机器学习方法系列之一：基于单特征因子的隐马尔科夫模型在商品期货上的应用

报告日期： 2016 年 12 月 29 日

李晓辉 助理分析师（金融工程）

从业资格号： F3022611

Tel: 8621-63325888

Email: xiaohui.li@orientfutures.com

★摘要：

在这篇报告中就着重讨论了这个被市场广议的隐马尔科夫模型是怎样的，如何把它应用到期货市场当中来，以及如何更好利用它进行择时预测以获取到更大的收益。

★样本内回测和样本外回测：

1. 使用“平均预期”的预测方式比“最大概率”的预测方式得到的回测结果更好

2. 使用“窗口数据”和“非窗口数据”预测时预测差异率基本在 4%-10%左右

3. 避免使用到未来数据，在使用“窗口数据”时，窗口的长度在超过 10 天之后，对预测结果的影响就不再变化

4. 对于固定的样本内数据集来说，状态数的增加会微微提高预测结果，但是这种提升效果不够特别显著，反而增加计算复杂度，故一般选取状态数为 6-8 之间

5. 通过修正的 Box-Cox 转换方法使得特征因子的统计分布更接近高斯分布，但是对预测结果却并无明显改善

6. 样本外的回测结果显示，对于超买超卖类的反趋势特征因子，它们的回测结果表现更优秀

★窗口滚动式回测：

7. 窗口的长度一般在 150 左右会使得回测结果更好一些

8. 特征因子的选取而呈现明显不同，会有特有的几个值表现得更好

9. 刻画了价格变动速率或波动特性的特征因子的回测结果表现更好

10. 对于其他品种的测试，焦炭、棕榈油等品种在不同特征因子回测中均表现优秀

金融工程

目录

1、背景介绍.....5
1.1、机器学习中的模式识别.....5
1.2、历史“重演”与择时预测.....5
2、隐马尔科夫模型.....6
2.1、模型简单介绍.....6
2.2、在金融市场中的应用.....8
3、历史回测.....9
3.1、数据准备.....9
3.2、样本内回测.....10
3.3、问题讨论.....16
3.4、样本外回测.....21
3.5、窗口滚动式回测.....24
4、其他品种回测结果.....27
5、关于回测设置问题的讨论.....28
6、总结及展望.....32

图表目录

图表 1：离散马尔科夫过程示意图 6
图表 2：隐马尔科夫模型示意图 7
图表 3：HMM 训练步和预测步流程框架示意图 10
图表 4：EM 算法找寻 HMM 最优参数的迭代收敛过程 11
图表 5：隐状态转移矩阵热力图 12
图表 6：特征因子统计分布 12
图表 7：各隐状态对应特征因子和次日收益率的统计结果 12
图表 8：训练参数之后各个隐状态对应的因子和对应次日收益率的统计分布 13
图表 9：标记在收盘价序列上的隐状态序列 14
图表 10：各个隐状态多头复利策略回测结果 14
图表 11：各隐状态多头策略累计复利净值 15
图表 12：理想的样本内回测多空策略的复利净值 15
图表 13：时间序列各时点上每个隐状态的发生概率 16
图表 14：预测准确率随截断窗口长度的变化 17
图表 15：预测准确率随隐状态数的变化 17
图表 16：夏普比率随隐状态数的变化 18
图表 17：风险收益比随隐状态数的变化 18
图表 18：特征因子 pct_chg 的统计分布 19
图表 19：特征因子 pct_chg 转换之后的统计分布 19
图表 20：原始特征因子 pct_chg 的 QQ 图 20
图表 21：转换后的特征因子 pct_chg 的 QQ 图 20
图表 22：特征因子 pct_chg 在进行转换前后回测净值对比 20
图表 23：特征因子 pct_chg 的样本外回测 21
图表 24：不同特征因子样本外回测得到平均预测结果 21
图表 25：不同特征因子样本外回测得到平均预测结果 22
图表 26：6 个平均表现最好的特征因子样本外回测净值表现 23
图表 27：6 个平均表现最好的特征因子样本外回测指标统计 23
图表 28：窗口滚动式回测流程示意图 24
图表 29：ROC 的滚动预测结果随窗口长度的变化 25
图表 30：pct_chg 的滚动预测结果随窗口长度的变化 25
图表 31：pct_chg 的滚动预测结果随窗口长度的变化 25
图表 32：ROC 的滚动预测结果随隐状态数的变化 26
图表 33：pct_chg 的滚动预测结果随隐状态数的变化 26

图表 34：不同特征因子窗口滚动式回测得到预测结果.....26
图表 35：不同品种 RC 的滚动预测结果.....27
图表 36：不同品种 ROC 的滚动预测结果.....27
图表 37：不同品种 pct_chg 的滚动预测结果.....27
图表 38：不同品种 vol_ratio 的滚动预测结果.....27
图表 39：RC 的滚动预测结果随止损比例的变化.....28
图表 40：ROC 的滚动预测结果随止损比例的变化.....28
图表 41：pct_chg 的滚动预测结果随止损比例的变化.....28
图表 42：vol_ratio 的滚动预测结果随止损比例的变化.....28
图表 43：RC 的滚动预测结果随手续费比例的变化.....29
图表 44：ROC 的滚动预测结果随手续费比例的变化.....29
图表 45：pct_chg 的滚动预测结果随手续费比例的变化.....29
图表 46：vol_ratio 的滚动预测结果随手续费比例的变化.....29
图表 47：RC 的滚动预测结果随滑点数的变化.....30
图表 48：ROC 的滚动预测结果随滑点数的变化.....30
图表 49：pct_chg 的滚动预测结果随滑点数的变化.....30
图表 50：vol_ratio 的滚动预测结果随滑点数的变化.....30
图表 51：RC 的滚动预测结果随杠杆比例的变化.....31
图表 52：ROC 的滚动预测结果随杠杆比例的变化.....31
图表 53：pct_chg 的滚动预测结果随杠杆比例的变化.....31
图表 54：vol_ratio 的滚动预测结果随杠杆比例的变化.....31

1、背景介绍

1.1、机器学习中的模式识别

谷歌 Deep Mind 团队开发的人工智能围棋程序 Alpha Go 战胜世界冠军李世石之后，其棋力水平甚至被职业棋手评定为超过了职业九段的水平，等级分也超过了曾排名世界第一、曾对其小觑的柯洁。这让我们不由惊叹基于深度学习的人工智能程序竟已能够做到如此让人意想不到的结果。引发了全世界范围内对人工智能燃起浓厚兴趣的深度学习是什么呢？事实上，深度学习只是机器学习领域的一个方面，是层次更深的神经网络学习，一般的简单的神经网络可能只包含输入层、隐含层和输出层，而深度学习的隐含层数目更多，并且结构更复杂，通过自顶而下地逐层训练使得学习达到更优的结果。谷歌已经开放了深度学习系统 TensorFlow 以方便更多的人对深度学习进行认识了解和使用。

这些机器学习领域的“新宠”也都更多地被应用到语音识别、自然语言理解等工业领域上。然而除了以神经网络为基础的机器学习方法，更早被业界在语音识别研究中所用到的经典的机器学习模型包括隐马尔科夫模型、动态时间规整算法等。隐马尔科夫模型是将系统内在的状态变化过程假设为满足一阶马尔科夫性质，并由外部的可被观测的一系列观测值推测出隐状态变化过程，而动态时间规整算法则是分析一段被观测的时序片段在历史时间序列上面的规整距离，如果距离越小，说明该片段与历史上的某片段时序较为相似，那么我们可以预计未来一段时间内也较可能仍旧有相似性。总的来说，这两种方法都是认为系统存在着某种可能在未来再次发生的模式，我们所对未来的预测其实是基于对历史数据的“模式识别”。

1.2、历史“重演”与择时预测

有效市场假说认为市场是有效的，那么在这种市场当中，价格会及时地反映出市场中的所有信息，而对市场的任何预测都是无效的没有意义的。但是这只是一种“理想”的市场状态，因为任何的信息（预期）的传递过程总会具有一定的延滞性，这使得价格并非反映出了所有的信息（预期），而是在其“理想”的价值附近震荡。但是现代金融理论认为市场具有分形结构，而分形特性在物理里面往往是指系统所具有的标度不变性，可能是时间标度也有可能是空间标度。对于时间标度来说，最被人所熟知的一个指标 Hurst 指数其实反映的就是一个时间序列的时间统计相关性，在物理学当中 Hurst 指数其实反映也是一随机过程的时间扩散系数，如果指数接近 0.5，就意味着自由（随机）扩散，在 0.5 到 1 之间说明该过程具有长程相关性，并且具有持续性，属于超扩散过程，若指数在 0 到 0.5 之间则说明该过程具有反持续性，属于次扩散过程，并且具有一定的突变性。大量的研究工作指出 A 股市场的 Hurst 指数大于 0.5，具有长程相关性，也就是说市场在不同时间尺度下可能和历史具有一定的相关性。所以从分形市场的角度来看，市场中所谓发生的历史“重演”其实不是完全的复制而仅仅是指与历史相似。

其实我们经常会拿现在的市场状况和历史上类似的状况相比，比如常常会有人报道历次大跌之后市场的走势，或者历年国庆、春节后的市场走势。桥水基金创始人 Ray Dalio 更是以研究历史出名，他最近发文认为如果要拿历史对标当今的话，最类似的就是 1935-1945 年之间的美国。但实际上历史所发生重演并不是完完全全的翻版和复制，因为不可能存在着和历史上某节点的一样的市场背景条件，所以“重演”并不等于重复。

其实我们拿历史的数据对比现在，也只是为了做参考，意义在于通过对历史上多次发生的这种大概率事件进行对比归纳，发现可能存在的共性特征，从而通过学习总结出对未来的预测和应对措施。

基于此，如果把模式识别方法应用到金融市场当中，其实核心的思路就是通过寻找和发现历史上较为相似的情形来对比如今的市场并预测未来的走势。但是这样的方法毕竟是基于历史的预测，而历史不代表未来。我们也是基于技术分析的思想，通过模式识别的方法，来寻找出可能存在的大概率发生的规律。

2、隐马尔科夫模型

2.1、模型简单介绍

隐马尔科夫模型（Hidden Markov Model, HMM）是被广泛使用在自然语言处理领域的一种统计学习模型，简单地讲就是对于某系统来说，其随时间的离散演变过程是具有隐含未知状态的马尔科夫过程。安德烈·马尔科夫于 1906 年最早定义了这样的一个随机过程，并模拟生成一条状态序列，而该状态序列在任何时刻的预期状态都只和上一时刻的状态信息相关，这一特性也被称为马尔科夫性质。马尔科夫链（Markov Chain）即描述了具有马尔科夫性质的时间和状态都是离散的序列。假设$\{X_t\}$是时间离散的一阶马尔科夫过程所对应的状态集合（Status Set），而由一个状态X_i到下一个状态X_{i+1}存在着转移概率，并且这种转移概率具有独立性，只与紧接的上一个时刻的状态推算得到，那么该过程就可以得到如下的描述（如图表 1 所示），

$$P(X_{t+1}=x \mid X_1=x_1, X_2=x_2, \cdots, X_t=x_t)=P(X_{t+1}=x \mid X_t)$$

即意味着任意时刻的系统状态x的条件概率分布仅仅是由其上一时刻的系统状态所决定的。因此也将马尔科夫性质称为无记忆性，尽管X_i受且仅受X_{i-1}的影响，是X_{i-1}的函数，但是实际上X_i却已经包含了位于其之前状态所包含的所有信息。假设马尔科夫过程的状态序列所对应的真实状态集合为$\{x_1, x_2, \cdots, x_n\}$，并且不同状态之间的转移概率矩阵$\mathbf{A}$（$n \times n$ 阶）是不随时间变化的，那么任一时刻状态可被描述为$X_t=\mathbf{A}^{t-1}X_1$，所以真正决定整个离散马尔科夫过程的是初始状态和转移概率矩阵。

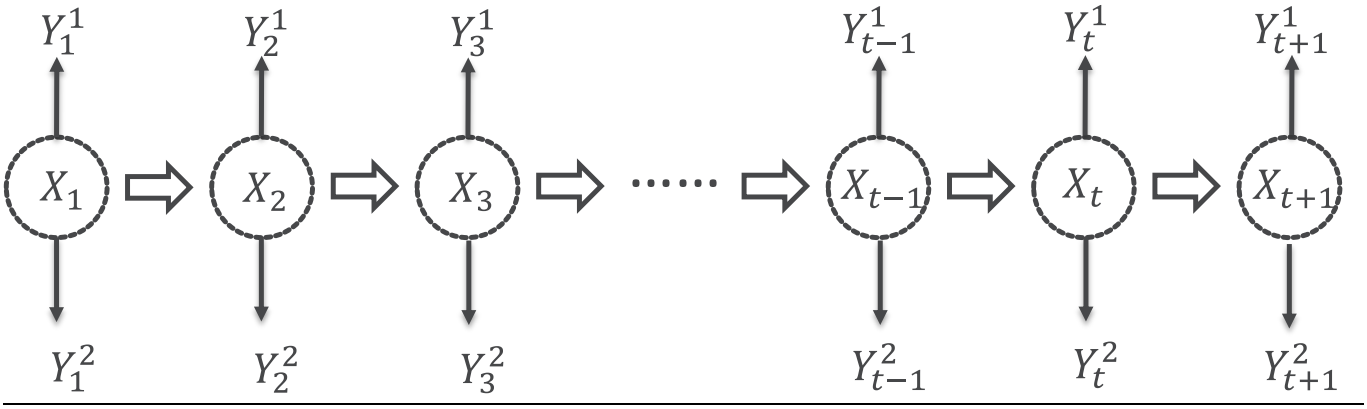
图表 1：离散马尔科夫过程示意图

资料来源：东证衍生品研究院

马尔科夫过程最开始也是被用于进行语音识别、词性标注、甚至是进行人力资源供给预测（这里的前提是状态的变化具有一定独立性）。我们获取到某个人说的一段语音，假设他说的每个字可以被称为一个状态，语音识别做的就是根据声学信号自动地标注出这一段话的每个字。而如果这段语音中带有拼读错误或者有一些其他的杂音，那么我们应怎么正确地推测出说话者所表达的意思呢？其实隐马尔科夫模型就是在假设系统满足马尔科夫性质，但是其状态集合则是非显性的，或者是无法进行直接观测的也或者是无

法具象化的，比如一段语音的“真实”意思，而我们却可以得到该系统随时间演变过程中的一系列观测值，比如这段语音的频率信号等。也就是说，隐马尔科夫过程描述的就是由一个由隐藏的马尔科夫链生成不可观测状态的随机序列，外界观测者是无法直接观测到该序列所对应的真实状态，而只知道各状态所产生的对应观测值。

图表 2：隐马尔科夫模型示意图



资料来源：东证衍生品研究院

假设隐马尔科夫模型的隐状态序列为 $X=\{X_i\}$，而其对应的状态数目为 n，状态集合则为 $\{x_1,x_2,\cdots,x_n\}$，不同状态转移概率矩阵（Transition Probability Matrix）为

$$\mathbf{A}=(a_{ij})_{n\times n}=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

其中 $a_{ij}=P(X_t=x_j \mid X_{t-1}=x_i)$，是上一时刻状态为 x_i 时下一时刻状态变为 x_j 的概率，这个概率是马尔科夫性质的体现，因为状态的转移概率只和其前一时刻的状态相关。假设观察任意状态 x_i 时，并且产生了 k 类具有不同意义的的观测指标，第 k 类观测指标所得的观测值集合（Observation Set）为 $\{y_1^k,y_2^k,\cdots,y_m^k\}$，共有 m 种不同的观测结果。那么我们便可以写出不同隐状态下，分别得到第 k 类观测指标的 m 种观测结果的观测概率矩阵（Emission Probability Matrix），也被称为混淆矩阵，

$$\mathbf{B}^k=(b_{ij}^k)_{n\times m}=\begin{pmatrix} b_{11}^k & \cdots & b_{1m}^k \\ \vdots & \ddots & \vdots \\ b_{n1}^k & \cdots & b_{nm}^k \end{pmatrix}$$

其中 $b_{ij}^k=P(Y_t^k=y_i^k \mid X_t=x_j)$，表示在隐状态为 x_j 的时候，观测得到 y_i^k 的概率。需要注意的是，在实际应用时我们常常假设观测值 y_i^k 的概率分布具有高斯分布的特点。除此之外，我们还应知道初始的状态是如何的，故假设的初始状态概率向量是

$$\boldsymbol{\pi}=(\pi_i)_n$$

其中 π_i 表示 t=1 时状态为 x_i 的概率。图表 2 展示了隐马尔科夫模型的示意图。

前文已经提到了一个马尔科夫过程是由初始状态和转移概率矩阵两者共同决定的，而一个隐马尔科夫过程则是由初始状态、转移矩阵和初始状态概率向量。转移矩阵 $\mathbf{A}$ 与初始状态概率向量 $\boldsymbol{\pi}$ 共同决定了所谓隐藏着的马尔科夫链，而观测矩阵 $\mathbf{B}^k$ 则确定了如何从隐状态得到第 k 类的观测指标，$\lambda=(\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ 被称为 HMM 三要素，三者共同决定了一个隐马尔科夫过程。

不管是语音识别还是手写识别领域，如果拿隐马尔科夫模型来进行模式识别的话，常常遇到评估、学习和解码这三种问题。评估问题也被称作概率计算问题主要考虑的是对于一个给定的隐马尔科夫模型其生成一个给定的观察序列的概率是多少，解决这类问题比较适合使用前向算法。学习问题是如果已知观察序列，如何知道最可能的模型参数，这种问题实际上就是参数估计的学习过程，更适合用 Expectation-Maximization（EM）算法（Baum 于 1972 年提出来 EM 算法的一个特例，前向-后向算法，也被称为 Baum-Welch 算法）。还有一种问题是关于解码的，也被称为预测问题，主要考虑的是到底是什么样的隐状态序列最有可能与已知生成的观测序列所匹配，解决这类问题应该使用 Viterbi 算法，这类问题也常见于词性标注或者分词。

2.2、在金融市场中的应用

HMM 最早是在 20 世纪 60 年代由 L. E. Baum 提出的，Baum 在后来加入了一家基金公司 Monemetrics，而这家公司则是日后大名鼎鼎的文艺复兴科技公司的前身。但是把隐马尔科夫模型“发扬光大”的应用则是进行语音识别，李开复曾在其博士毕业论文中提出来基于 HMM 构建的一个用于语音识别的系统 Sphinx，并被《商业周刊》评为 1988 年美国最重要的科技发明。尽管如今语音识别领域可能更多地采用了深度学习的方法，但 HMM 也算是对这一领域的研究做出巨大贡献。

刚才提到的文艺复兴科技公司可谓是基金投资领域的传奇，通过进行量化投资连续 27 年的投资回报率高于股神巴菲特。市场上不少人猜测其创始人西蒙斯和他所招募的“小伙伴们”——数学教授或语音识别专家，最早在外汇交易上使用了隐马尔科夫模型。受此启发，也越来越多的人尝试在金融市场当中使用 HMM 来对市场进行预测。

如果把隐马尔科夫模型应用到金融市场的话，那么我们需要考虑的其实主要是两种问题：第一是对 HMM 的参数估计（学习），第二是基于观测值的状态预测（解码）。已某交易所内交易的商品期货为考察对象的话，假设该商品的内在状态为牛市涨、跌和震荡已经熊市中涨、跌和震荡，所以不同的隐状态反映的不仅仅是表面的涨跌情况不一，而且也可能反映了市场环境的不同。如果只是仅仅知道它的交易价格，实际上仍然无法知道它所处的隐状态。实际上，交易价格、成交量等截面特征因子信息序列都是该商品在被交易时被观察到的表观“现象”，我们可以通过这些特征因子序列来刻画商品的交易状态，就如同物理学中对粒子状态的描述需要知道参考系中的粒子的坐标 x 和动量 p 一样。因此有了多维的观测值序列，而商品期货的隐状态序列便可以通过 HMM 这个“黑匣子”得到。首先，我们将一段 k 维的观测值时间序列通过 EM 算法，采用极大似然的方法进行参数估计，估计出 HMM 最可能的参数组合。然后根据未来的观测值序列，利用已训练好的 HMM 对未来的状态进行预测。

但是需要注意的是在金融市场中应用隐马尔科夫模型也是具有其局限性的。第一是因为

隐马尔科夫模型的首要假设是马尔科夫性质假设，

$$P(X_{t+1} \mid X_t, X_{t-1}, \cdots, X_1, Y_t, Y_{t-1}, \cdots, Y_1) = P(X_{t+1} \mid X_t)$$

也就是说隐马尔科夫模型的本质仍然是马尔科夫链，但实际上金融市场中交易对象的状态往往不具有严格马尔科夫性质，实际上每个隐状态之间的转变并非独立的，而且也不是概率不变的。第二，隐马尔科夫模型还有一个重要假设是，输出独立（观测独立）假设，即不同维度的观测值之间、不同时刻的观测值之间应是相互独立、互不影响的，

$$P(Y_t \mid X_t, X_{t-1}, \cdots, X_1, Y_{t-1}, \cdots, Y_1) = P(Y_t \mid X_t)$$

但实际上根据金融物理中的统计结论我们知道，不同时刻的观测值相互之间具有一定的长程相关性，因此观测值并不是严格独立的。第三，隐马尔科夫模型在进行参数估计、预测解码的时候，假设了观测值序列满足正态分布的条件，但实际情况是不同维度的观测值并不都能够一定满足正态分布。以收盘价计算得到的对数收益率序列为例，其分布其实为截断列维分布，相较正态分布具有“尖头胖尾”的特点。因此，尽管这种方法因为其自身的假设从而具有它自有的局限性，但是我们在这篇报告中要探讨的内容更多的是模式识别方法的应用思路，而不是应用的结果。

3、历史回测

3.1、数据准备

1. 测试对象

我们以国内商品市场上日均交易量最大、交易最为活跃的螺纹钢为隐马尔科夫模型择时预测的回测对象。

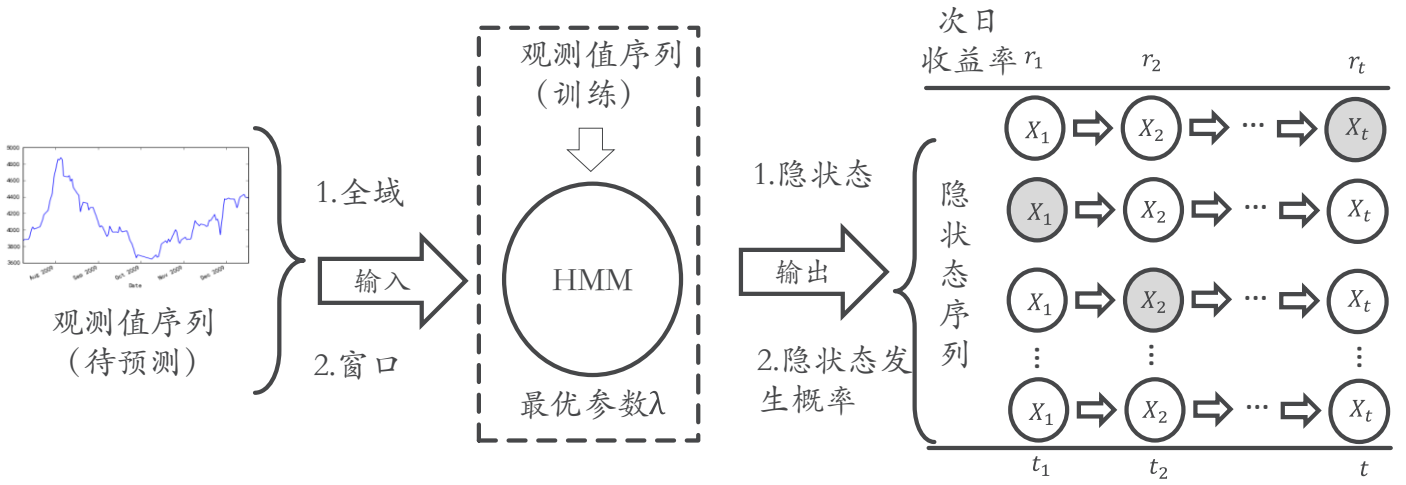
2. 特征因子

主力合约数据从其上市 2009 年 3 月 27 日开始，截止到 2016 年 10 月底为止。而特征数据（即观测值数据）则不仅仅包含了开盘价、最高价、最低价、收盘价、成交量和持仓量等基本行情指标，同样地也包含了从 Wind 当中所获取到的一些技术分析类的指标，总计 43 个指标，其中每个指标均以 Wind 给出的默认参数计算得到。这 43 个指标可以被看做是特征提取之后所得到的特征因子。事实上每一类的特征因子都可以看做是螺纹钢隐状态序列所对应的观测值，而观测值又会反映出隐状态的变化。在进行回测之前我们还需要把特征因子中含有空值的数据部分剔除掉，所以实际得到的总样本数据范围是 2009 年 6 月 11 日至 2016 年 10 月 21 日，共计 1790 组 43 维截面因子数据。我们首先以其中一个因子“pct_chg”(即当日相较前日结算价的涨跌幅的百分值*100) 作为 HMM 隐状态的观测值，来观察一下 HMM 是如何被应用在金融市场中的价格预测的，并接下来对几个问题进行更细致的分析与讨论。

我们需要首先假设它的隐状态总数为 6 个。其实任何人都无法真正知晓实际有多少个隐状态，而且我们也无法得知每个隐状态的具体指代什么内容。对于 HMM 来说，设定了隐状态总数 N 之后，模型就会通过最大似然估计算法用 N 个整数对 HMM 隐状态序列进行标记。一旦得到了对 HMM 的最优参数估计，我们便可以根据 HMM 所反映出来的特点分析得到这 6 个隐状态可能指代的内容是什么。

大体上，我们利用 HMM 进行回测可以分为三大步，第一步训练模型得到参数组合，第二步利用已训练好的模型对待预测的观测值序列进行解码，得到最可能的隐状态序列及相应的发生概率，并通过与次日收益率之间的关系计算得到对未来的预测（涨和跌，用+1 和-1 表示），而第三步是根据预测情况进行回测，计算出回测评价指标的结果。这里比较重要的是，也是和 HMM 相关的是前两步（如下图所示），我们在进行第三步之前可以先由前两步得到预测结果，然后与实际涨跌情况进行比较，计算出预测准确率。对于第三步来说，回测框架是独立而且固定的，因此回测过程并不会对预测情况造成反馈影响，而且历史回测可能会因回测期内的数据而产生不同的回测结果，从而可能得到差异较大的回测指标评价结果，因此这里我们就着重考量预测准确率，以该指标来对 HMM 的训练估计步和解码预测步进行分析评价。

图表 3：HMM 训练步和预测步流程框架示意图



资料来源：Wind，东证衍生品研究院

在得到了对次日的涨跌预测之后，我们用螺纹钢的历史数据进行了相关的回测，在得到日度回测结果之后会根据上面表中的指标进行评价。具体地，回测规则如下：

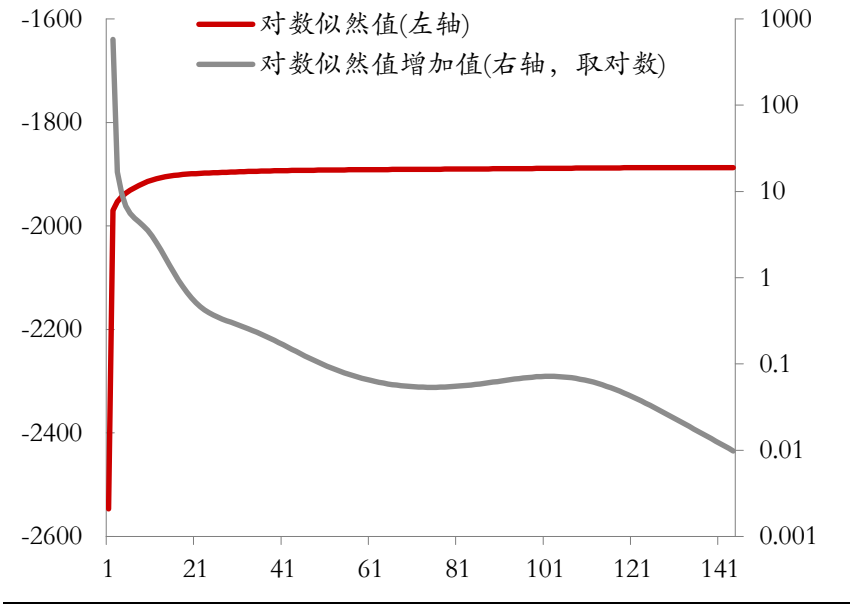
1. 可以做多和做空，如产生信号以次日开盘价进行开仓、平仓
2. 复利投资，并固定杠杆比例约为 3 倍（使用资金比例 30%，保证金比例 10%）
3. 交易手续费为万分之一，暂未考虑滑点
4. 止损：相对于开仓价累计浮亏超 0.5%
5. 预测方向与持仓相同时不操作，相反时会先平仓再进行反向开仓

3.2、样本内回测

我们将全样本的历史数据按 2015 年前后，分为样本内和样本外。首先，我们对样本内的数据进行 HMM 方法的回测。

第一步：训练

图表 4：EM 算法找寻 HMM 最优参数的迭代收敛过程



资料来源：Wind，东证衍生品研究院

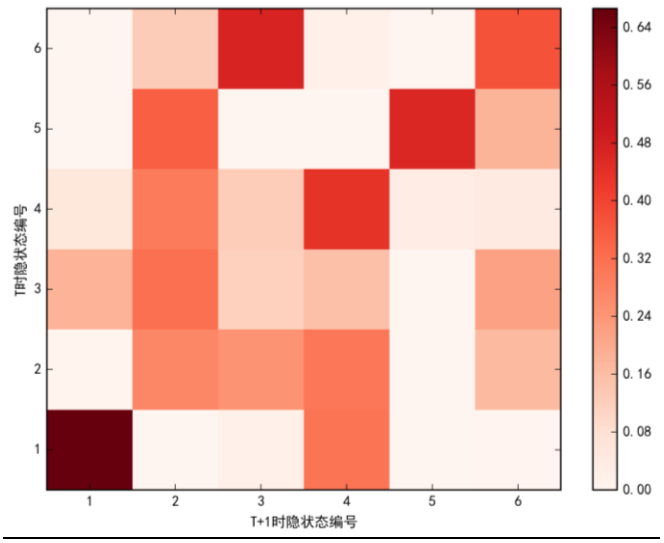
HMM 的训练过程，简单地说，其实就是利用观测值序列推测出可能的最优参数组合，这个过程用到的是 Expectation-Maximization 算法。EM 算法其实是求解极大似然估计的一种方法，它的特点是可以使用缺损的、截断的等不完整数据进行参数估计计算。简单地说，EM 算法主要包括两步，expectation 步和 maximization 步，通过迭代的方式最大化完整数据的对数似然值的期望，最终得到最大化的不完整数据的对数似然值。这个迭代过程其实是一个逐次逼近、逐步收敛的过程。首先，可以随机地初始化参数，计算出每个训练样本的可能结果的概率，以确定出目前在这组参数下的最可能状态，然后再由样本对当前参数进行修正，并再次计算出修正之后的参数及状态发生概率，这样，通过迭代的方式不断使似然值得到收敛，同时参数也逐渐地逼近于真实值。

我们利用单个特征因子百分比涨跌幅作为唯一的观测值，尝试对螺纹钢主力合约的隐马尔科夫链进行极大似然估计。我们设置收敛阈值为 0.01，即当某一次迭代之后所对数似然值的增加值小于 0.01，迭代过程才会停止。下图展示了 EM 算法在进行参数估计的时候，对数似然值随着迭代次数的变化过程，共计迭代 144 次。我们可以看到，迭代过程会很快地使对数似然值增加，但是增加的幅度却随迭代次数不断地降低，直到增加幅度低于所设的阈值，这时我们才认定此时得到的参数已经非常接近于真实参数值了。

在训练之后，我们便得到了 HMM 在样本内数据上的各个参数，包括初始时各隐状态的概率向量 π，不同隐状态之间的转移概率矩阵 A 和从各隐状态观测到不同观测值的概率矩阵 B。图表 5 展示了转移概率矩阵的热力图，其中纵轴代表 T 时刻的隐状态，而横轴代表 T+1 时刻的隐状态，颜色越深表示这两种隐状态之间的转变概率越大。我们也发现转移矩阵并非对称阵，两个隐状态互为前后的转移概率并不相同，这也间接反映了 HMM 的马尔科夫性质。而对于混淆矩阵 B 来说，由于观测值是连续变化的，而不是有限数的离散值，这从特征因子的原始统计分布（图表 6）就可以看到。在通过 HMM 标记了隐

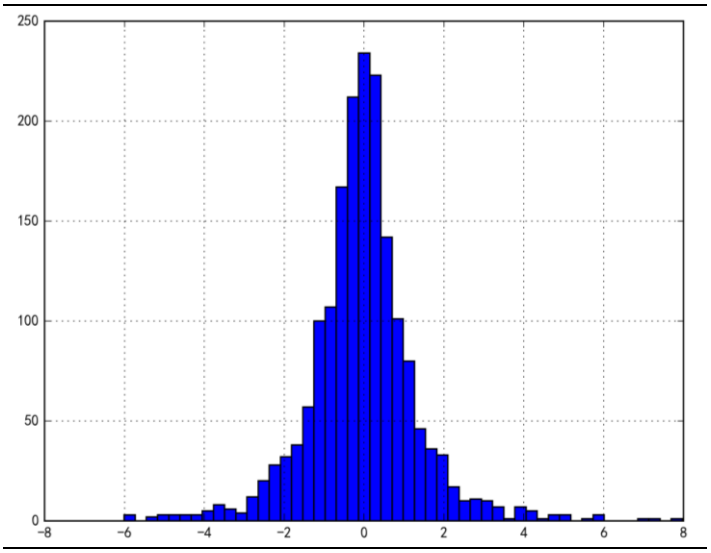
状态之后，各隐状态对应的特征因子的分布如何呢？图表 7 左侧列的 6 幅图展示了各个隐状态所对应的特征因子的统计分布，实际意义是间接说明了在各个隐状态下观测到不同观测值的概率，而如果把他们合并在一起其实就是图表 6 中的统计分布图。有了这些统计分布，我们现在可以大致猜测出这些隐状态可以被分为两类，一类对应着标准差较小，分布较为集中的特征因子，具体为编号 2、3、6 的隐状态；而另一类则是标准差较大，分布较广的特征因子，具体为编号 1、4、5 隐状态。而图中右侧列的 6 幅图则分别展示了这 6 个不同的隐状态各自所对应的次日收益率的统计分布，他们的平均值和标准差如下表所示。

图表 5：隐状态转移矩阵热力图



资料来源：Wind，东证衍生品研究院

图表 6：特征因子统计分布



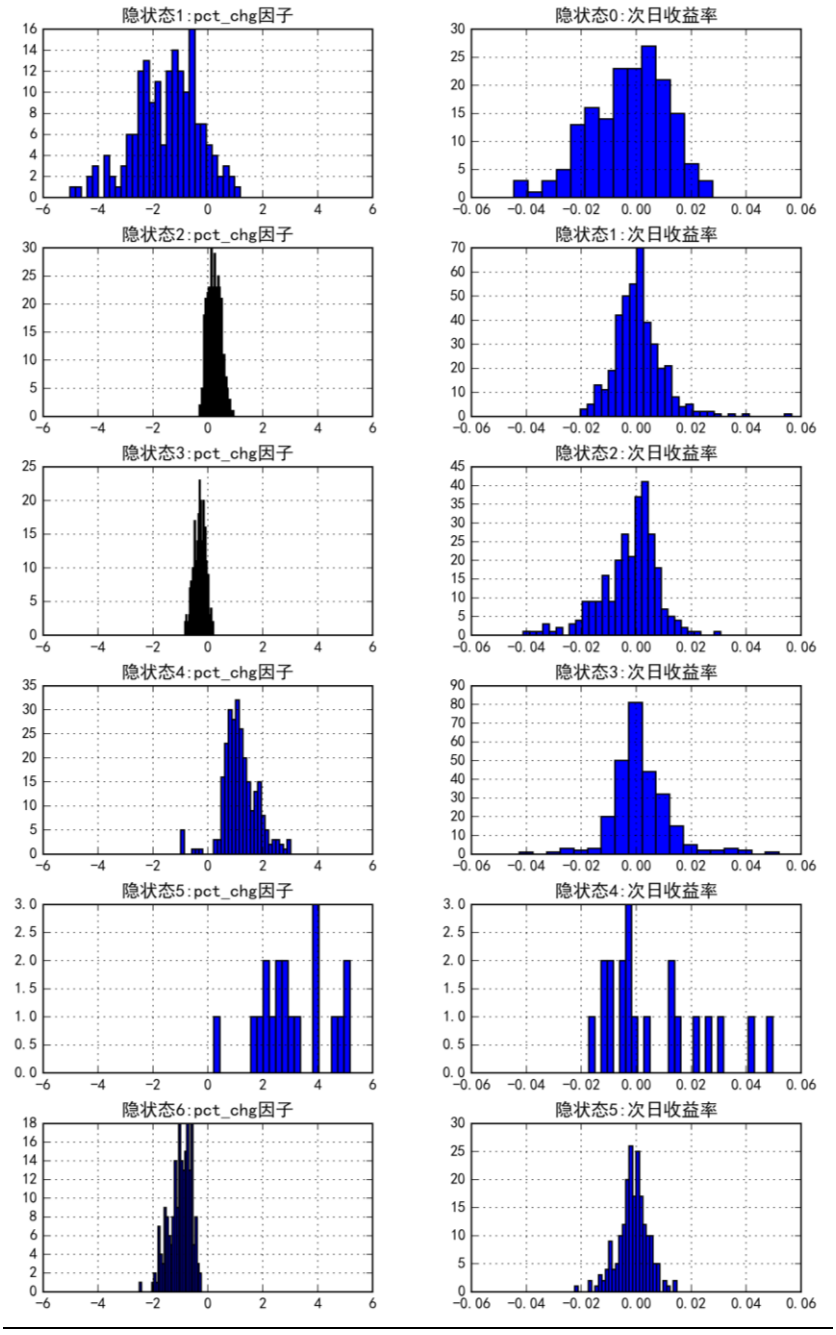
资料来源：Wind，东证衍生品研究院

图表 7：各隐状态对应特征因子和次日收益率的统计结果

隐状态编号	pct_chg 因子平均值	pct_chg 因子标准差	次日收益率平均值	次日收益率标准差
1	-1.2394	1.7008	-0.0022	0.0166
2	0.2308	0.0955	0.00083	0.0089
3	-0.2739	0.0890	-0.0024	0.0104
4	0.8199	0.6303	0.00189	0.0123
5	2.8291	2.8412	0.00658	0.0185
6	-0.8871	0.2839	-0.0013	0.0056

资料来源：Wind，东证衍生品研究院

图表 8：训练参数之后各个隐状态对应的因子和对应次日收益率的统计分布

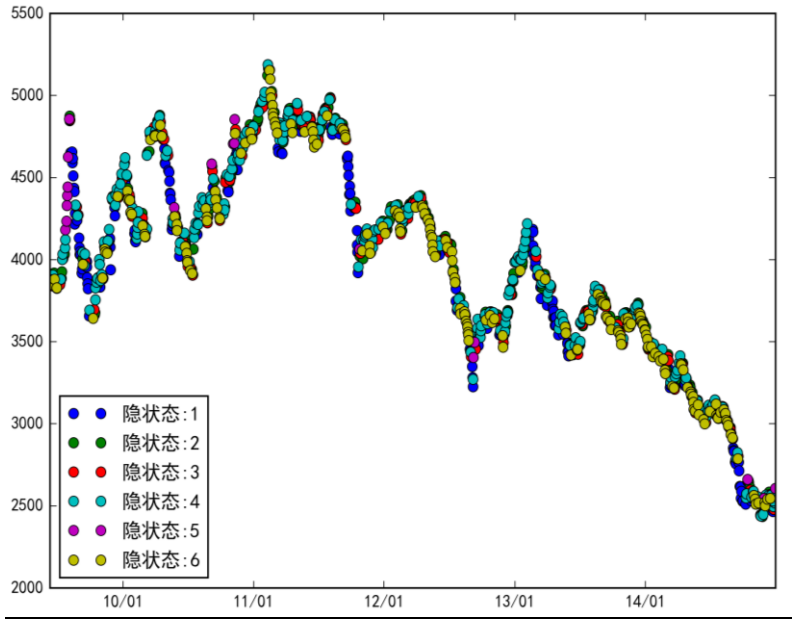


资料来源：Wind，东证衍生品研究院

第二步：预测

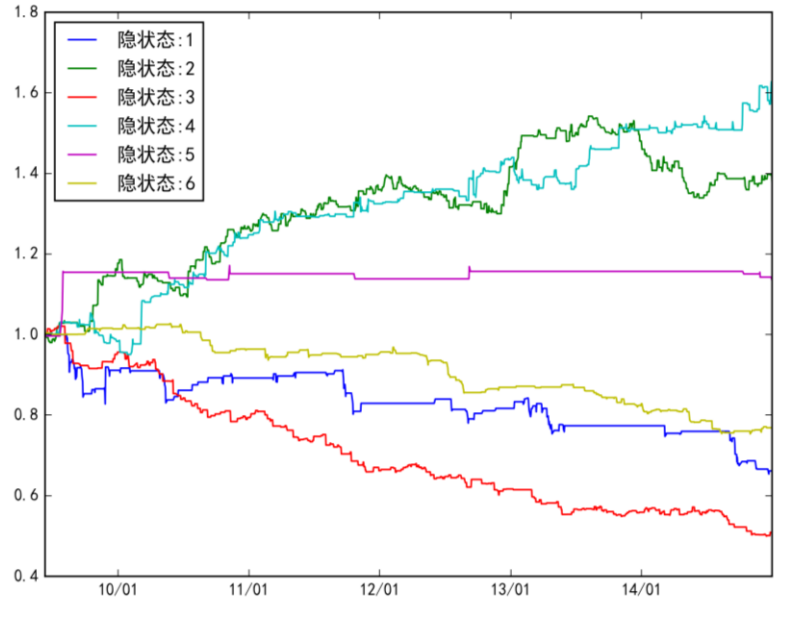
我们已经通过 EM 算法的训练之后得到了模型的最适参数，如果再把样本内数据作为输入数据扔给训练好的模型之后，便可得到对样本内数据的解码预测。这个解码的过程事实上是寻找最可能的隐状态序列，目的是最大化在某隐状态序列条件下得到目前已知

图表 9：标记在收盘价序列上的隐状态序列



资料来源：Wind，东证衍生品研究院

图表 10：各个隐状态多头复利策略回测结果



资料来源：Wind，东证衍生品研究院

的观测值序列的概率。这样看来如果采用穷举法也可以得到该结果，但是如果序列长度为 N，隐状态数为 6 的话，那么共计有 6^N 种可能性，穷举法会非常耗时耗力。而《数学之美》中提到解决这种解码问题的一种最经典的算法，Viterbi 算法。该算法是一种动态规划算法，使用递归方式减少计算复杂度。我们这里便用这种算法解码观测值序列得到最可能的隐状态序列。图表 9 展示了在通过 pct_chg 因子训练并解码后得到的隐状态序

列后，在螺纹钢的收盘价序列上标记出对应的隐状态。我们在前面曾得到了每个隐状态所对应的次日收益率的分布（图表 8），但是隐状态对应的具体涨跌情况仍然不知。所以，我们这里假设一种多头策略，即只在出现某编号的隐状态时，才对次日的收益率进行复利相乘，其他时刻不进行操作。通过这种多头策略，我们便可以得到在样本内区间各个隐状态的具体涨跌情况，如图表 10 所示。

第三步：回测

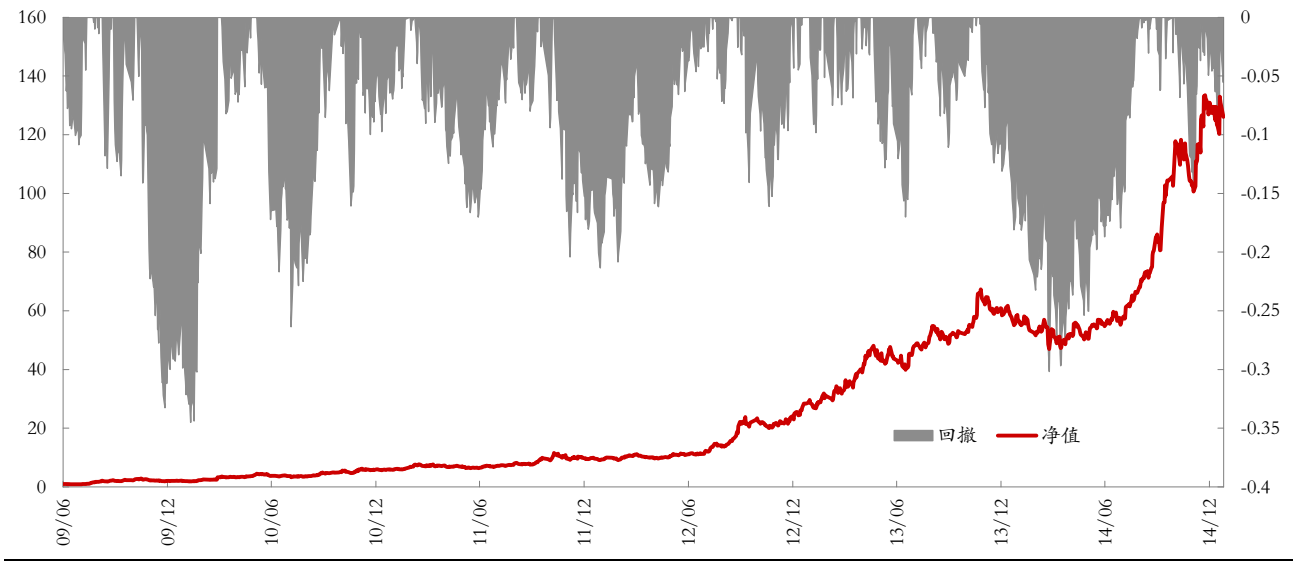
图表 11 给出了各隐状态多头策略的次日平均收益率、次日收益率平均值与标准差的比值和复利计算得到的累计净值。结合图表 10，我们认为在当前状况下，编号为 4、2、6 的隐状态比较适合做多，而编号为 6、1、3 的隐状态则比较适合做空。在知道了各个隐状态对应的涨跌情况之后，我们便可以按照前文给出的规则，得到一种理想的多空策略回测结果，如图表 12 所示，该策略得到预测准确率为 53.2%，年化收益率 146.4%，夏普率 3.12，收益最大回撤比 4.24。

图表 11：各隐状态多头策略累计复利净值

隐状态编号	次日收益率平均值	次日收益率平均值/标准差	复利计算的累计净值
4	0.0019	0.154	1.627
2	0.0008	0.094	1.378
5	0.0066	0.355	1.136
6	-0.0013	-0.228	0.769
1	-0.0022	-0.134	0.662
3	-0.0024	-0.226	0.509

资料来源：Wind，东证衍生品研究院

图表 12：理想的样本内回测多空策略的复利净值



资料来源：Wind，东证衍生品研究院

但是，为何我们认为这样的多头策略是理想的呢？这是因为在进行解码预测的时候，我们其实用到了未来数据。如果我们仔细分析 Viterbi 算法的话，会发现这个算法最大特点

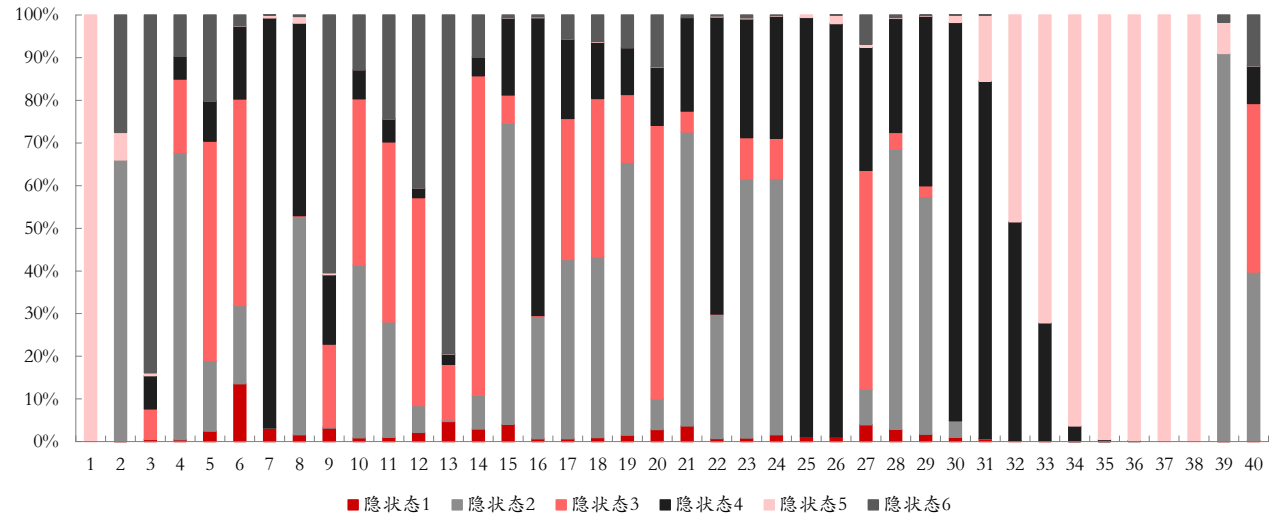
是会考察观测值序列的“上下文”，也就是说实际上用该算法进行解码的时候，不仅仅和输入的某时刻观测值序列的数值之前的数据有关，而且也和之后的数据有关。所以，对于同时包含 T 时刻的数据、T 时刻前后长度不相同的两段观测值序列，在使用参数固定的同一个 HMM 进行解码时，得到的 T 时刻对应的最可能的隐状态会有可能不一样。Viterbi 算法中的“上下文”极大地影响了各时刻隐状态概率的计算。

3.3、问题讨论

现在我们来讨论一下，目前为止我们在使用 HMM 的过程中遇到的几个重要问题。

问题 1：HMM 预测出了隐状态序列，如何得到对下一日涨跌情况的预测？

图表 13：时间序列各时点上每个隐状态的发生概率



资料来源：Wind，东证衍生品研究院

我们之前所采用的预测方式是在 HMM 训练之后对样本数据进行预测得到最可能的隐状态序列，然后计算观测具有同一编号的隐状态所对应的次日收益率复利累积值的表现，如果为正就认为该状态适合做多，如果为负就认为该状态适合做空。但是实际上 HMM 预测所给出的隐状态序列，是取的每个时点上发生概率最大的那个隐状态。通过这种多头策略的方式给出对未来的涨跌判断是否一定有效呢？图表 13 展示了待预测时间序列前 40 个时点，HMM 用 Viterbi 算法预测出来的 6 个隐状态的发生概率，柱子越长代表该颜色对应的隐状态的概率越大，而 HMM 给出的隐状态序列实际上默认地采用了各个时点上具有最大概率的隐状态。HMM 预测出来的隐状态实际上只包含了最大概率隐状态的信息，那么是不是可以采用平均预期的方式进行预测呢？这种预测方式是直接计算每个时点上对未来收益率的平均期望值，即用 6 个状态的发生概率 $P_i(t)$ 分别乘以这 6 个状态对应的未来收益率平均值 $\hat{r}_i$，再加和起来就得到对次日收益率的平均预期，

$$R(t)=\sum_{i=1}^{n}P_i(t)\cdot\hat{r}_i(t)$$

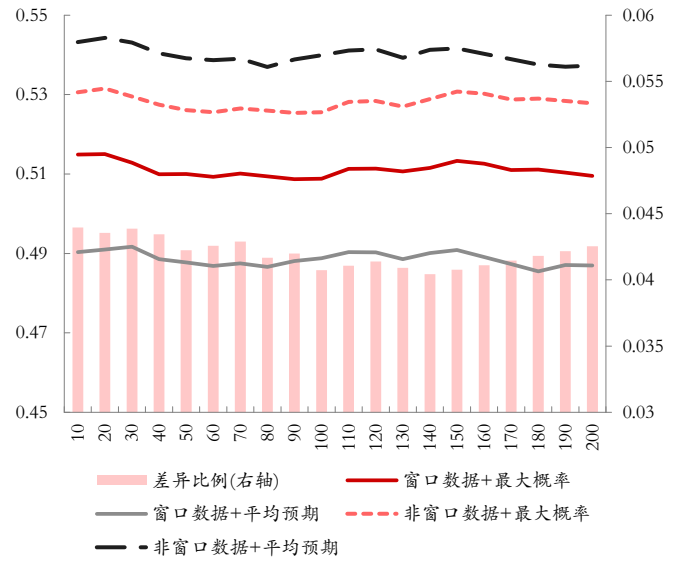
$$\hat{r}_i(t)=\frac{\sum_{\Delta t=1}^{5}e^{-\Delta t/5}\cdot r_i(t+\Delta t)}{\sum_{\Delta t=1}^{5}e^{-\Delta t/5}}$$

为了能够对每个隐状态对应的未来收益率进行更为准确的刻画，这里的 $\hat{r}_i$ 在计算的时候，我们采用了对未来 5 日内的实际收益率的加权平均，其中权重是按照时间跨度 e 指数衰减，即与当前时间点越远权重越低。如果最终得到的平均预期 R 为正的话，那么就对次日预测上涨，而如果为负的话，那么就预测次日下跌。因此，我们便有了两种对次日涨跌进行预测的方式，一种我们称之为“最大概率”方式，HMM 在解码之后得到了实际为各个时点上概率最大的隐状态序列，然后根据与实际次日涨跌情况的对应，给出适合做多的状态组和做空的状态组，之后预测得到的新状态就依据做多组合做空组来进行判断；而第二种方式我们称为“平均期望”，就是通过计算出每时点的次日收益率的平均期望，依据其正负来对次日涨跌进行判断。但是，这两种方式孰好孰坏，我们会在后面几个问题的讨论中进一步分析。

问题 2：如何解决 Viterbi 算法中对未来数据的使用问题？

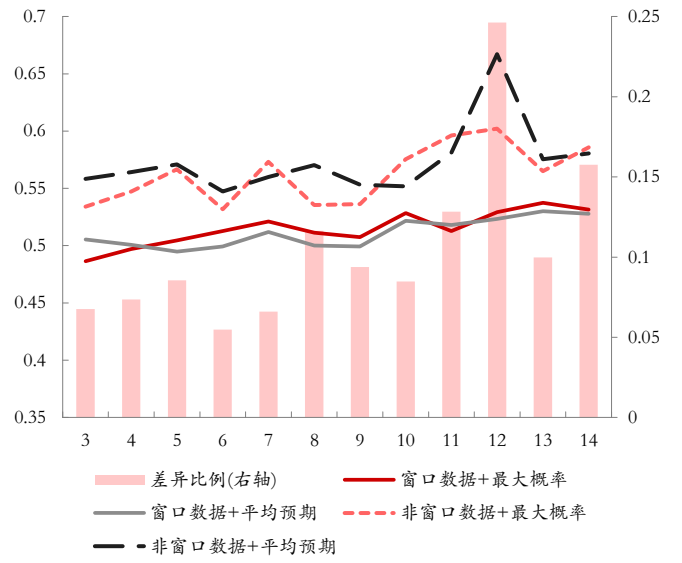
Viterbi 算法计算复杂度低、速度快，虽然好用，但是对于历史回测问题来说，算法中的未来函数问题有可能使得我们在实际应用中得到“不可能”得到的结果。所以，如果我们把全部待预测的数据一次性扔给 HMM 的话，它解码出来的最可能的隐状态序列其实是不正确的。那么为了能够避免未来数据的使用，我们尝试用基于每个回测日期最近一段时间的窗口截断数据，通过 HMM 进行解码得到这段数据的隐状态序列，然后取该序列的最后一个值作为对下一日的预测，也即是说我们只用了包含待预测观测值和该观测值之前的数据进行解码。但对于观测值之前的数据的长度取多久合适以及这种方法和原来的结果的差异有多大这两个问题，我们也做了相关的计算，结果如下面的图表所示。

图表 14：预测准确率随截断窗口长度的变化



资料来源：Wind，东证衍生品研究院

图表 15：预测准确率随隐状态数的变化



资料来源：Wind，东证衍生品研究院

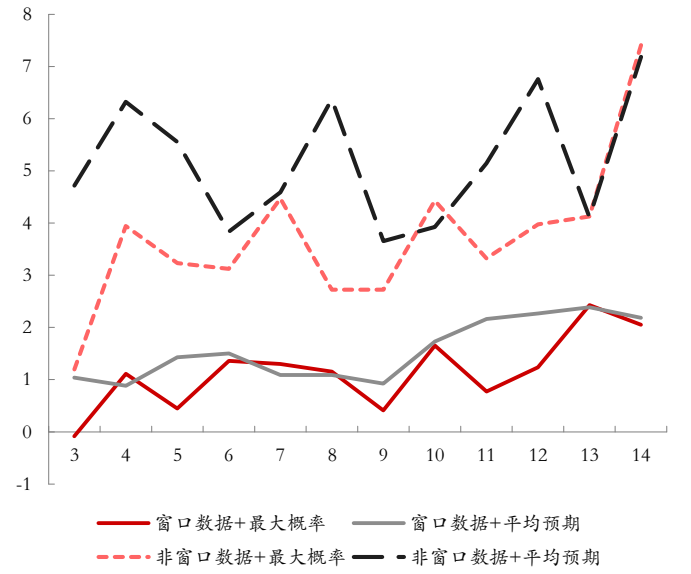
“窗口数据”和“非窗口数据”以在输入给 HMM 的待预测数据有没有包含未来数据为区分，前者只截取预测点之前的数据，而后者则包含了未来数据。“最大概率”和“平均预期”则是在前文问题 1 当中提到的两种预测方式，由此则可以构成 4 种不同的预测组合。我们发现在避免采用未来数据时，会使得模型的预测准确率下降，这一点也说明

了未来数据确实能够使得结果“好看”一些，但是却不切实际。而且在选择不同长度的窗口截断数据时该不同的预测组合基本都保持不变，使用未来数据与否的涨跌预测值的差异率（预测不相同的数量/预测总数）也基本稳定在 0.04-0.045 之间。最后，我们也看到，在使用未来数据时，采用“平均预期”的预测方式能够提高预测准确率，而不使用未来数据时，这种预测方式则会使得结果显得更加“随机”。

问题 3：HMM 中隐状态数目的大小对预测有什么影响？

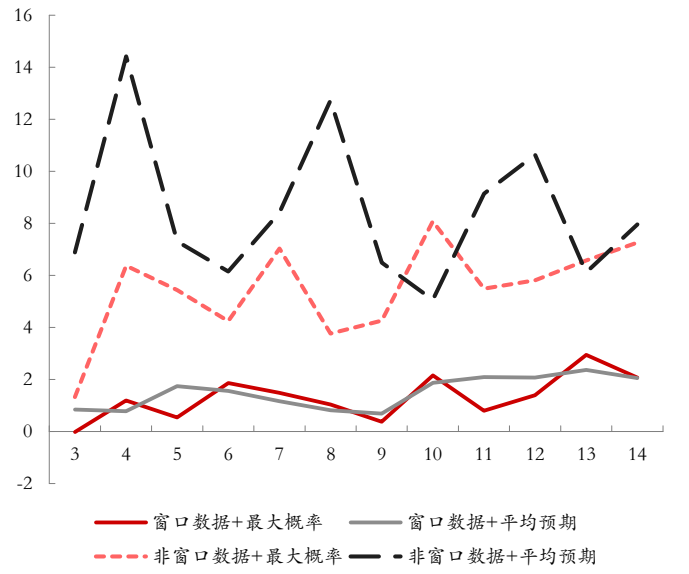
如图表 15 所示，同样地，我们也分了四种预测组合，来分析不同大小的隐状态数对预测准确率的影响。整体来看，不管采用哪种预测组合，都有隐状态数越高其预测准确率越高的特点，这说明隐状态数的增加，HMM 能够更好更准确地识别观测序列从属于哪种隐状态，进而提高对待预测序列的解码准确度。另一方面，和图表 14 结果一致的是，对于没有使用到未来数据的两个预测组合来说，他们的预测准确度相对更低一些。另外，我们这里得出的结论，也可以从夏普比率、风险收益比等回测指标随隐状态的变化情况（如图表 16 和 17 所示）中得到佐证。虽然提高隐状态数能够提高预测准确率，而且也同样提高了回测夏普比率、风险收益比，但是计算复杂度也提高了，因此造成运行时间偏慢。为此，我们认为在进行回测的时候选取隐状态数 6、7、8 是比较合适的。

图表 16：夏普比率随隐状态数的变化



资料来源：Wind，东证衍生品研究院

图表 17：风险收益比随隐状态数的变化



资料来源：Wind，东证衍生品研究院

问题 4：特征因子的统计分布特性是否对预测结果具有显著影响？

值得注意的是，HMM 假设了每个维度的观测值序列都应具有正态分布特性，但是实际上不同的观测值都很难满足这一条件，几乎都无法呈现高斯分布的特点。但是我们却可以通过某些转换方法把非正态分布的观测序列转换成正态分布序列或近似正态分布的序列，然后我们对比一下转换之后的观测序列是否能够有效提高 HMM 的预测准确率。

通常情况下，在统计学中我们会采用 Box-Cox 方法进行数据转换，假设被转换数据序列

为$\{y\}$，那么转换之后的新序列$\{y'\}$应满足如下的式子，其中λ是转换参数，

$$y'=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda}, \lambda \neq 0\\ \log(y), \lambda = 0\end{cases}$$

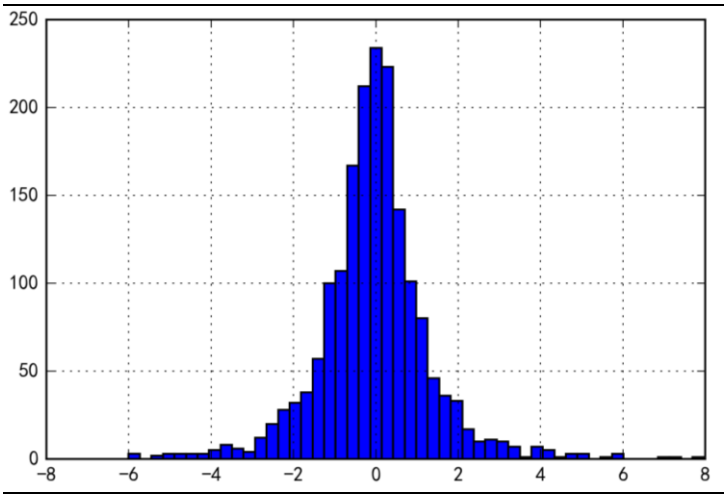
但是在使用该方法的过程中我们会发现它的局限性，它要求被转换的数据一定为正。而我们得到的特征因子序列，比如涨跌幅等，往往会有负数值出现，那么就不能直接地使用 Box-Cox 方法进行转换。为此，我们在该方法的基础上提出了修正版的 Box-Cox 转换方法，其形式如下，

$$y'=\begin{cases}\dfrac{(y+1)^{\lambda}-1}{\lambda}, \lambda \neq 0, y \geq 0\\ -\dfrac{(-y+1)^{\lambda}-1}{\lambda}, \lambda \neq 0, y < 0\\ \log(y+1), \lambda = 0, y \geq 0\\ \log(-y+1), \lambda = 0, y < 0\end{cases}$$

新的转换方法也同样满足原始 Box-Cox 转换中所要求的几个特点，（1）保持了被转换序列中的次序大小，即转换函数应是一个递增函数，这样才不会改变原始数据的顺序，而改变的只是数据之间的距离；（2）转换函数是连续的、可导的，不存在“特殊点”。最后，我们就采用新的转换方法将样本数据中的各个特征因子都进行一次转换。如果某观测数据 y 的绝对值非常大时，转换因子选取不当就很容易使得转换后的结果几乎完全相等。所以，我们在对数据进行转换之前会首先对原始数据进行标准化处理。需要注意的是，这种方法只能使得原始的非正态分布数据集更为接近正态分布，而并不能一定保证就是正态分布。

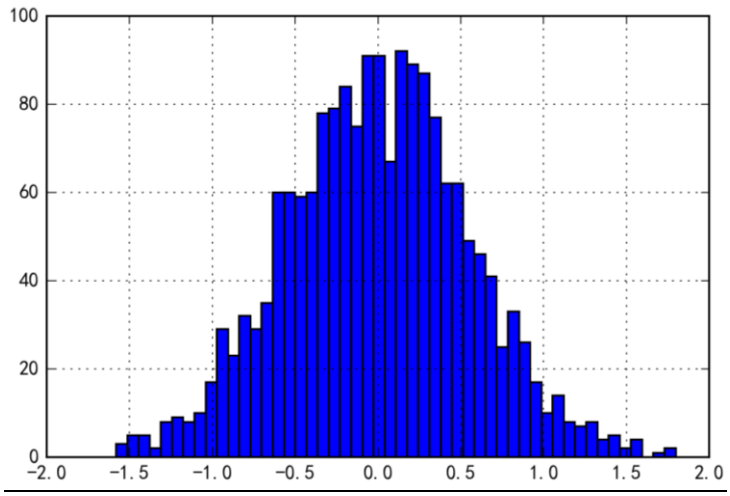
图表 18-21 展示了原始特征因子 pct_chg 统计分布和 Quantile-Quantile Plot（QQ 图），以及用修正 Box-Cox 法转换得到的该因子的统计分布和 QQ 图。我们发现如果λ=0.075，在经过转换之后该因子更接近于正态分布，再用 Shapiro-Wilk 法(W 检验)和 D′ Agostino 法(D 检验)分别得到的 p 值为 0.66 和 0.74，均无法拒绝是正态分布的空假设。

图表 18：特征因子 pct_chg 的统计分布



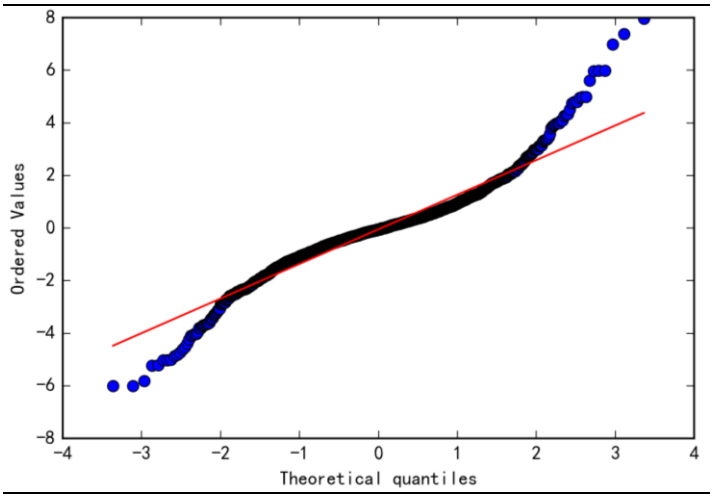
资料来源：Wind，东证衍生品研究院

图表 19：特征因子 pct_chg 转换之后的统计分布



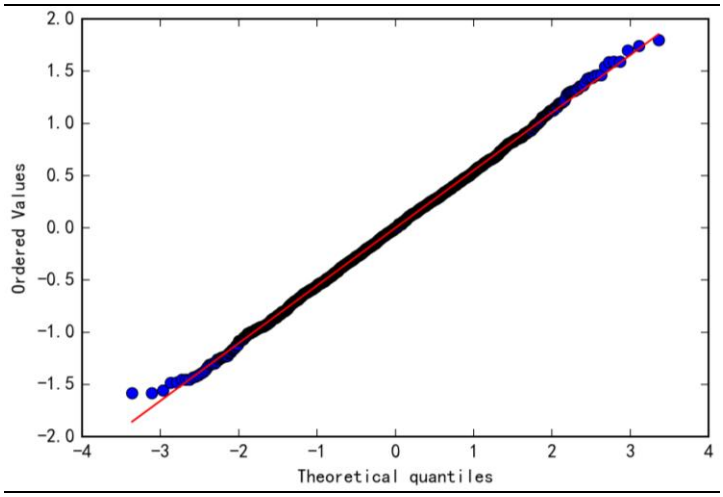
资料来源：Wind，东证衍生品研究院

图表 20：原始特征因子 pct_chg 的 QQ 图



资料来源：Wind，东证衍生品研究院

图表 21：转换后的特征因子 pct_chg 的 QQ 图



资料来源：Wind，东证衍生品研究院

图表 22：特征因子 pct_chg 在进行转换前后回测净值对比

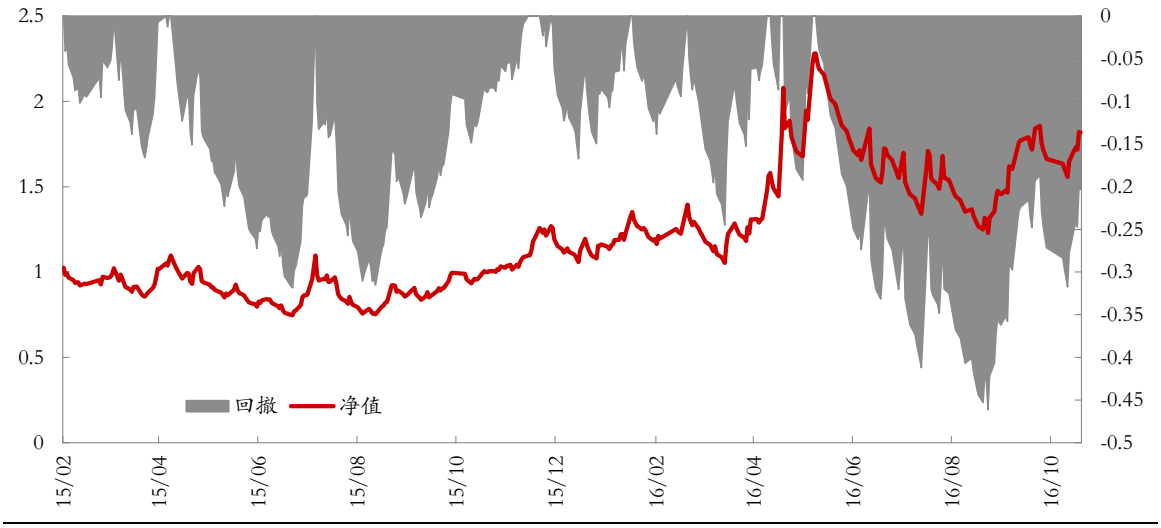
资料来源：Wind，东证衍生品研究院

而在我们使用前文中提到的“窗口数据”和“平均预期”的预测方法组合，对转换前后的特征因子 pct_chg 进行了回测，实际结果如图 22 所示，经过转换的数据并没有获得更好的回测结果。该结论说明了，在 HMM 模型进行预测时如果采用高斯分布的观测值数据，实际也并不能对预测起到更好的作用。我们认为原因可能是因为，尽管满足了输入数据应具有高斯分布特性的假设，但是却仍然无法改变“观测独立性”这个假设，因为即使做了修正 Box-Cox 法的转换，却没有改变观测序列的时间相关性，而甚至使得原始数据被抹去了一些重要的信息，致使回测结果不如转换前。但是净值走势比较相近，又说明了这种无法轻易被去除的时间相关性才是对于预测结果的影响最大的因素。因此，我们认为特征因子是否具有高斯分布对于预测结果的影响不大，我们在后面的回测中均只采用因子原始数值。

3.4、样本外回测

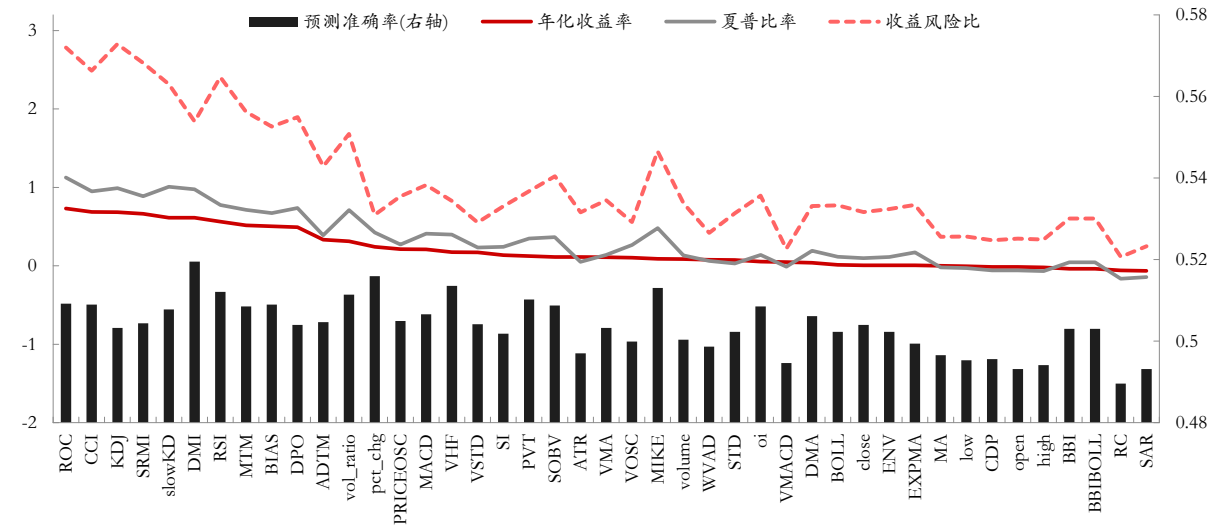
我们之前的回测都是基于样本内数据进行的，实际应该拿样本内训练好的模型，对样本外的数据集进行回测，以此来检验 HMM 是否在预测涨跌方面的有效性。我们在这里对特征因子 pct_chg 的回测依然是基于隐状态数为 6 的样本内训练之后得到的 HMM，并且也采取“窗口数据”和“平均预期”的预测方式组合进行预测。从图 23 的回测净值曲线可以看出，对于单个特征因子 pct_chg 相比于在样本内的回测结果，样本外的回测效果会稍微差一些。在 3 倍杠杆下，从 2015 年初到 2016 年 10 底，获得了 43.5%的年化收益率，夏普比率 0.70，收益最大回撤比 0.94，而预测准确率为 52.0%。

图表 23：特征因子 pct_chg 的样本外回测



资料来源：Wind，东证衍生品研究院

图表 24：不同特征因子样本外回测得到平均预测结果



资料来源：Wind，东证衍生品研究院

为了对比不同因子在样本区间上面的预测有效性，我们将原始的总样本数据按照长度等分为 6 个片段，随机抽取其中任意两个不同片段，将时间轴较早的作为样本内数据集用以训练，而将时间轴较晚的则作为样本外数据集用以预测，共计 15 种组合。然后我们对 43 个特征因子均在这 15 个组合上进行了样本内训练和样本外的预测，并得到了其在样本外数据上面的平均预测结果，如图表 24 和 25 所示。

图表 25：不同特征因子样本外回测得到平均预测结果

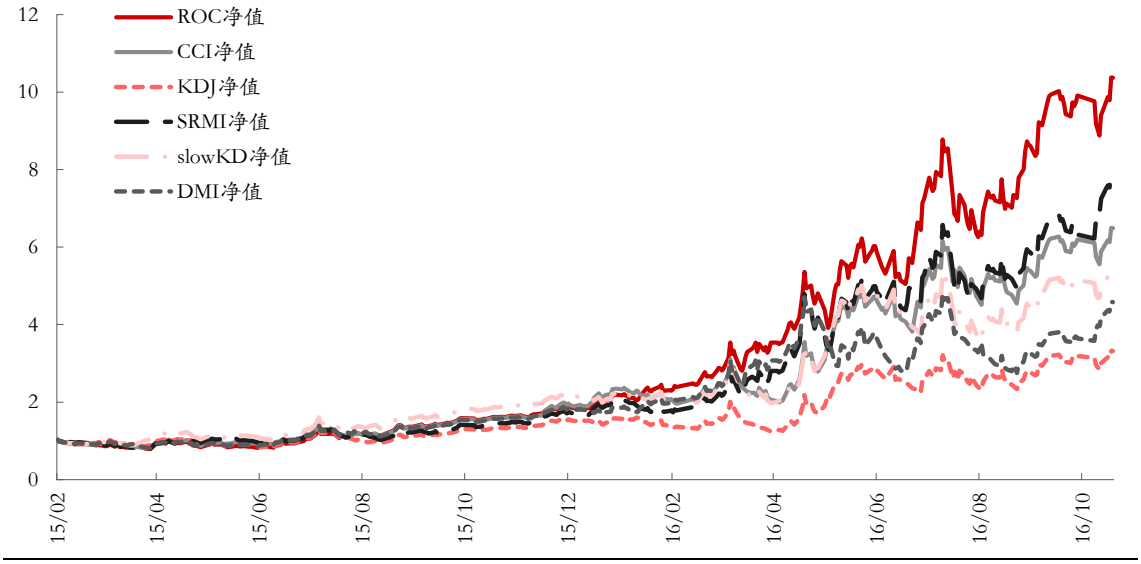
特征因子	年化收益率	夏普比率	收益最大回撤比	预测准确率	特征因子	年化收益率	夏普比率	收益最大回撤比	预测准确率
open	-1.38%	-0.058	0.346	49.32%	MA	0.22%	-0.019	0.370	49.65%
high	-2.02%	-0.066	0.336	49.41%	MACD	20.99%	0.410	1.030	50.66%
low	-0.50%	-0.029	0.372	49.53%	MIKE	9.02%	0.481	1.461	51.31%
close	0.78%	0.097	0.686	50.40%	MTM	51.74%	0.715	1.965	50.85%
volume	8.56%	0.134	0.799	50.03%	PRICEOSC	21.33%	0.271	0.887	50.49%
oi	5.37%	0.139	0.895	50.85%	PVT	12.48%	0.348	0.950	51.02%
pct_chg	24.36%	0.424	0.649	51.60%	RC	-5.71%	-0.165	0.115	48.96%
ADTM	33.28%	0.389	1.269	50.47%	ROC	73.02%	1.127	2.782	50.92%
ATR	11.29%	0.052	0.683	49.70%	RSI	56.19%	0.775	2.406	51.21%
BBI	-3.86%	0.045	0.604	50.30%	SAR	-6.34%	-0.142	0.249	49.32%
BBIBOLL	-3.86%	0.045	0.604	50.30%	SI	13.77%	0.243	0.761	50.18%
BIAS	50.55%	0.672	1.775	50.90%	SOBV	11.40%	0.367	1.141	50.88%
BOLL	1.39%	0.116	0.772	50.23%	SRMI	66.34%	0.888	2.586	50.44%
CCI	68.84%	0.948	2.487	50.90%	STD	7.38%	0.029	0.670	50.23%
CDP	-1.30%	-0.058	0.327	49.56%	VHF	17.54%	0.398	0.828	51.36%
DMA	3.90%	0.192	0.761	50.61%	VMA	10.98%	0.140	0.835	50.32%
DMI	61.26%	0.976	1.841	51.95%	VMACD	4.78%	-0.010	0.217	49.46%
DPO	49.13%	0.738	1.895	50.40%	VOSC	10.40%	0.266	0.560	49.99%
ENV	0.70%	0.112	0.724	50.23%	VSTD	17.13%	0.234	0.553	50.42%
EXPMA	0.61%	0.172	0.776	49.94%	WVAD	7.71%	0.064	0.420	49.87%
KDJ	68.33%	0.991	2.827	50.32%	vol_ratio	31.32%	0.711	1.682	51.14%
slowKD	61.27%	1.008	2.314	50.78%					

资料来源：Wind，东证衍生品研究院

整体来看，尽管不同特征因子的预测准确率都在 50%左右，但如果我们统计预测值（涨和跌）的游程的话，其实会发现 HMM 所给出的预测序列的游程比随机出来的涨跌序列的游程要长，也从而反映出尽管预测准确率接近 50%但是其并不是严格随机的。此外，如果按照样本外回测结果中的年化收益率进行排名的话，我们发现排名较靠前的特征因子的预测准确率稍比排名靠后的因子高，夏普比率、收益最大回撤比这些直接和预测结果相关的指标也同样如此。在不同样本外数据上面的平均预测表现较好的特征因子有，ROC、CCI、KDJ、SRMI、slowKD、DMI 等能够得到年化收益率较高、夏普接近于 1 或以上且收益最大回撤比达到 1.5 以上。比较有意思的是，这 6 种特征因子无一例外均

是反趋势类的指标，通过不同的方法对市场中的超买、超卖状态进行刻画描述。这个结论可能与我们日常所理解的结果相悖，因为对于大部分的策略来说，都是以趋势追踪为主，而且长期来看期货市场中价格也往往表现出明显的趋势特性。但是我们也分析了 HMM 模型为何能够筛选出反趋势指标表现趋势，该策略大部分能够赚钱的时间点基本上都是靠这种反转指标提前做出反向的预测操作而赚钱的，所以胜率一般不高。

图表 26：6 个平均表现最好的特征因子样本外回测净值表现



资料来源：Wind，东证衍生品研究院

图表 27：6 个平均表现最好的特征因子样本外回测指标统计

指标	ROC	CCI	KDJ	SRMI	slowKD	DMI
交易日日数	417	417	417	417	417	417
交易次数	54	92	104	85	85	55
累计收益率	936.86%	548.94%	231.81%	695.47%	449.79%	357.97%
年化收益率	310.98%	209.62%	106.43%	250.16%	180.10%	150.82%
年化波动率	63.31%	64.30%	61.37%	65.65%	62.72%	61.79%
夏普比率	4.87	3.22	1.70	3.77	2.83	2.40
最大回撤率	-28.73%	-35.87%	-38.17%	-30.33%	-35.08%	-42.76%
收益最大回撤比	10.83	5.84	2.79	8.25	5.13	3.53
获胜次数	14	19	18	16	19	7
失败次数	40	73	86	69	66	48
胜率	25.93%	20.65%	17.31%	18.82%	22.35%	12.73%
赔率	10.35	8.14	7.20	9.90	7.31	14.88
预测准确率	54.92%	54.20%	51.80%	54.92%	54.68%	54.68%

资料来源：Wind，东证衍生品研究院

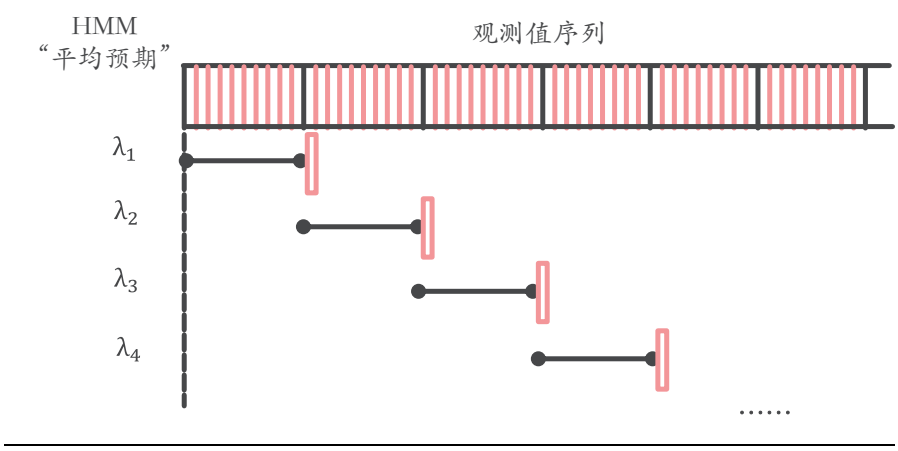
为了和图表 23 所示的特征因子 pct_chg 的样本外回测结果进行对比，我们将这 6 个平均表现最好的特征因子也在 2015 年之后的数据集上进行样本外回测（2015 年之前的数据

集用来进行训练)，回测方式依然是 3 倍杠杆，每笔交易如果相比于开盘价亏损 0.5%则退出平仓。图表 27 展示了这 6 个特征因子的回测指标统计结果，均能够取得很好的效果。

3.5、窗口滚动式回测

尽管我们进行了样本内、样本外的回测，但是毕竟市场是无时不在变化着的，因此我们又再次进行了窗口滚动式的回测方法。这个方法与前文提到的“窗口数据”是不一样的，那里虽然使用的是窗口截断数据，但是 HMM 模型却依然是基于样本内数据集得到的，而且在样本外预测时该模型也不会发生改变。而实际情况应该是，所被训练的 HMM 也应该是时时被更新的，用最近一段时间内的数据训练得到新的 HMM 和它最优的参数组合 λ，并以最新模型和最新的数据对下一天进行预测，该过程可以用如图表 28 所示的示意流程图来简单表示，对次日的预测我们仍继续采用前文中提及的“平均预期”的方法，也就是利用各个隐状态的发生概率作为权重，与各个隐状态在这一段时间上的未来收益率相乘，得到对下一天收益率的平均预期。

图表 28：窗口滚动式回测流程示意图

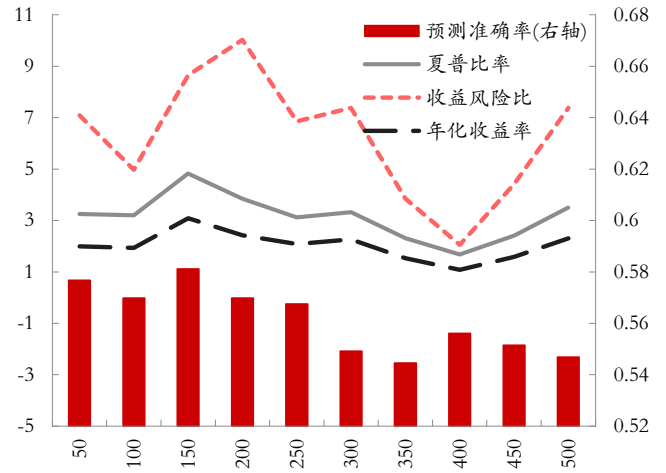


资料来源：东证衍生品研究院

问题 1：滚动窗口的长度的影响如何?

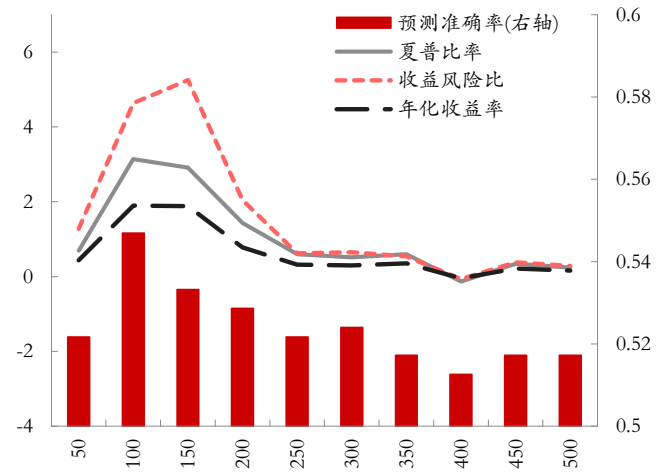
我们分别对前文中得到的表现很好的特征因子之一 ROC 和一般的特征因子 pct_chg 进行了窗口滚动式的回测，他们的回测结果如图表 29 和 30 所示。从结果中可以看到，如果窗口长度选择在 150 天左右是最合适的。图表 31 则展示了这两个特征因子在滚动窗口长度为 150 天时，滚动预测得到的结果，采用的隐状态数为 6，回测方式仍然是 3 倍杠杆。ROC 因子预测准确率达到 58%，收益最大回撤比为 8.65，而 pct_chg 因子的预测准确率则为 53.3%，收益最大回撤比为 5.25。

图表 29：ROC 的滚动预测结果随窗口长度的变化



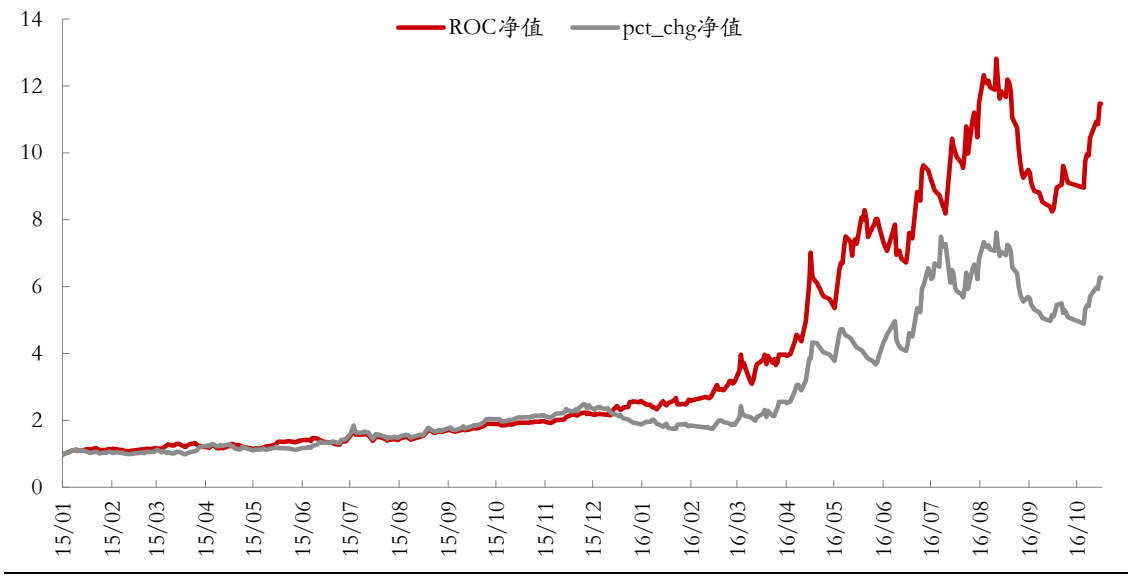
资料来源：Wind，东证衍生品研究院

图表 30：pct_chg 的滚动预测结果随窗口长度的变化



资料来源：Wind，东证衍生品研究院

图表 31：pct_chg 的滚动预测结果随窗口长度的变化



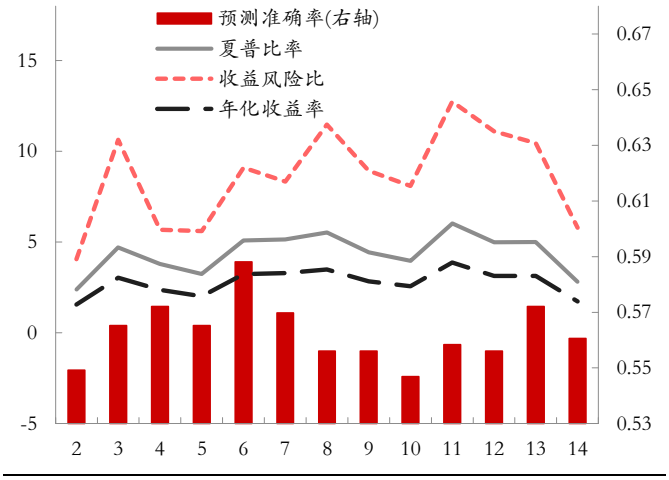
资料来源：Wind，东证衍生品研究院

问题 2：滚动窗口预测时最优的隐状态数是多少？

当然还有一点需要考虑的是，HMM 中隐状态数的大小对这种窗口滚动式的回测结果的影响。我们在样本内回测的时候曾得出了隐状态数目越高，结果预测准确率就会越高，但是那是对于一段固定的样本内数据而言。实际上，HMM 模型在进行预测时也应时时被更新，那么此时隐状态大小对整体的预测效果就不一定仍满足上面的结论了。为此，我们针对特征因子 ROC 和 pct_chg 分别画出了他们在隐状态数发生变化时的预测结果的变化情况，这里采用的窗口长度是问题 1 中得到的最优长度 150 天，并且仍然采取 3 倍杠杆进行回测，结果如图表 32 和 33 所示。结果显示，特征因子 ROC 和 pct_chg 均有一些特有的隐状态数使用窗口滚动这种方式得到的回测结果表现较好。但是，由于这种变化具有非规律性，且基本上在隐状态数取 6 时，特征因子的回测表现都不错，因此我们

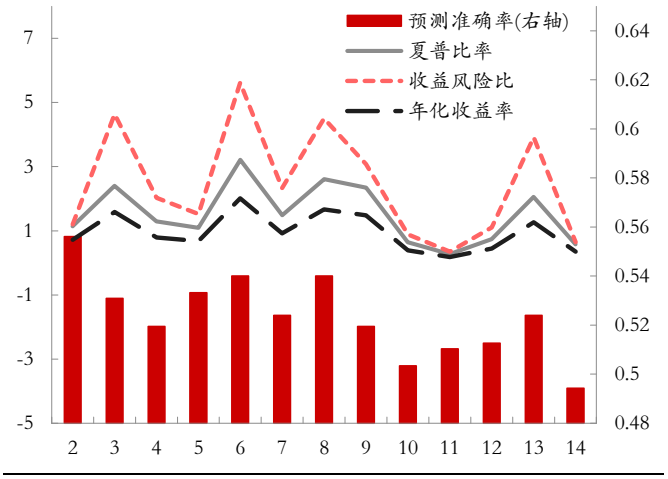
在之后的窗口滚动回测中也都确定隐状态数为 6。

图表 32：ROC 的滚动预测结果随隐状态数的变化



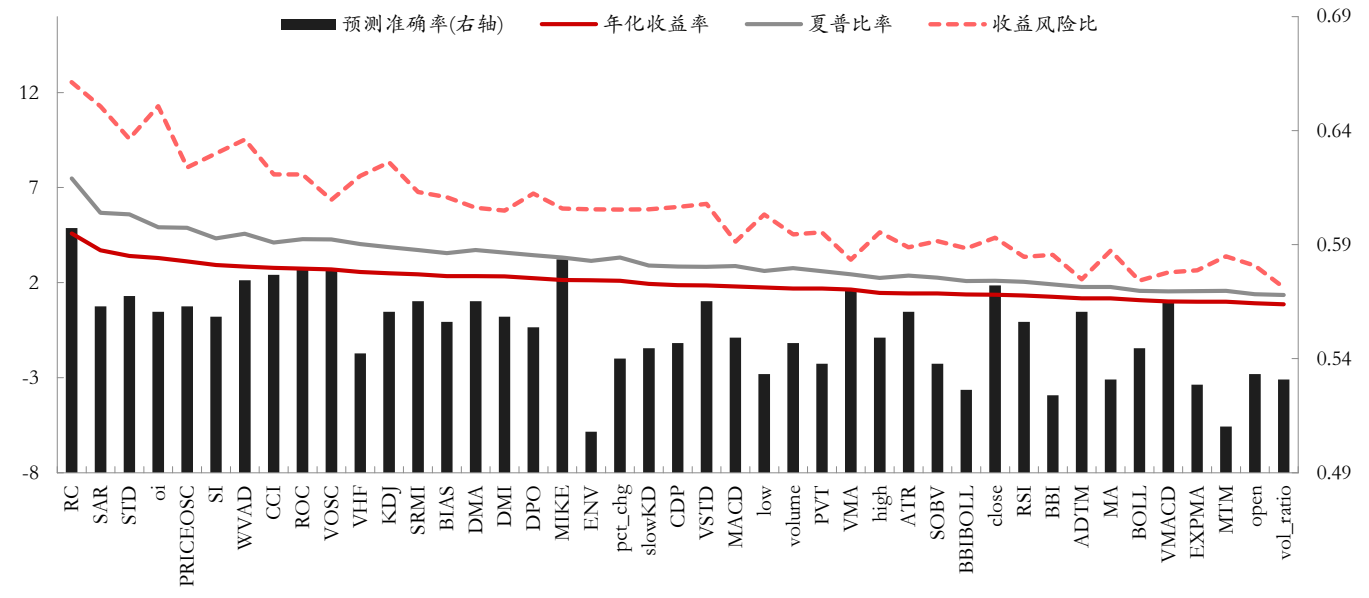
资料来源：Wind，东证衍生品研究院

图表 33：pct_chg 的滚动预测结果随隐状态数的变化



资料来源：Wind，东证衍生品研究院

图表 34：不同特征因子窗口滚动式回测得到预测结果



资料来源：Wind，东证衍生品研究院

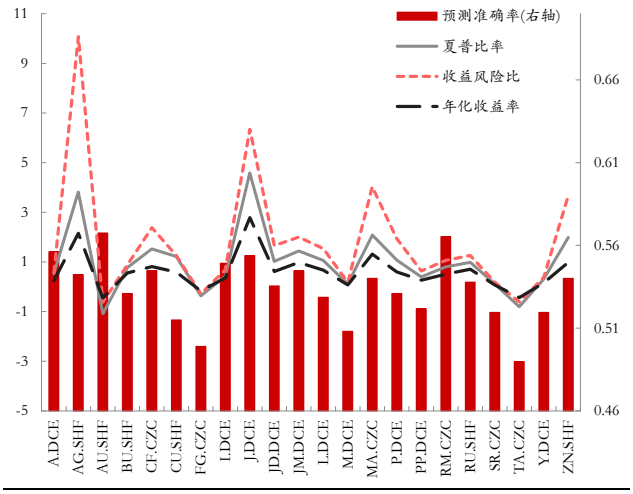
问题 3：各个特征因子在窗口滚动回测时的表现如何？

根据前两个问题，我们在讨论其他各个特征因子的窗口滚动回测结果时，首先确定了最优的移动窗口长度为 150 天，而隐状态数都统一为 6 个。我们在此基础上，再次对 43 个不同的特征因子也都进行了滚动窗口的回测，各因子的回测结果按照年化收益率排序如图 34 所示。我们发现回测表现较好的几个因子是 RC、SAR、STD、oi、PRICEOSC、SI 等，有意思的是除了 oi，其他这些指标大都反映了价格变动速率、波动率等情况，事实上也是一种反趋势的震荡型指标。排名在这些因子之后，则是前文中在样本外回测平

均表现很好的那一些特征因子。但是排名在靠后的特征因子则大都是一些反映价格趋势的因子，比如 MA、MACD、MTM 等。

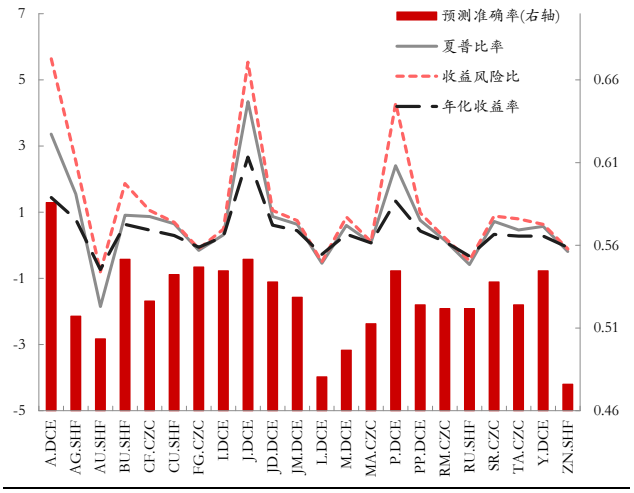
4、其他品种回测结果

图表 35：不同品种 RC 的滚动预测结果



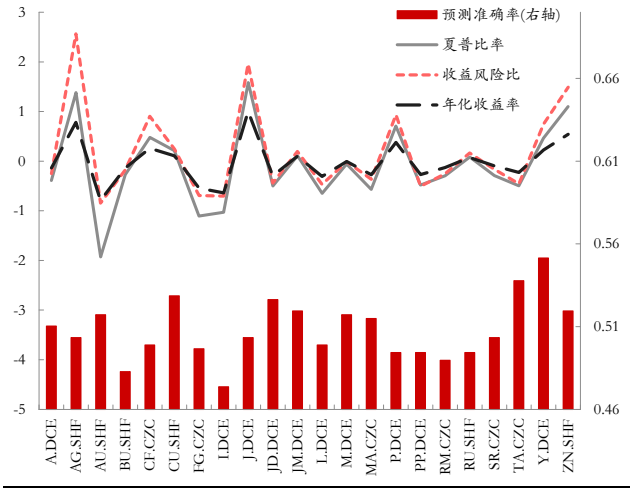
资料来源：Wind，东证衍生品研究院

图表 36：不同品种 ROC 的滚动预测结果



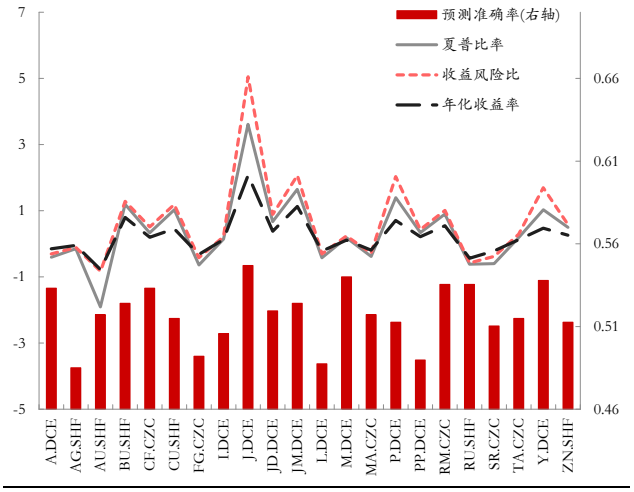
资料来源：Wind，东证衍生品研究院

图表 37：不同品种 pct_chg 的滚动预测结果



资料来源：Wind，东证衍生品研究院

图表 38：不同品种 vol_ratio 的滚动预测结果



资料来源：Wind，东证衍生品研究院

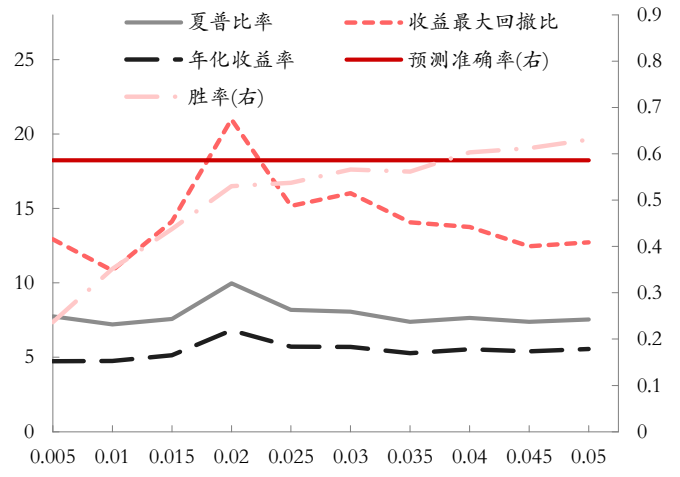
前文所有的分析都是基于螺纹钢这个目前在国内市场成交最为活跃的品种，为了探讨用 HMM 进行择时预测的品种适用性，我们也挑选了一些其他的成交量较为活跃、且上市在 2015 年之前的共计 22 个期货品种，并也采用如前文所介绍的更为接近实际情况的窗口混动式的回测方法对这 22 个品种的主力合约进行了回测，结果如图表 35-38 所示。我们分别用图表 34 中表现各异的四个特征因子，RC、ROC、pct_chg 和 vol_ratio 来对其他 22 个品种进行窗口滚动式的预测。从结果上看，表现较好的品种有焦炭（J.DCE）、棕榈油（P.DCE）等品种。

5、关于回测设置问题的讨论

因为前文中我们曾假定了回测框架的基本规则设置是固定不变的，止损点位设置为累计浮亏 0.5%，手续费比例为 0.0001，滑点为 0，而杠杆比例则为 3。但是对于实际情况来说，固定的回测条件可能稍显简单且过于单一，无法应对复杂多变的市场环境。因此，我们在给出了相同的预测序列，通过更改回测条件，考察不同的回测设置规则（主要是止损比例、手续费比例、滑点数和杠杆比例）对回测结果产生的直接影响。这里我们仍然以螺纹钢为例，对其主力合约分别用 ROC、RC、pct_chg、vol_ratio 这 4 个在图表 34 中表现有好有坏的特征因子来进行窗口滚动式回测。

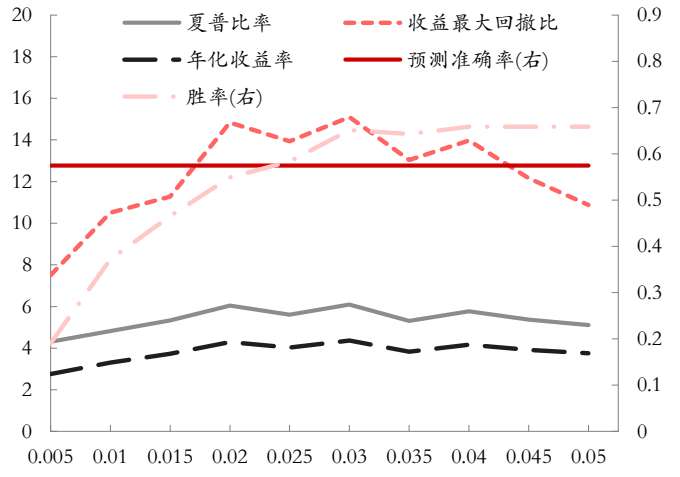
1. 止损比例

图表 39：RC 的滚动预测结果随止损比例的变化



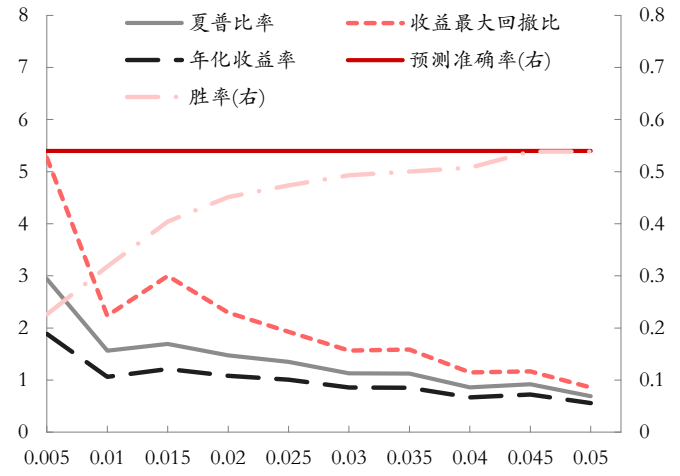
资料来源：Wind，东证衍生品研究院

图表 40：ROC 的滚动预测结果随止损比例的变化



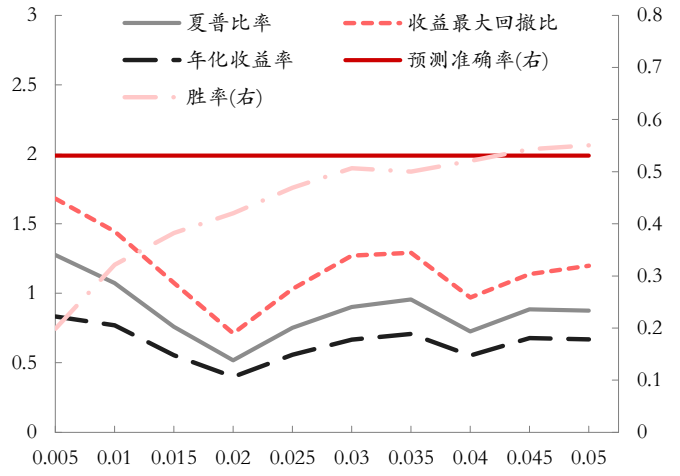
资料来源：Wind，东证衍生品研究院

图表 41：pct_chg 的滚动预测结果随止损比例的变化



资料来源：Wind，东证衍生品研究院

图表 42：vol_ratio 的滚动预测结果随止损比例的变化



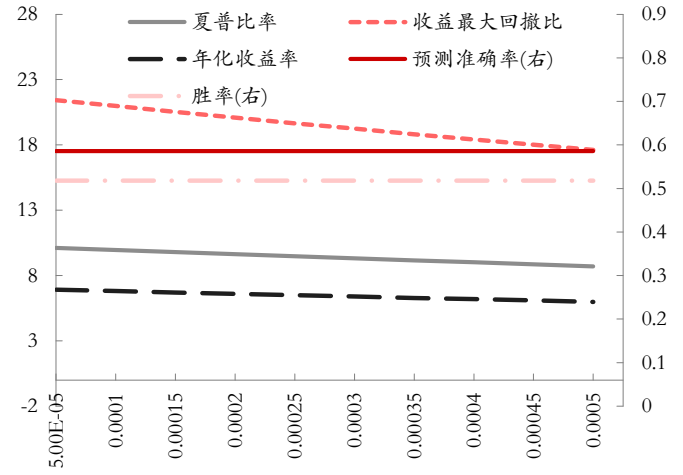
资料来源：Wind，东证衍生品研究院

从图表 39-42（3 倍杠杆回测）可以看出，回测中止损线位置的设置对回测结果的影响差异巨大。对于不同特征因子得到的滚动回测结果，设置不同的触发止损命令的条件，回测结果的统计指标表现不一。当然由于是同一个因子得到的同一个回测结果，因此各图

中的预测准确率都是一条直线（直线位置因使用的特征因子不同而不同）。但是胜率（收益为正的次数除以总交易次数）则都会随着止损线的“宽松”，而逐渐升高，这说明了前文中得到的低胜率、高赔率的回测结果，往往是因为止损次数较多而导致的。但是，并不是放松了止损触发的容易程度就一定会导致收益不好或者好，从图表 39 和 42 可以看到同样的 2%的止损位，RC 因子得到的回测结果较好，而 vol_ratio 因子得到的结果反而很差，所以关于止损位的设置因人而异，需要我们根据选取的特征因子的具体情况而定。

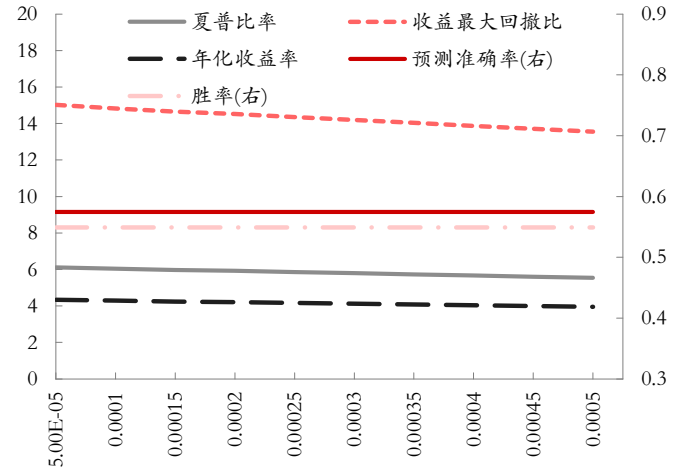
2. 手续费比例

图表 43：RC 的滚动预测结果随手续费比例的变化



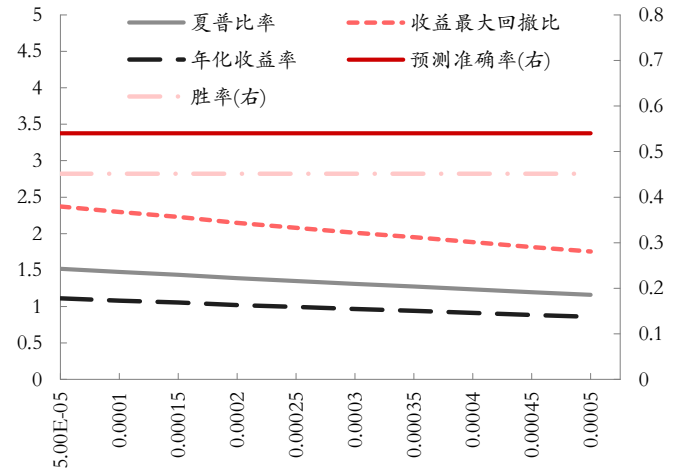
资料来源：Wind，东证衍生品研究院

图表 44：ROC 的滚动预测结果随手续费比例的变化



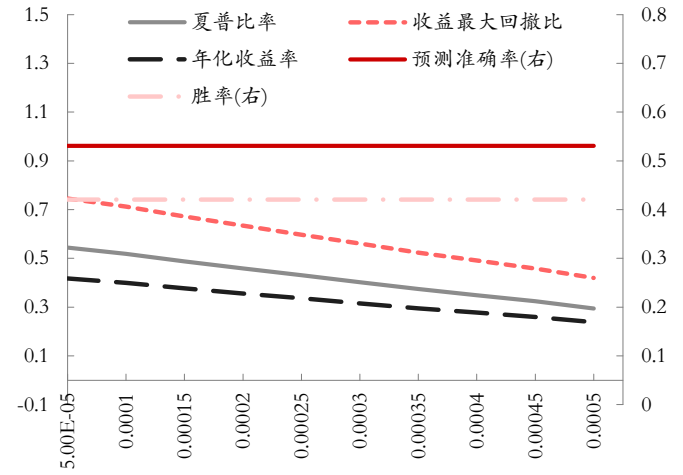
资料来源：Wind，东证衍生品研究院

图表 45：pct_chg 的滚动预测结果随手续费比例的变化



资料来源：Wind，东证衍生品研究院

图表 46：vol_ratio 的滚动预测结果随手续费比例的变化



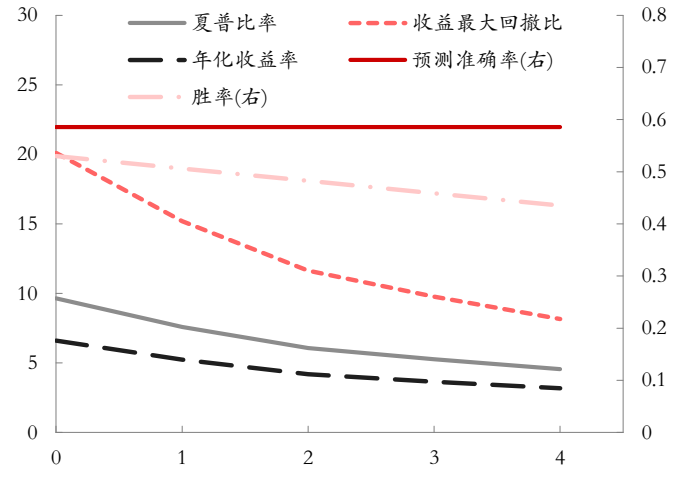
资料来源：Wind，东证衍生品研究院

手续费比例大小的设置对回测结果的影响其实正是如大家所预知的一样，手续费越高，回测收益越低，但是这种影响所导致的变化程度并不高，从图表 43-46（3 倍杠杆回测）也可以看出年化收益率、收益最大回撤比和夏普比率都是下降地比较缓慢。而对于胜率

来说，由于并不影响开仓与平仓的时点，所以交易次数及获胜次数也会基本保持不变，因此胜率在图表也就保持为一条直线。

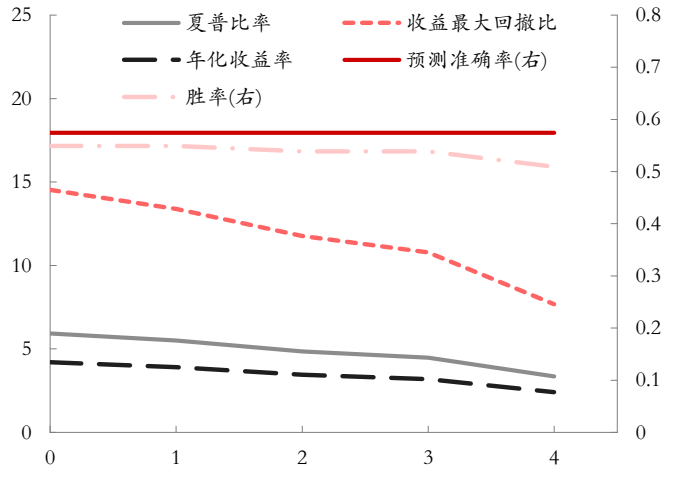
3. 滑点数

图表 47：RC 的滚动预测结果随滑点数的变化



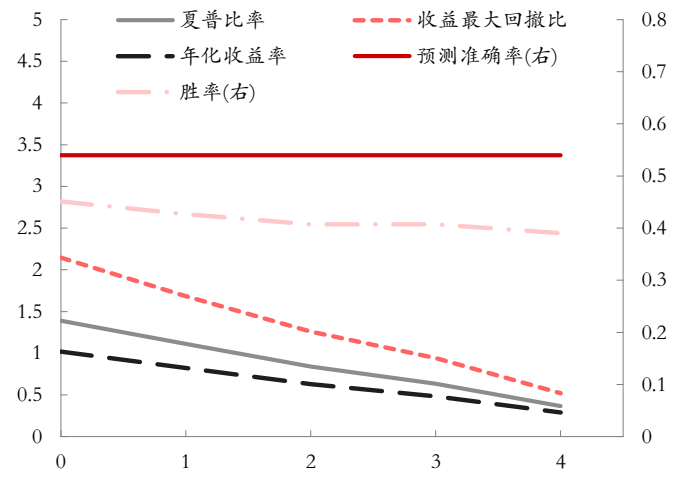
资料来源：Wind，东证衍生品研究院

图表 48：ROC 的滚动预测结果随滑点数的变化



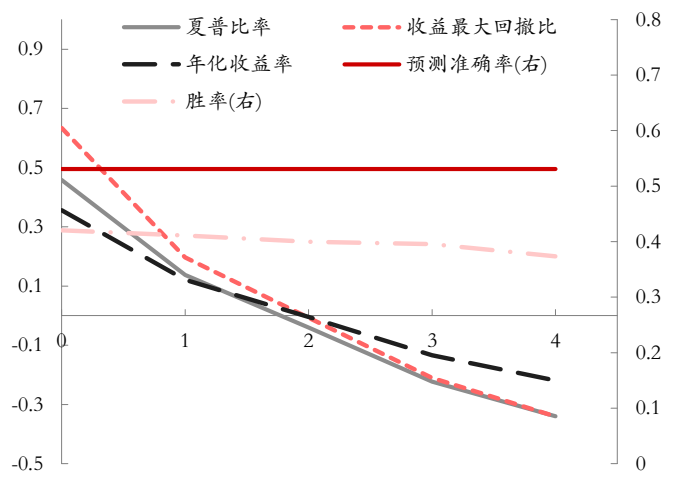
资料来源：Wind，东证衍生品研究院

图表 49：pct_chg 的滚动预测结果随滑点数的变化



资料来源：Wind，东证衍生品研究院

图表 50：vol_ratio 的滚动预测结果随滑点数的变化

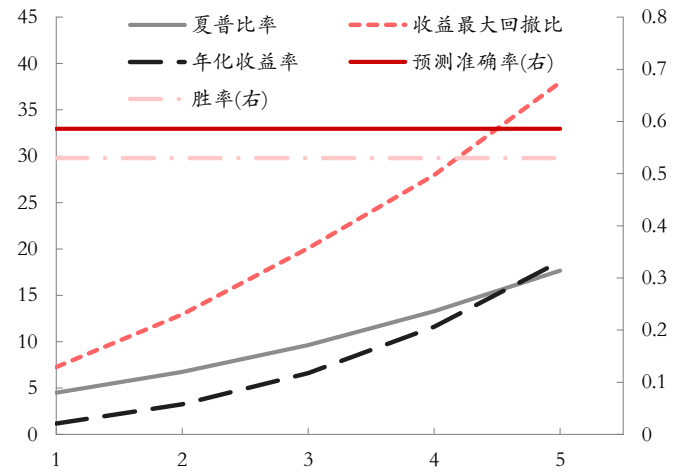


资料来源：Wind，东证衍生品研究院

我们之前的回测均没有设置滑点，但是对于现实情况来说，更受关注的应该是滑点数的影响，这可以直接反应策略在实施时的攫取套利机会的有效程度。滑点数其实是一种市场冲击成本的反映，对于交易不够活跃的市场来说，滑点数应相应增加，相反，滑点数则较小。我们从图表 47-50（3 倍杠杆回测）可以看到，随着滑点数增加，年化收益率、夏普比率和收益最大回撤比均较明显地快速下降。

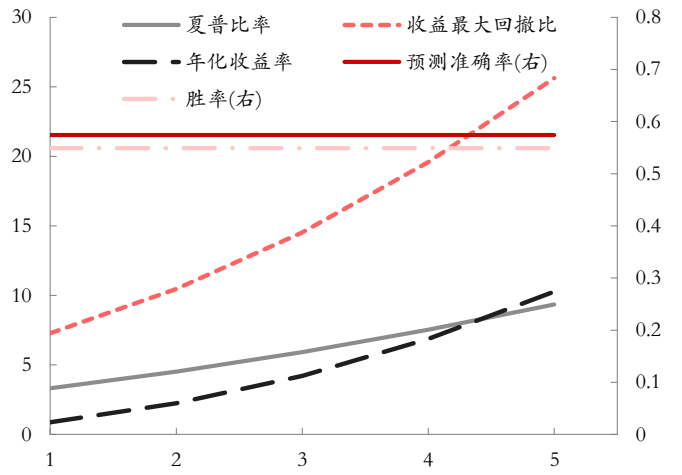
4. 杠杆比例

图表 51：RC 的滚动预测结果随杠杆比例的变化



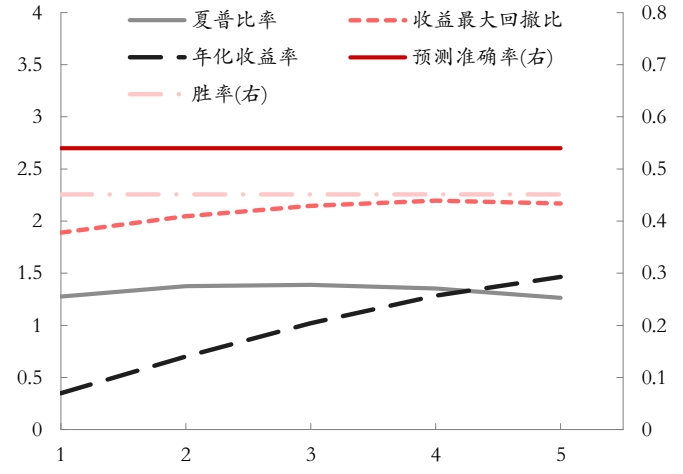
资料来源：Wind，东证衍生品研究院

图表 52：ROC 的滚动预测结果随杠杆比例的变化



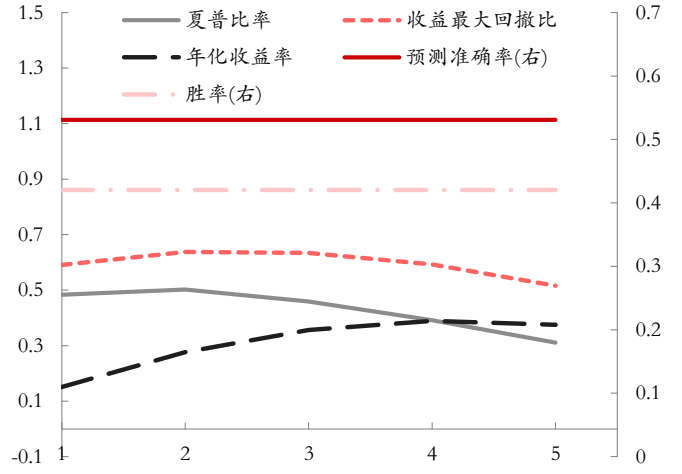
资料来源：Wind，东证衍生品研究院

图表 53：pct_chg 的滚动预测结果随杠杆比例的变化



资料来源：Wind，东证衍生品研究院

图表 54：vol_ratio 的滚动预测结果随杠杆比例的变化



资料来源：Wind，东证衍生品研究院

最后我们看一下杠杆比例的影响，而杠杆比例其实只是反映了对收益和风险的放大倍数，它的改变也并不会影响交易的时点，因此从图表 51-54（有止损，止损比例 2%）中可以看到胜率也是维持不变的。但是对于表现相对优异的两个特征因子 RC 和 ROC 来说，它们的滚动预测结果的收益表现会随着杠杆比例增加而变得更好，对于 pct_chg 和 vol_ratio 这两个排名较靠后的因子，它们的收益表现则在一定范围内能随着杠杆比例增加而改善，但是超过该范围，可能会因风险（最大回撤）的过度放大而降低收益表现。

总的来说，这里我们探讨了一些回测规则的影响，由于我们的回测框架是统一的，比如要求产生信号的次日开盘进行交易、允许做多和做空、允许反转交易等，但是对于更细节的问题，比如具体应该设置怎么样的止损比例、手续费、滑点数和杠杆比例，其实还是应和策略实施时的具体情况相匹配，以能够获得更多更稳定的收益。

6、总结及展望

从尤金·法玛提出了有效市场假说之后，曾有人认为这个假说或许就应该是一条普适的“真理”，事实上和该假说等价的无套利原理也成了大多定价理论的基础，这就好比物理学家从看似杂乱无规律的观测数据中发现了一则决定了内在机制的定律一样。但是，真实市场确实有效，但不是时时有效也并非处处有效。我们认为市场有效性的体现，好比大自然的自我恢复能力的体现，在一定阈值内并且无外力干扰下，市场总会来回震荡式地回归地其理性的位置。但是就像物理学中所说的非平衡态一样，该市场和大自然的回复过程具有一定的弛豫时间，那么也就是说市场总会在其出现在非平衡位置的时候具备可套利的机会，尽管我们也知道这种套利机会一旦被发现和利用了，便更能促使它逐渐磨灭并消失。既然如此，一旦我们具备了足够的能力以观察到这种非平衡的状态，不管是低频还是高频。因此，也不难理解为何在 Alpha Go 战胜李世石之后有人遐想把 Alpha Go 用在金融市场中，但说到底其实也就是机器学习在金融市场中的应用。

作为机器学习中模式识别方法中的一种重要模型，隐马尔科夫模型曾被传闻是詹姆斯·西蒙斯的量化基本最重要的赚钱工具之一。我们在这篇报告中就着重讨论了这个被市场广议的隐马尔科夫模型是怎样的，如何把它应用到期货市场当中来，以及如何更好利用它进行择时预测以获取到更大的收益。

隐马尔科夫模型其实是一种模式识别能力很强的方法，我们尽管不知道隐状态具体指代什么样的状态，但是我们却可以通过对未来收益走势的统计得到当前隐状态对涨和跌的判断。从结果上来看，隐马尔科夫模型对于样本内的隐状态的标记准确率很高，样本内的回测结果也都比较好。但是样本外的结果就大不如样本内的结果好，预测准确率基本也都在 50%左右。这种近 50%的预测准确率看似随机，但实际并不是具有较强的随机性，因为我们得到的预测序列的游程要比普通随机序列的游程更长一些。

目前很多关于 HMM 的策略便浅尝辄止在样本内的预测，对样本外的预测也或多或少有些问题。因为 HMM 进行预测的 Viterbi 算法会和输入数据的“上下文”有关，所以一次性输入全部的样本外待测数据和一段一段地输入待测数据所得到的结果是不一样的。如若使用这种“窗口数据”就意味着使用了未来函数进行预测。我们经过统计发现，这两种方式得到的预测差异率基本在 4%-10%左右，所以这也导致了实际使用 HMM 进行预测的效果更差。除此之外，我们也讨论了两种预测方式，一种是“最大概率”，另一种是“平均预期”。HMM 预测实际得到的是每个时点上各个隐状态的发生概率，HMM 实际解码得到的隐状态序列正是对应着最大发生概率的隐状态，第一种方法就是用“最大概率”对应的隐状态进行预测。而第二种方法是利用各隐状态的发生概率作为权重，再计算得到对未来收益率的“平均预期”，然后如果该平均预期收益为正就预测涨，为负就预测跌。结果我们发现“平均预期”的预测方法比“最大概率”的方法稍好一些，但区别不太大。另外，除了样本内和样本外的预测，我们也提出了一种更切实际的窗口滚动式的预测，统计发现了更为合适的窗口长度以及隐状态数。

汇总一下我们在商品市场上研究应用 HMM 时得到的结论：

1. 使用“平均预期”的预测方式比“最大概率”的预测方式得到的回测结果更好

2. 使用“窗口数据”和“非窗口数据”预测时预测差异率基本在 4%-10%左右

3. 为避免使用到未来数据，在使用“窗口数据”时，窗口的长度在超过 10 天之后，对预测结果的影响就不再变化

4. 对于固定的样本内数据集来说，状态数的增加会微微提高预测结果，但是这种提升效果不够特别显著，反而增加计算复杂度，故一般选取状态数为 6.-8 之间

5. 通过修正的 Box-Cox 转换方法使得特征因子的统计分布更接近高斯分布，但是对预测结果却并无明显改善

6. 样本外的回测结果显示，对于超买超卖类的反趋势特征因子，它们的回测结果表现更优秀

7. 在窗口滚动式的回测中，窗口的长度一般在 150 左右会使得回测结果更好一些

8. 在窗口滚动式回测中，隐状态数会因特征因子的选取而呈现明显不同，会有特有的几个值表现得更好

9. 在窗口滚动式回测中，刻画了价格变动速率或波动特性的特征因子，它们的回测结果表现更好

10. 对于其他品种的测试，焦炭、棕榈油等品种在不同特征因子回测中均表现优秀

尽管得到了这些结论，但是我们也需要注意的是该模型也确实存在它独有的缺点，我们在前文中也强调了，HMM 要求隐状态序列具有马尔科夫性质以及观测独立等条件，但是实际的金融数据具有明显的相关性，并无法保证交易对象的状态只受上一时刻决定，而且也无法保证观测数据的独立性。但如果抛去这些假设不说，总体来说我们认为 HMM 还是有它的应用意义的，但还是需要有进一步的改进或讨论。

1. 对 HMM 的训练所使用的 EM 算法，虽然能够快速收敛，但是比较容易陷入局部最优的局面。以后可改进的是，可以循环多次进行训练，观察收敛次数和收敛点是否发生很明显的变化。

2. 我们目前主要讨论的是单因子（一维的观测值序列）对预测的影响，之后可以将观测值扩展到多维，并讨论原始多维数据与降维去噪正交化的数据对预测结果的区别。

3. 由于 HMM 的解码过程不仅仅给出了各状态的发生概率，还给出了对输入数据片段解码的似然值，因此可以探讨从似然值入手，寻找和当前片段似然值较相近的历史其他片段，并以此进行预测。这样的预测方式特别类似动态时间规整方法对市场的预测，因此可以再对这两种方法进行一些对比讨论。

期货走势评级体系（以收盘价的变动幅度为判断标准）

走势评级	短期（1-3 个月）	中期（3-6 个月）	长期（6-12 个月）
强烈看涨	上涨 15%以上	上涨 15%以上	上涨 15%以上
看涨	上涨 5-15%	上涨 5-15%	上涨 5-15%
震荡	振幅-5%-+5%	振幅-5%-+5%	振幅-5%-+5%
看跌	下跌 5-15%	下跌 5-15%	下跌 5-15%
强烈看跌	下跌 15%以上	下跌 15%以上	下跌 15%以上

上海东证期货有限公司

上海东证期货有限公司（简称东证期货）是东方证券股份有限公司全资子公司，注册资本达10亿元，系国内四家期货交易所的结算会员。

东证期货专注于金融期货和商品期货的研究与服务，提供权威、及时的研发产品服务和投资策略；专注于信息技术的创新，创建安全、快捷的交易通道，开发多样化、个性化的交易系统；专注于构筑全面的风险管理和客户服务平台。

东证期货管理团队管理经验丰富，业绩出众，在业内享有盛誉。人才管理及激励机制完善，公司拥有硕士学历以上人员占比30%，具有海外证券和期货经历的高端人才占比10%。

2010年，东证期货发展迅猛，成绩斐然，成为业内进步最快、最受瞩目的期货公司之一。2011年初，东证期货荣获2010年度中国金融期货交易所年度会员金奖，同时获投资者教育奖、客户管理奖、技术管理奖和功能发挥奖等四项单项大奖；荣获上海期货交易所优胜会员第七名，铜、橡胶和燃料油三项企业服务奖；荣获大连商品交易所优秀会员第九名；东证期货研究所荣获大连商品交易所、和讯网第二届全国“十大期货研发团队”农产品团队全国第二名、化工团队全国第五名；荣获郑州商品交易所行业进步奖等。

东证期货全年无风险事故，充分体现了公司稳健经营，稳步发展的经营宗旨。

分析师承诺

李晓辉

本人具有中国期货业协会授予的期货执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接接收到任何形式的报酬。

免责声明

东证期货研究所

地 址： 上海市中山南路 318 号东方国际金融广场 2 号楼 22 楼
联系人：梁爽
电 话： 8621-63325888-1592
传 真： 8621-33315862
网 址： www.orientfutures.com
Email： research@orientfutures.com