法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

隐马尔科夫模型实践

邹博

主要内容

- 实现中文分词
  - 根据语料训练
  - 对新文件分词
  - 副产品：编码转换
- 高斯分布隐马尔科夫模型
  - 标记值为离散分布，观测值为连续分布
- 股价数据提取隐特征
- 数据处理的应用：电流强度的整流
  - GMHMM
- 开源库：Jieba分词、hmmlearn

中文分词

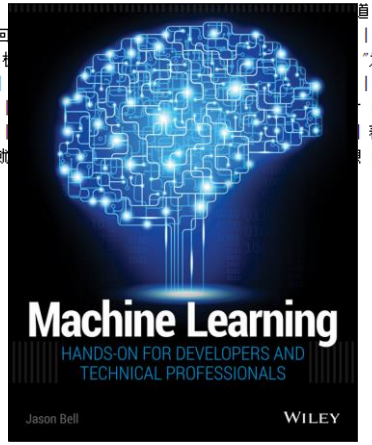
```
if __name__ == "__main__":
    pi, A, B = load_train()
    f = file(".\\text\\novel.txt")
    data = f.read()[3:].decode('utf-8')
    f.close()
    decode = viterbi(pi, A, B, data)
    segment(data, decode)
```

bug HMM

Console | Frames | Variables | Watches

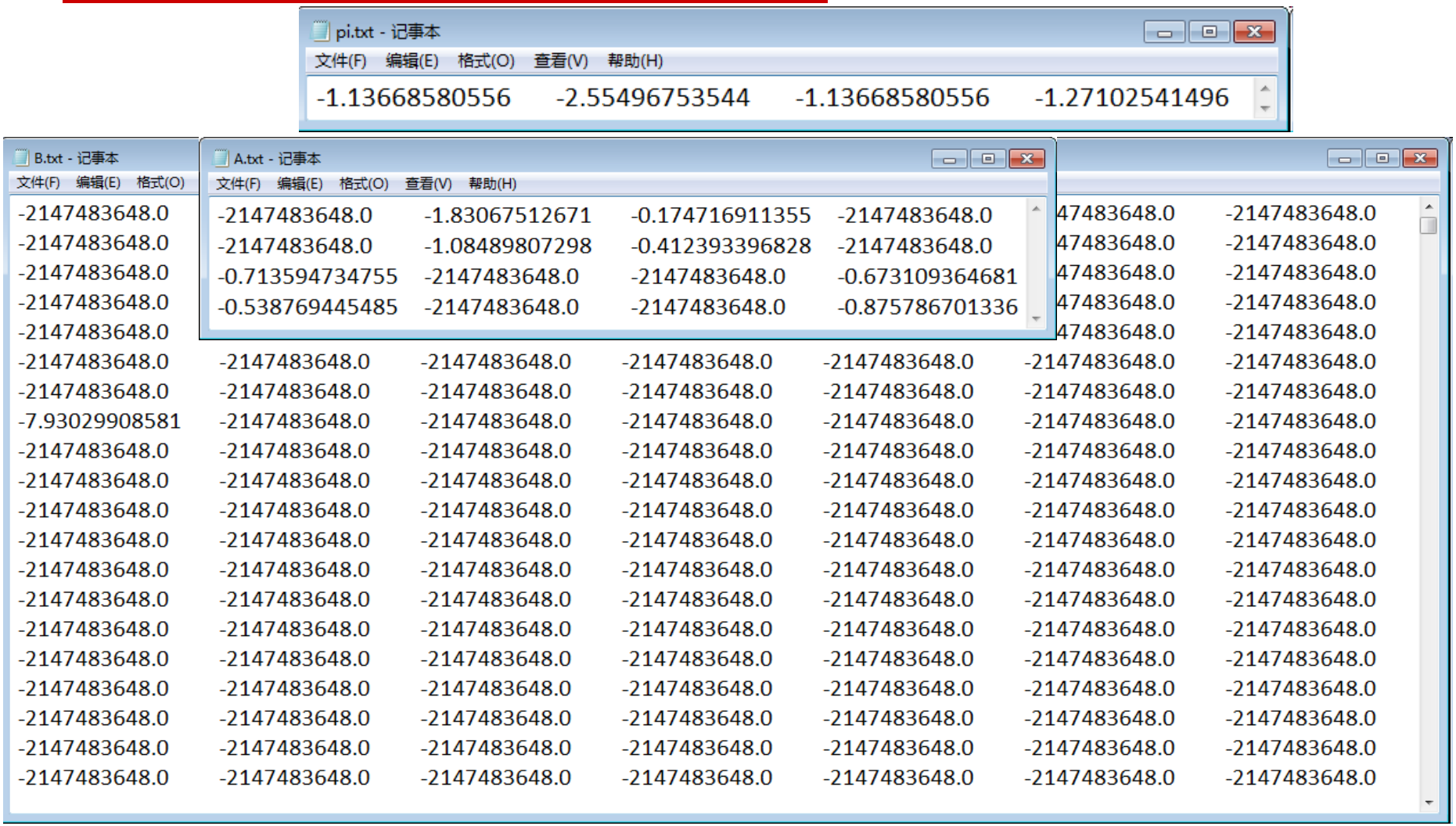
```
Connected to pydev debugger (build 139.1001)
我 | 与 | 地坛 |
史 | 铁生 |
一 |
我 | 在 | 好 | 几 | 篇 | 小说 | 中 | 都 | 提到 | 过 | 一 | 座 | 废弃 | 的 | 古园 | , | 实际 | 就 | 是 | 地坛 | 。 | 许多 | 年前 | 旅游业 | 还 | 没有 | 开
地坛 | 离 | 我家 | 很 | 近 | 。 | 或者 | 说 | 我家 | 离 | 地坛 | 很 | 近 | 。 | 总之 | , | 只好 | 认为 | 这 | 是 | 缘分 | 。 | 地坛 | 在 | 我出 | 生前 |
它 | 等待 | 我 | 出生 | , | 然后 | 又 | 等待 | 我活 | 到 | 最狂 | 妄 | 的 | 年龄 | 上 | 忽地 | 残废 | 了 | 双腿 | 。 | 四百 | 多 | 年里 | , | 它 | 一面
自 | 从 | 那个 | 下午 | 我 | 无意 | 中进 | 了 | 这 | 园子 | , | 就 | 再 | 没长 | 久 | 地离 | 开过 | 它 | 。 | 我 | 一 | 下子 | 就 | 理解 | 了 | 它 | 的 |
两 | 条 | 腿 | 残废 | 后 | 的 | 最初 | 几年 | , | 我找 | 不 | 到 | 工作 | , | 找 | 不到 | 去路 | , | 忽然 | 间 | 几乎 | 什么 | 都 | 找 | 不到 | 了 | ,
除 | 去 | 几 | 座 | 殿堂 | 我 | 无法 | 进去 | , | 除去 | 那 | 座 | 祭坛 | 我 | 不 | 能 | 上去 | 而 | 只能 | 从 | 各个 | 角度 | 张望 | 它 | , | 地坛 | 的
剩 | 下 | 的 | 就 | 是 | 怎样活 | 的 | 问题 | 了 | , | 这 | 却 | 不 | 是 | 在 | 某 | 一个 | 瞬间 | 就 | 能 | 完全 | 想透 | 的 | 、 | 不 | 是 | 一次性 |
二 |
我 | 才 | 想到 | , | 当年 | 我 | 总是 | 独自 | 跑 | 到 | 地坛 | 去 | , | 曾 | 经给 | 母亲 | 出 | 了 | 一个 | 怎样 | 的 | 难题 | 。 |
她 | 不 | 是 | 那种 | 光会 | 疼爱 | 儿子 | 而 | 不 | 懂得 | 理解 | 儿子 | 的 | 母亲 | 。
有 | 一 | 回 | 我摇 | 车出 | 了 | 小院 | ; | 想 | 起 | 一件 | 什么 | 事 | 又 | 返身 | 回
有 | 一次 | 与 | 一个 | 作家 | 朋友 | 聊天 | , | 我问 | 他学 | 写作 | 的 | 最 | 初动 |
在 | 我 | 的 | 头 | 一 | 篇 | 小 | 说 | 发表 | 的 | 时候 | , | 在 | 我 | 的 | 小 | 说 |
只 | 是 | 到 | 了 | 这 | 时候 | , | 纷纭 | 的 | 往事 | 才 | 在 | 我 | 眼前 | 幻现 | 得 |
摇 | 着 | 轮椅 | 在 | 园中 | 慢慢 | 走 | , | 又 | 是 | 雾霭 | 的 | 清晨 | , | 又 | 是 |
曾 | 有 | 过好 | 多 | 回 | , | 我 | 在 | 这园 | 子里 | 呆得 | 太久 | 了 | , | 母亲 | 就
```

前言 |
数据 | , | 数据 | , | 数据 | ! | 想 | 必在
等 | 媒介 | 的 | 持续 | 冲击 | 下 | , | 人们
的 | 洗礼 | 。 | 现实 | 需求 | 推动 | 了 | 对
这些 | 数据 | 来 | 自于 | 社交 | 媒体 | 、 |
“ | 物联网 | ”) | 、 | 传感器 | 等 | 任何
大 | 多 | 数数 | 据 | 挖掘 | 的 | 宣传 | 着 |
数据 | 洪 | 水(| data flood) | 的 | 预言 |
数据 | , | 硬件 | 推销 | 人员 | 会进 | 一步
能够 | 满足 | 处理 | 速度 | 的 | 要求 | 。 |
对 | 的 | , | 但 | 是 | 我们 | 值得 | 停下 |
务 | 进行 | 适当 | 的 | 再 | 认识 | 。 |
近 | 年来 | , | 数据 | 挖掘 | 和 | 机器 | 学习 | 在 | 我们 | 周围 | 持续 | 火爆 | , | 各种 |
媒体 | 也 | 不断 | 推送 | 着 | 海量 | 的 | 数据 | 。 | 仔细 | 观察 | 就 | 能 | 发现 | , | 实
际 | 应用 | 中 | 的 | 那些 | 机器 | 学习 | 算法 | 与 | 多 | 年前 | 并 | 没有 | 什么 | 两样 |
; | 它们 | 只 | 是 | 在 | 应用 | 的 | 数据 | 规模 | 上 | 有些 | 不同 | 。 | 历数 | 一 | 下 |
产生 | 数据 | 的 | 组织 | , | 至少 | 在 | 我 | 看来 | , | 数目 | 其实 | 并 | 不 | 多 | 。 |
无非 | 是 | Google | 、 | Facebook | 、 | Twitter | 、 | NetFlix | 以及 | 其 | 他 | 为数 | 不
多 | 的 | 机构 | 在 | 使用 | 若 | 干学 | 习算法 | 和 | 工具 | , | 这些 | 算法 | 和 | 工具 |
使 | 得 | 他们 | 能够 | 对 | 数据 | 进行 | 测试 | 分析 | 。 | 那么 | , | 真正 | 的 | 问题 |
是 | : | “ | 对于 | 其 | 他人 | , | 大数 | 据 | 框架 | 下 | 的 | 算法 | 和 | 工具 | 的 | 作用
| 是 | 什么 | 呢 | ? | ”|
我承认 | 本书 | 将 | 多 | 次 | 提及 | 大 | 数据 | 和 | 机器 | 学习 | 之间 | 的 | 关系 | , |
这 | 是 | 我 | 无法 | 忽视 | 的 | 一个 | 客观 | 问题 | ; | 但 | 是 | 它 | 只 | 是 | 一个 | 很
| 小 | 的 | 因素 | , | 终极 | 目标 | 是 | 如何 | 利用 | 可用 | 数据 | 获取 | 数据 | 的 | 本质



Jason Bell. *Machine Learning: Hands-On for Developers and Technical Professionals*. Wiley.2014

HMM参数训练


pi.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
-1.13668580556 -2.55496753544 -1.13668580556 -1.27102541496
B.txt - 记事本
文件(F) 编辑(E) 格式(O)
A.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
-2147483648.0 -1.83067512671 -0.174716911355 -2147483648.0
-2147483648.0 -1.08489807298 -0.412393396828 -2147483648.0
-0.713594734755 -2147483648.0 -2147483648.0 -0.673109364681
-0.538769445485 -2147483648.0 -2147483648.0 -0.875786701336
-7.93029908581

HMM中文分词

```python
def viterbi(pi, A, B, o):
    T = len(o)   # 观测序列
    delta = [[0 for i in range(4)] for t in range(T)]
    pre = [[0 for i in range(4)] for t in range(T)]  # 前一个状态    # pre[t][i]：t时刻的i状态，它的前一个状态是多少
    for i in range(4):
        delta[0][i] = pi[i] + B[i][ord(o[0])]
    for t in range(1, T):
        for i in range(4):
            delta[t][i] = delta[t-1][0] + A[0][i]
            for j in range(1,4):
                vj = delta[t-1][j] + A[j][i]
                if delta[t][i] < vj:
                    delta[t][i] = vj
                    pre[t][i] = j
            delta[t][i] += B[i][ord(o[t])]
    decode = [-1 for t in range(T)]  # 解码：回溯查找最大路径
    q = 0
```

18.3.HMM | 18.2.Segmentation | 18.HMM

```
C:\Python27\python.exe D:/Python/18.2.Segmentation.py
我 | 与 | 地坛 |
史 | 铁生 |
一 |
我 | 在 | 好 | 几 | 篇 | 小说 | 中 | 都 | 提到 | 过 | 一 | 座 | 废弃 | 的 | 古园 | ， | 实际 | 就 | 是 | 地坛 | 。 | 许多 | 年前 | 旅游业 | 还 | 没有 | 开展 |
地坛 | 离 | 我家 | 很 | 近 | 。 | 或者 | 说 | 我家 | 离 | 地坛 | 很 | 近 | 。 | 总之 | ， | 只好 | 认为 | 这 | 是 | 缘分 | 。 | 地坛 | 在 | 我出 | 生前 | 四百
它 | 等待 | 我 | 出生 | ， | 然后 | 又 | 等待 | 我活 | 到 | 最狂 | 妄 | 的 | 年龄 | 上 | 忽地 | 残废 | 了 | 双腿 | 。 | 四百 | 多 | 年里 | ， | 它 | 一面 | 剥蚀
自 | 从 | 那个 | 下午 | 我 | 无意 | 中进 | 了 | 这 | 园子 | ， | 就 | 再 | 没长 | 久 | 地离 | 开过 | 它 | 。 | 我 | 一 | 下子 | 就 | 理解 | 了 | 它 | 的 | 意图
两 | 条 | 腿 | 残废 | 后 | 的 | 最初 | 几年 | ， | 我找 | 不 | 到 | 工作 | ， | 找 | 不到 | 去路 | ， | 忽然 | 间 | 几乎 | 什么 | 都 | 找 | 不到 | 了 | ， | 我
除 | 去 | 几 | 座 | 殿堂 | 我 | 无法 | 进去 | ， | 除去 | 那 | 座 | 祭坛 | 我 | 不 | 能 | 上去 | 而 | 只能 | 从 | 各个 | 角度 | 张望 | 它 | ， | 地坛 | 的 | 每
剩 | 下 | 的 | 就 | 是 | 怎样活 | 的 | 问题 | 了 | ， | 这 | 却 | 不 | 是 | 在 | 某 | 一个 | 瞬间 | 就 | 能 | 完全 | 想透 | 的 | 、 | 不 | 是 | 一次性 | 能够 |
二 |
我 | 才 | 想到 | ， | 当年 | 我 | 总是 | 独自 | 跑 | 到 | 地坛 | 去 | ， | 曾 | 经给 | 母亲 | 出 | 了 | 一个 | 怎样 | 的 | 难题 | 。 |
她 | 不 | 是 | 那种 | 光会 | 疼爱 | 儿子 | 而 | 不 | 懂得 | 理解 | 儿子 | 的 | 母亲 | 。 | 她 | 知道 | 我心 | 里 | 的 | 苦闷 | ， | 知道 | 不 | 该 | 阻止 | 我出
有 | 一 | 回 | 我摇 | 车出 | 了 | 小院 | ； | 想 | 起 | 一件 | 什么 | 事 | 又 | 返身 | 回来 | ， | 看见 | 母亲 | 仍站 | 在 | 原地 | ， | 还 | 是 | 送 | 我 | 走
有 | 一次 | 与 | 一个 | 作家 | 朋友 | 聊天 | ， | 我问 | 他学 | 写作 | 的 | 最 | 初动 | 机是 | 什么 | ？ | 他想 | 了 | 一会 | 说 | ： | ″为 | 我 | 母亲 | 。 | 为
在 | 我 | 的 | 头 | 一 | 篇 | 小 | 说 | 发表 | 的 | 时候 | ， | 在 | 我 | 的 | 小 | 说 | 第一 | 次 | 获奖 | 的 | 那些 | 日子 | 里 | ， | 我真 | 是 | 多么 | 希望
```

Jieba分词

```
rainHMM.py × | 18.2.Segmentation.py × | 18.3.jieba_intro.py × | 18.4.GMHMM.py ×
 -coding:utf-8-

nport sys
nport jieba
nport jieba.posseg


f __name__ == "__main__":
   reload(sys)
   sys.setdefaultencoding('utf-8')
   f = open('.\\text\\18.novel.txt')
   str = f.read().decode('utf-8')
   f.close()

   seg = jieba.posseg.cut(str)
   for s in seg:
       # print s.word, s.flag,
       print s.word, '|',
```

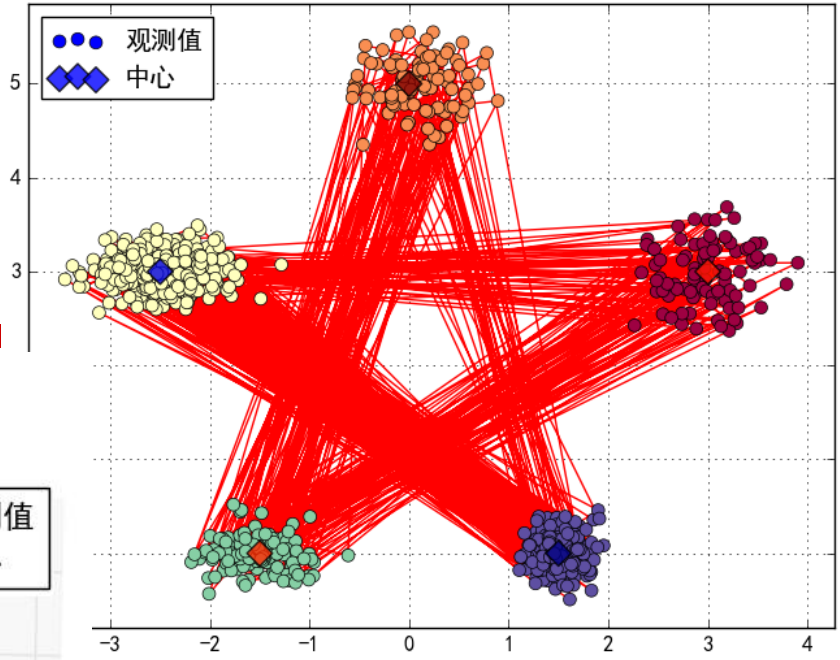
```
18.3.HMM | 18.3.jieba_intro | 18.HMM
 我 | 与 | 地坛 |
| 史铁生 |
| 一 |
| 我 | 在 | 好几篇 | 小说 | 中 | 都 | 提到 | 过 | 一座 | 废弃 |Loading model cost 0.419 seconds.
Prefix dict has been built succesfully.
 的 | 古园 | , | 实际 | 就是 | 地坛 | 。 | 许多年 | 前 | 旅游业 | 还 | 没有 | 开展 | , | 园子 | 荒芜 | 冷落 | 得 | 如同 | 一片 | 野地 | , | 很少 | 被 | 人 | 记起 | 。
| 地坛 | 离 | 我家 | 很 | 近 | 。 | 或者说 | 我家 | 离 | 地坛 | 很 | 近 | 。 | 总之 | , | 只好 | 认为 | 这 | 是 | 缘分 | 。 | 地坛 | 在 | 我 | 出生 | 前 | 四百多年 | 就
| 它 | 等待 | 我 | 出生 | , | 然后 | 又 | 等待 | 我 | 活到 | 最 | 狂妄 | 的 | 年龄 | 上 | 忽地 | 残废 | 了 | 双腿 | 。 | 四百多年 | 里 | , | 它 | 一面 | 剥蚀 | 了 | 古
| 自从 | 那个 | 下午 | 我 | 无意 | 中 | 进 | 了 | 这 | 园子 | , | 就 | 再 | 没 | 长久 | 地 | 离开 | 过 | 它 | 。 | 我 | 一下子 | 就 | 理解 | 了 | 它 | 的 | 意图 | 。 |
| 两条腿 | 残废 | 后 | 的 | 最初 | 几年 | , | 我 | 找 | 不到 | 工作 | , | 找 | 不到 | 去路 | , | 忽然 | 间 | 几乎 | 什么 | 都 | 找 | 不到 | 了 | , | 我 | 就 | 摇 | 了
| 除去 | 几座 | 殿堂 | 我 | 无法 | 进去 | , | 除去 | 那 | 座 | 祭坛 | 我 | 不能 | 上去 | 而 | 只能 | 从 | 各个 | 角度 | 张望 | 它 | , | 地坛 | 的 | 每 | 一棵树 | 下 |
| 剩下 | 的 | 就是 | 怎样 | 活 | 的 | 问题 | 了 | , | 这 | 却 | 不是 | 在 | 某 | 一个 | 瞬间 | 就 | 能 | 完全 | 想透 | 的 | 、 | 不是 | 一次性 | 能够 | 解决 | 的 | 事 |
| 二 |
| 我 | 才 | 想到 | , | 当年 | 我 | 总是 | 独自 | 跑 | 到 | 地坛 | 去 | , | 曾经 | 给 | 母亲 | 出 | 了 | 一个 | 怎样 | 的 | 难题 | 。 |
| 她 | 不是 | 那种 | 光 | 会 | 疼爱 | 儿子 | 而 | 不 | 懂得 | 理解 | 儿子 | 的 | 母亲 | 。 | 她 | 知道 | 我 | 心里 | 的 | 苦闷 | , | 知道 | 不该 | 阻止 | 我 | 出去 | 走
| 有 | 一回 | 我 | 摇车 | 出 | 了 | 小院 | ; | 想起 | 一件 | 什么 | 事 | 又 | 返身 | 回来 | , | 看见 | 母亲 | 仍 | 站 | 在 | 原地 | , | 还是 | 送 | 我 | 走时 | 的 |
| 有 | 一次 | 与 | 一个 | 作家 | 朋友 | 聊天 | , | 我 | 问 | 他 | 学 | 写作 | 的 | 最初 | 动机 | 是 | 什么 | ? | 他 | 想 | 了 | 一会 | 说 | : | " | 为 | 我 | 母亲 | 。
```

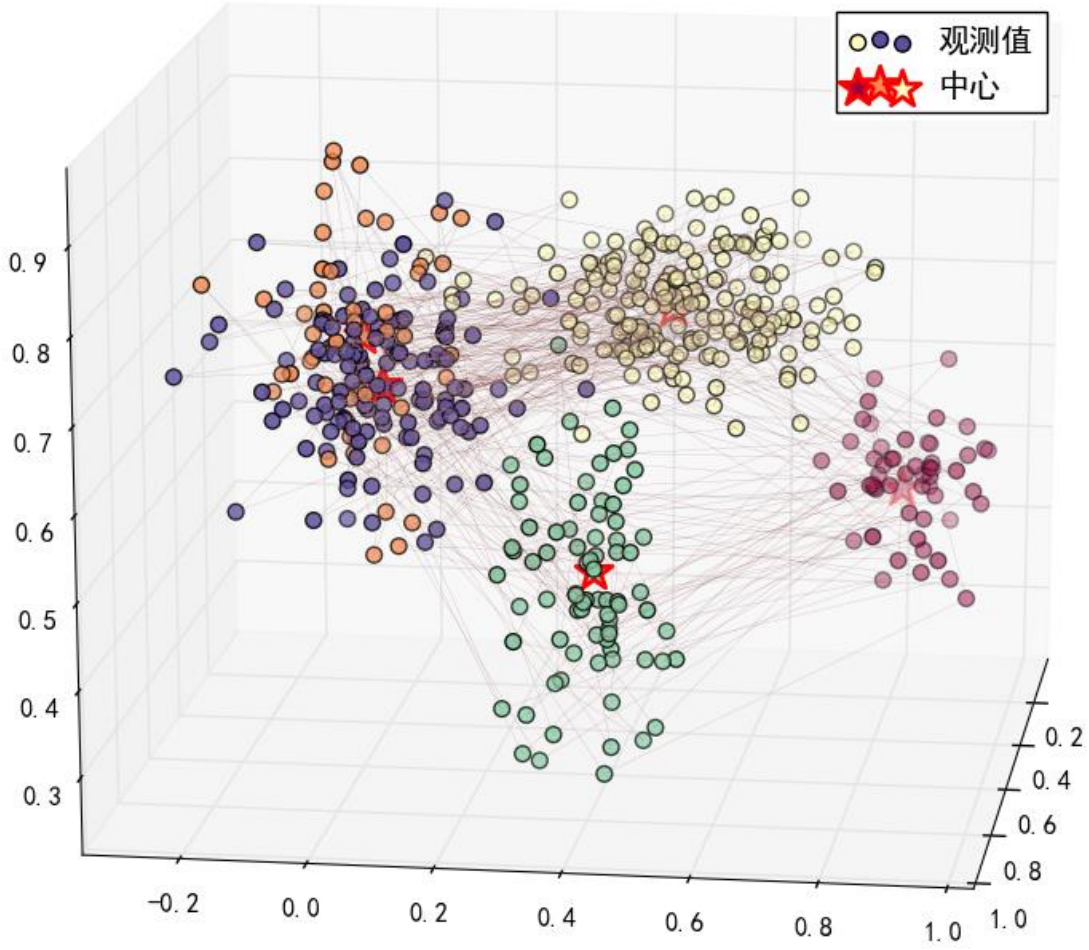
Hmmlearn的安装

```
D:\Python\Package>pip install hmmlearn-0.2.0-cp27-cp27m-win32.whl
Processing d:\python\package\hmmlearn-0.2.0-cp27-cp27m-win32.whl
Installing collected packages: hmmlearn
Successfully installed hmmlearn-0.2.0
```

GMHMM

GMHMM参数估计和类别判定





GMHMM参数估计

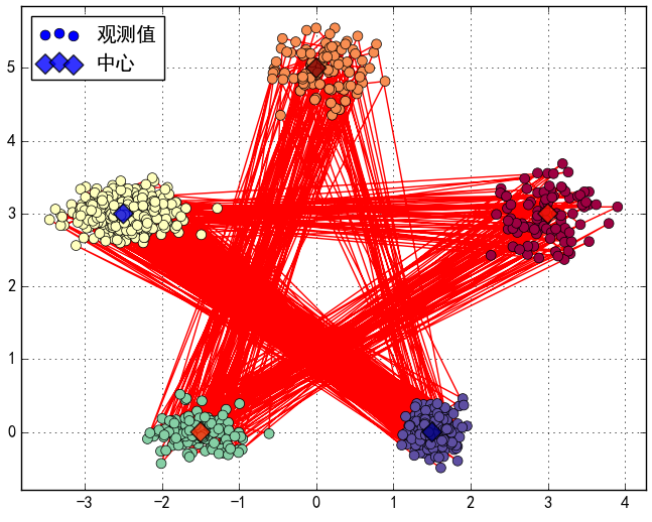
初始概率： [0.19356424 0.25224431 0.21259213 0.19217803 0.14942128]
转移概率：
[[0.25822029 0.   0.35651955 0.38526017 0.   ]
 [0.   0.34669639 0.   0.6067387   0.04656491]
 [0.04868208 0.   0.46521279 0.   0.48610513]
 [0.3825259   0.31237801 0.   0.30509609 0.   ]
 [0.   0.09539815 0.62865435 0.   0.2759475]]
均值：
[[3.   3.]
 [0.   5.]
 [-2.5 3.]
 [-1.5 0.]
 [1.5 0.]]
方差：
[[[0.12 0.]
  [0.   0.09]]

 [[0.12 0.]
  [0.   0.09]]

 [[0.12 0.]
  [0.   0.03]]

 [[0.09 0.]
  [0.   0.03]]

 [[0.03 0.]
  [0.   0.03]]]



估计初始概率： [0. 0. 1. 0. 0.]
估计转移概率：
[[0.24444444 0.   0.43333333 0.32222222 0.   ]
 [0.   0.36082474 0.   0.60824742 0.03092784]
 [0.03406326 0.   0.47688564 0.   0.48905109]
 [0.43902439 0.27642276 0.   0.28455285 0.   ]
 [0.   0.10071942 0.6294964   0.   0.26978417]]
估计均值：
[[2.98641153 2.97594103]
 [0.09781242 5.00394771]
 [-2.47643196 2.99259797]
 [-1.51986115 -0.0035412]
 [1.50315967 -0.00746037]]
估计方差：
[[[0.11979558 0.01093522]
  [0.01093522 0.09896496]]

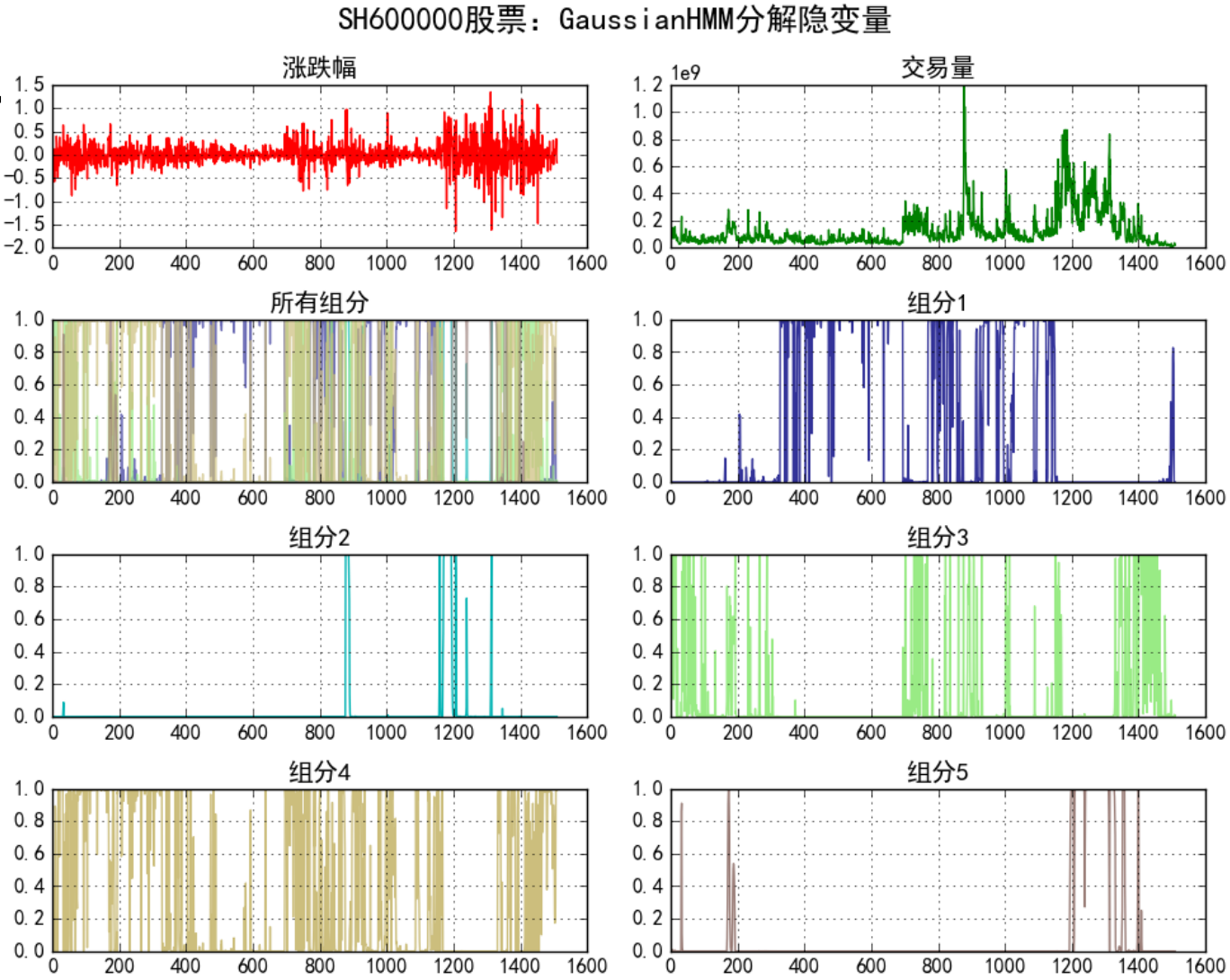
 [[0.10760117 0.00087227]
  [0.00087227 0.07097137]]

 [[0.11128863 0.00142049]
  [0.00142049 0.02646752]]

 [[0.09187351 -0.00410475]
  [-0.00410475 0.03027345]]

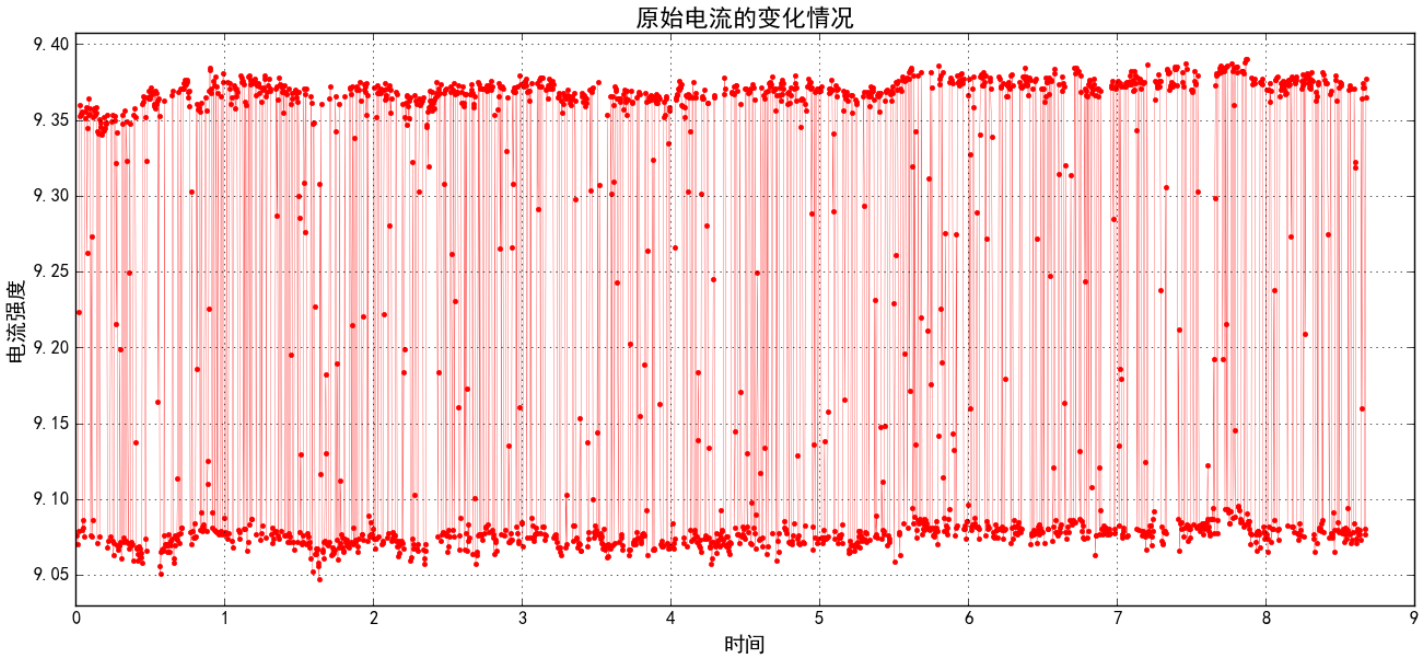
 [[0.02501027 0.00066473]
  [0.00066473 0.02779045]]]

提取特征


SH600000股票：GaussianHMM分解隐变量
涨跌幅
交易量
1e9
所有组分
组分1
组分2
组分3
组分4
组分5

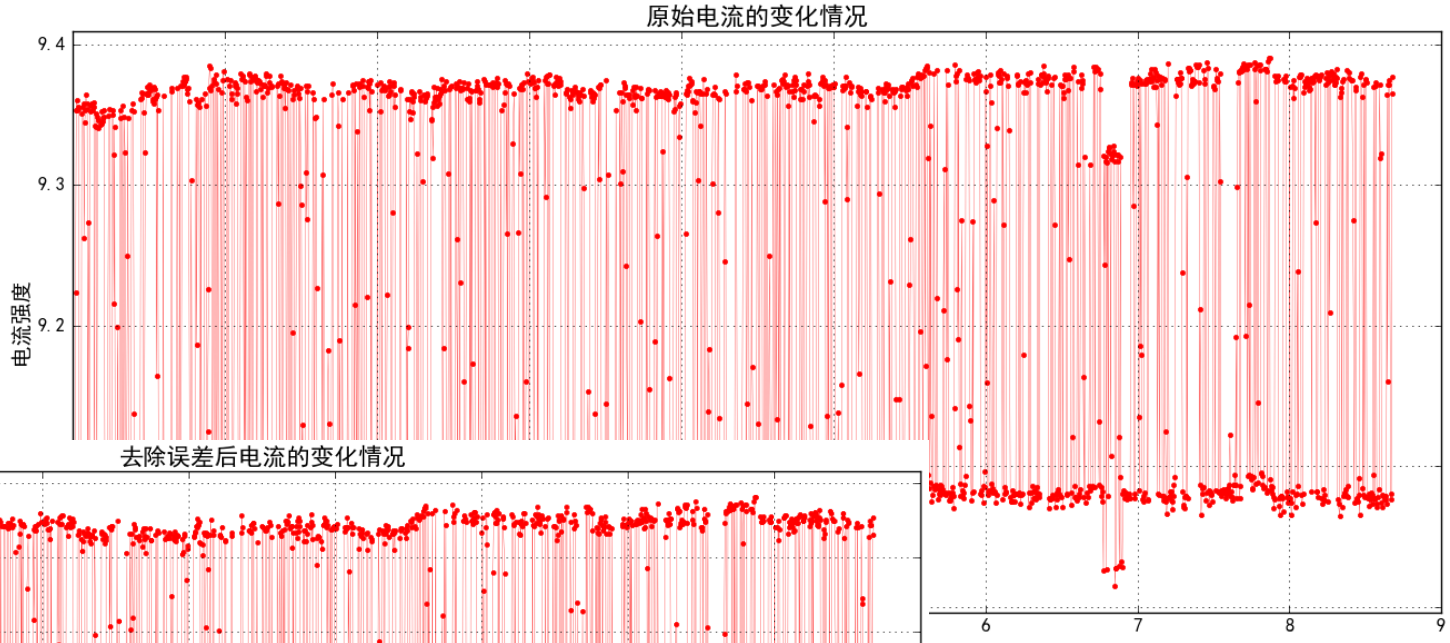
电流数据的校正

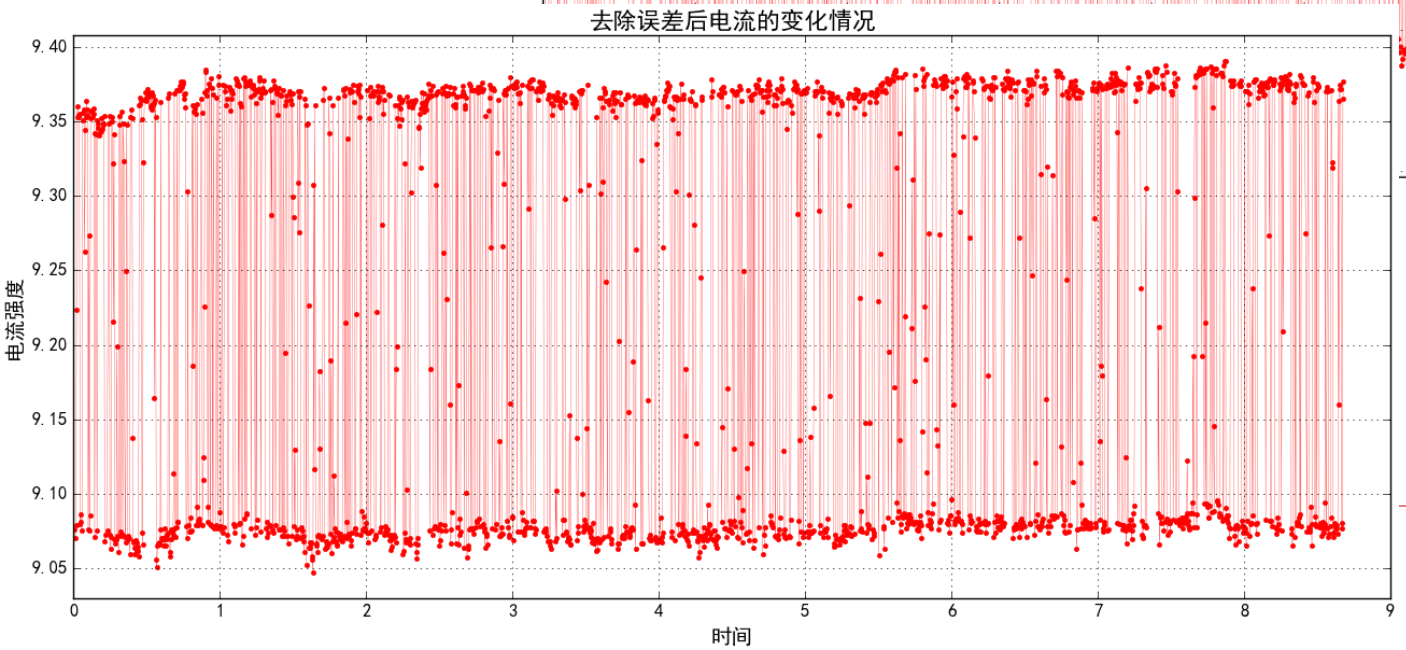
- 现有电路的测量电流数据如右图(部分)，由于电路的系统误差，其电流强度如下图所示。试对其进行整流，得到规则电流。



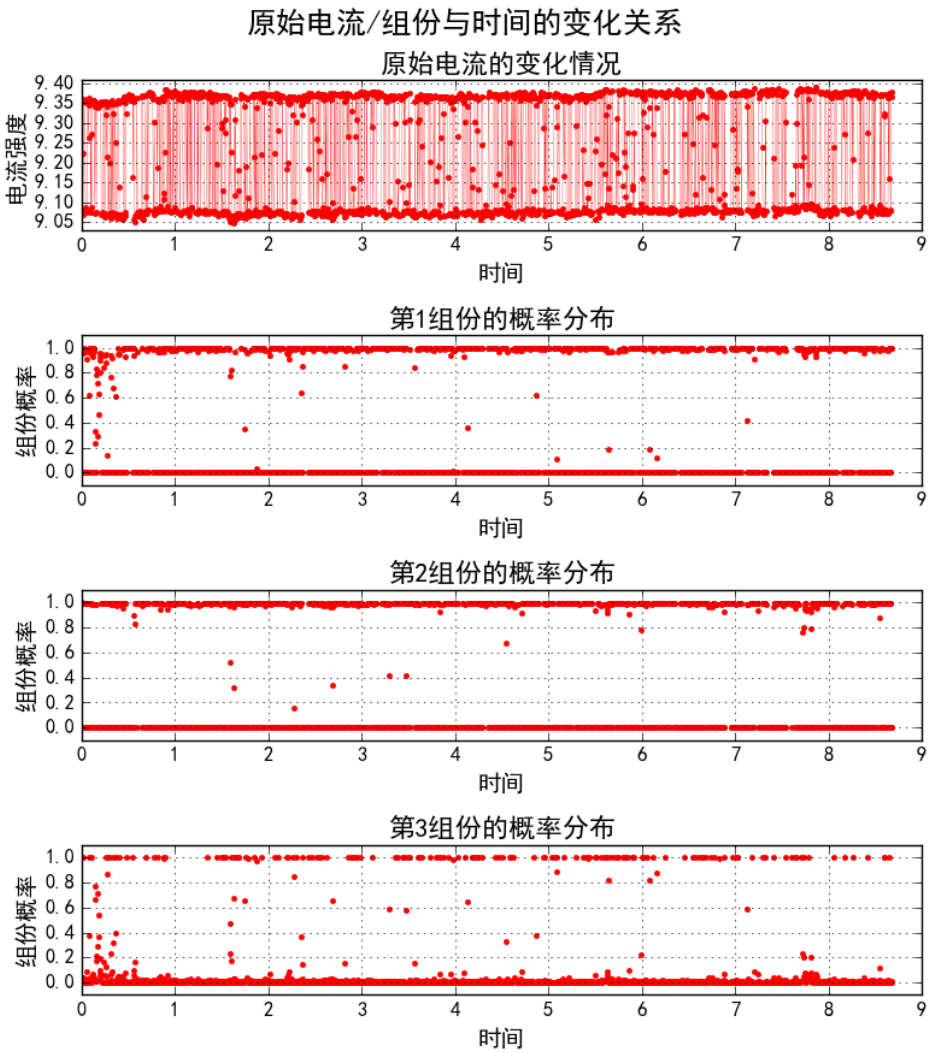
	A	B
1	Time	Current
2	0.00677	0.000009076
3	0.01269	0.000009070
4	0.01665	0.000009079
5	0.02069	0.000009224
6	0.02465	0.000009353
7	0.02865	0.000009360
8	0.03278	0.000009353
9	0.03669	0.000009354
10	0.04074	0.000009355
11	0.04465	0.000009080
12	0.04878	0.000009086
13	0.0527	0.000009081
14	0.05823	0.000009075
15	0.06068	0.000009351
16	0.06638	0.000009356
17	0.07269	0.000009358
18	0.07674	0.000009263
19	0.08069	0.000009345
20	0.08469	0.000009353
21	0.08865	0.000009364
22	0.09269	0.000009358
23	0.09669	0.000009354
24	0.10065	0.000009076
25	0.10517	0.000009273
26	0.10863	0.000009075
27	0.11269	0.000009086
28	0.11668	0.000009352
29	0.12079	0.000009357
30	0.12463	0.000009358
31	0.12957	0.000009353
32	0.13665	0.000009358
33	0.14078	0.000009355
34	0.14469	0.000009071
35	0.14865	0.000009342
36	0.15265	0.000009341
37	0.15665	0.000009351
38	0.16065	0.000009076
39	0.16469	0.000009349
40	0.16871	0.000009346
41	0.17281	0.000009340
42	0.18086	0.000009343
43	0.18865	0.000009344
44	0.19269	0.000009352

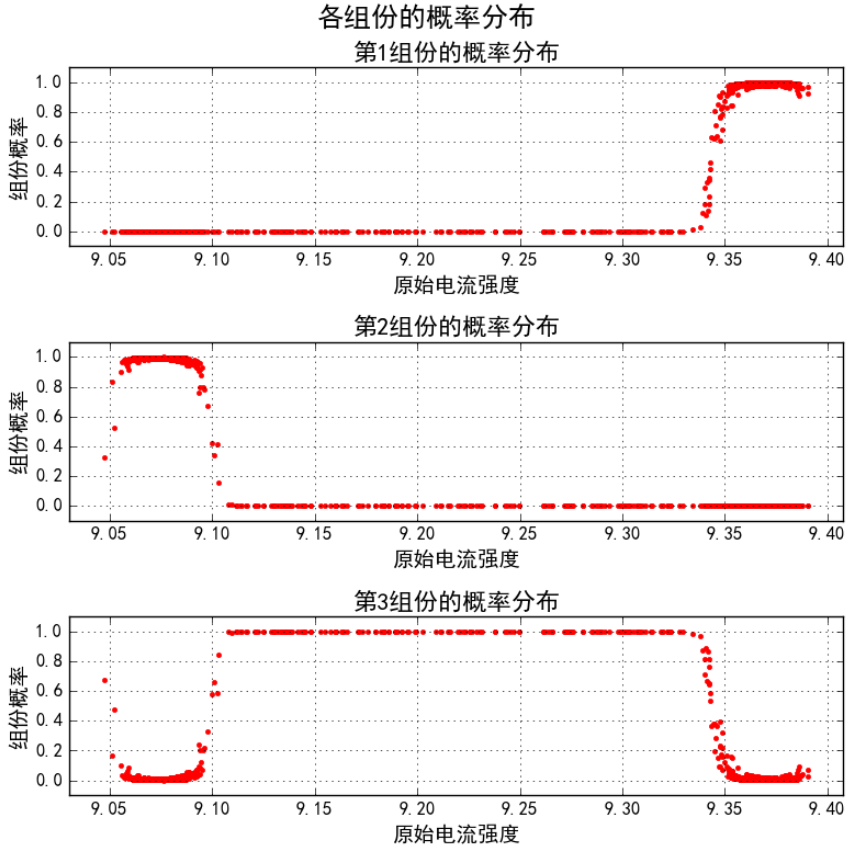
去除明显的异常



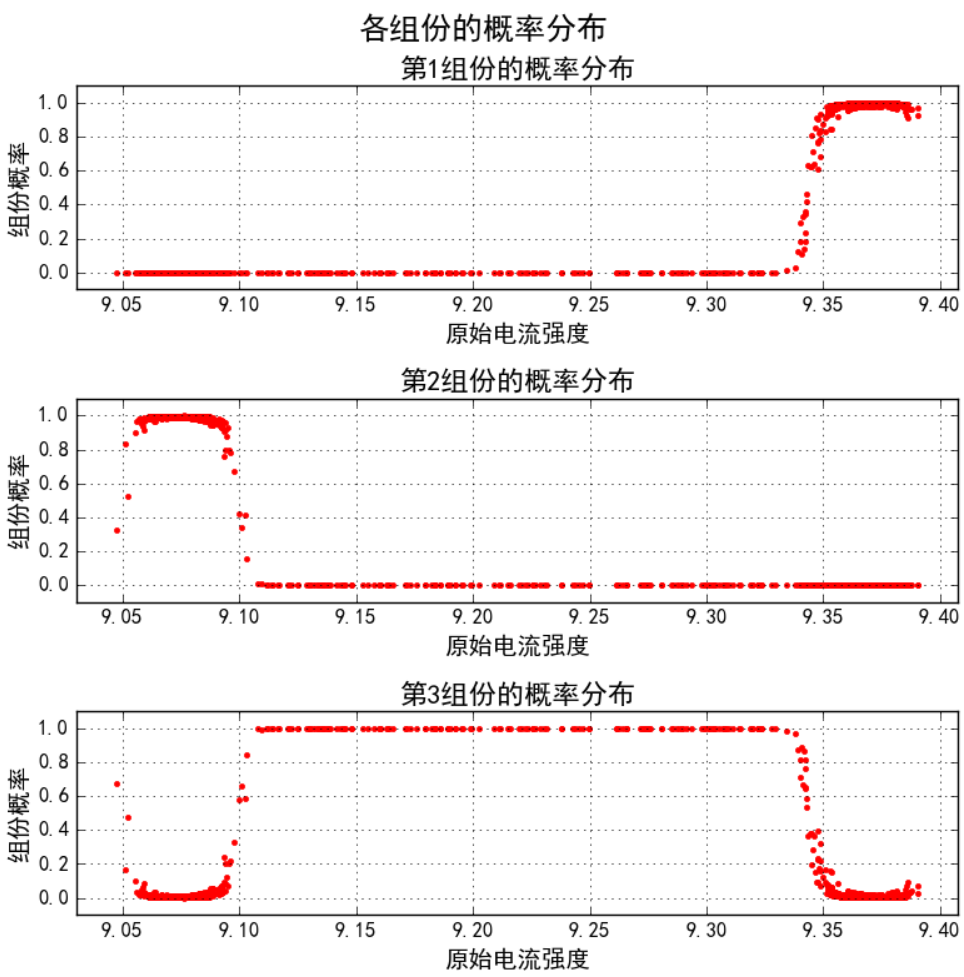


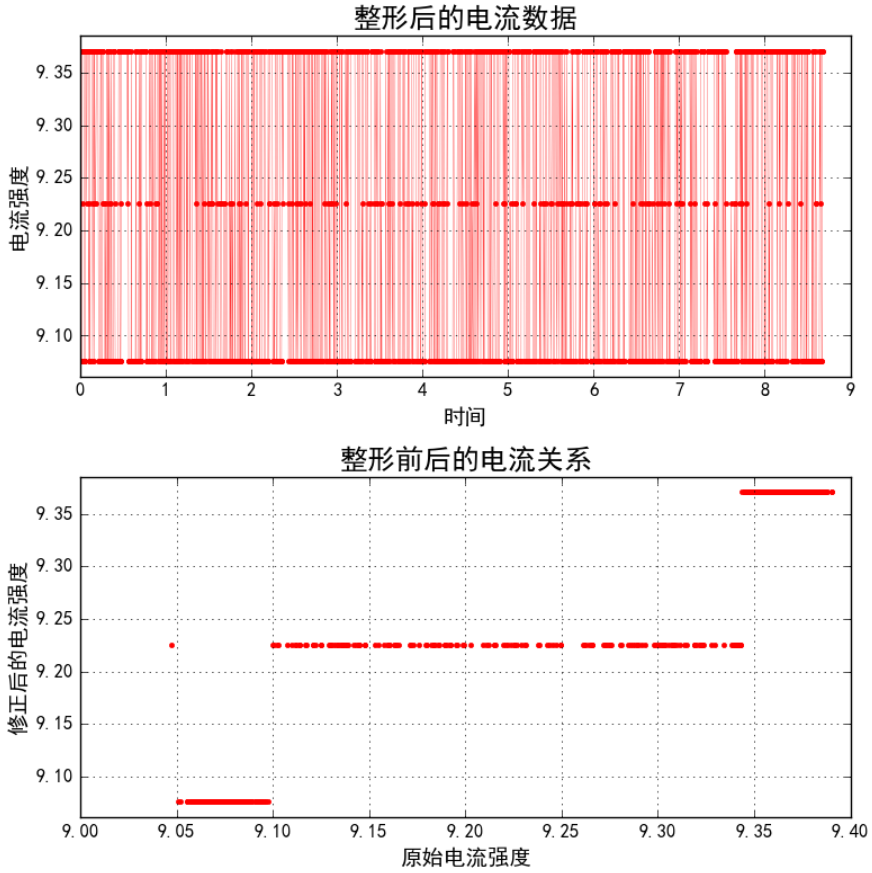
HMM隐状态特征分解





根据隐状态做整流





Code

```
n_components = 3
data = pd.read_excel(io='Current.xls', sheetname='Sheet1', header=0)
# data['Current'] = MinMaxScaler().fit_transform(data['Current'])
data['Current'] *= 1e6

# 去除明显的异常值
data_clean(False)

x = data['Time'].reshape(-1, 1)
y = data['Current'].reshape(-1, 1)
model = hmm.GaussianHMM(n_components=n_components, covariance_type='full', n_iter=10)
model.fit(y)
components = model.predict_proba(y)
components_state = model.predict(y)
components_pd = pd.DataFrame(components, columns=np.arange(n_components), index=data.index)
data = pd.concat((data, components_pd), axis=1)
print 'data = \n', data

plt.figure(num=1, facecolor='w', figsize=(8, 9))
plt.subplot(n_components+1, 1, 1)
plt.plot(x, y, 'r.-', lw=0.2)
plt.ylim(extend(y.min(), y.max()))
plt.grid(b=True, ls=':')
plt.xlabel(u'时间', fontsize=14)
plt.ylabel(u'电流强度', fontsize=14)
plt.title(u'原始电流的变化情况', fontsize=16)
for component in np.arange(n_components):
    plt.subplot(n_components+1, 1, component+2)
    plt.plot(x, data[component], 'r.')
    plt.ylim((-0.1, 1.1))
    plt.grid(b=True, ls=':')
    plt.ylabel(u'组份概率', fontsize=14)
    plt.xlabel(u'时间', fontsize=14)
    plt.title(u'第%d组份的概率分布' % (component+1), fontsize=16)
plt.suptitle(u'原始电流/组份与时间的变化关系', fontsize=18)
plt.tight_layout(pad=1, rect=(0, 0, 1, 0.96))
```

hmmlearn参考文献

- 安装包：
  - https://pypi.python.org/pypi/hmmlearn
- Github代码：
  - https://github.com/hmmlearn/hmmlearn
- 文档：
  - http://hmmlearn.readthedocs.io/en/latest/tutorial.html

我们在这里

- http://wenda.ChinaHadoop.cn
  - 视频/课程/社区
- 微博
  - @ChinaHadoop
  - @邹博_机器学习
- 微信公众号
  - 小象
  - 大数据分析挖掘

感谢大家！

恳请大家批评指正！