哈爾濱工業大學

人工智能数学基础实验报告

题　　目	数据拟合
学　　院	计算机科学与技术
专　　业	人工智能
学　　号	2021112845
学　　生	张智雄
任 课 教 师	刘绍辉

哈尔滨工业大学计算机科学与技术学院

2023.5

实验一:数据拟合

1、 实验内容或者文献情况介绍

1.1 数据拟合的背景

数据拟合(Data Fitting)是一种重要的统计学和机器学习方法，能够广泛应用于数据分析、预测、模型推断和决策支持等领域。

具体而言，是指在已知的数据集合中，通过建立数学模型来获取数据的整体特征、描述和预测数据的趋势和规律的分析方法。面对繁多且冗杂的原始数据时，数据拟合能够一定程度上解释数据间的复杂关系，并对未知数据进行预测和推断。

1.2 数据拟合的方法

数据拟合通常能够通过结合拟合、回归和神经网络等方法来求解：

- **拟合：**最小化观测数据与（线性、多项式、指数、对数等）模型预测值之间的误差来确定最佳拟合曲线或函数。
- **回归：**通过一定的参数估计（最大似然、梯度下降等），建立一个能够解释或预测因变量的模型，以揭示自变量与因变量之间的关系。
- **神经网络：**能够处理高维数据和复杂的非线性关系，在函数逼近、回归分析和分类等任务上具有良好的拟合效果。
- **其他方法：**如支持向量机、稀疏拟合方法、时间序列拟合等方法也能用于解决一些具体问题中的数据拟合。

1.3 实验内容

分别使用 ransac 方法和最小二乘方法对直线和曲线来进行拟合。

1.3.1 直线拟合

- 根据直线方程$ax + by + c = 0$，产生随机点(x_i, y_i)，增加随机噪声成为$(x_i + n_i, y_i + m_i), i = 1,2, \ldots$，根据这些点，拟合直线 $ax + by + c = 0$中的参数。
- 如果有一系列平行直线$ax + by + c_1 = 0$，$ax + by + c_2 = 0$，$ax + by + c_3 = 0$，然后对直线上的点添加类似的噪声，拟合这些平行直线的参数。

1.3.2 曲线拟合

- 假设给定函数$y = asin\omega 1t + bsin\omega 2t$ (可以自定义函数)，从中生成n个点 $(t_i, \hat{y_i}), i = 1,2, \ldots$，假设点$\hat{y_i} = y_i + n_i, y_i = asin\omega_1 t_i + bsin\omega_2 t_i$， 其中$n_i$是服从高斯分布或者拉普拉斯分布的随机噪声，然后用拟合、回归和神经网络的方法来求解模型f，使得模型$f(t, w)$拟合原函数的误差最小。
- 尝试添加一些外点，观测拟合效果，并提出改进方法。

2、 算法简介及其实现细节

2.1 算法简介

最小二乘法(Least Squares Method)是一种常用的数学优化方法，用于拟合数据和估计模型参数。它通过最小化观测值与模型预测值之间的残差平方和来找到最优解，实现伪代码如下：

ALGORITHM 1 Least Squares Fitting (最小二乘拟合)

1：**input** X ←数据点集，Y ←观测值集，$iter$ ←迭代次数；
2：构建线性方程组的矩阵形式$XA = B$；
3：**while** $i < iter$ **do**
4：　　计算残差平方和$loss = \left|Y_{predict} - Y_{target}\right|^2$；
5：　　利用梯度下降等方法调整模型参数X，最小化$loss$；
6：**end while**
7：返回模型参数向量X；

RANSAC(Random Sample Consensus)是一种鲁棒性较强的拟合方法，用于估计数学模型参数，对于含有异常值或噪声的数据集具有较好的适应性。RANSAC通过迭代的方式，从数据集中随机选择一部分样本进行拟合，并根据预定义的阈值判断样本是否符合拟合模型，从而筛选出符合拟合模型的内点集合。
具体实现伪代码如下：

ALGORITHM 2 Random Sample Consensus(RANSAC 拟合)

1：**input** X ←数据点集，Y ←观测值集，$iter$ ←终止阈值；
2：内点集合$S \leftarrow \emptyset$，内点数$m \leftarrow 0$
3：随机选择最小样本集s拟合模型X，$T \leftarrow$ 内点数阈值（终止阈值）；
4：**while** $i < iter$ **do**
5：　　计算所有数据点到模型的距离d，$S \leftarrow points(d <$ 距离阈值$t)$
6：　　计算一致集S中的内点数m'
7：　　**if** $m' > T$ **do**：更新模型X，**then goto** 10；
8：　　**else**　　：重新随机选择s，拟合模型X
9：**end while**
10：返回模型参数向量X，即为满足最大一致集S的模型；

其中，距离阈值t一般可以通过经验选取，但如果测量误差服从$N\sim(0,\sigma^2)$，则点到模型几何距离的平方服从$\chi^2(n)$分布，可由$p_{\chi^2(n)}(x)$公式计算。

$$p_{\chi^2(n)}(x)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-x/2}x^{n/2-1}(0<x<\infty);0(\text{其他})$$

而终止阈值T一般没有固定的规则，常常根据内点比例的估计值，如果内点数目与一致集大小相当时停止。

2.2 实现细节

2.2.1 直线拟合

对于直线的一般方程$ax+by+c=0$，可以等价变形为斜率方程$y=kx+b$。

- 对于最小二乘法，设观测样本数为N，则残差平方和$loss$为

$$loss=\sum_{n=1}^{N}(kx_i+b-y_i)^2$$

为最小化$loss$，分别对k,b求偏导使其等于0，可得如下微分方程组，

$$\begin{cases}\dfrac{\partial(loss)}{\partial k}=k\sum_{n=1}^{N}{x_i}^2+b\sum_{n=1}^{N}x_i-\sum_{n=1}^{N}x_i\,y_i=0\\[2ex]\dfrac{\partial(loss)}{\partial b}=k\sum_{n=1}^{N}x_i-\sum_{n=1}^{N}y_i+Nb=0\end{cases}$$

由此可以解得模型参数k,b如下，具体代入数据点信息即可得到拟合直线。

$$\begin{cases}k=\dfrac{\sum_{n=1}^{N}x_i\sum_{n=1}^{N}y_i-N\bar{x}\,\bar{y}}{(\sum_{n=1}^{N}x_i)^2-N\bar{x}^2}\\[2ex]b=-\dfrac{\sum_{n=1}^{N}{x_i}^2\sum_{n=1}^{N}y_i-(\sum_{n=1}^{N}x_i)^2\bar{y}}{(\sum_{n=1}^{N}x_i)^2-N\bar{x}^2}\end{cases}$$

- 而对于 RANSAC 方法，初始随机选择样本集J作为内点拟合直线（基于最小二乘法或其他方法，本文选用最小二乘法），通过距离阈值t划分内点和外点，更新内点集合重复拟合，直至拟合模型合理或达到迭代次数为止。其中点(x_0,y_0)距离到直线$y=kx+b$的距离d计算公式为：

$$d=\frac{|kx_0-y_0+b|}{\sqrt{1+k^2}}$$

2.2.2 曲线拟合

对于给定函数$y=asin\omega 1t+bcos\omega 2t$，假定拟合时函数形式未知，实验采用最小二乘法和 RANSAC 方法对其进行多项式拟合。

- 对于最小二乘法，设观测样本数为N，多项式函数形式为$y(x,a)=a_m x^m+a_{m-1}x^{m-1}+\cdots+a_1x^1+a_0=\sum_{j=1}^{m}a_jx^j\}$，可形式化为一个最优化问题：

$$\min E(w), E(w)=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,a)-t_n\}^2$$

而$y(x,a)$又可写成矩阵相乘形式为

$$\begin{bmatrix} x_1^m & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_N^m & \cdots & 1 \end{bmatrix}\begin{bmatrix} a_m \\ \vdots \\ a_0 \end{bmatrix}=\begin{bmatrix} y_m \\ \vdots \\ y_0 \end{bmatrix}$$

为方便表示可记为$X\cdot A=Y$，其中X为范德蒙德行列式(Vandermonde determinant)，则残差平方和$loss$为，

$$loss=\|X\cdot A-Y\|^2$$

则目标函数为$\min(loss)$，将$loss$展开得到，

$$loss=(X\cdot A-Y)^T\cdot(X\cdot A-Y)=(A^T\cdot X^T-Y^T)\cdot(X\cdot A-Y)$$
$$=A^TX^TXA-A^TX^TY-Y^TXA+Y^TY=A^TX^TXA-2A^TX^TY+Y^TY$$

由于$loss$是A的函数，则对A求导$\frac{\partial(loss)}{\partial A}=\mathbf{0}$即求得模型参数$X$，$\frac{\partial(loss)}{\partial A}$具体为，

$$\frac{\partial(loss)}{\partial A}=\frac{\partial(A^TX^TXA-2A^TX^TY+Y^TY)}{\partial A}=2X^TXA-2X^TY=0$$

解得$A=(X^TX)^{-1}X^TY$

将数据点信息代入上述公式即可求得多项式拟合模型。

- 而对于 RANSAC 方法，初始随机选择样本集J作为内点拟合曲线（基于最小二乘法或其他方法，本文选用最小二乘法），通过距离阈值t划分内点和外点，更新内点集合重复拟合，直至拟合模型合理或达到迭代次数为止，其中点(x_0,y_0)距离可以通过计算点与多项式函数f在函数值上的差异来衡量。

3、 实验设置及结果分析（包括实验数据集）

实验数据来源于给定函数曲线的随机采样，并加入服从正态分布的噪声，对于直线产生随机点$(x_i,y_i),i=1,2,\ldots$，增加随机噪声成为(x_i+n_i,y_i+m_i)；对于曲线从中生成n个点$(t_i,y_i),i=1,2,\ldots$，增加随机噪声成为(t,y_i+n_i)。函数拟合过程中，函数形式及参数对于拟合器均未知。

3.1 单直线拟合

给定测试函数$2x-y-6=0$，在$[0,10]$区域随机均匀采样 100 组样本点，增加服从正态分布$N\sim(0,1)$的噪声，分别使用最小二乘法和 RANSAC 方法进行拟合，结果如下。

最小二乘法拟合结果$k = 1.94083, \mathrm{b} = -5.84156$

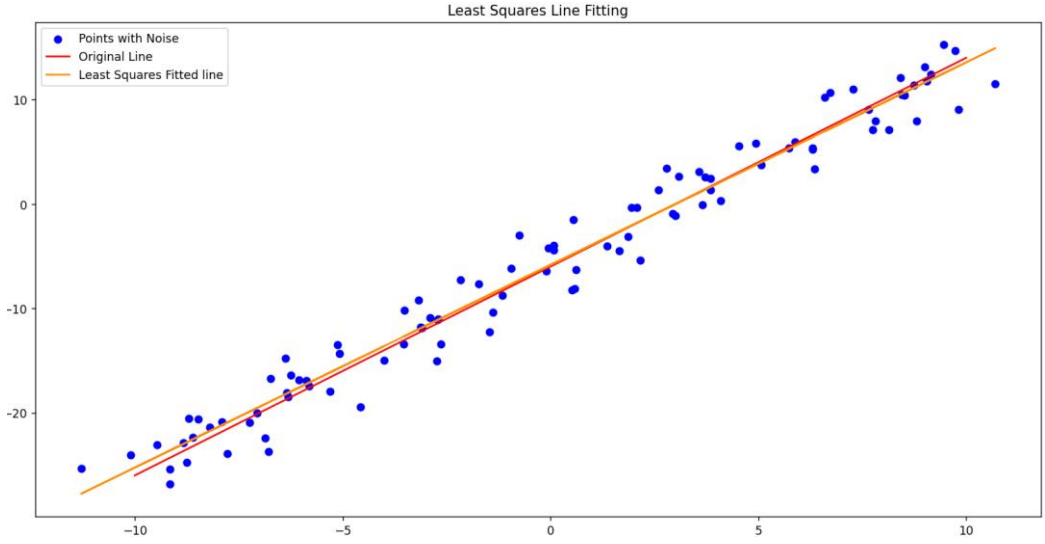

Figure 1　最小二乘法拟合直线

（蓝点为带噪声的采样点，红线为参考直线，橙线为拟合直线）

RANSAC 法拟合结果$k = 1.94083, \mathrm{b} = -5.84156$，分析 RANSAC 法与最小二乘法结果几乎一致的原因为：此时无外点干扰，且 RANSAC 时以最小二乘法为基础模型进行拟合的。

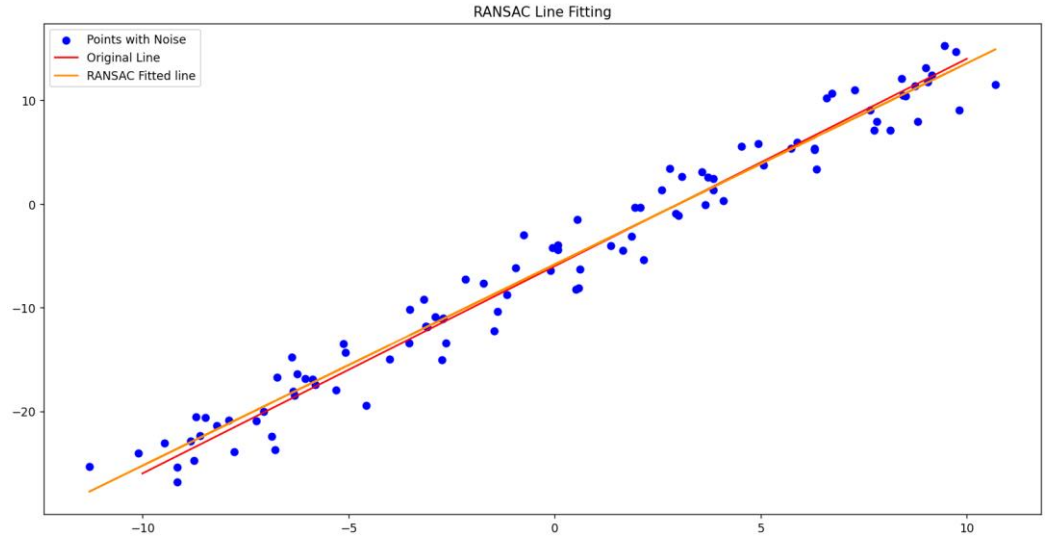

Figure 2 RANSAC 法拟合直线

3.2 一组平行直线的拟合

对于一组平行直线，抽样数据点集混杂，直接拟合机器难以辨别。因而需要首先将样本点进行分类处理。

查阅资料发现 RANSAC 算法还可用于分类拟合问题，其基本思想是通过迭代的方式从数据集中随机选择一部分数据点，嵌入多类别 SVM 或多类别逻辑回归等算法构建模型。而对于内点的选择以及迭代优化上与单直线情况基本一致。

实验针对平行直线$2x+3y-10=0, 2x+3y=0, 2x+3y+10=0$随机抽样，而后进行分类拟合，得到结果如下。

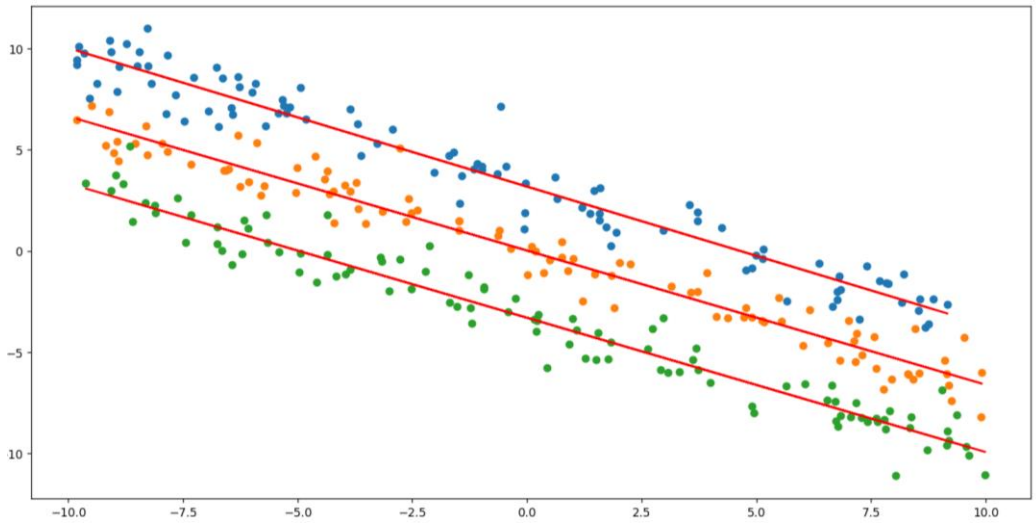

Figure 3 RANSAC 分类拟合平行直线

（不同颜色的点代表 RANSAC 对来自不同直线的点的分类结果）

3.3 曲线拟合

给定测试函数$y = sint + 0.5cos2t$，在[0,10]区域随机均匀采样 100 组样本点，增加服从正态分布$N\sim(0,0.1)$的噪声，分别使用最小二乘法和 RANSAC 方法进行多项式拟合，实验发现，当多项式最高项次数大于 8 时，拟合效果较好；同时最高次数超过 15 时会发生过拟合现象，具体结果如下。

在最小二乘法中，部分不拟合的原因可能为：多项式拟合的局限性，以及该数据区域的噪声干扰不均匀导致的。

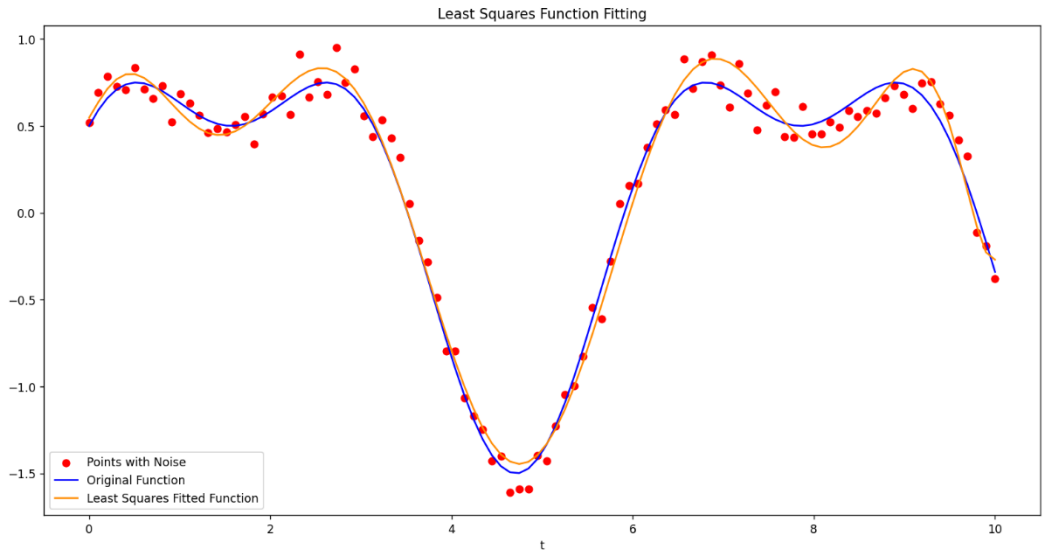

Figure 4 最小二乘法拟合曲线（degree=10）

（红点为带噪声的采样点，蓝线为参考曲线，橙线为拟合曲线）

而在 RANSAC 法中，对比上图，左侧部分拟合数据高于图 4 中对应区域，而右侧区域拟合效果好于图 4，原因推测为 RANSAC 算法舍弃了此区域机器认为的“干扰”点，因而导致拟合结果存在不一致的情况。

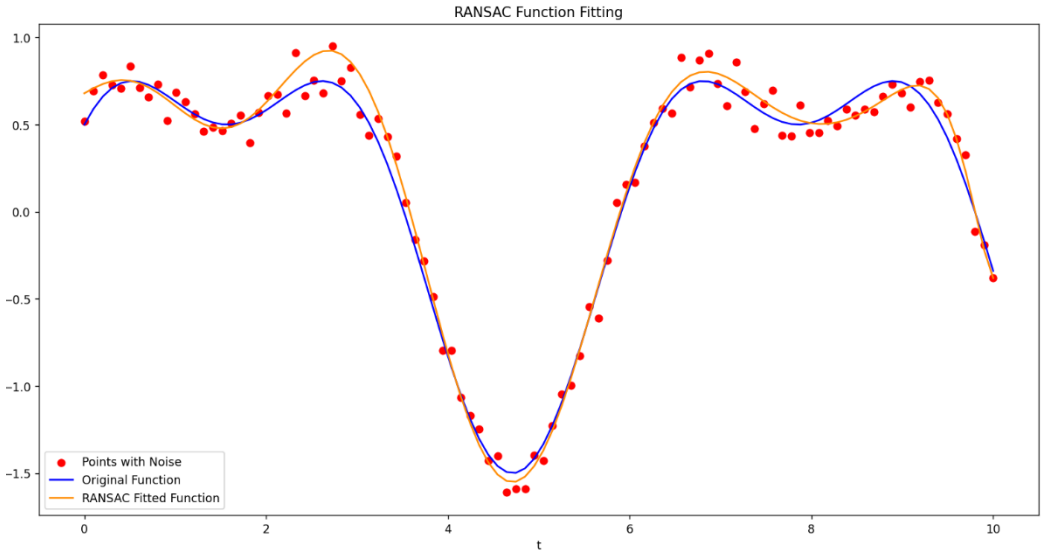

Figure 5 RANSAC 法拟合曲线（degree=12）

同时，实验过程中，还尝试了利用 MLPRegressor 建立多层感知器回归模型，此种模式不对模型形式进行明确假设的情况下进行数据拟合，提供了更大的灵活性，但模型的可解释性较差，对数据的依赖性强，需要一定数量的数据和一定深度的网络结构才能得到较为拟合的结果。

实验发现，以 ReLU 函数为激活函数，当模型结构为 9 层，每层 100 个神经元时，可以得到较好的拟合结果，但在数据集右侧仍有极不拟合的部分，可以考虑设置L_2正则项加以约束，以减少过度拟合的影响，提高模型的泛化能力，具体结果如下。

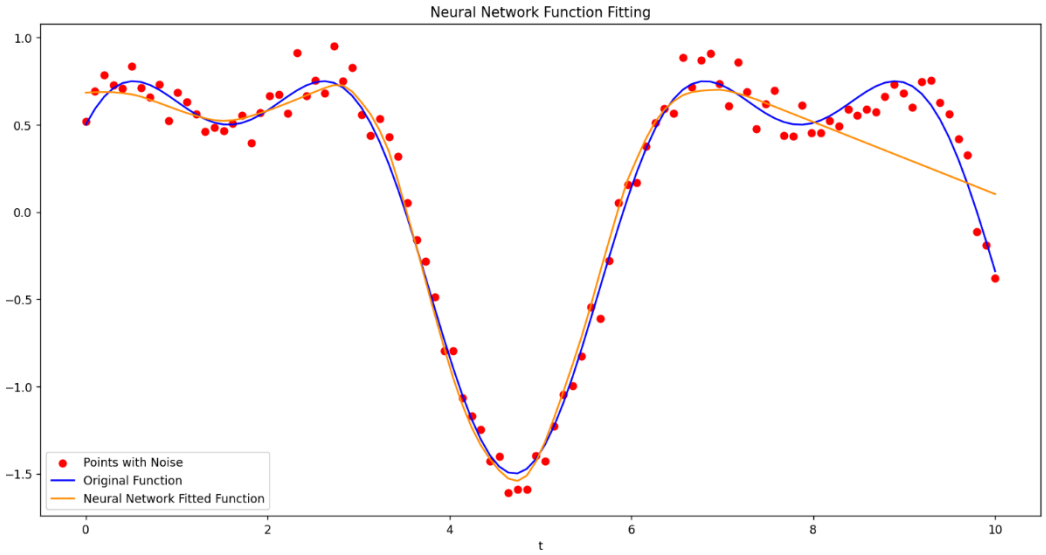

Figure 6 神经网络拟合曲线（未加正则项）

设置L_2正则项后，结果好了一点点，但是可能由于数据抽样点少且不均匀的缘故，拟合的效果还是比较差，数据受干扰的影响较大。

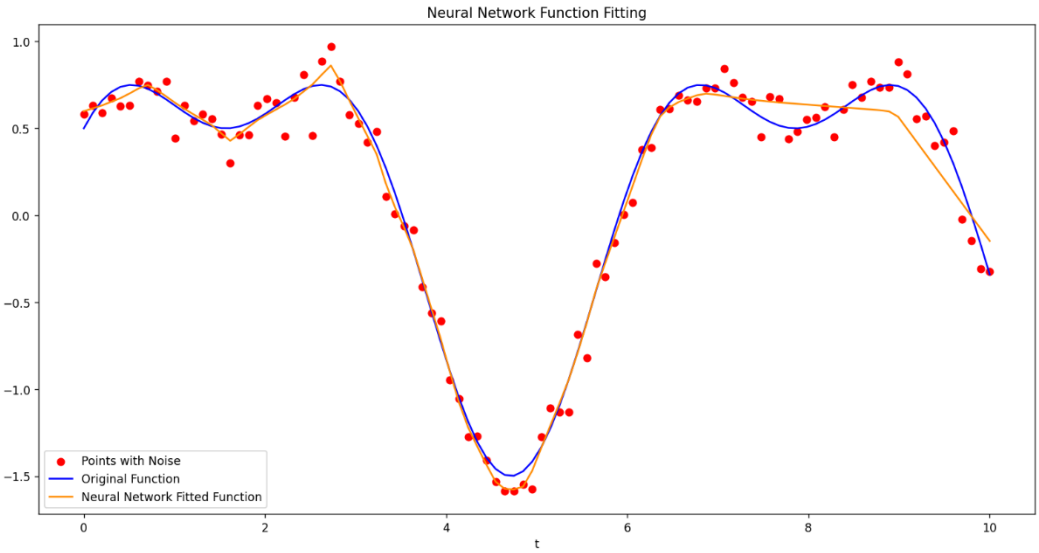

Figure 7 神经网络拟合曲线（加正则项）

3.4 鲁棒性检测

针对上述直线和曲线的拟合，加入 15 个外点，即误差很大的点，观测算法对外界干扰的鲁棒性。

3.4.1 直线检测

为便于观测，向数据点集中增加 15 个明显高于函数值的外点，设置噪声水平$N\sim(0,1)$，具体结果如下。

对于最小二乘法，直线的拟合受到外点的影响斜率k变小，而截距b变大，拟合直线与参考直线有明显差距，

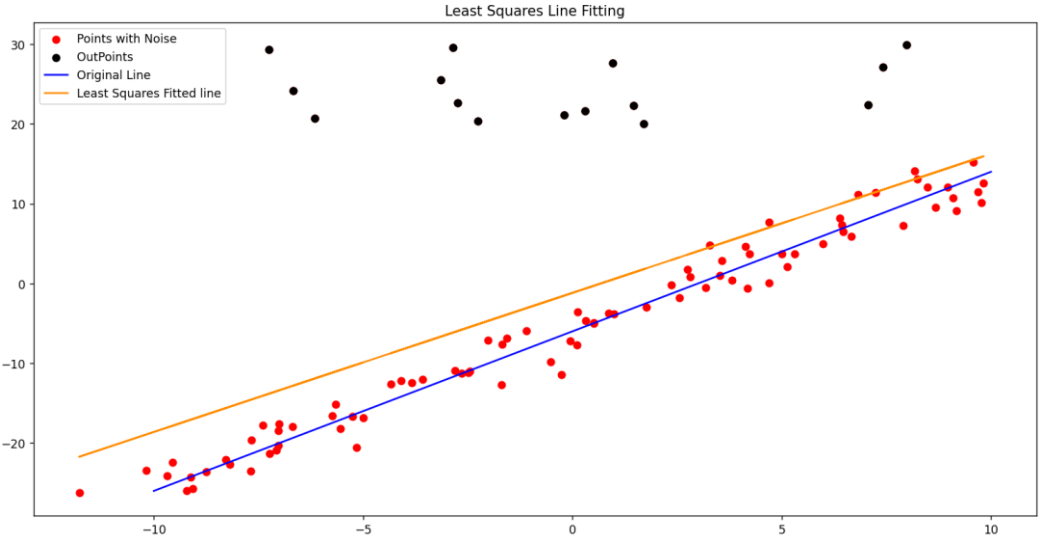

Figure 8 最小二乘法拟合直线（含外点）

（黑点为外点，红点为带噪声的采样点，蓝线为参考直线，橙线为拟合直线）

而 RANSAC 方法拟合效果基本不受影响，因为算法本身已经删除了偏差较大的数据点，具有较好的鲁棒性。

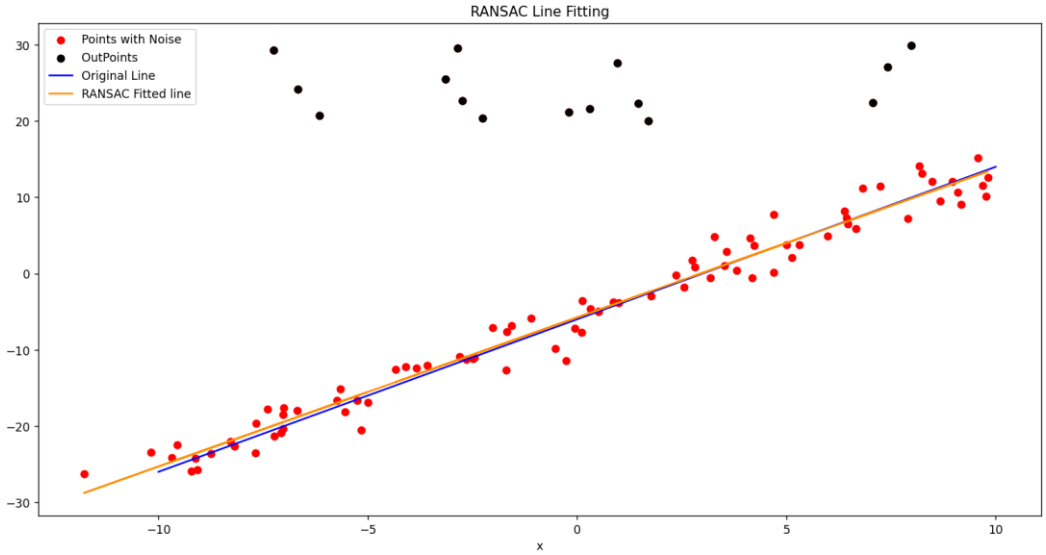

Figure 9 RANSAC 法拟合直线（含外点）

3.4.2 曲线检测

随机加入 15 个偏差较大的外点后，设置噪声水平$N\sim(0,0.1)$，实验发现，由于外点和算法本身的随机性，拟合效果具有一定波动性，本节主要讨论的是经过重复实验的一般化情况，具体拟合结果如下。

最小二乘法受外点影响较大，许多区域都受外点影响发生了偏移，尤其是在外点集中分布的数据域，拟合效果明显下降。

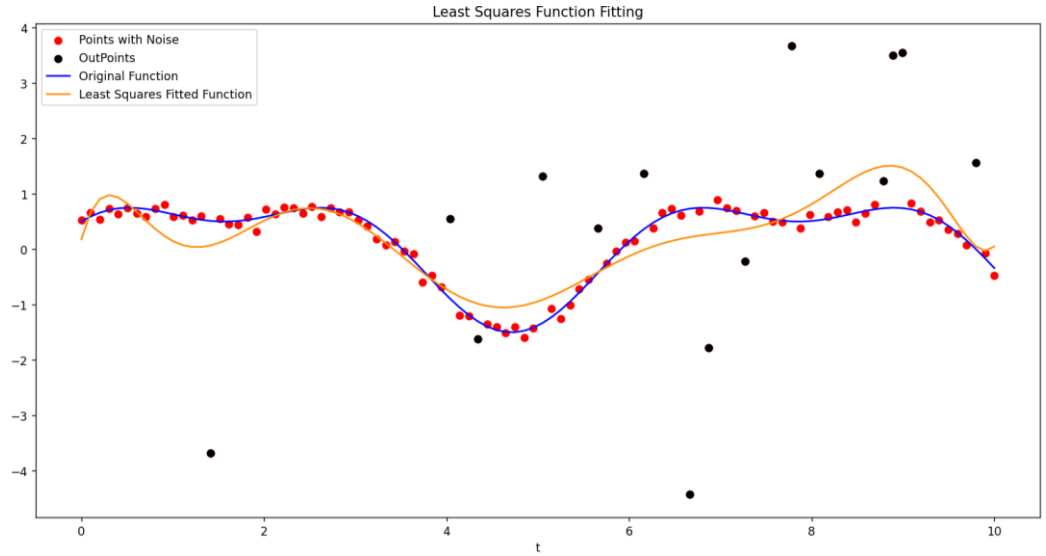

Figure 10 最小二乘法拟合曲线（含外点）

（黑点为外点，红点为带噪声的采样点，蓝线为参考曲线，橙线为拟合曲线）

而 RANSAC 方法拟合效果在大部分数据点较好，基本不受外点的影响。

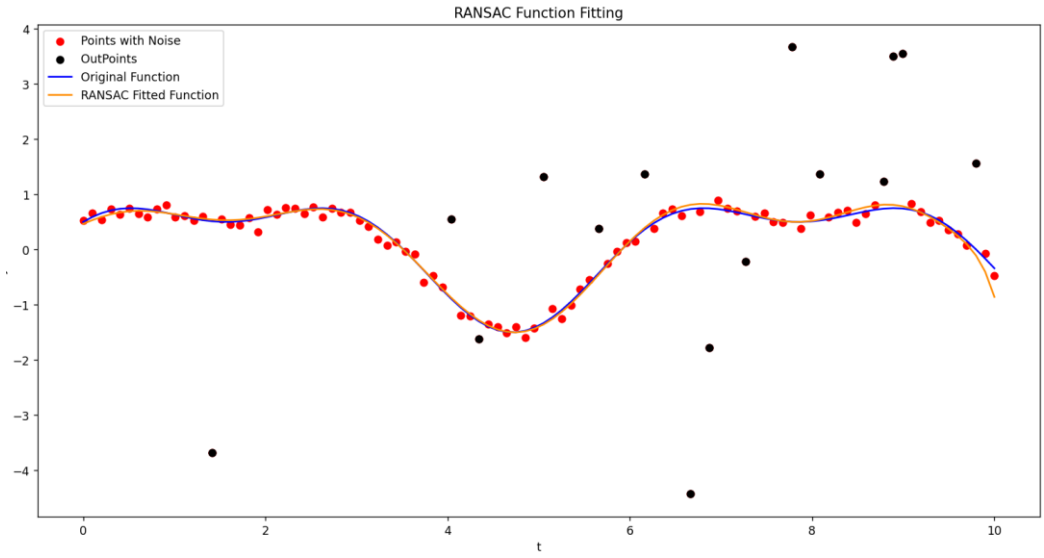

Figure 11 RANSAC 法拟合曲线（含外点）

而神经网络对外点的干扰尤为敏感，推测可能此数据段数据点较稀疏，网络不能正确区分内外点而学习其特征的原因导致某个点出现陡增陡降的情况，函数极为不规则，在外点集中分布的数据域拟合结果较差。

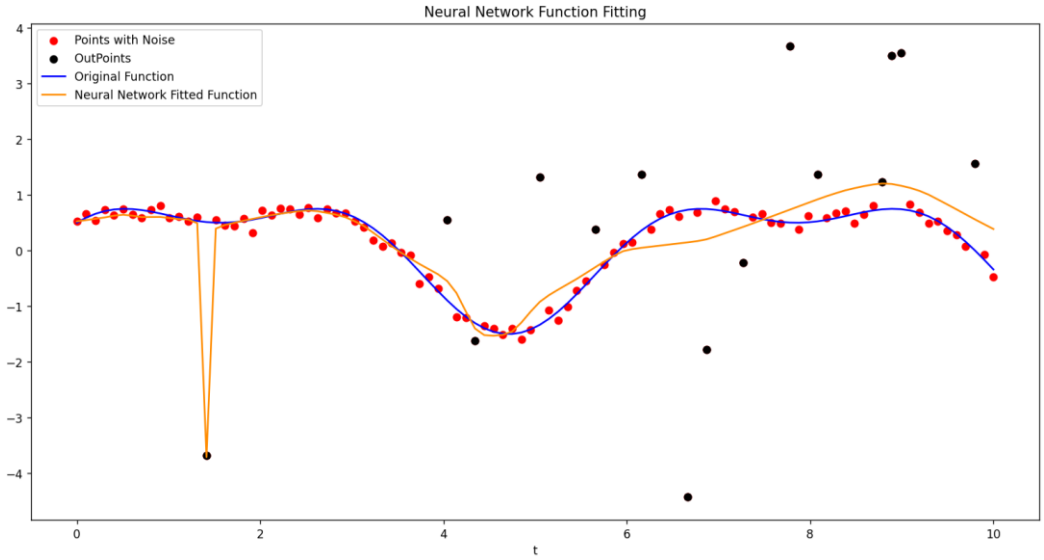

Figure 12 神经网络拟合曲线（含外点）

3.5 针对外点的优化

为减少外点对拟合的干扰，提高模型的鲁棒性，除上述所述的 RANSAC 算法，还可以通过性质更加好的函数形式以及增加正则项的方式实现优化。

3.5.1 给定更加优秀的函数形式

可以直接设置函数形式为$y = asin\omega_1 t + bcos\omega_2 t$，利用 curve_fit 方法，对函数中的参数初始化，迭代寻找效果较好的拟合参数，具体结果如下。

但此种方法受函数形式和参数初值的影响较大，在实际应用中需要足够的先验知识进行辅助。但在未知数据点的预测表现优于上述几种方法，在实际分析中一般首先对实际情况进行分析，建立合理的数学模型后进行数据拟合，能大大提高预测的可靠性。

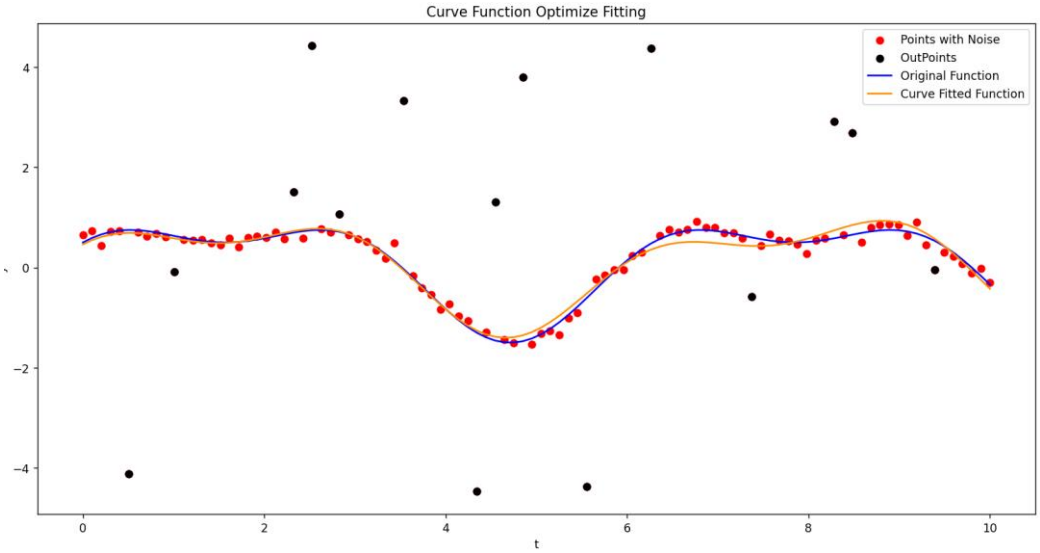

Figure 13 给定函数形式拟合曲线（含外点）

3.5.2 增加正则项

还可通过增加正则化的方式对近似模型进行约束，一般包括L_1正则项和L_2正则项。添加正则项后，问题可以形式化为，

$$\min \tilde{E}(w), \tilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, a) - t_n\}^2 + \lambda L_{regular}$$

其中，λ控制正则项与误差项的均衡程度。

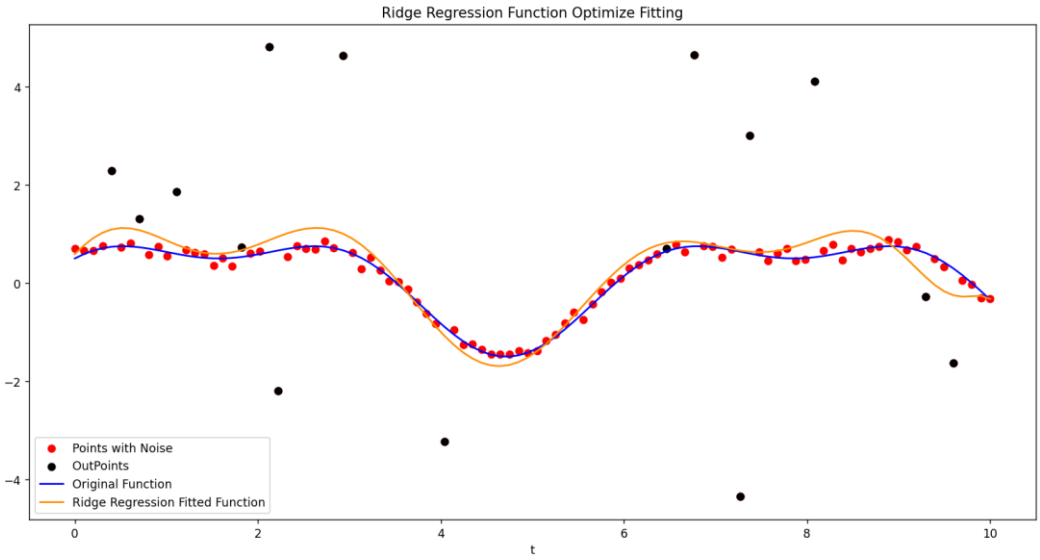

Figure 14 岭回归拟合曲线（含外点）

L_1正则项指在损失函数中添加模型参数的L_1范数作为正则化项，使得模型参数w中的一部分变为 0，从而实现特征选择和模型稀疏性。

$$L_1 = \lambda \|w\|_1$$

而L_2正则项是指在损失函数中添加模型参数的L_2范数作为正则化项，惩罚模型参数的平方和，减小模型参数的幅度，减少模型的过拟合风险。

$$L_2 = \lambda \|w\|_2^2$$

在此模型中参数较少，使用L_1正则意义不大，因而选择利用L_2正则项的岭回归(Ridge Regression)算法，能够实现对模型进行平滑处理，具体结果如上图 14。

3.6 结果分析

上述拟合过程已通过作图简单直观分析，接下来从数据层面评价上述各种方法的拟合效果。采用均方根误差$\boldsymbol{Root-Mean-Squre(RMS)}$进行误差评价

$$E_{RMS} = \sqrt{2E(w^*)/N}$$

其中，较小的E_{RMS}值表示拟合效果较好，误差较小；较大的E_{RMS}值表示拟合效果较差，误差较大。

3.6.1 直线的误差评价

对于上述直线$2x - y - 6 = 0$，在有外点和不加外点干扰情况下，最小二乘法和 RANSAC 方法拟合结果的E_{RMS}计算如下

E_{RMS}	无外点	有外点
最小二乘法	0.284278	5.452107
RANSAC 方法	0.284275	0.459128

Table 1 直线误差评价

可以发现与上述 3.1 节和 3.4.1 节所得结果基本一致，最小二乘法和 RANSAC 方法在无干扰时拟合效果几乎一致；在有干扰时，RANSAC 方法的鲁棒性明显优于最小二乘法。

3.6.2 多项式的次数选择

对于测试函数$y = sint + 0.5cos2t$的多项式拟合次数选择，实验测试了$degree = 1{\sim}15$的情形，比较了不同$degree$下最小二乘和 RANSAC 法的E_{RMS}拟合误差，据此选择合适的拟合多项式次数。

实验发现，对于最小二乘法，随$degree$的增大，E_{RMS}呈现下降趋势，而当$degree > 10$时，E_{RMS}处于较低水平，且变化较为平缓；而对于 RANSAC 方法，随$degree$的增大，E_{RMS}呈现先上升后下降的趋势，而当$degree > 11$时，E_{RMS}处于较低水平，且变化较为平缓。

因而最终实验选取的是$degree = 10\sim12$的多项式拟合，在保证拟合效果的同时尽量避免过拟合现象的发生。

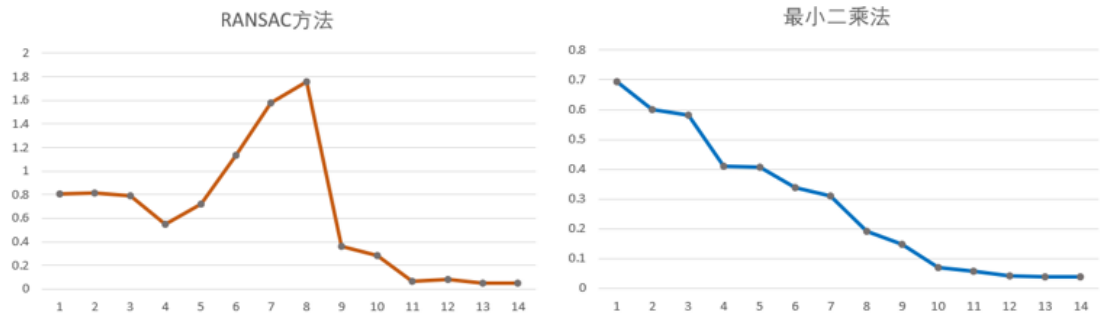

Figure 15 多项式次数对误差的影响

（左为 RANSAC 方法，右为最小二乘法；横轴为 degree，纵轴为 RMS 误差）

3.6.3 拟合方法的比较

对于加入外点情形，对上述实验使用到的不同拟合方法进行误差评价。具体结果如下表。可以看到 RANSAC 的效果仍然是最优的，岭回归和曲线拟合相较于最小二乘都有不同程度的优化。

	最小二乘	RANSAC	神经网络	岭回归	曲线拟合
RMS	0.500301	0.231138	0.658387	0.479458	0.391720

Table 2 拟合方法的比较（含外点）

4、 实验结论

实验完成了对单一直线、一组平行直线和曲线在有外点和无外点干扰情况下的多种方法拟合，对拟合方法的鲁棒性进行了检验，通过 RMS 误差评估了各种拟合方法的拟合效果，可得到如下结论。

1) 相较于最小二乘法，RANSAC 算法对于异常值或噪声的数据集具有较好的适应性。二者在无较大干扰时的拟合能力近似，而对于有“外点”的干扰情形，RANSAC 模型的鲁棒性更好。

2) 但由于噪声和算法本身的随机性，拟合效果可能出现不一致的情况，甚至出现预测值无穷大的情况。本报告中均针对重复实验得到的一般情况进行分析。

3) 实验所做拟合缺乏对函数本身的数学分析，更多依赖于抽样点本身的数据分布特征，在抽样数据域上拟合的误差较小，但会在未知数据点的预测上出现较大误差。

4) 如果能在拟合前确定函数形式，仅做参数的收敛拟合，效果会更好，同时收敛速度也会更快。但拟合效果受初值影响具有一定随机性，可能最终得到局部最优解而非全局最优解。

5) 可以通过增加正则项的方式对近似模型进行约束，实现特征选择和对模型的平滑处理，但是系数的选择需要多次实验。

6) 利用神经网络的方式可以对图形进行非参数拟合，不对模型形式进行明确假设，灵活性较强，但缺乏可解释性，且对数据的依赖较强，容易过拟合，需要较大数据样本才能得到较为稳定的拟合结果。

7) 还可尝试 Huber 回归等模型，不同模型在不同曲线的拟合上具有不同的效果，需根据实际情况选择拟合模型。

5、 参考文献

[1] 深入理解 L1、L2 正则化 https://zhuanlan.zhihu.com/p/29360425

[2] 最小二乘公式推导 https://blog.csdn.net/qq_45717425/article/details/120665970

[3] curve_fit 拟合方法 https://blog.csdn.net/qq_43403025/article/details/

[4] 数据拟合：直线拟合-多项式拟合 https://blog.csdn.net/qq_34777600/article/details/79501932

[5] Motulsky H J, Ransnas L A. Fitting curves to data using nonlinear regression: a practical and nonmathematical review[J]. The FASEB journal, 1987, 1(5): 365-374.

[6] Strang G. Linear algebra and learning from data[M]. Cambridge: Wellesley -Cambridge Press, 2019.

[7] 李航. 统计学习方法[M]. 清华大学出版社, 2012.