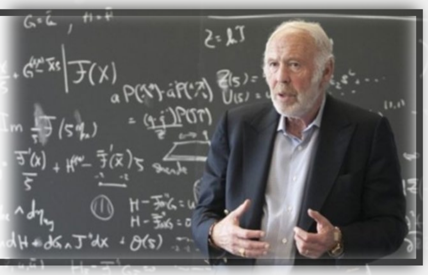
基于隐马尔可夫链（HMM）的金融市场交易策略设计

蔡玮钦 516120910098

王奕能 516120910101

王振宇 5141619041

背景介绍



- Renaissance & Medallion（文艺复兴科技和大奖章）
- 由Simons带领一群物理学家和数学家碰撞在一起
- 1989年到2008年的年化收益达到35.6%。在全球金融危机的08年，大部分对冲基金都亏损，而大奖章的return高达80%。
- 成立初期的创始人中，鲍姆等人提出了广泛应用在语音识别等领域的HMM模型和鲍姆-威尔士算法，用来确定不可确知的变量可能出现的概率。
- 人大的一位教授14年也写了一本书——*解密复兴科技：基于隐蔽马尔科夫模型的时序分析方法*。

原理部分

隐马尔可夫模型

确定模型参数

确定最佳历史隐状态序列

- 隐马尔可夫模型
  - 基本假设
  - 决定参数
- 确定模型参数
  - Baum-Welch算法
- 确定最佳历史隐状态序列
  - Viterbi算法

隐马尔可夫模型

- 定义:
  - $\{Xn\}$ 是一个普通的马尔可夫链,每当马尔可夫链进入状态 j 时，以概率 p(s|j)给出信号s,用 $\{Sn\}$ 表示信号序列，如果隐状态Xn不可观测，而信号Sn可以被观测，则得到一个隐马尔可夫模型。
- 基本假设:
  - X_n 仅和 X_{n-1} 有关，即 $P(X_n|S_n, Xn-1, S_{n-1}, \ldots, X_1,S_1)=P(X_n|X_{n-1})$
  - S_n 仅和 X_n 有关，即 $P(S_n|X_n,S_{n-1},X_{n-1},\ldots,S_1,X_1)=P(S_n|X_n)$

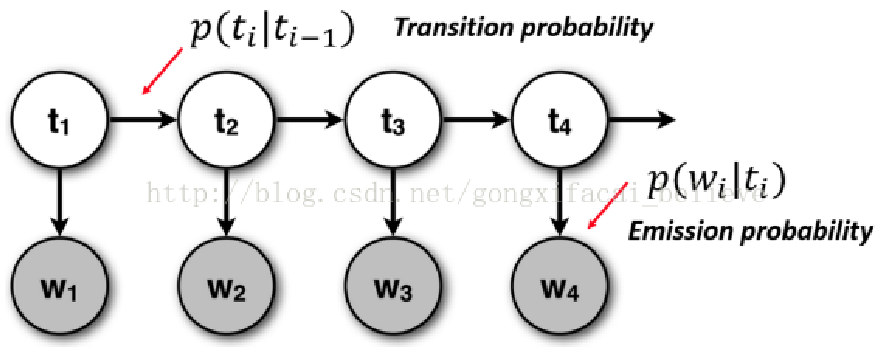
隐马尔可夫模型

- 参数：
  - {Xn} 的初始状态概率分布 π

初始状态概率分布	状态 1	状态 2	……	状态 n	……
概率	π_1	π_2	……	π_n	……

  - {Xn} 的转移矩阵 A和模型的观测矩阵 B

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{n1} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{pmatrix} \quad \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{n1} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nn} \end{pmatrix}$$



α_{ij} 表示状态i转移到状态j的概率
β_{ij} 表示不可观测的状态处在 i 时给出可观测信号 j 的概率

确定模型参数

- 假设股票/期货市场存在着某些不可观测的隐状态，这些状态满足马尔可夫链构成{*Xn*}。每天可观测到的价格走势作为状态的信号 {*Sn*}。
- 该模型中，初始概率分布 π，转移矩阵 A，观测矩阵 B 都未知，仅可观测到信号。需要估计参数λ(π, A,B) 。

确定模型参数

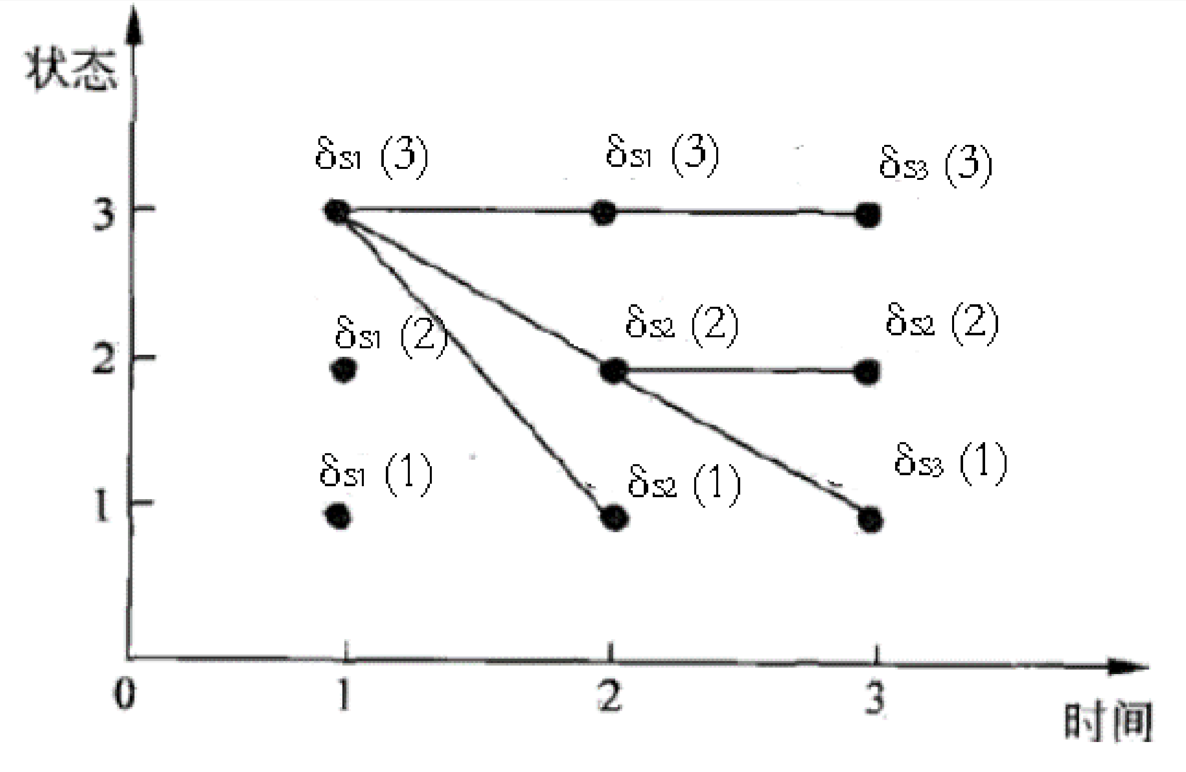
- Baum-Welch 算法：
  - 估计 λ 并使 $P(\{Sn\}|\lambda)$ 最大，采用对数极大似然估计。
  - $\bar{\lambda}$ 为当前估计参数,计算机循环迭代得到使 Q 最大的 λ，即最佳估计参数。

$$Q(\bar{\lambda},\lambda)=\sum_{\{X_n\}}(ln(P(\{X_n\},\{S_n\}|\lambda))P(\{X_n\}|\{S_n\},\bar{\lambda})$$

VITERBI算法

- 最大化状态路径的概率：
  - Max(P(Xn,Xn-1,…X1| λ, Sn,Sn-1,…S1))
- 1初始化，对每个i
  - t=1时， $\delta_{S_1}(i)=\pi_i\ \beta_{jS_1}$
- 2递推(k>1),对每个i,j
  - t=k时， $\delta_{S_k}(i)=\max(\delta_{S_{k-1}}(j)\ \alpha_{ji}\ \beta_{iS_k})$
- 3终点
  - 取最大的$\delta_{S_n}(i)$
- 4逆向找in

对j取max



$\delta_{S_k}(i)$表示第k步观测到Sk且处在状态i的最大路径概率

ETF实证部分

数据来源

交易策略

策略表现

- 基于前述的模型，我们构建了相应的交易策略。
- 其中，在ETF上的应用不算成功，但在期货价差上能有较好表现。主要在价差部分介绍具体策略。

变量定义

50ETF（510050）2010-2019年数据

变量名	定义
open, high, low, close	开/高/低/收盘价
ret1	$\text{close}_t/\text{close}_{t-1} - 1$
ret5	$\text{close}_t/\text{close}_{t-5} - 1$
ATR	Average True Range
RSI	Relative Strength Index
OBV	On Balance Volume
MFI	Money Flow Index

$$ATR(n)_t = \frac{1}{n}\sum_{i=1}^{n} TR_{t-i+1}$$

Source:Wind

样本内标注

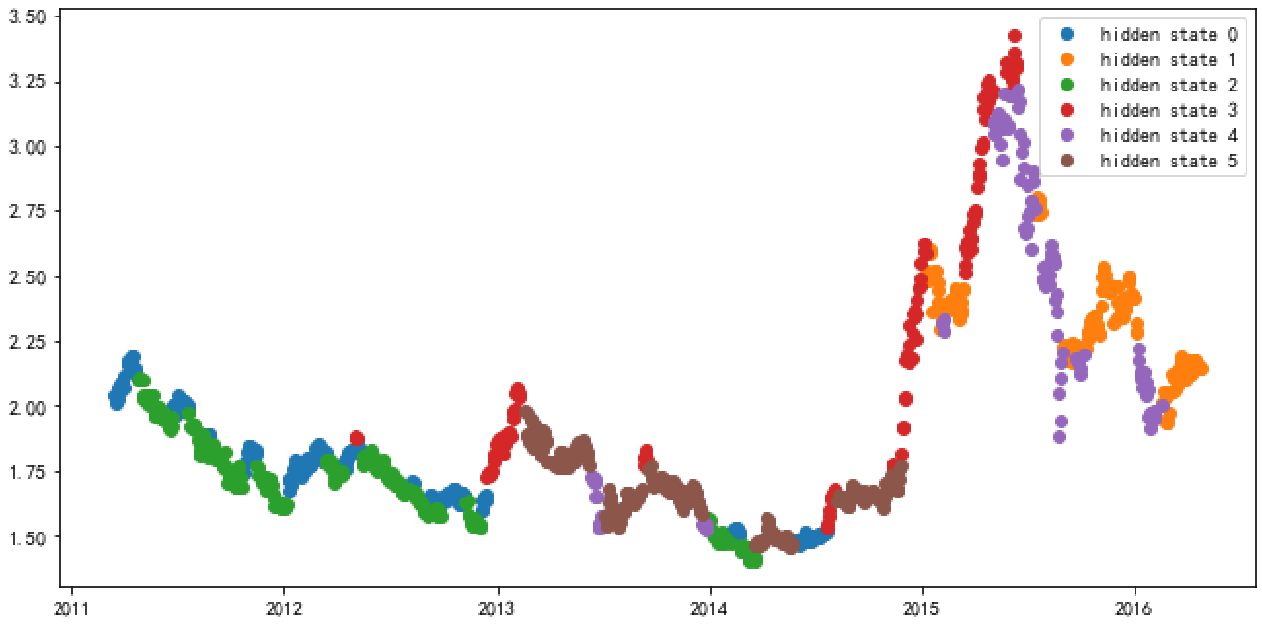
- 首先对数据进行样本内标注，查看各个隐状态之间是否有区分度
- 取隐状态个数=6
- 例如：

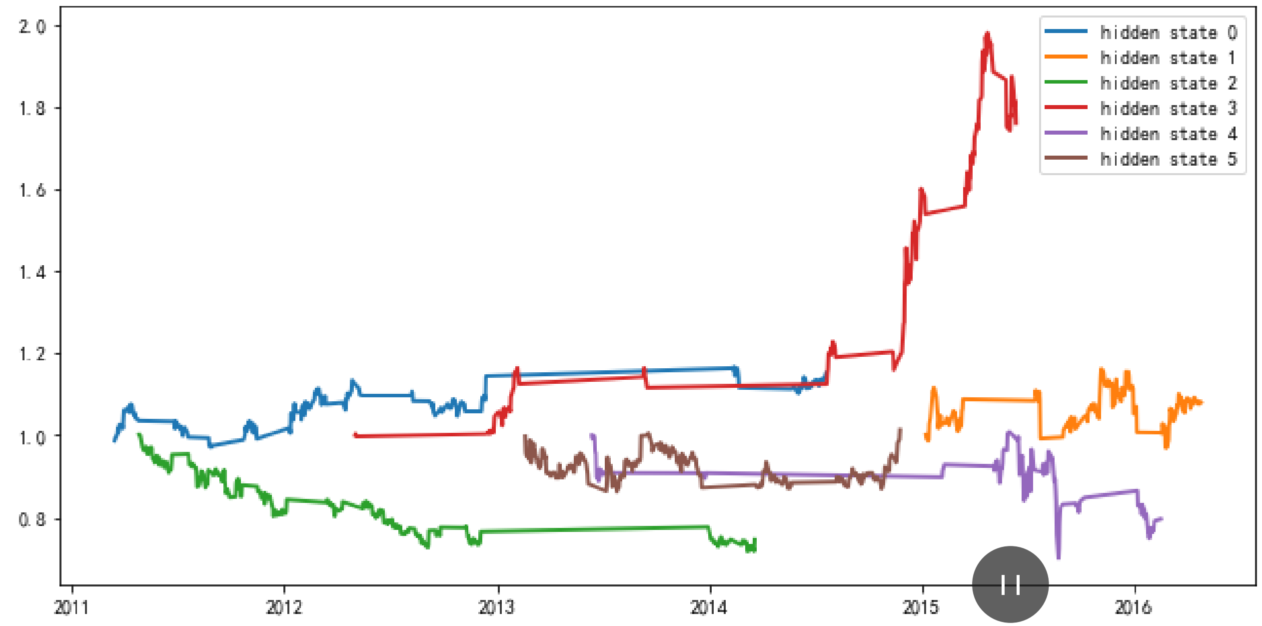
  Hidden state 0: 震荡

  Hidden state 2: 熊市

  Hidden state 3: 牛市

  Hidden state 4: 熊市



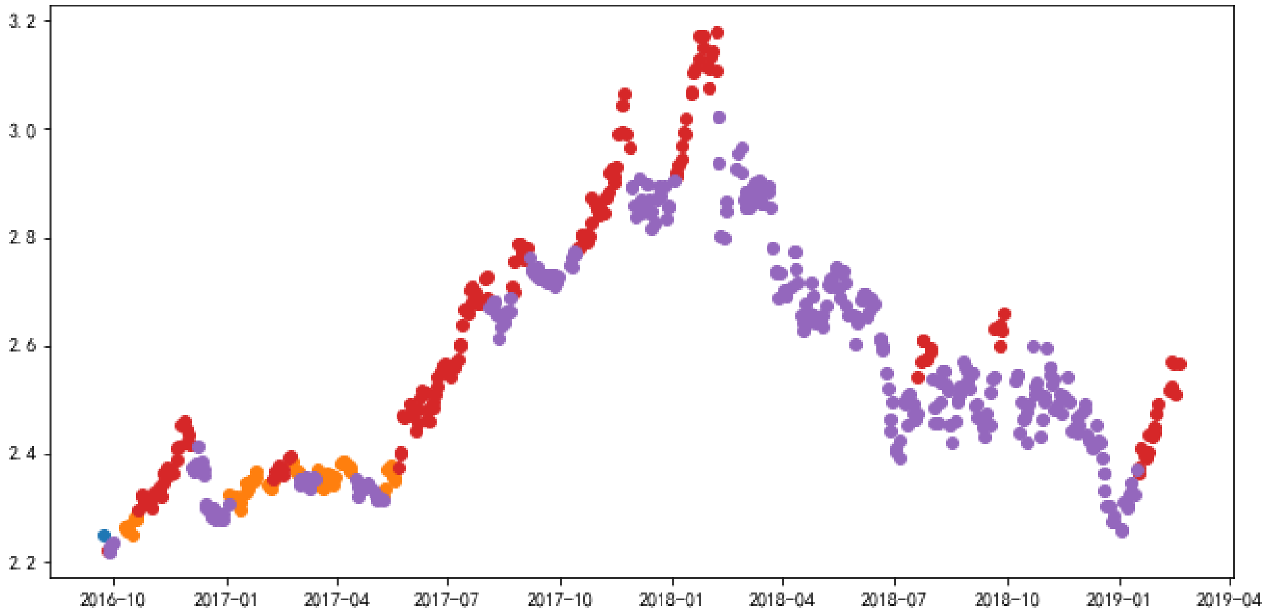


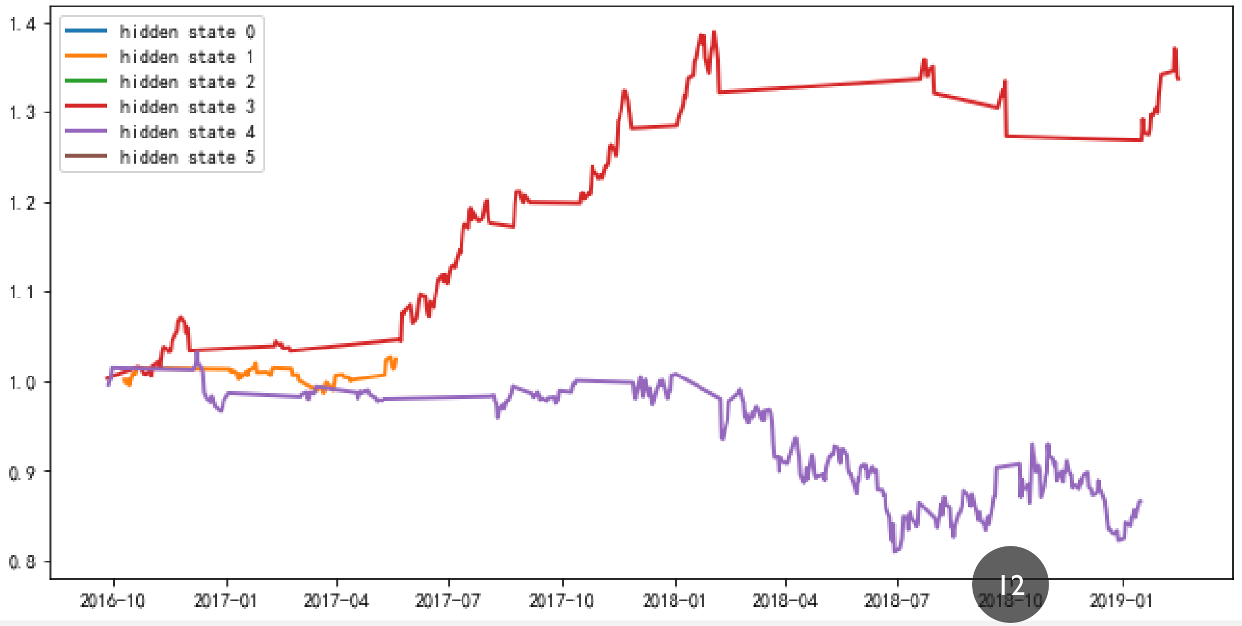
样本外标注

- 样本外数据看似良好地延续了隐状态对应样本内的特征
- 例如：

  Hidden state 3: 牛市

  Hidden state 4: 熊市
- 将两者一对冲就可以用来交易？ No





遇到的问题

- 尽管样本外各种隐状态有区分度，但实际交易中不可能预知未来的可观测序列。
- 因此，上述方法至多是一个伪回测
  - 中信期货《隐马尔可夫模型商品期货应用初探》一文很睿智地使用了上述方法。
  - 尝试使用其他策略但效果不佳

期货实证部分

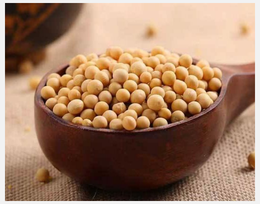
数据来源

交易策略

回测表现

- 实证中我们发现股票类价格序列不平稳：
  - 其统计特征不易于延续
  - 隐马尔科夫模型相当依赖于统计特征
- 转而寻找较为平稳的时间序列：
  - 例如 高相关性期货品种价差序列
- 商品期货价差序列数据维度较小：
  - 一阶矩很容易获得
  - 但不存在直接的交易量等数据

数据来源



- 选择豆一和豆粕两个相关系数较高的品种（近一年两者相关系数0.802）

豆一和豆粕合约价格走势图（日频，2005/01 - 2019/03）

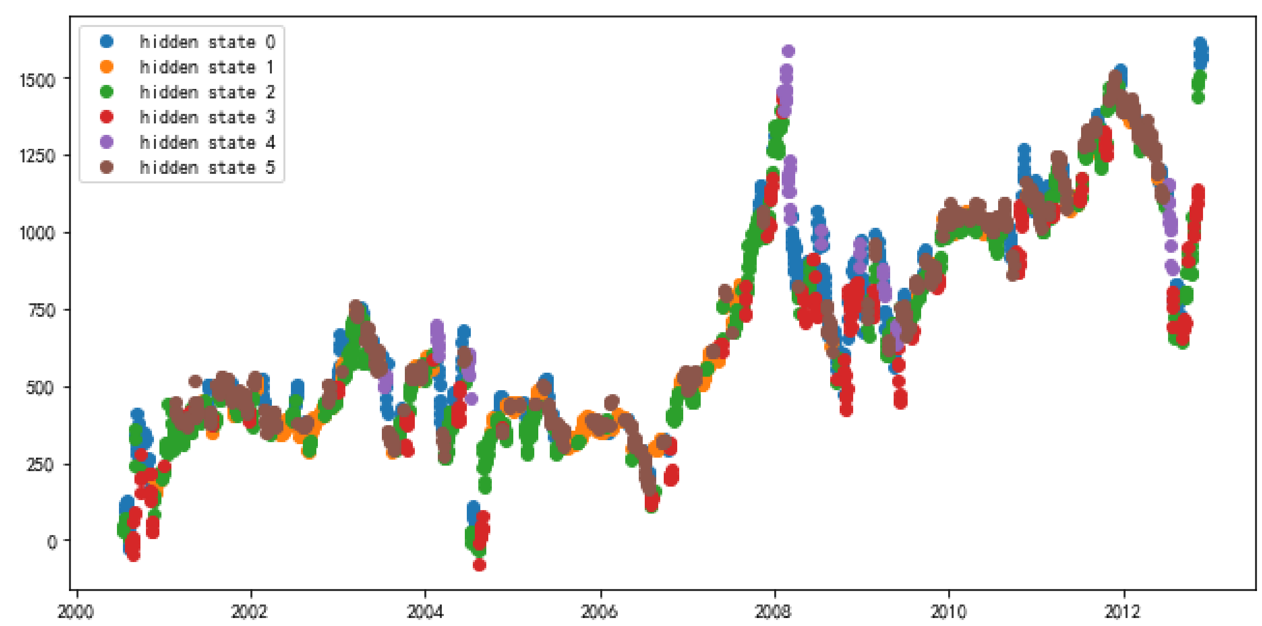
- *以上数据均来自Wind数据库*

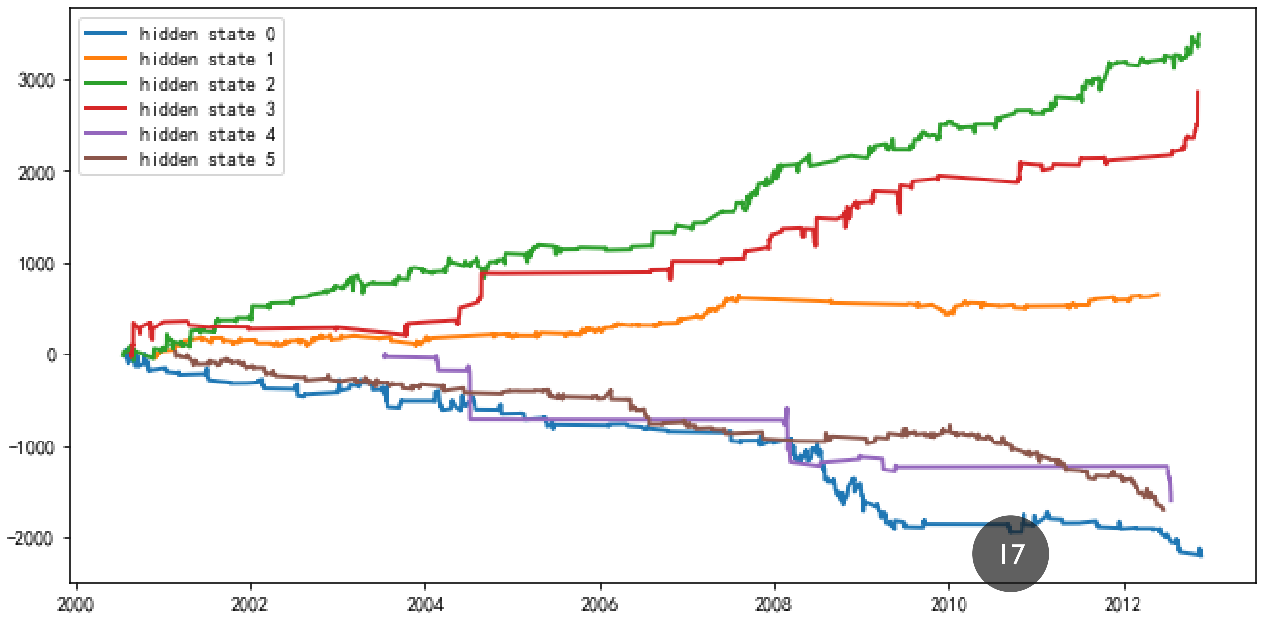
变量定义

变量名	定义
close	豆粕close - 豆一close
ret1	$\mathrm{close}_t/\mathrm{close}_{t-1} - 1$
ret5	$\mathrm{close}_t/\mathrm{close}_{t-5} - 1$
ret10	$\mathrm{close}_t/\mathrm{close}_{t-10} - 1$
ret20	$\mathrm{close}_t/\mathrm{close}_{t-20} - 1$

数据初探

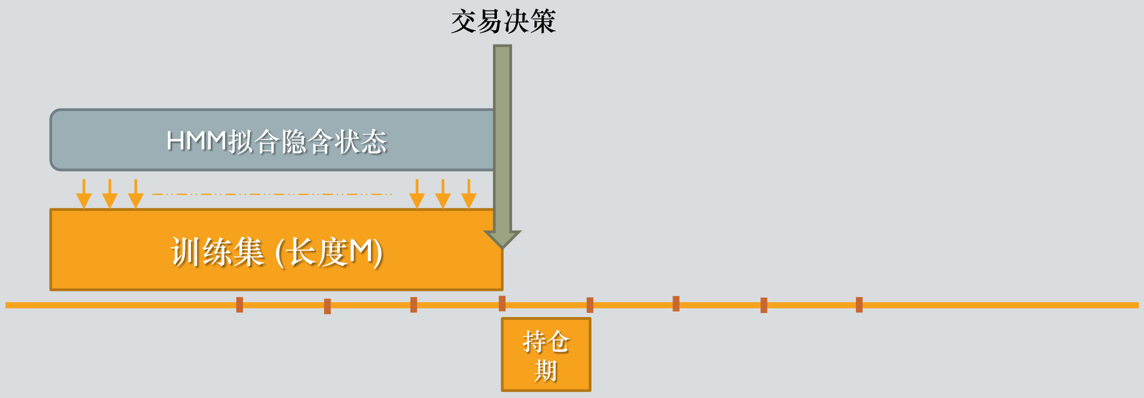
- 首先对数据进行样本内标注，查看各个隐状态之间是否有区分度
- 利用 *ret1, ret5, ret10* 作为观测值对隐状态标记的结果较好

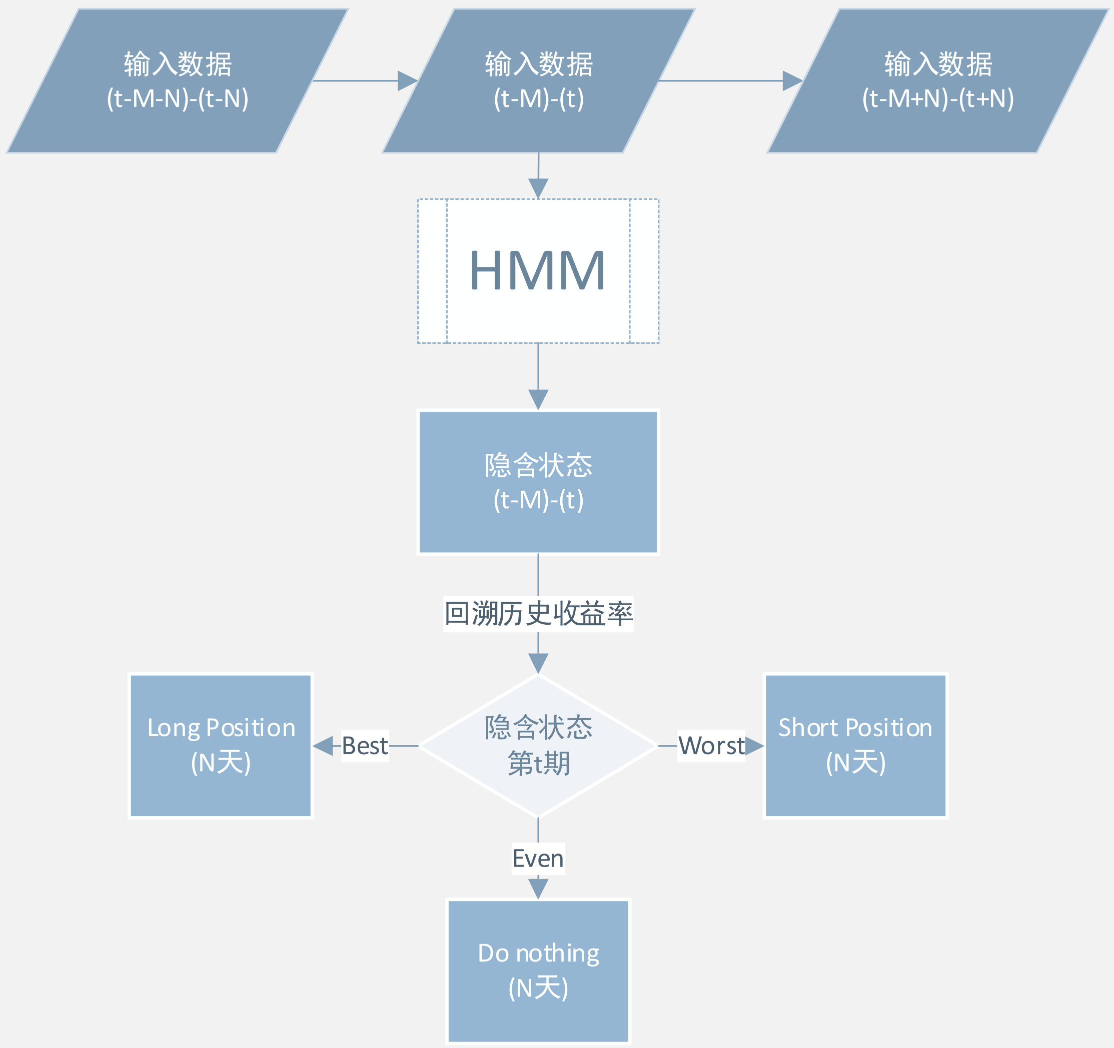




交易策略

- 策略滚动回测流程示意图





交易策略

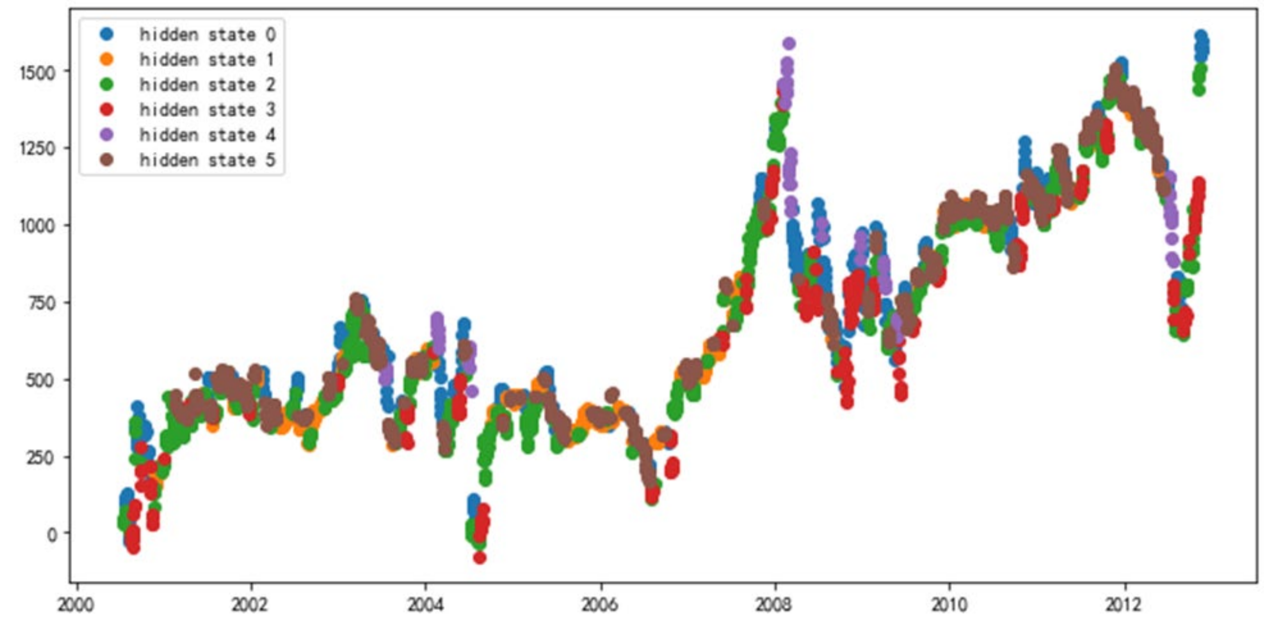
- 采用滚动回测方式：使用过去M天的观测值序列进行参数估计，用最后一天的隐含状态来预测后N天的状态。
- 一个典型的 Embedded Markov Transition Matrix 如下。可以看到下一阶段转移到别处的概率都较小，留在原处的概率最大。

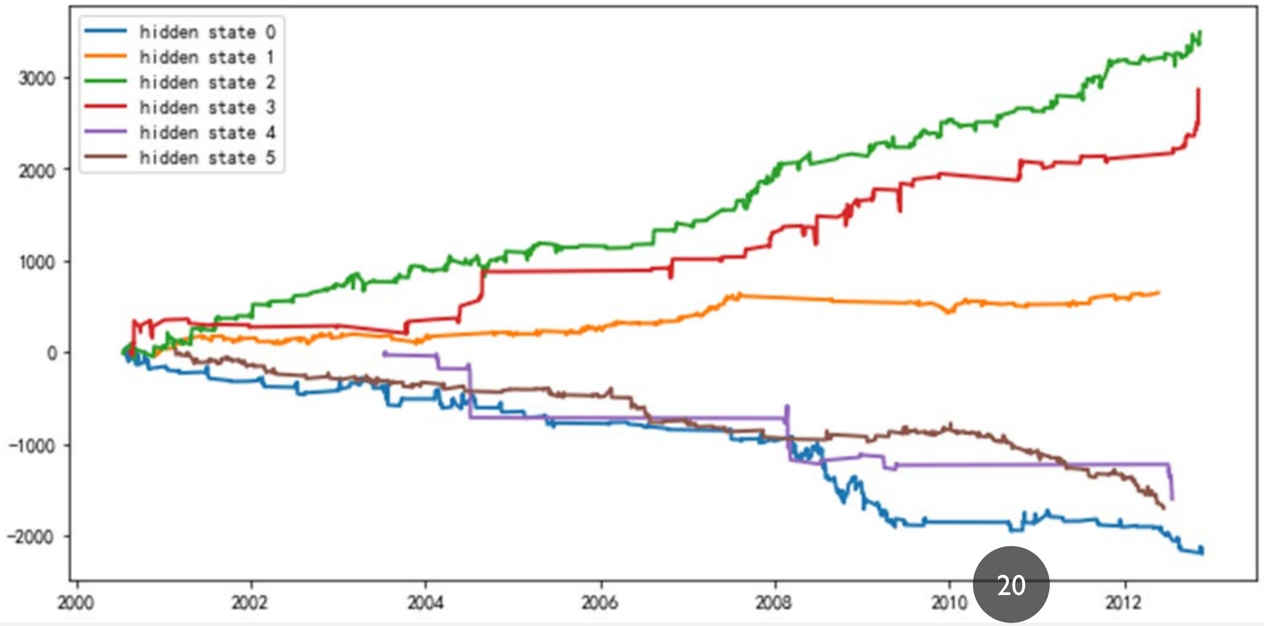
	0	1	2	3	4	5
0	0.676906	0.14084	0.0159308	0.069996	0.00202478	0.094302
1	0.156863	0.599827	0.00189536	0.0403659	0.0237836	0.177265
2	0.0203786	0.0436632	0.719506	0.197854	8.60321e-17	0.0185988
3	0.180515	0.023688	2.9031e-06	0.750891	0.0397787	0.00512427
4	0.0251405	0.0254936	0.00777064	0.0297438	0.783649	0.128203
5	0.0264203	0.221896	0.0096795	0.0122587	0.0796568	0.650088

- 因此，根据状态转移矩阵预测到的下一期隐状态即为当期隐状态。

交易策略

- 隐状态对应含义不确定，且起始状态也不确定，无法直接判断在各个隐状态下应该如何交易。
- 需要借助样本内各个隐状态对应的收益来判断交易策略。
- 例如，右图中程序可以自动判断在 *hidden state 2* 应当做多，在 *hidden state 0* 应当做空。
- 选取2013-2019年的数据做回测
- 交易频率较低，手续费可以忽略不计

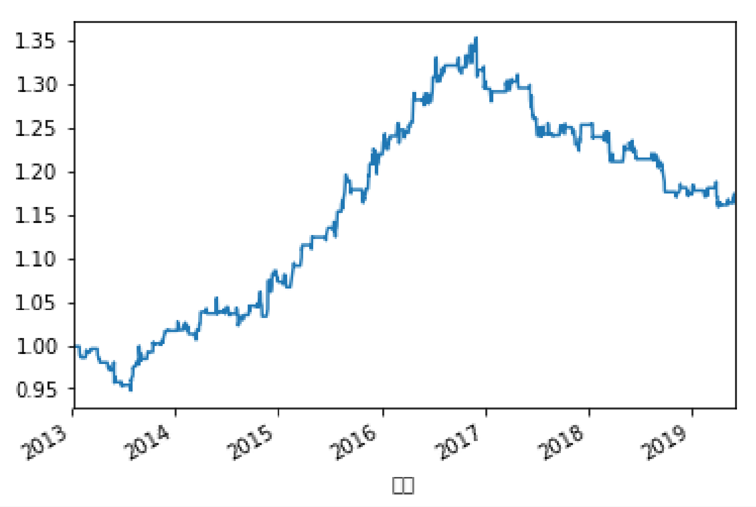




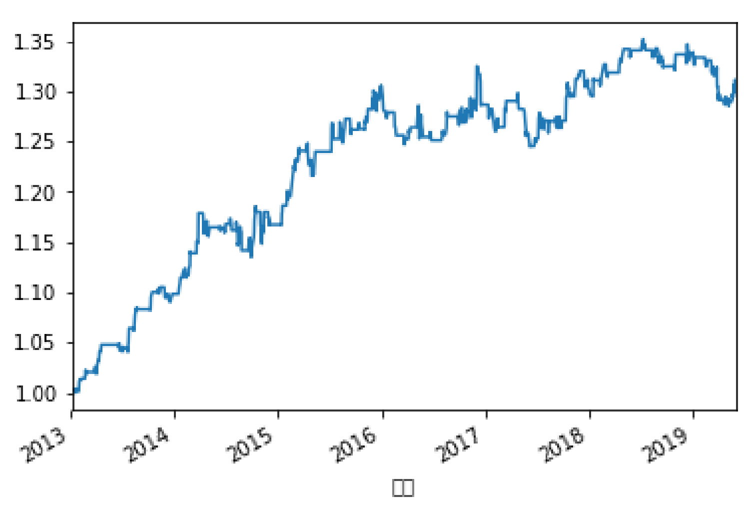
策略表现

2013-2019年 M=2000 N=5

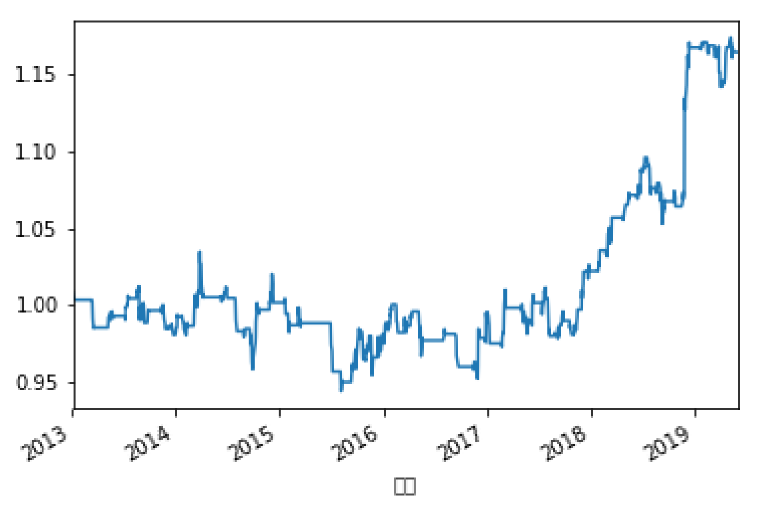
ret1, ret5



★ ret1, ret5, ret10

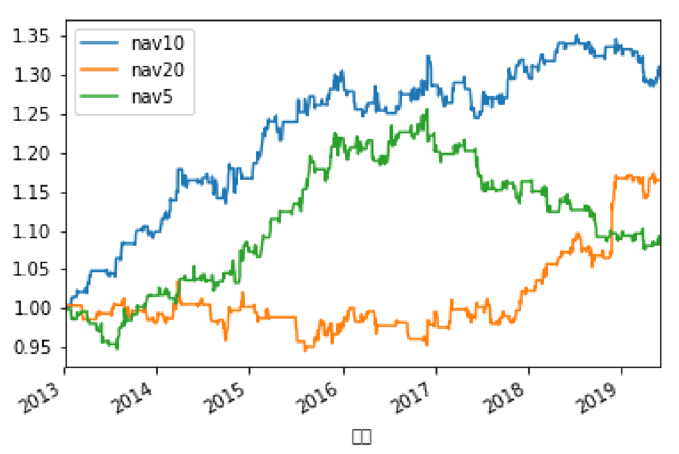


ret1, ret5, ret10, ret20



策略表现

2013-2019年 M=2000 N=5



	ret5	ret10	ret20
总收益	9.42%	30.72%	15.40%
胜率	47.65%	55.46%	57.50%
最大回撤	-14.39%	-6.03%	-8.75%
夏普比率	0.31	0.91	0.47

总结与展望

具有稳定的收益

可以用于交易

模型解释能力不足

- 实证结果表明HMM模型可以被应用于交易策略的设计
  - 具有较稳定的收益
  - 独创性地应用于商品价差
- 但模型仍存在一些不足：
  - 属于机器学习一类的模型
    - 模型解释力不足
    - 需要非常大量的训练数据

总结与展望

更多品种的金融资产

配合期权策略

利用时间序列模型

- 由于时间有限，我们的模型还有很多有待拓展的地方:
  - 配合期权交易策略:

    有些隐状态对应于市场震荡，此时卖出Strangle/Straddle期权组合会有效提高收益和资金利用率
  - 配合时间序列模型:

    事实上用GARCH一类的模型来预测平稳的时间序列很有优势，可以用多个隐状态同时构建多个GARCH，依据隐状态来选择时间序列模型。

总结与展望

输入数据的处理

交易策略

- 由于时间有限，我们的模型还有很多有待拓展的地方：
  - 输入数据：

    可以加入机构持仓量

    遇到高维数据可以采用PCA方法降维
  - 交易策略上：
    - 尝试更多频率
    - 尝试止损止盈

THANKS

Q&A

APPENDIX

另一种交易策略方法

- 利用确定的模型参数 (π,A,B) ，和窗口的观测信号 {*Sn*} ，求窗口后一天的信号的分布 。
- 直接根据估计的信号制定交易策略。
  - 如估计下一天的信号为上涨[3%,5%]，买入股票。

另一种交易策略方法（续）

- 前向概率： $F_n(j) = P(S^n = s^n, X_j)$

  S^n 表示前 n 个信号的随机向量，s^n 表示某个信号向量。

  Fn(j) 可以由以下递推得到：

  $F_1(j) = \pi_j \beta_{js_1}, F_n(j) = \beta_{js_n} \sum_i F_{n-1}(i)\alpha_{ij}$

- 窗口最后一天的隐状态概率分布：

  $P(X_n = j | S^N = s_n) = \frac{P(X_n=j, S^N=s_n)}{P(X_n=j)} = \frac{F_n(j)}{\sum_i F_n(j)}$

- 预测第n+1天的隐状态分布，取条件于 Xn

  $P(X_{n+1} = j | s^n) = \sum_i \alpha_{ij} P(X_n = i | s_n)$

- 预测第n+1 天的收益率分布，取条件于Xn+1

- $P(S_{n+1} = j | s^n) = \sum_i \beta_{ij} P(X_{n+1} = i)$

该步相当耗时…

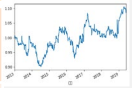
参数敏感性测试 – 总收益率

多最好的一个隐状态，空最差的一个隐状态

	M=1000	M=2000	M=3000
N=1	7.36%	5.76%	-4.68%
N=5	0.00%	**30.72%**	21.56%

多最好的两个隐状态，空最差的一个隐状态

	M=1000	M=2000	M=3000
N=1	1.76%	13.55%	8.39%
N=5	-9.89%	**21.79%**	-18.25%



HMM程序

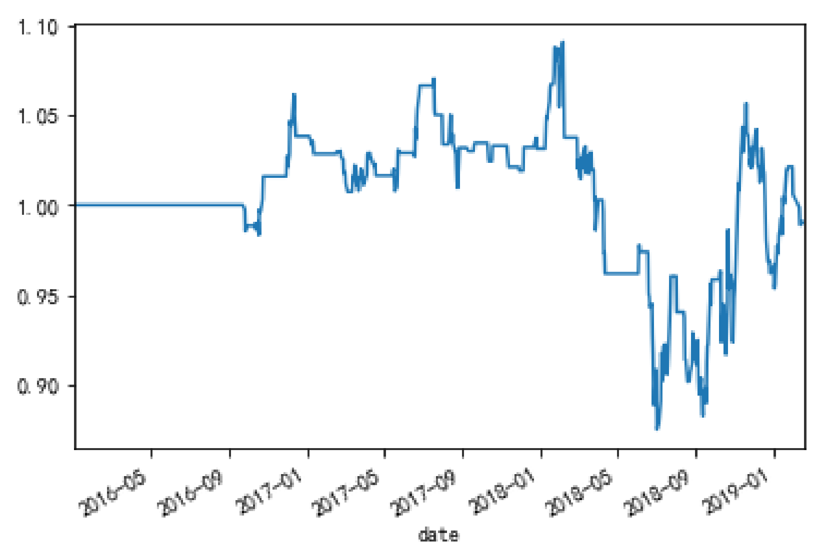
```python
# 选取全样本前半部分作为训练集，后半部分作为测试集
# 对于可观测序列进行隐马尔科夫链建模

trainX = A[:3000]
testX = A[3000:]
model = GMMHMM(n_components=6, n_iter=2000).fit(trainX)
hidden_states = model.predict(testX)
```

状态标注作图程序

```
#%% ax0：样本内的隐状态标注， ax1：样本内各个隐状态对应的收益曲线
_, axes = plt.subplots(2,1,figsize=(11,12))
for i in range(model.n_components):
  pos = (hidden_states==i)
  axes[0].plot(data.iloc[pos]['close'],'o’, label='hidden state %d'%i,lw=2)
  axes[1].plot(data['ret1'].shift(-1).iloc[pos].apply(np.exp).cumprod(),
        label='hidden state %d'%i,lw=2) # 复合收益率
axes[0].legend(); axes[1].legend() # 添加图例
```

ETF实证部分 - 滚动回测



- 将期货实证部分的策略运用于ETF序列上，发现HMM模型完全没有头绪地在进行交易。
- 我们将其归因为数据不平稳。

参考文献

- S. Ross, *Introduction to Probability Models*
- 朱民 等，*解密复兴科技：基于隐蔽马尔科夫模型的时序分析方法。*
- 中信期货，*隐马尔可夫模型商品期货应用初探*
- 兴业证券，*股指期货交易策略系列报告之三： 基于隐马尔科夫链的交易策略*
- *唐灵儿，基于HMM-GARCH 模型的期权定价研究*
- Lawrence R. Rabiner "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE 77.2, pp. 257-286, 1989.
- Jeff A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models.", 1998.

成员分工

- 蔡玮钦、王振宇
  - HMM模型部分（数学原理和算法）
- 王奕能
  - 实证部分（程序和回测）