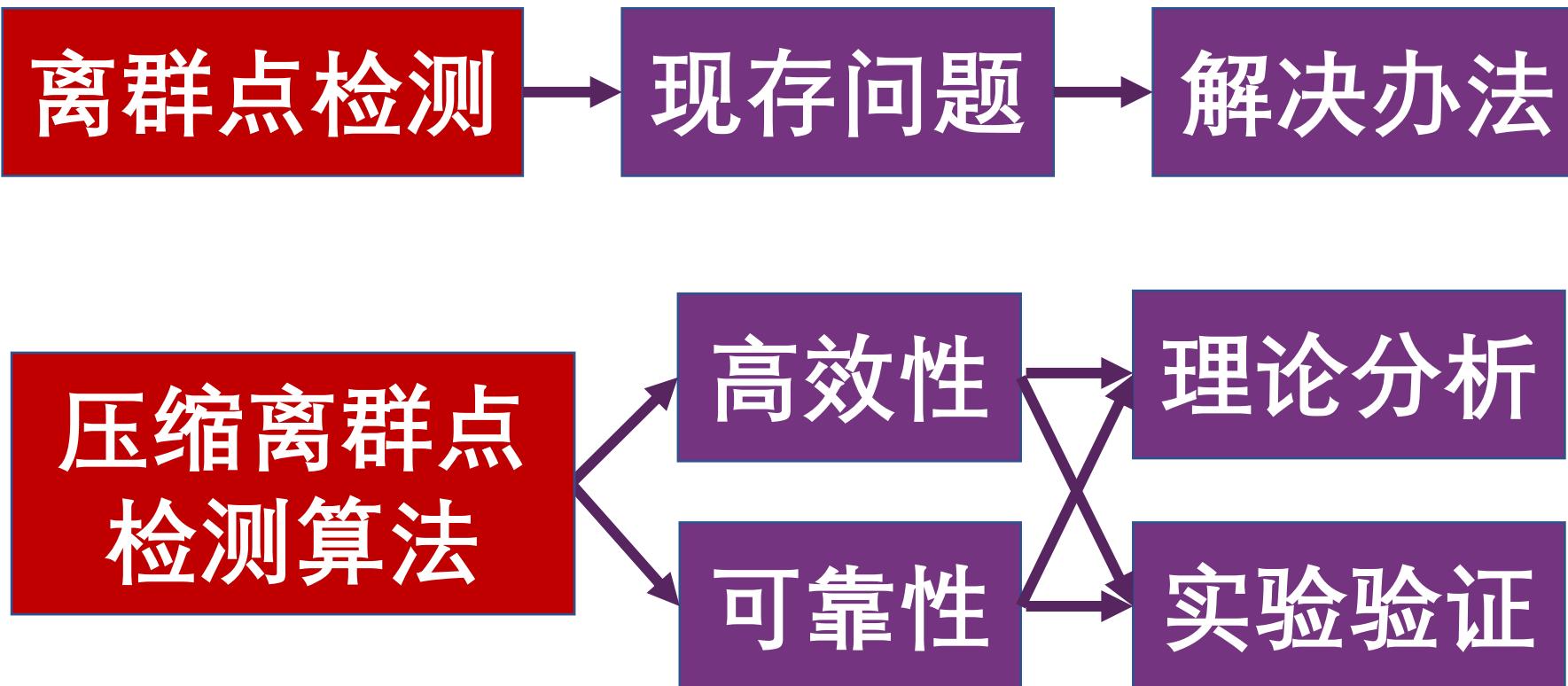
lkz18@mails.tsinghua.edu.cn

liukunzan.github.io

2020年8月13日



# 研究小结



- 项目背景
- 算法提出
- 理论分析
- 实验结果
- 整理总结