回归模型预测混凝土抗压强度

组员（按字母序）：bgy zms zxy zzh

本次任务基于混凝土有关的1030条数据集进行回归预测实验，我们首先基于现有土木工程理论基础进行了变量组合，获得具有实际意义的物理量，然后使用了线性回归，支持向量机回归与高斯过程回归三种模型对强度进行预测，使用交叉验证方式评估模型，并基于Friedman检验与Nemenyi后续检验验证模型优度，最后给出最优模型的验证结果，并对本次实验进行总结。

# 任务重述

混凝土是土木工程中最重要的材料。 而混凝土的抗压强度是年龄和其他因素的高度非线性函数。这些影响包括水泥，高炉矿渣，粉煤灰，水，高效减水剂，粗骨料和细骨料等等。一些学者已经通过人工神经网络进行函数的拟合，试图建立合适的模型预测混凝土的抗压强度，并取得了有效的结果。

本次实验将在给出的一个含9个变量的数据集进行回归预测。其中，前8个是可能会影响混凝土抗压强度的变量，第9个是混凝土的抗压强度值，共1030组数据。本次实验将基于以上数据，使用各种回归模型进行对抗压强度的预测，并比较模型的合适程度。

# 数据处理和数据清洗

数据集中的数据格式如下所示：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cement (component 1)(kg in a m^3 mixture) | Blast Furnace Slag (component 2)(kg in a m^3 mixture) | Fly Ash (component 3)(kg in a m^3 mixture) | Water (component 4)(kg in a m^3 mixture) | Superplasticizer (component 5)(kg in a m^3 mixture) | Coarse Aggregate (component 6)(kg in a m^3 mixture) | Fine Aggregate (component 7)(kg in a m^3 mixture) | Age (day) | Concrete compressive strength(MPa, megapascals) |

每一个变量都有可能对混凝土强度造成一定的影响，为了更加方便地看出各个变量对强度的影响，每一个单独的属性和混凝土强度对应的散点图在图II.1中标出，下面将给出混凝土抗压强度影响因素的理论分析，然后对每个变量或组合变量进行分析。

## 混凝土抗压强度影响因素的理论分析

影响混凝土强度的因素主要有以下几个：水泥强度，水灰比，骨料，外加剂，养护工作等。

水泥的强度和水灰比是决定混凝土强度的最主要因素。水泥是混凝土中的胶结组分，其强度的大小直接影响混凝土的强度。在配合比相同的条件下，水泥的强度越高，混凝土强度也越高。当采用同一水泥（品种和强度相同）时，混凝土的强度主要决定于水灰比；在混凝土能充分密实的情况下，水灰比愈大，水泥石中的孔隙愈多，强度愈低，与骨料粘结力也愈小，混凝土的强度就愈低。反之，水灰比愈小，混凝土的强度愈高。混凝土的抗压强度与水灰比和水泥强度之间符合以下近似关系：

式中，C表示每立方米混凝土中的水泥用量，W—每立方米混凝土中的用水量，是混凝土抗压强度；表示水泥的实际强度；是经验系数，与骨料品种等有关，其数值需通过试验求得。

骨料的表面状况影响水泥石与骨料的粘结，从而影响混凝土的强度。碎石表面粗糙，粘结力较大；卵石表面光滑，粘结力较小。因此，在配合比相同的条件下，碎石混凝土的强度比卵石混凝土的强度高。骨料的最大粒径对混凝土的强度也有影响，骨料的最大粒径愈大，混凝土的强度愈小。

在混凝土中掺入外加剂，可使混凝土获得早强和高强性能，混凝土中掺入早强剂，可显著提高早期强度；掺入减水剂可大幅度减少拌合用水量，在较低的水灰比下，混凝土仍能较好地成型密实，获得很高的28d强度。在混凝土中加入掺合料，可提高水泥石的密实度，改善水泥石与骨料的界面粘结强度，提高混凝土的长期强度。因此，在混凝土中掺入高效减水剂和掺合料是制备高强和高性能混凝土必需的技术措施。

混凝土在正常养护条件下，其强度将随着龄期的增加而增长。最初的7～14天内，强度增长较快，28天以后增长缓慢，龄期延续很长，混凝土的强度仍有所增长。

## 各变量和抗压强度相关性分析

各变量和混凝土的抗压强度的散点图如图II.1 所示。

可以看出，水泥强度和抗压强度呈明显的正相关关系，高炉渣基本呈正相关关系，灰密度基本呈反相关关系，其余散点比较混乱，看不出明显的关系。

除了各变量独立的散点关系，由于水灰比在专业领域内的重要的参考性，还需要对水灰比这一特定的物理量进行分析，除此以外，由于粗粒度和细粒度骨料的绝对值可能无法直接得到一定关系，同时画出粗粒度/细粒度这一统计量于抗压强度的散点图，观察是否存在显著关系。两个物理量与抗压强度的散点关系如图II.1 最后两张图所示。尽管在图上基本也看不出关联性，但仍然将其加入到回归变量中。

很显然，使用以上10个变量（8个给出的变量，加上2个给出变量的组合）对抗压强度进行预测，一定会有一些不必要的因素，例如，从图II.1 中可以看出，在很大范围内，灰密度对强度似乎并不会有影响，同理，每一个粗粒度骨料看起来也对强度影响不大，为了更好地刻画影响因素，需要对得到的10个数据进行进一步的处理，希望获得一个比较直观的结果。

## 主成分分析（PCA）

为了获得更加显著的影响因素，将对以上10个变量进行主成分分析，由于在理论分析中，抗压强度主要与水泥强度，水灰比，骨料，温度，外加剂，龄期有关，因此在进行主成分分析时，应该将10个变量变换为6个左右的主成分。表II.1 是在对全体数据进行线性回归模型下的，主成分个数为5-9的情况下的误差统计，当主成分为6个时，相关误差统计达到最小。尽管在其他模型下，可能更多的主成分效果会更好，但6个主成分在特征空间一定领域内是一个局部最优值，本实验将基于6个主成分进行回归分析。变换后的主成份数据集记为，变换前的10个变量数据矩阵记为，则进行主成分分析后，两者的关系为：

其中，C是主成分参数矩阵，该式表示主成分由各初始统计量进行线性组合后得到。

图II.3 给出了主成分分析后，各主成分和抗压强度的散点关系。经过主成分分析后的6个变量比10个初始变量表现出更加明显的相关关系。在下面的所有回归模型中，都将使用主成分进行回归。

|  |  |  |
| --- | --- | --- |
| 主成分个数对线性回归模型误差的影响 | | |
| 主成分个数 | R方 | RMSE |
| 9 | 0.72 | 7.57 |
| 8 | 0.72 | 7.54 |
| 7 | 0.72 | 7.51 |
| **6** | **0.72** | **7.48** |
| 5 | 0.38 | 11.19 |

表II.1 主成分个数对线性回归模型误差的影响

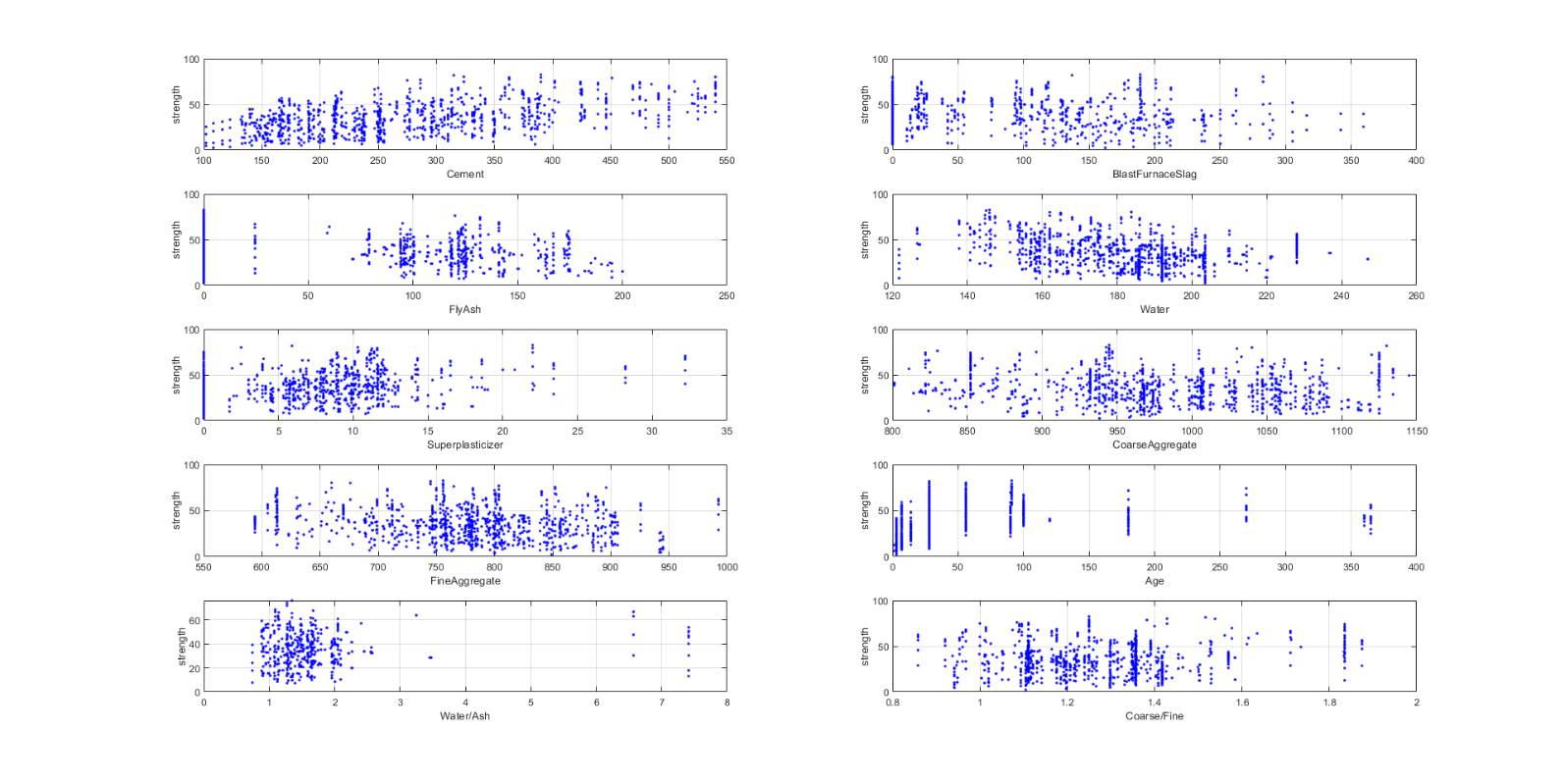


Figure II.1 各变量（及其组合）与强度的散点关系图

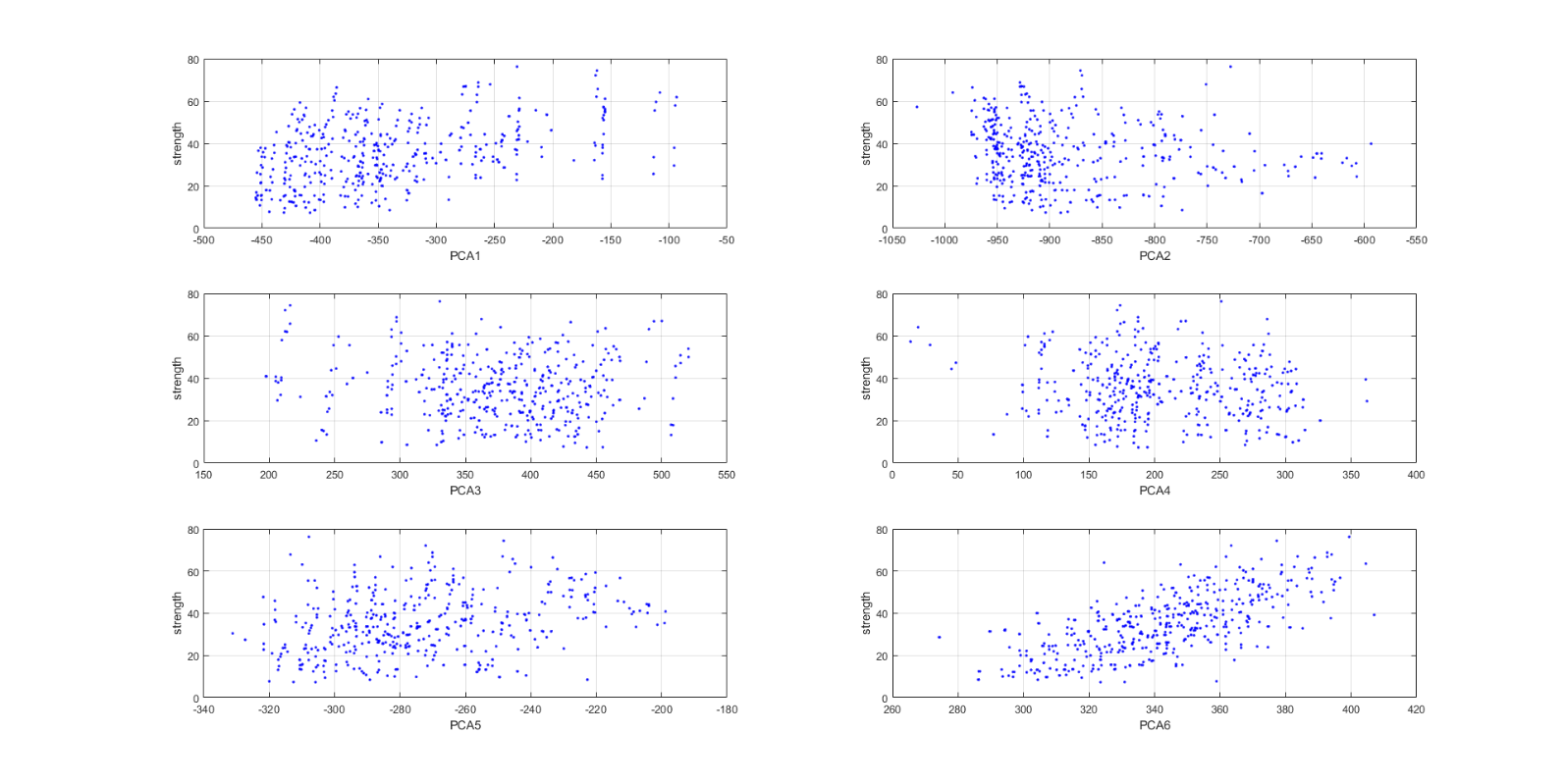


Figure II.2 各主成分与抗压强度散点图

# 回归模型分析与建立

## 线性回归模型

线性回归模型公式表达如下：

其中，参数a，b满足使下列表达式最小化：

等价于最小化下列表达式：

显然这是一个关于x的凸函数，对上式求导，令导数为0，得到

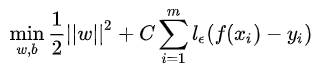
令b满足过样本中心点时，上式绝对成立，则得到方程

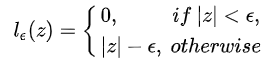
求解上述两个方程，则得到了相关参数，即得到目标线性模型。运用回归模型，只要采用的模型和数据相同，通过标准的统计方法可以计算出唯一的结果。但对于非线性系统的表现通常不会很好。

## 支持向量机回归模型

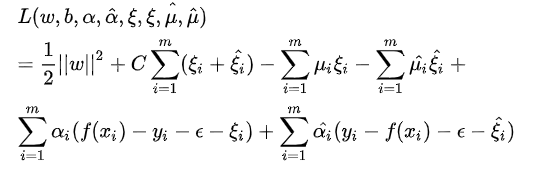
支持向量机回归公式表达式如下：

其中，参数a，b满足下列表达式：

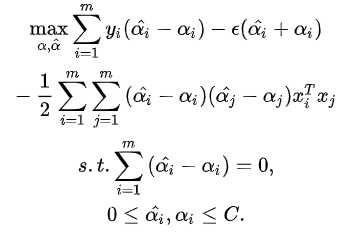




类似于解支持向量机过程，引入松弛因子，转化为拉格朗日函数:



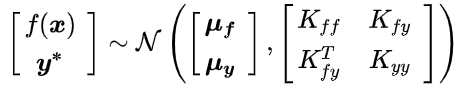
该函数也是一个凸函数，基于凸优化方法，找到其对偶问题，令其满足KKT条件，得到以下方程式：



基于SMO算法即可解出参数值。直观地理解，支持向量就是对最终的的计算起到作用的样本(α &gt; 0 \alpha\gt0α>0)。但是加入了一个ϵ 不敏感的区域，即，误差小于这个值时是允许的，不敏感区域形同一个“管道"。 “管壁”上的为边界支持向量，位于“管道”之外的为非边界支持向量。因此，该模型对具有一定抖动程度的数据具有很好的适应性。

## 高斯过程回归模型

高斯过程回归模型公式表达式如下：



带入高斯过程表达式中:

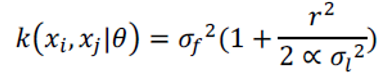


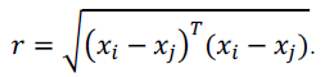
反解得到：





其中，是相应的协方差函数，即核函数，核函数在高斯过程回归中非常重要，一个恰当的核函数可以极大地提高高斯过程回归的准确率。在本次回归任务中，取：





由于GPR模型是概率模型，即是一种贝叶斯方法，因此有了训练好的模型，不仅可以预测目标值，还可以得到整个函数的分布，并计算其预测区间。因此，高斯过程对于非线性系统具有较高的准确性。

# 模型回归结果分析

## 交叉验证结果分析

交叉验证可以“充分利用”有限的数据找到合适的模型参数，防止过度拟合。交叉验证方法如下：将数据集平均分成N份，选其中的N-1份作为训练集，剩余的1份作为验证集，用以上得到的N种情况训练得到模型的参数。

为了评估模型的表现，引入两个评估指标：

### RMSE（Root Mean Squard Error）均方根误差

均方根误差表示如下:

该值越大，表示误差越大，一般来说，该值越小越好。

### (R square) R方

R方衡量模型的正确率，其表达式如下：

其中 ,表示预测值，表示实际值，表示实际值的平均。一般来说，该值越大，说明拟合效果越好。

表IV.1 给出了三种不同的回归模型的分析，同时图IV.1 给出了三种回归模型R方的箱线图，高斯过程回归模型表现较好。

|  |  |  |  |
| --- | --- | --- | --- |
| R方统计 | | | |
| 折数 | LR | SVR | GPR |
| 1 | 0.66 | 0.81 | 0.88 |
| 2 | 0.72 | 0.82 | 0.91 |
| 3 | 0.71 | 0.80 | 0.89 |
| 4 | 0.72 | 0.86 | 0.87 |
| 5 | 0.71 | 0.78 | 0.92 |

|  |  |  |  |
| --- | --- | --- | --- |
| RMSE统计 | | | |
| 折数 | LR | SVR | GPR |
| 1 | 7.85 | 6.18 | 5.24 |
| 2 | 7.18 | 6.03 | 4.42 |
| 3 | 8.18 | 6.48 | 4.53 |
| 4 | 7.79 | 4.76 | 5.05 |
| 5 | 7.03 | 6.97 | 3.90 |

表IV.1 误差分析表

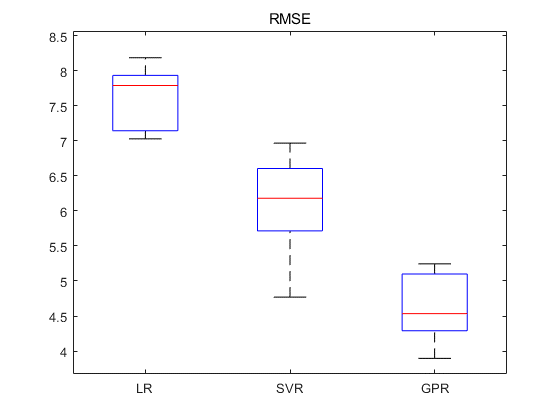
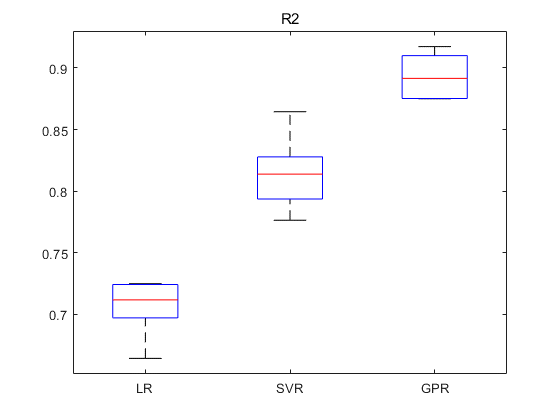


Figure IV.1 误差统计箱线图

## 假设检验验证模型优度

### Friedman检验

Friedman检验是用来判断某一模型在回归的结果上是否能显著得优于其他模型。这是一种非参数检验方法，利用模型在回归任务上表现的“序”来进行检验。考虑有M个回归模型，在N折交叉验证的情况下相当于我们有N个数据集和验证集，在这些验证集上表现最好的模型排序为1，表现最差的模型排序为M，若两个模型表现相同则平分排序。

令表示模型j在第i折交叉验证上的表现排序，其中

令表示模型j的平均表现排序，这个平均序可以表示为

检验统计量的计算方式如下

对检验统计量进行一些修正，得到新的检验统计量

这个检验统计量符合自由度及的F分布。根据统计数据的值计算检验统计量的值，并且与F分布的临界值进行比较。如果较大，则表示模型之间的性能有显著差异。

### Nemenyi后续检验

Nemenyi后续检验是用来比较模型的回归表现的，它同样是一种非参数检验方法。如果一个模型的回归表现较好，那么它的平均序将会比另一个模型的平均序高于一个判别阈值以上。这个判别阈值的计算方法如下

#### 其中，是一个取决于显著性水平的值，M是回归模型的数目，N是数据集的个数。

### 结果验证

对上文中的3种模型型，分别进行5折交叉验证，得到了如下结果：

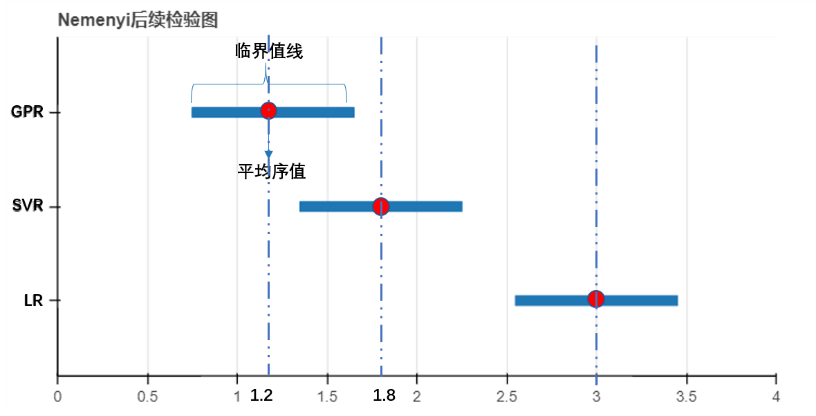
|  |  |  |  |
| --- | --- | --- | --- |
| R方排序 | | | |
| 折数 | LR | SVR | GPR |
| 1 | 3 | 2 | 1 |
| 2 | 3 | 2 | 1 |
| 3 | 3 | 2 | 1 |
| 4 | 3 | 1 | 2 |
| 5 | 3 | 2 | 1 |

表V.1 R方排序表

首先进行Friedman检验，计算检验统计量的值，在显着性水平的情况下，自由度为2和8的F分布值为11.04。而检验统计量的值计算为21，明显大于4.46。因此，原假设被拒绝，高斯模型在该回归问题上表现十分优异。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| Friedman检验 | | | | |
| FF观测值 | 不同显著性水平α下F(2,8)的值 | | | |
| 0.1 | 0.05 | 0.025 | 0.01 |
| 21 | 3.11 | 4.46 | 7.57 | 8.65 |

再进行Nemenyi后续检验，在显著性水平的情况下，的值为2.344，计算判别阈值为0.9376，故高斯回归模型的表现显著好于线性回归模型，而与支持向量机回归模型的差距不显著，若调整显著性水平，则也能得到二者表现结果显著的结果。

****

# 最优模型回归结果

由于在第一部分已经对数据进行了相关分析，本报告所有回归均建立在主成分上，下图为在主成分上进行高斯过程回归的预测值和实际值的散点图。

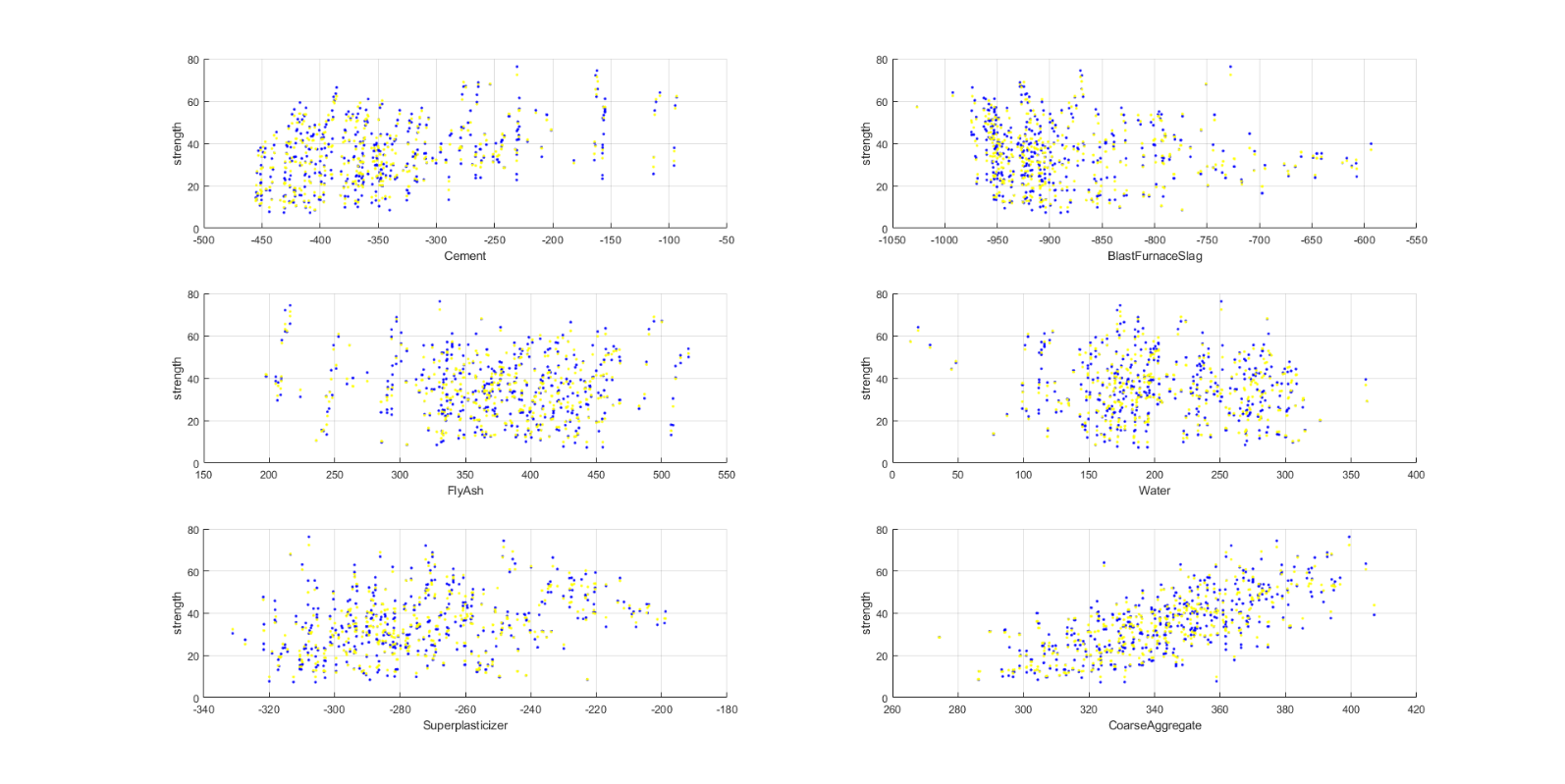


Figure V.1 各主成分真实值-预测值散点图

图中可看出，拟合情况和现实情况基本符合，具有一定的可信性。图V.2 是各主成分预测值-真实值的残差图，从图中可以看出，每一个主成分视角下的误差都基本符合平均分布，且误差被控制在[-5,+5]左右。图V.3 给出了预测值-真实值的比对图，所有散点基本贴近y=x，以上所有结果共同说明该回归模型的有效性。

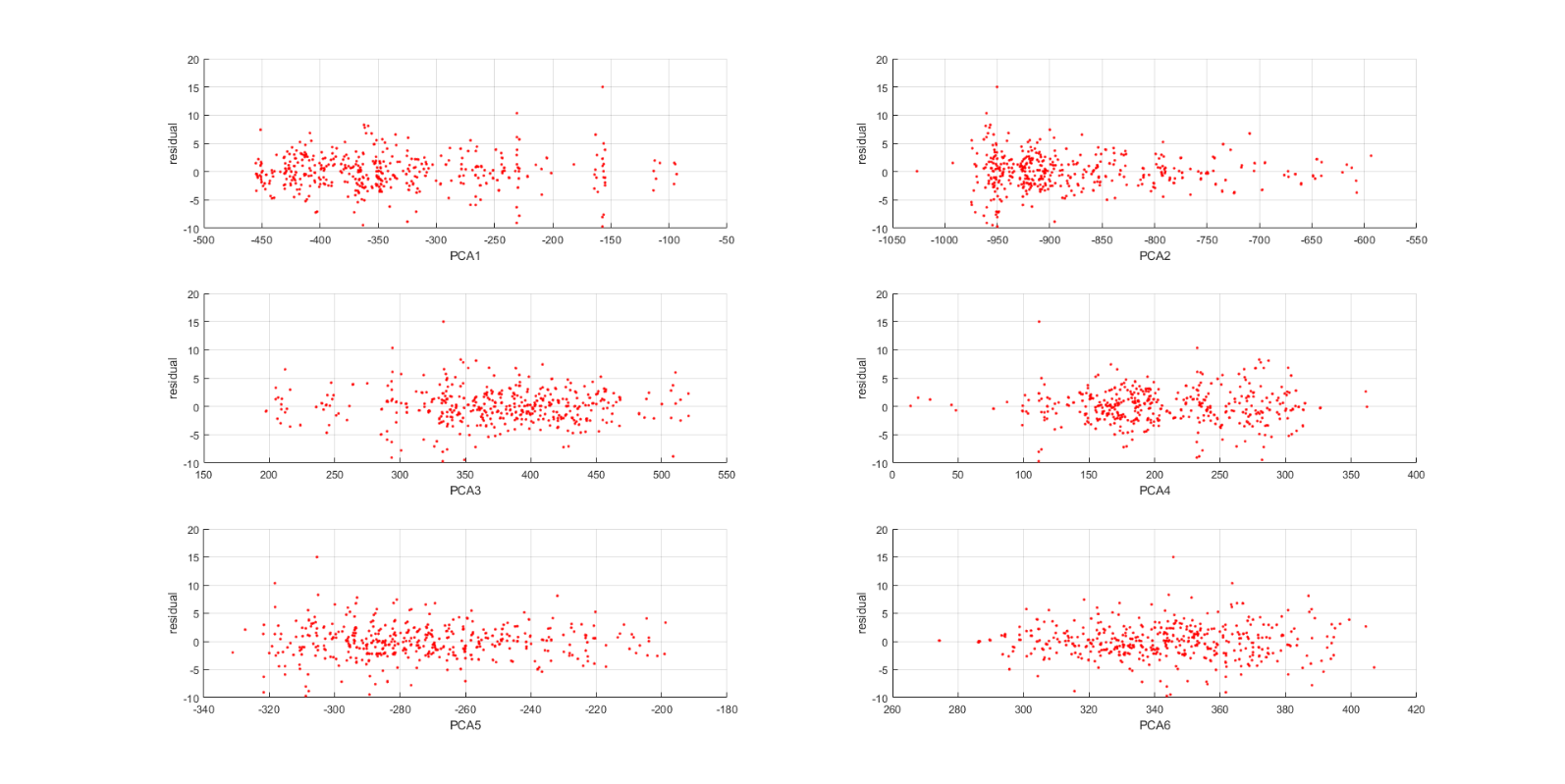


Figure V.2 各主成分真实值-预测值残差图

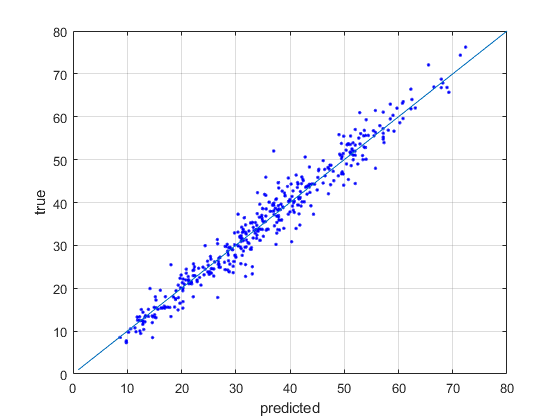


Figure V.3 真实值-预测值散点图

# 总结

本次实验使用维数较高的一组数据，在一个高度非线性关系上进行了回归实验，为了得到令人置信的成果，本次实验中，多种模型被测试，也尝试过多种特征提取方法，最终6主成分+高斯过程回归取得了理想的效果。回归是非常具有理论性的统计学方法，使用回归得到的模型，在可解释性，可预测性和对模型进行进一步分析上有着绝对的优势。在本次实验中，有以下两点值得指出。

### 根据既有理论框架进行建模

回归模型虽然强大，但如果获得一个令人满意的模型，离不开对数据的人工的分析和敏锐的数据直觉。如果对数据不加分析，直接在所有给出的变量中套用回归模型，得到的效果并不会太好。在本次实验中，混凝土抗压强度影响因素有一定的理论支持，基于此理论提取相应特征，才能让回归达到更好的效果。

### 选取最合适的模型

本次实验选用的三种模型都非常有代表性：线性回归模型是最广泛使用的模型，其简洁优雅的表达式能够在高维空间取得较好的效果。支持向量机模型是经典的机器学习方法，支持向量机回归模型非常适用于在一定误差允许范围内的预测，比如在此题中，预测的结果不需要非常精确，只要在工程中保持在一个区间内即可，那么采用该模型效果会比较好。而高斯过程回归模型应为核函数的灵活性非常适用于非线性系统和高维系统，因此在本题中，该模型表现最好。总结起来就是，没有最好的模型，只有最合适的模型。