**1 PCA introduction**

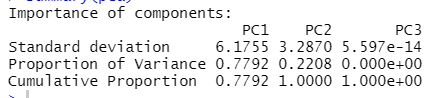
这部分我就不写了，可以找一些详细的背景资料

**2 train model**

我们使用PCA的方式进行特征工程，并希望与单变量HMM模型、多变量HMM模型进行比对。基于前两次作业，我们考虑Sub\_metering\_1、 Sub\_metering\_2、 Sub\_metering\_3对Global\_active\_power的影响，并基于此建立单变量、多变量模型。

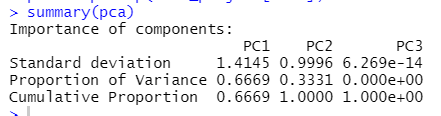
**2.1 PCA method**

此部分，我们对PCA进行特征提取。经过summary的输出如下：



可以看到第一主成分累计比例为77.92%，到第二主成份占比为100%。因此，对于此分析来说，第三主成分完全没有必要，前两个主成分就能涵盖原始的三个变量。这也是我们使用前两个主成分来训练PCA多变量模型的原因【这里是遇到的一个问题，就是对于PCA多变量模型，纳入几个因子是合理的】。

通过查看发现，进行PCA之前最好进行scale的处理（中心化处理），这样能过够去除不同量纲带来的影响【这里是遇到的第二个问题，需要进行scale处理】。因此，我们对纳入的变量进行scale处理后，再次使用PCA提取特征。经过summary的输出如下：



从上述结果中可以看到，第一主成分占比为66.69%，到第二主成分累计占比已经达到100%。这与前面的分析结论一致。通过两次对比也可以发现，进行scale之后的第一主成分占比有所下降，这是因为进行中心化后，出去了量纲对主成分提取的影响。

在进行特征工程提取后，我们建立相关HMM模型并且对比模型效果。

**2.2 train model**

这部分我们需要建立四个模型，并且对比个模型的性能。我们对模型进行如下声明：

时间窗：2018年每周四早上6点到9点

单变量模型：自变量Sub\_metering\_1

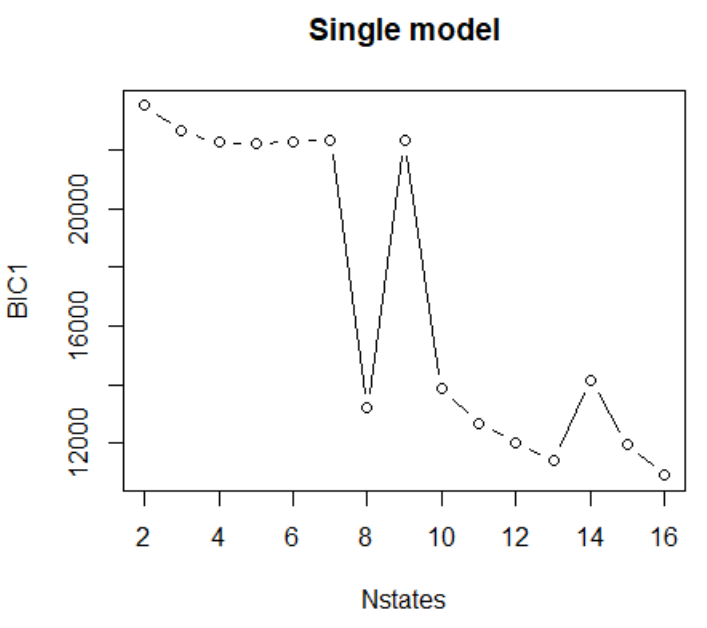
多变量模型：Sub\_metering\_1、 Sub\_metering\_2、 Sub\_metering\_3

PCA单变量模型：Sub\_metering\_1、 Sub\_metering\_2、 Sub\_metering\_3三个变量中的第一主成分

PCA多变量模型：Sub\_metering\_1、 Sub\_metering\_2、 Sub\_metering\_3三个变量中的前两个主成分

对于四种模型，我们再2-16个状态中进行尝试，利用BIC来判断最佳状态数。

单变量模型中，从下图中可以看到Nstaets应该选择13



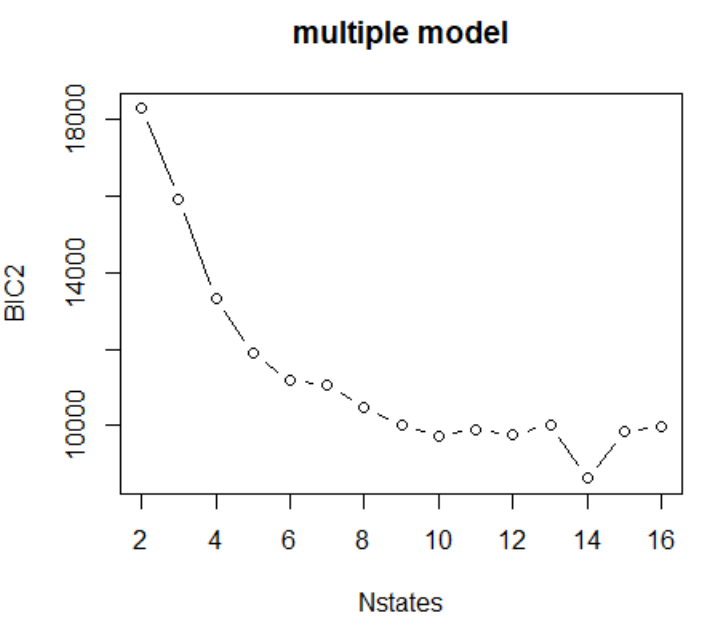
Convergence info: 'maxit' iterations reached in EM without convergence.

'log Lik.' -4820.04 (df=194)

AIC: 10028.08

BIC: 11425.23

多变量模型中，从下图可以看到Nstates应该选择12



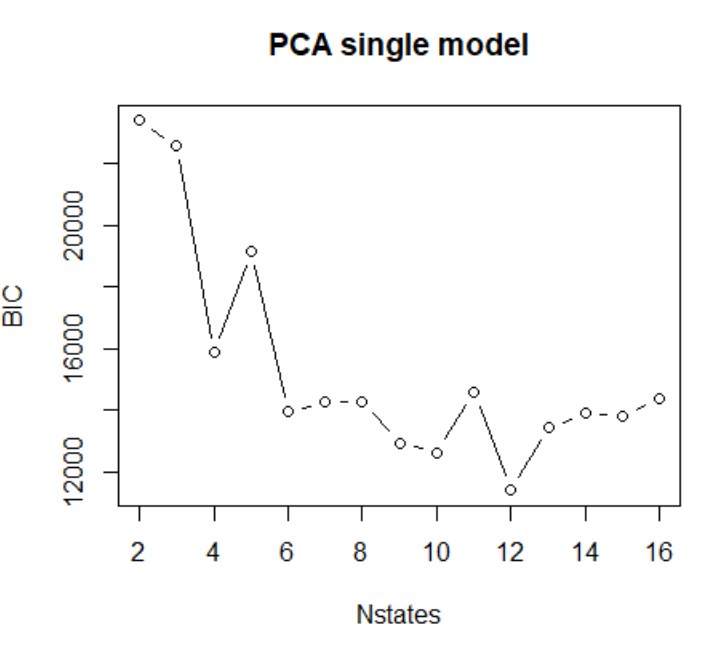
Convergence info: 'maxit' iterations reached in EM without convergence.

'log Lik.' -3151.739 (df=251)

AIC: 6805.478

BIC: 8613.131

PCA单变量模型中，Nstates也应该选择12



converged at iteration 479 with logLik: -4935.291

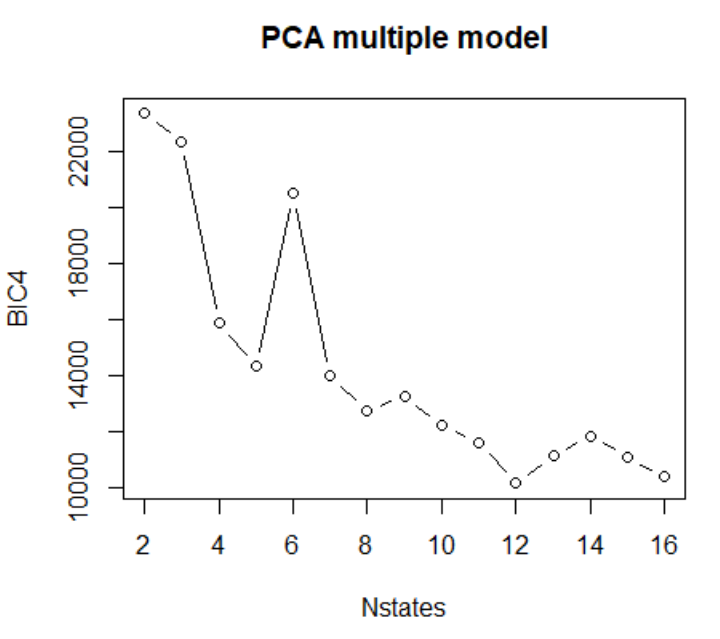
Convergence info: Log likelihood converged to within tol. (relative change)

'log Lik.' -4935.291 (df=167)

AIC: 10204.58

BIC: 11407.28

PCA单变量模型中，Nstates也应该选择12



Convergence info: 'maxit' iterations reached in EM without convergence.

'log Lik.' -4252.643 (df=179)

AIC: 8863.286

BIC: 10152.41

再找到模型最佳状态数后，我们对模型进行训练并且再测试集上进行了测试。

**2.2 analysis model result**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 训练集log like | 测试集log like |  |
| 单变量模型 | -4311.387 | -1628.297 |  |
| 多变量模型 | -4652.606 | -1522.963 |  |
| PCA单变量模型 | -5659.18 | -2023.379 |  |
| PCA多变量模型 | -5394.926 | -1965.557 |  |

**3 Anomaly Detection**

随后，我们利用四个模型再三个异常数据上进行检测，结果如下。

1. 横向来看，各个模型再不同的异常数据之上拟合效果有差异。整体来看data2的异常程度大于data3, data3的异常程度大于data1.
2. 纵向比较：从模型上来看，PCA单变量HMM模型的拟合程度低于单变量HMM模型，PCA多变量HMM模型的拟合程度低于多变量HMM模型。但是从整体效果上看，拟合程度相差不大。但是PCA在高维数据情况下可以大幅降低模型训练的数据量，减少模型训练时间。

|  |  |  |  |
| --- | --- | --- | --- |
|  | Anomaly data 1 | Anomaly data 2 | Anomaly data 3 |
| 单变量模型 | -9095.435 | -9247.775 | -9626.247 |
| 多变量模型 | -8720.659 | -8771.323 | -8226.645 |
| PCA单变量模型 | -10240.26 | -10760.32 | -10348.57 |
| PCA多变量模型 | -8709.89 | -9053.312 | -8709.039 |