RANSCA

陈烁龙

2022 年 8 月 8 日

目录

插图

表格

摘要

RANSCA 算法是一种基于概率的模型构建手段。其相较于最小二乘法，能够在数据集存在较多粗差或者误差数据的情况下，重构处正确的模型。

关键词：RANSCA，最小二乘，直线拟合

1 概述

只要是通过传感器获取的数据，都不可避免的存在误差。如果误差非常完美的满足正态分布，则使用最小二乘可以很好的构建出模型。但是很多时候，误差分布差异较大，甚至存在较多粗差 (错误) 数据，这时传统的最小二乘法就会失效。

当然，通过均值漂移或者方差膨胀进行的最小二乘，可以弥补传统最小二乘的不足。但是实现较为复杂，效率较低，参数不易控制。再者，若基于高斯-牛顿等数值优化方法，可以定义相应的核函数，以抑制误差较大项对损失函数的贡献。

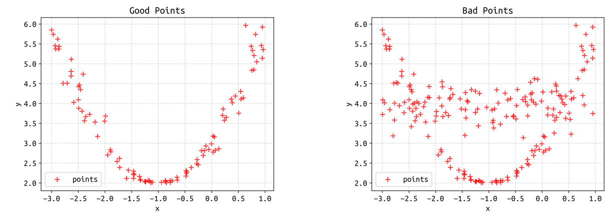
本文主要介绍 RANSCA 方法，即随机采样一致性算法，对于该种问题有比较良好的解决效果。另外，本文也会使用高斯牛顿和最小二乘方法作为对照，说明三者的不同。

2 问题

线有一条位于某平面的抛物线，为获取其函数表达式，对其进行了多次采样 (即通过某种设备测量曲线上点的二维位置)。由于设备的精度有像，加之测量环境较差，得到了比较差的测量结果，如下图所示：

现要求给出该曲线的表达形式，即下表达式中的参数：

$$y = ax^2 + bx + c$$



(a) 不存在粗差 (b) 存在粗差

图 1: 源数据

2.1 最小二乘

设测得的点集为 $PC = \{p_1, p_2, ..., p_n\}$，且 $p_i = (x_i, y_i)$。由于存在测量误差，所以点的坐标值不会完全符合真实曲线的函数表达式。我们记：

$$v_i = ax_i^2 + bx_i + c - y_i$$

对于所有的点，我们可以列出以下表达式：

$$\begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ & \cdots & \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}$$

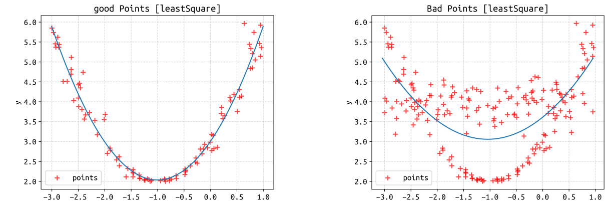
即：

$$V = BX - l$$

求解线性方程组：

$$X = (B^T B)^{-1} B^T l$$

即可获得解。



(a) 不存在粗差 (b) 存在粗差

图 2: 最小二乘

2.2 高斯-牛顿法

首先写出我们的损失函数：

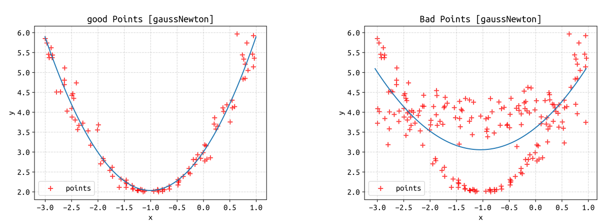
$$Error = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(ax_i^2 + bx_i + c - y_i)$$

对于单个误差项 e_i，我们对待求参数求解雅可比矩阵：

$$\begin{cases} \frac{\partial e_i}{\partial a} = x_i^2 \\ \frac{\partial e_i}{\partial b} = x_i \\ \frac{\partial e_i}{\partial c} = 1 \end{cases} \rightarrow j_i = \begin{pmatrix} x_i^2 \\ x_i \\ 1 \end{pmatrix}$$

对于每一个数据，我们都求解误差和雅可比矩阵，而后求和。最后求解线性方程组迭代得到结果：

$$\begin{cases} H = \sum_{i=1}^{n}(j_i j_i^T) \\ g = \sum_{i=1}^{n}(-j_i e_i) \\ H\Delta X = g \end{cases}$$



(a) 不存在粗差 (b) 存在粗差

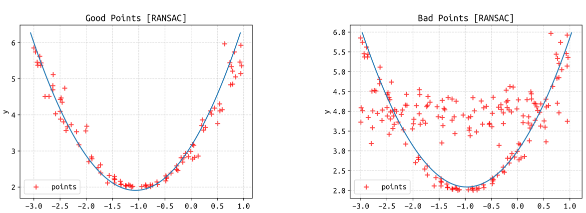
图 3: 高斯-牛顿法

2.3 RANSCA

RANSAC 算法假设数据中包含正确数据和异常数据 (或称为噪声)。该算法核心思想就是随机性和假设性，随机性是根据正确数据出现概率去随机选取抽样数据，根据大数定律，随机性模拟可以近似得到正确结果。假设性是假设选取出的抽样数据都是正确数据，然后用这些正确数据通过问题满足的模型，去计算其他点，然后对这次结果进行一个评分。

对于当前问题，该算法的步骤表述为：

1. 给出迭代次数和容许的误差阈值。容许的误差阈值即我们可以接受的，点到拟合的曲线的容许偏差。该值需要根据具体问题给出；
2. 选择三个点，计算出一个初始的二次曲线模型；
3. 根据当前拟合得到的模型和容许误差，选择出符合点 (inliers)；
4. 基于满足容许误差的数据点，再次拟合模型；
5. 基于新的模型，计算每个符合点的残差。而后计算平均残差和作为该模型的衡量数值 (越小越好)。这里我们用点在 y 方向上到该二次曲线的距离来进行衡量；
6. 重复迭代。迭代完成后，返回平均残差和最小的模型。



(a) 不存在粗差 (b) 存在粗差

图 4: RANSCA