第 8 章 基于信息的预测

【本章导读】信息预测在主动管理投资中尤为重要，一组信息的质量如何判定？信息对于投资的影响是否显著？这都是我们在获得一组与金融资产数据相关信息时需要回答的问题。优秀的信息会对投资组合的超额收益产生正向的影响，但不良信息对于投资组合的构建没有任何帮助。本章将介绍处理信息和评判信息的方法，灵活运用这些方法有助于在提取因子特征中获取优势。

8.1 引言

主动管理本质就是预测。在前面的几章中，我们已经对于预测的源头——信息做出了一系列的说明，很多基本面的波动，技术面的价格与波动率变化都可以看作信息。第五章的例子引入了对 ROE 因子的简单处理，为读者第一次介绍如何利用这些信息。

怎样有效地处理和分析这些信息，并将其转化为阿尔法，是信息处理要解决的关键问题。本章通过从原始信息中提炼精炼预测的方式，来评估单个信息和复数信息对于投资组合收益率的影响。

本章还将讲述一些提取信息的方法，以用于金融数据的时序预测，包括传统的统计模型和机器学习方法。读者会看到，当 OLS 方法不再是构建信息的主要方法后，利用新方法构建的新因子有的原远比传统因子优秀。

关于因子的讨论还将继续。本章将继续以沪深 300 为参考标的，以时间序列信息预测和投资组合收益的关系。

8.2 基本概念

对于收益率的预测可以分为三类

（1）朴素预测：就是一致预期收益率，它是无信息（或未获信息）情形下的预测。对于一个量化策略来说，朴素预测的收益率就是投资者选择的基准组合的收益率。例如，选择沪深 300 作为策略基准，朴素预测的收益率就是沪深 300 在策略频率上的风险调整收益率。

（2）原始预测：表示主动投资经理所获信息的原始形式　例如：收入预测，某种盈利估计、买卖建议等。这类原始预测变量在量级和单位上可能不同，不能直接对超常收益率预测。

（3）精炼预测：通过基本预测公式将原始预测转化为“精炼的预测”输出的预测值与超额收益率具有相同的形式（和单位），是原始预测 上经过信息调整的预测，蕴含原始预测中的有效信息。例如，经过 Z 值化处理的 ROE，根据研报数量给定的 0-1 变量等。

8.3 预测经验法则

8.3.1 预测基本公式

拥有了信息之后，我们可以在基准收益的基础上，针对原始的预测向量做出条件收益率的预测。

$$E\{\boldsymbol{r}|\boldsymbol{g}\} = E\{\boldsymbol{r}\} + Cov\{\boldsymbol{r},\boldsymbol{g}\}\cdot Var^{-1}\{\boldsymbol{g}\}\cdot(\boldsymbol{g} - E\{\boldsymbol{g}\})$$

其中$\boldsymbol{r}$：超额收益率向量（N 只资产）；

$\boldsymbol{g}$：原始预测向量（K 个预测）；

$E\{\boldsymbol{r}\}$：朴素预测；

$E\{\boldsymbol{g}\}$：原始预测向量的期望值；

$E\{\boldsymbol{r}|\boldsymbol{g}\}$：基于信息的预测；基于信息$g$的条件预期超额收益率；；

$Cov\{\boldsymbol{r},\boldsymbol{g}\}$：表示$\boldsymbol{r}$和$\boldsymbol{g}$的协方差矩阵；

$Var\{\boldsymbol{g}\}$：表示$\boldsymbol{g}$的方差；

可证明该公式是具有最小均方误差的线性无偏估计，即最优线性无偏估计。

更为广义的，假定朴素预测产生的基准收益率为$\boldsymbol{b}$，利用的原始预测向量为$\boldsymbol{g}$，针对原始预测向量的处理矩阵为 A，一般线性估计可以写为：

$$\boldsymbol{r}(\boldsymbol{g};\boldsymbol{b},\boldsymbol{A}) = \boldsymbol{b} + \boldsymbol{A}\cdot\boldsymbol{g}$$

估计误差为：

$$\boldsymbol{q} = \boldsymbol{r} - \boldsymbol{r}(\boldsymbol{g};\boldsymbol{b},\boldsymbol{A})$$

均方误差为：

$$\mathrm{MSE}\{\boldsymbol{b},\boldsymbol{A}\} = \mathrm{E}\{\boldsymbol{q}^T\cdot\boldsymbol{q}\} = \mathrm{E}\{\sum_n q_n^2\}$$

其中，关于$E\{\boldsymbol{r}\}$（朴素预测）的确定方法是，给定基准组合 B，那么朴素（一致）预测为：

$$E\{\boldsymbol{r}\} = \beta\mu_B$$

β是关于业绩基准组合的贝塔；　μ_B是业绩基准的一致预期超额收益率

在明确了朴素预测的确定方法后，基于基本预测公式，我们可以构建某个信息对于预期收益率产生的影响。由于信息导致的预期收益率变动是预测收益于基准收益之差。

$$\Phi = E\{\boldsymbol{r}|\boldsymbol{g}\} - E\{\boldsymbol{r}\} = Cov\{\boldsymbol{r},\boldsymbol{g}\}\cdot Var^{-1}\{\boldsymbol{g}\}\cdot(\boldsymbol{g} - E\{\boldsymbol{g}\})$$

也称为精炼预测(refined forecast)　即为预期的超常收益率。

进一步，由观察信息所引起的对一致预期收益的调整可以表示为：

$$\begin{aligned}&\Phi = E\{\boldsymbol{r}|\boldsymbol{g}\} - E\{\boldsymbol{r}\}\\ &= Cov\{\boldsymbol{r},\boldsymbol{g}\}\cdot Var^{-1}\{\boldsymbol{g}\}\cdot(\boldsymbol{g} - E\{\boldsymbol{g}\})\\ &= \mathrm{Std\{r\}}\cdot\frac{Cov\{\boldsymbol{r},\boldsymbol{g}\}}{\mathrm{Std\{r\}Std\{g\}}}\cdot\frac{(\boldsymbol{g} - E\{\boldsymbol{g}\})}{\mathrm{Std\{g\}}}\\ &= \mathrm{Std\{r\}}\cdot IC\cdot Scores\end{aligned}$$

$\mathrm{Std\{r\}}$表示超额收益率的标准差，即波动率，$Scores = \frac{(\boldsymbol{g}-E\{\boldsymbol{g}\})}{\mathrm{Std\{g\}}}$是信息的标准化得分:减去期望后除以标准差,$IC$是收益率与信息的相关系数，即信息系数，反映信息 g 的预测能力。

针对所得的推导，我们可以将观察信息所引起的对一致预期收益的调整视为充分利用信息的精炼预测，这个预测是利用信息产生的阿尔法之和：

$$\textit{精炼预测} = \textit{波动率}\cdot IC\cdot\textit{标准分值}$$

8.3.2 单信息源分析

8.3.2.1 二元模型分析

对于单个信息源，假定一个运用场景，我们希望在获知了某个资产收益率期望和波动率的前提下，通过单个信息源预测该资产下一个季度的收益率。在二元分析中，我们假设：

（1）已知目标资产的收益率是怎样产生的（在本例中，通过多个因子的线性组合产生）；

（2）已知目标资产的预测变量是怎样产生的（在本例中，通过多个与收益率相关的因子和多个不与收益率相关的因子线性组合产生）

设季度超额收益率的期望值$E\{\boldsymbol{r}\}=1.5\%$，等同于 6%的年预期超额收益率，设季度波动率为 9%，等同于 18%的年波动率。

该资产的收益率表达式如下：

$$\mathrm{r}=1.5\%+\theta_1+\theta_2+\cdots+\theta_{81}$$

1.5%是收益率的期望值，也是资产确定的预期收益。θ_1，θ_2 …θ_{81}是影响资产预期收益率的因素，也是收益率中的不确定部分，θ 两两相互独立，且等概率取值+1%或-1%。因此每个θ 的期望值为 0，标准差为 1%，则 r 的标准差为 9%，同季度波动率 9%一致。但是，我们无法观察到单个θ 的值，只能观察到总体 r。

接下来我们进行原始预测。在一般的预测场景下，在每个投资周期末才可以观察到收益值，但是必须在期初对其进行预测。预测的表达式为：

$$\mathrm{g}=2\%+\theta_1+\theta_2+\theta_3+\eta_1+\eta_2+\eta_3+\cdots\eta_{13}$$

原始预测具有 2%的期望和 4%的标准差，变量$\theta_1,\theta_2,\theta_3$是收益率的要素，是预测者在期初获得有关收益率的部分信息。 η_1,η_2,η_3 …η_{13}是原始预测中额外的不确定性的部分，与收益率无关。假设η_i都是等概率取值为+1%或-1%，每个θ_i是相互独立的，且与η_i也是相互独立的。原始预测可以看作是一些有用信息和无用信息的组合。可以把θ_i看作是信号单位，把η_i看作是噪音单位。

根据预测的表达式，预测者一共共有 16 个单位的信息，3 个是信号，13 个是噪音。然而不幸的是，原始预测只能看到总和 g，但无法区分表达式中的信号和噪音。

在本例中，g 就是用于预测收益率的唯一信息，利用该信息，我们可以求得 IC 和 g 的标准分值。

$$IC=\frac{Cov\{\boldsymbol{r},\boldsymbol{g}\}}{\mathrm{Std\{r\}Std\{g\}}}$$

$Cov\{\boldsymbol{r},\boldsymbol{g}\}$就是 r 和 g 之间共有元素的协方差之和$Cov\{\boldsymbol{r},\boldsymbol{g}\}=3\times(1\%)^2$

$$IC=\frac{3\times(1\%)^2}{9\%\times4\%}=0.0833$$

利用基本预测公式得到的精炼预测为：

$$\Phi=E\{\boldsymbol{r}|\boldsymbol{g}\}-E\{\boldsymbol{r}\}=\mathrm{Std\{r\}}\cdot IC\cdot Scores=9\%\times0.0833\times\frac{g-2\%}{4\%}$$

8.3.3.2 回归分析

在二元模型中，我们假设了解收益率和预测变量（原始预测）的产生过程。但是现实中，我们并不了解。因此，必须基于已经拥有的可获知数据对其产生过程作出推断，或依靠经验和直觉在已知的信息基础上进行推断。如果在获得了信息和收益时间序列数据的情况下，我们一般通过回归分析来精炼原始预测。

考虑 T 个时间单位上的预测变量 g(t)及实现收益率 r(t)的时间序列　令m_r和m_g为 r 和 g 的样本均值，$Var\{\boldsymbol{r}\}, Var\{\boldsymbol{g}\}, Cov\{r,g\}$为样本方差和协方差，使用时间序列回归作为精炼原始预测的工具。

$$\mathrm{r(t)} = c_0 + c_1 g(t) + \varepsilon_t$$

利用 r 和 g 的序列数据进行 OLS 回归，c_0，c_1的最小二乘估计是

$$c_1 = \frac{\mathrm{Std\{r\}} \times Corr\{\boldsymbol{r},\boldsymbol{g}\}}{\mathrm{Std\{g\}}}$$

$$c_0 = m_r - c_1 m_g$$

定义信息 g 的标准分为：

$$\frac{(\boldsymbol{g} - m_g)}{\mathrm{Std\{g\}}}$$

精炼预测的公式为：

$$\Phi = E\{\boldsymbol{r}|\boldsymbol{g}\} - E\{\boldsymbol{r}\} = c_0 + c_1 g(T+1) - m_r$$

$$= m_r - c_1 m_g + c_1 g(T+1) - m_r$$

$$= c_1\left[g(T+1) - m_g\right]$$

$$= \frac{\mathrm{Std\{r\}} \times Corr\{\boldsymbol{r},\boldsymbol{g}\}}{\mathrm{Std\{g\}}}\left[g(T+1) - m_g\right]$$

$$= \mathrm{Std\{r\}} \cdot Corr\{r,g\} \cdot z(T+1)$$

$$= 波动率 \times IC \times 标准分$$

8.3.3 双信息源分析

前面的场景都是只有一个信息源 g 的，而现实场景中，多个信息源的场景更为常见。

假设一种资产收益有两种预测$g_1 g_2$，相关系数为ρ_{12}

$$g = [g_1 \quad g_2] \qquad \rho_g = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix}$$

从而有

$$Var\{\boldsymbol{g}\} = \mathrm{Cov\{g,g\}} = \begin{bmatrix} Std\{g_1\} & 0 \\ 0 & Std\{g_2\} \end{bmatrix} \times \rho_g \times \begin{bmatrix} Std\{g_1\} & 0 \\ 0 & Std\{g_2\} \end{bmatrix}$$

收益率 r 与两个预测$g_1 g_2$之间协方差矩阵包括 2 个信息系数

$$\mathrm{Cov}\{\mathrm{r},\mathrm{g}\} = Std\{r\} \times [IC_1 \quad IC_2] \times \begin{bmatrix} Std\{g_1\} & 0 \\ 0 & Std\{g_2\} \end{bmatrix}$$

$$IC_1 = \mathrm{Corr}\{\mathrm{r}, g_1\} \qquad IC_2 = \mathrm{Corr}\{\mathrm{r}, g_2\}$$

根据精炼预测的基本公式，有：

$$\Phi = E\{\boldsymbol{r}|\boldsymbol{g}\} - E\{\boldsymbol{r}\}$$

$$= Cov\{\boldsymbol{r}, \boldsymbol{g}\} \cdot Var^{-1}\{\boldsymbol{g}\} \cdot (\boldsymbol{g} - E\{\boldsymbol{g}\})$$

$$= Std\{r\} \times [IC_1 \quad IC_2] \times \rho_g^{-1} \begin{bmatrix} \frac{1}{Std\{g_1\}} & 0 \\ 0 & \frac{1}{Std\{g_2\}} \end{bmatrix} \times \begin{bmatrix} g_1 - E\{g_1\} \\ g_2 - E\{g_2\} \end{bmatrix}$$

则可简化为：

$$\Phi = E\{\boldsymbol{r}|\boldsymbol{g}\} - E\{\boldsymbol{r}\} = \mathrm{Std}\{\mathrm{r}\} \cdot H \cdot Z = \mathrm{Std}\{\mathrm{r}\} \cdot IC_1^* \cdot Z_1 + \mathrm{Std}\{\mathrm{r}\} \cdot IC_2^* \cdot Z_2$$

其中$\mathrm{H} = [IC_1 \quad IC_2] \times \rho_g^{-1} = IC^T \times \rho_g^{-1} = [IC_1^* \quad IC_2^*]$

标准化评分$Z = \left[\frac{g_1 - E\{g_1\}}{Std\{g_1\}} \quad \frac{g_2 - E\{g_2\}}{Std\{g_1\}}\right]^T = [Z_1 \quad Z_2]^T$

正交化以后，可以求得修正的 IC 值得表达式

$$IC_1^* = \frac{IC_1 - \rho_{12} IC_2}{1 - \rho_{12}^2} \qquad IC_2^* = \frac{IC_2 - \rho_{12} IC_1}{1 - \rho_{12}^2}$$

8.4.4 实例

在实践中应用精炼预测超额收益的经验法则时，主要衡量的参数涉及：波动率，信息系数,得分。现实中为了方便估计这几个参数，通常假定，残差波动率的估计是可能的。

在没有充足的历史信息来确定原始预测的 IC 时，使用不太精确但已被检验过的先验估计。

表 8-1 给出了一些变量的预测水准

变量水准	IC 均值
良好的预测变量	0.05
优秀的预测变量	0.1
世界级的预测变量	0.15
错误的回测或者内幕交易	>0.2

表格 8-1

如表 8-2 所示的 ROE 因子，其 IC 均值为 0.087，是较为优秀的因子

	数值（季度统计）
IC 均值	0.087
IC 标准差	0.152
风险调整 IC（IR）	0.576

	数值（季度统计）
T 统计量（IC）	1.821
P 显著值（IC）	0.102
IC 偏度	0.266
IC 峰度	-1.068

表格 8-2 ROE 因子相关数值

因子的滚动 IC 也是一个衡量 IC 稳定性的重要性指标，如图 8-1 所示，ROE 因子的 22 日移动平均为 0.024，可见近期的 IC 值较以往的 IC 值偏高。

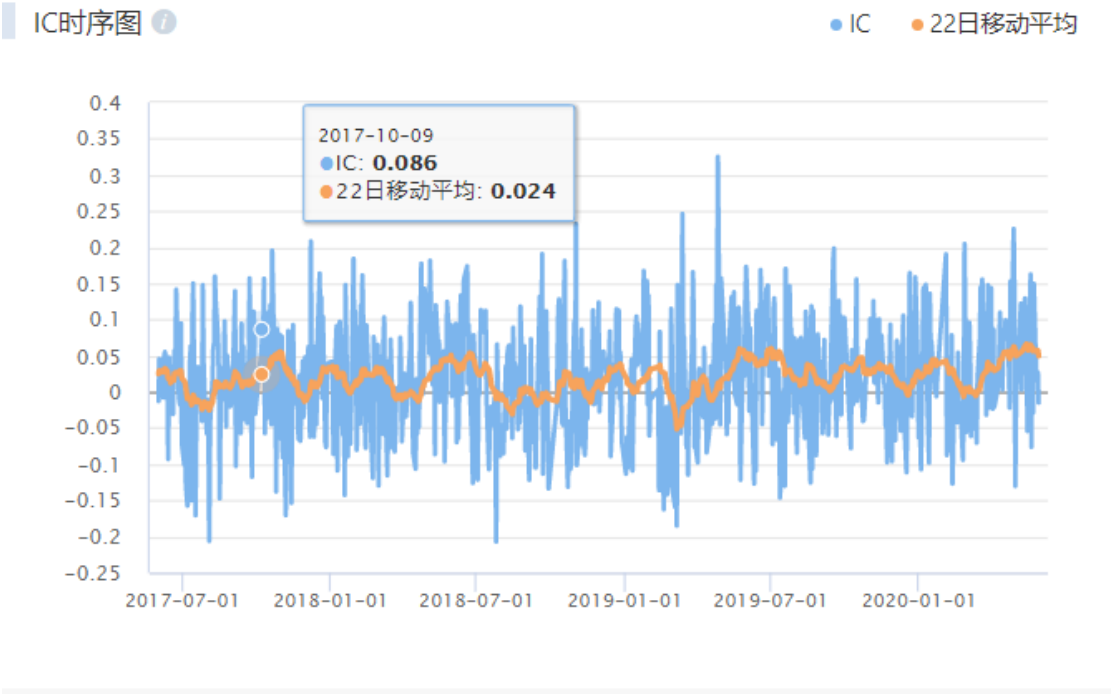


图 8-1 IC 历史滚动平均

8.4 连续时序预测

8.4.1 ARIMA 模型

ARIMA 模型全称为自回归移动平均模型(Autoregressive Integrated Moving Average Model,简记 ARIMA)，是由博克思(Box)和詹金斯(Jenkins)于 70 年代初提出的著名时间序列预测方法。

AR 模型指的是自回归模型，利用变量自身的值去做未来的预测。

一个 AR（P）模型的表达式如下：

$$y_t = \mu + \sum_{t=1}^{p} \gamma_i y_{t-i} + \epsilon_t$$

y_t是当前值，P 是阶数，γ_i是自相关系数

MA 模型指的是移动平均模型。指的是用变量估计的误差来预测未来

一个 MA（P）模型的表达式如下：

$$y_t = \mu + \sum_{t=1}^{p} \gamma_i \epsilon_{t-i} + \epsilon_t$$

在结合二者的基础上，对原先的 y 进行差分，就可以得到 ARIMA（p,q,d），其中 P 代表 AR 模型阶数，q 代表 MA 模型阶数，d 代表对于变量的差分次数。

$$y_t = \mu + \sum_{t=1}^{p} \gamma_i \epsilon_{t-i} + \sum_{t=1}^{p} \gamma_i y_{t-i} + \epsilon_t$$

8.4.2 GARCH 模型

在传统计量经济学模型中，干扰项的方差被假设为常数。但是许多经济时间序列呈现出波动的集聚性，在这种情况下假设方差为常数是不恰当的。

ARCH 模型将当前一切可利用信息作为条件，并采用某种自回归形式来刻划方差的变异，对于一个时间序列而言，在不同时刻可利用的信息不同，而相应的条件方差也不同，利用 ARCH 模型，可以刻划出随时间而变异的条件方差。

ARCH 模型有以下假定：

（1） 资产收益率序列的扰动a_t是序列不相关的，但是不独立。

（2） a_t的不独立性可以用其延迟值的简单二次函数来描述。

一个 ARCH（m）模型可以表示为：

$$a_t = \sigma_t \varepsilon_t \qquad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots \alpha_m a_{t-m}^2 \qquad \alpha_0 > 0; \forall \mathrm{i} > 0, \alpha_i \geqslant 0$$

其中，ε_t为均值为 0，方差为 1 的独立同分布（iid）随机变量序列，通常假定其服从标准正态分布。σ_t^2为条件异方差。

在 ARCH 模型的基础上，Bollerslev 在 1996 年又提出了广义的 ARCH 模型，也就是 GARCH 模型，用于预测资产收益率的波动率。

$$a_t = \sigma_t \varepsilon_t \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2$$

$$\alpha_0 > 0; \beta_j > 0 \forall \mathrm{i} > 0, \alpha_i \geqslant 0$$

上面的模型被称为 GARCH（m,s）,确定 m，s 的阶数需要借助经验或者某些信息准则。

8.4.3 Prophet 模型

Prophet 是基于传统计量经济模型的一种变体，其使用的是加性模型，也就是分别用不同的部分来拟合时间序列不同的趋势，叠加起来则是整个时间序列模型：

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

g(t)是趋势函数，表示时间序列上的非周期变化，在 Prophet 中有以下两种形式：

(1)当预测的标的具有饱和增长的特性（即增长到一定程度不再增长）

$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^T\delta)(t - (m + a(t)\gamma)}$$

C(t)为增长的上限，每一个 t 时间点，上限都是不同的。

（2）当预测的标的不具有饱和增长的特性

$$g(t) = (k + a(t)^T\delta)t + (m + a(t)\gamma)$$

（1）（2）式子中，$(k + a(t)^T\delta$代表增长速率来使得函数连续，$(m + a(t)\gamma$代表线性偏移，a(t)是一个 bool 函数，取值 1 或者 0。

s(t)用于刻画时间序列中的周期性（季节性）变化，其形式为傅立叶级数，如下所示：

$$s(t) = \sum_{n=1}^{N}(a_n\cos(\frac{2\pi nt}{P}) + b_n\sin(\frac{2\pi nt}{P}))$$

P 是时间周期，当 P=7 时刻画的是以周为周期。a_n和b_n是需要学习的参数。由傅立叶级数的性质可知，N 越大越能刻画变化多的周期性模式，默认使用 N=10 刻画以年为单位的周期性变化，N=3 刻画以周为单位的周期性变化。

h(t)即为 holday，其用于拟合节假日和特殊日期，比如中国的双十一、美国的黑五等，函数形式为：

$$h(t) = Z(t)\tau$$

Z(t)代表节日的时长，τ代表市场对于趋势的整体影响程度。

8.4.4 模型对比&案例

基于以上的连续时序模型，我们使用一个真实案例作为对比说明。

如表格 8-3，我们 选取了 2018 年至 2020 年上半年的沪深 300ETF（000300）涨跌幅作为研究对象。

日期	沪深 300 涨跌幅
2018-01-03	0.010207
2018-01-04	0.003337
2018-01-05	0.003651
2018-01-08	0.002900
2018-01-09	0.005844
2018-01-10	0.005367
2018-01-11	0.000909
2018-01-12	0.003578
2018-01-15	0.004616
2018-01-16	0.000771

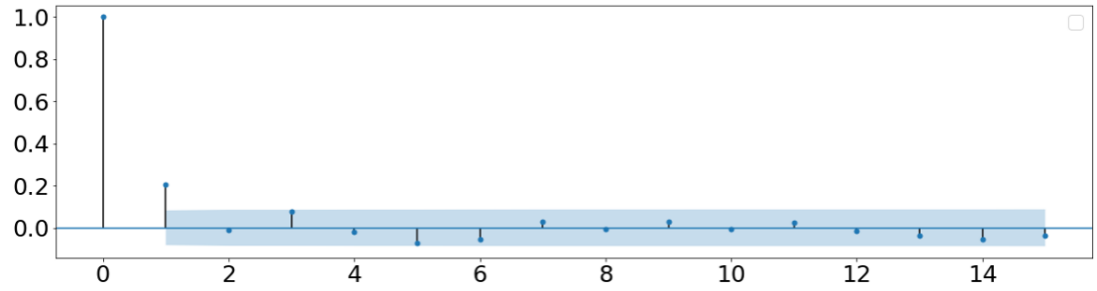
表格 8-3 沪深 300 涨跌幅时间序列

首先对于涨跌幅序列进行 ADF 平稳性检验：

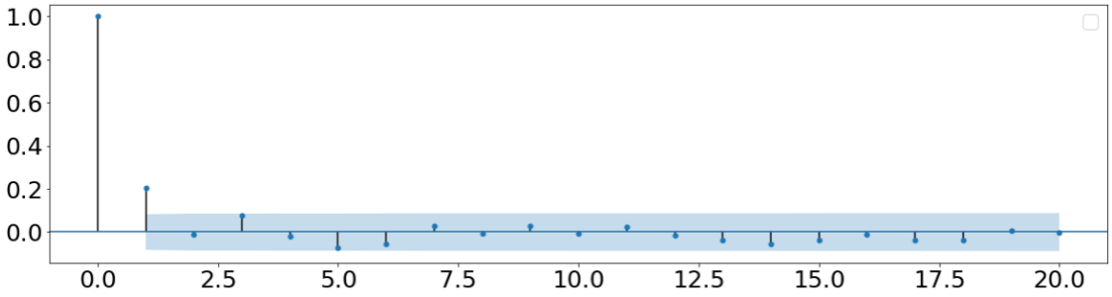
T 统计量	-11.900750522141337,
T 统计量 P 值	5.582753289859608e-22,
延迟阶数	2
利用样本数	,562
99%，95%，90%置信区间下的临界的 ADF 检验值	99%': -3.442039359113542； 95%: -2.8666965134862514； 90%: -2.5695162601790758

表格 8-4

ARMA 模型和 GARCH 模型都需要计算其自相关系数和偏自相关系数，因此首先计算 0-15 阶的自相关系数（如图 8-2），以及 0-20 阶的偏自相关系数（如图 8-3）



图表 8-2 自相关系数



图表 8-3 偏自相关系数

我们通过偏自相关和自相关系数的 lag 图，可以确定 ARMA 模型的阶数是[1,1],因为涨跌幅序列未经差分就已经平稳，所以 ARIMA 模型的阶数为(1,0,1)回归结果如表 8-5

	系数值	标准差	Z 值	P 显著性	2.5%分位点	97.5%分位点
截距项	-4.973e-06	0.001	-0.009	0.993	-0.001	0.001
Ar（1）	-0.4000	0.239	-1.677	0.094	-0.868	0.068
Ma（1）	0.6117	0.212	2.890	0.004	0.197	1.027

表格 8-5 arma 模型回归结果（观察确定阶数）

如果采用 AIC 信息准则，最小化原则进行定阶，则有 ARIMA 模型的阶数为(2,0,1)，得到的回归结果如表 8-6

	系数值	标准差	Z 值	P 显著性	2.5%分位点	97.5%分位点
截距项	-5.837e-05	0.001	-0.100	0.921	-0.001	0.001
Ar（1）	-0.7374	0.045	-16.221	0.000	-0.826	-0.648
Ar（2）	0.1516	0.044	3.455	0.001	0.066	0.238
Ma（1）	0.9787	0.016	59.399	0.000	0.946	1.011

表格 8-6 arma 模型回归结果（AIC 信息准则）

ARIMA 的表达式为：

$$r_t = -5.837\mathrm{e}-05 - 0.7374r_{t-1} + 0.1516r_{t-2} + 0.9787\varepsilon_{t-1} + \varepsilon_t$$

GARCH 模型是在 AR（p）模型的基础上再次估计波动率模型的结果，由 AIC 信息准则确定的 AR 阶数为 2，在此基础上，构建 GARCH 模型如表 8-7

Dep. Variable:	000300	R-squared:	0.041
Mean Model:	AR	Adj. R-squared:	0.038
Vol Model:	GARCH	Log-Likelihood:	1707.44
Distribution:	Normal	AIC:	-3402.87
Method:	Maximum Likelihood	BIC:	-3376.98

表格 8-7 GARCH 模型回归相关参数

收益率均值模型如表 8-8

	参数值	标准差	T 统计量	P 显著值	95.0%水平区间
截距项	2.1431e-04	5.010e-04	0.428	0.669	[-7.676e-04,1.196e-03]
ar[1]	0.2629	4.596e-02	5.720	1.063e-08	[0.173, 0.353]
ar[2]	-0.0426	4.793e-02	-0.888	0.375	[-0.136,5.139e-02]

表格 8-8 AR（2）模型参数

方差波动率模型如表 8-9

	参数值	标准差	T 统计量	P 显著值	95.0%水平区间

截距项	3.9430e-05	2.417e-06	16.312	8.176e-60	[3.469e-05,4.417e-05]
alpha[1]	0.1025	7.133e-02	1.437	0.151	[-3.734e-02, 0.242]
beta[1]	0.5976	6.017e-02	9.932	3.018e-23	[0.480, 0.716]

表格 8-9 GARCH **模型回归结果**

由回归方程可以得到 GARCH 均值模型和波动率模型的估计表达式：

$$r_t = 2.1431\text{e}-04 + 0.2629r_{t-1} - 0.0426r_{t-2} + \varepsilon_t$$

$$\sigma_t^2 = 3.9430\text{e}-05 + 0.1025a_{t-1}^2 + 0.5976\sigma_{t-1}^2 + \varepsilon_t$$

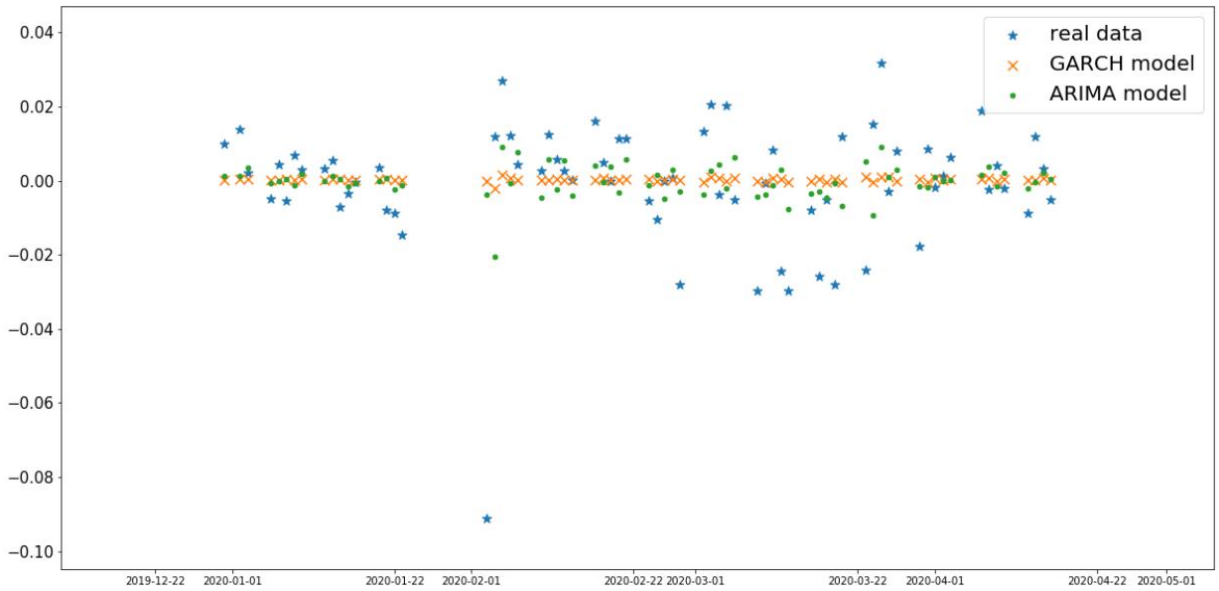
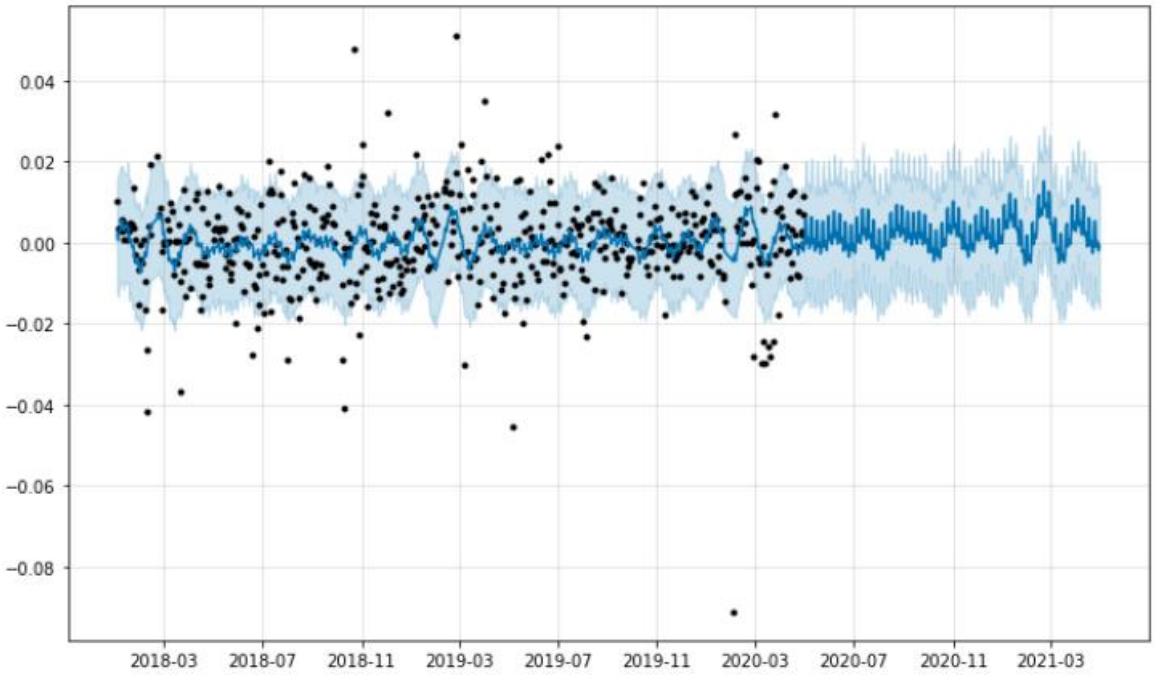


图 8-4 两种模型对于 2020 年的预测

两种模型对于 2020 年的测试集部分预测如图 8-4，可以看到在真实的预测坏境中，实际值的波动情况远远大于预测值的波动情况，而 GARCH 模型更能够预测出实际值的波动。

Prophet 模型可以增强对于周期性行情的预测，相对于其他的模型，其在预测波动行情的条件下表现更好，但 A 股极少表现出周期性的特质，预测的上界和下界，以及实际值预测值如图 8-5



图表 8-5 Prophet 模型预测情况

模型的均值在 prophet 中可以选择线性上涨或者 logistic 上涨，这里选择线性上涨，如图 8-6

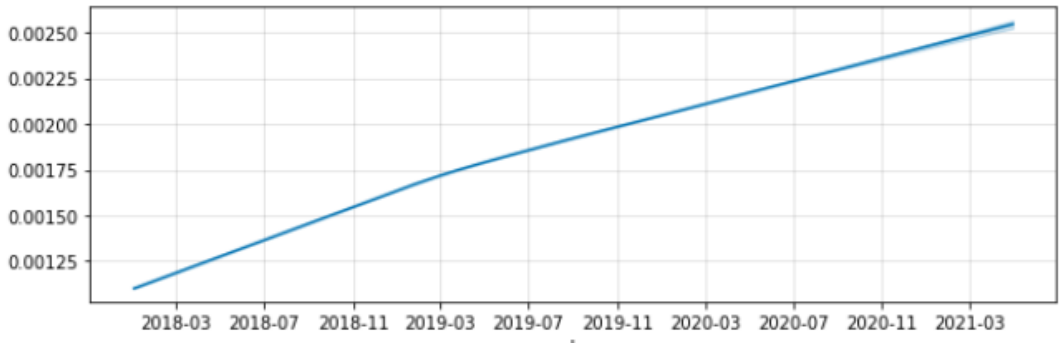


图 8-6 Prophet 模型线性上涨

由于星期的不同，各个星期有特异值，如图 8-7

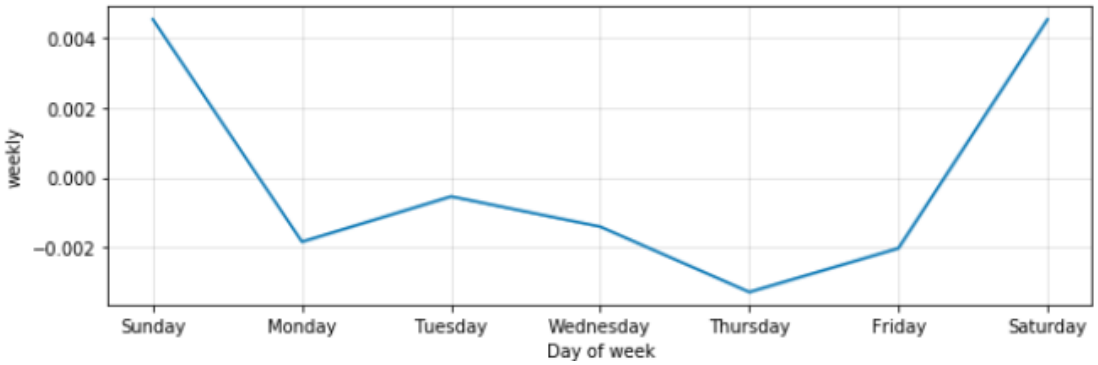
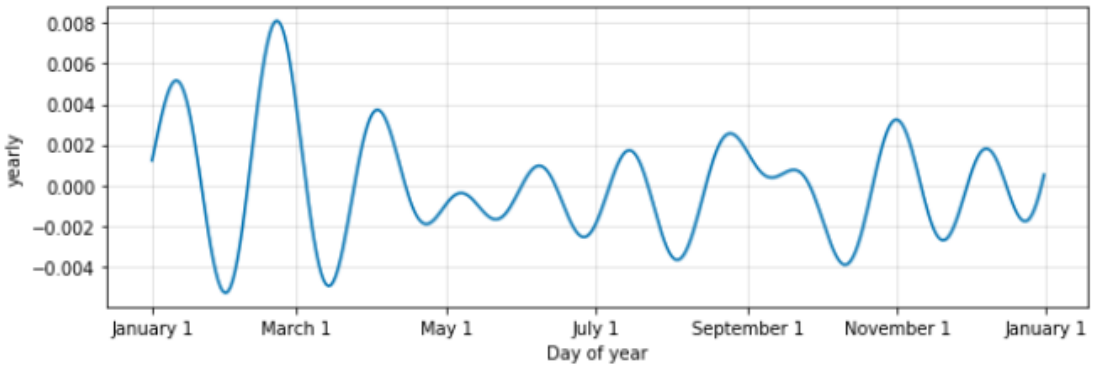


图 8-7 Prophet 模型星期特异值

在一年内的特异性也会发生变化



图表 8-8 Prophet 模型年线特异值

对于预测模型的评判。一般有三种方法，首先是,均方误差 MSE，衡量预测值与真实值之差的平方的均值，MAE 衡量预测值与真实值之差的绝对值的均值，符号正确率在择时模型中较为常用，衡量预测值与真实值符号一致的概率。

三种模型对于沪深 300 的预测情况如表 8-10

使用模型	均方误差 MSE	绝对平均误差 MAE	符号正确率
Garch	0.013%	0.679‰	53.9%
ARIMA	0.012%	0.621‰	55.5%
Prophet	0.012%	0.774‰	57.5%

表格 8-10

8.5 离散时序预测

8.5.1 马尔可夫链模型

马尔可夫链模型是描述一类重要的随机动态系统（过程）的模型。该过程是时间、状态均为离散的随机转移过程。

马尔可夫链是随机变量$X_1X_2X_3$的一个数列。这些变量的范围，即他们所有可能取值的集合，被称为“状态空间”，而X_n的值则是在时间 n 的状态。如果X_{n+1} 对于过去状态的条件概率分布仅是X_n的一个函数，也就是当前状态的取值只与上一个时期的状态相关。

$$P(X_{n+1} = x|X_1X_2X_3, \dots X_n) = P(X_{n+1} = x|X_n)$$

其中的 x 代表过程中的某个状态，恒等式代表马尔科夫性质。

总体来说，马尔科夫链是满足以下假设的马尔科夫过程：

（1）t+1 时刻系统状态的概率分布只与 t 时刻的状态有关，与 t 时刻以前的状态无关；

（2）从 t 时刻到 t+1 时刻的状态转移与 t 的值无关。一个马尔可夫链模型可表示为=(S，P，Q)，其中各元的含义如下：

1、S 是系统所有可能的状态所组成的非空的状态集，有时也称之为系统的状态空间，它可以是有限的、可列的集合或任意非空集。本文中假定 S 是可数集(即有限或可列)。用小写字母 i, j(或S_i, S_j)等来表示状态。

2、$\mathrm{P} = [P_{ij}]_{n\times n}$是系统的状态转移概率矩阵，其中$P_{ij}$表示系统在时刻 t 处于状态 i，在下一时刻 t+1 处于状态 i 的概率，N 是系统所有可能的状态的个数。对于任意 i∈s，有$\sum_{j=1}^{N} P_{ij} = 1$

3、$\mathrm{Q} = [q_1, q_1 \dots q_n]$是系统的初始概率分布，qi 是系统在初始时刻处于状态 i 的概率，满足$\sum_{i=1}^{N} q_i = 1$。

8.5.2 循环神经网络 RNN

在传统神经网络输入输出的基础上，添加一条自循环回路表征当前状态，使得当前状态会影响其未来状态，这就形成了循环神经网络。循环神经网络将前一时刻神经元的输出状态 ht-1，与当前时刻输入 xt，一起送入计算 ht。即有

$$h_t = f(h_{t-1}, x_{t)}$$

通过添加对于上一级神经元的状态函数，神经元具有了记忆功能，使得其能够处理时间序列的数据。

一个常见的 RNN 模型表示如下

$$h_t = \tanh(h_{t-1}w_{hh} + x_t w_{xh} + b_h) = \tanh([h_{t-1}, x_t]W + b_h)$$

$$y_t = softmax(w_{hy}h_t + b_y)$$

在 RNN 中，上一时刻的状态（记忆）与当前时刻的输入拼接成一维向量作为循环体的全连接层神经网络的输入。权重矩阵w_{hy}, w_{xh}, w_{hh}和偏置项$b_h b_y$在 t0 和 t1 时刻的循环体中是继承的，这也说明了 RNN 结构中的参数在不同时刻中也是一脉相承的。

8.5.3 长短期记忆网络 LSTM

RNN 第一次提出了针对于时间序列的神经网络模型，但是简单的 RNN 回路很容易产生梯度消失和梯度爆炸的现象。针对 RNN 模型，霍普里特 施米德胡贝在 1997 年提出了 LSTM 模型，作为 RNN 模型的优化和继承之作。

LSTM 模型增加了三个门回路，对于输入输出进行优化：

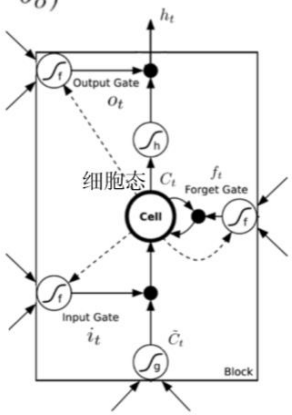
（1） 输入门：将输入的信息选择性的记录到细胞中（Cell）

（2） 遗忘门 ft：将细胞状态中的信息选择性的遗忘

（3） 输出门 ot：把细胞中的信息选择性的进行输出

一个完整的 LSTM 神经元如下图

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] \ + \ b_o\right)$$



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

图表 8-9

遗忘门公式如下

$$f_t = \sigma(W_f \times [h_{t-1}x_t] + b_f)$$

遗忘门读取 $h_{t-1}x_t$，输出一个规整到 0 到 1 之间的数值，作用到神经元中，用于决定上一个时间点的信息有多少被传递到这个神经元中。

输入门公式如下：

$$i_t = \sigma(W_i \times [h_{t-1}x_t] + b_i)$$

输入门读取 $h_{t-1}x_t$，同样输出一个规整到 0 到 1 之间的数值，用于决定当前时间点的信息有多少被传递到这个神经元中。

候选态公式如下

$$\tilde{C}_t = tanh(W_C \times [h_{t-1}x_t] + b_c)$$

输入门读取 $h_{t-1}x_t$，输出被筛选后的当前信息，并作用到神经元上。

最后两部分信息加总，得到传递到下一个神经元的信息和输出的信息：

神经元：

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t$$

输出门：

$$o_t = \sigma(W_o \times [h_{t-1}x_t] + b_o)$$

$$h_t = o_t \times \tanh(C_t)$$

8.5.4 门控循环单元网络 GRU

GRU 模型基本上市 LSTM 模型的简化版，其将遗忘门和输入门整合一番，变成了更新门，其他部分没有特大的改动，一个简单的模式图如图 8-10

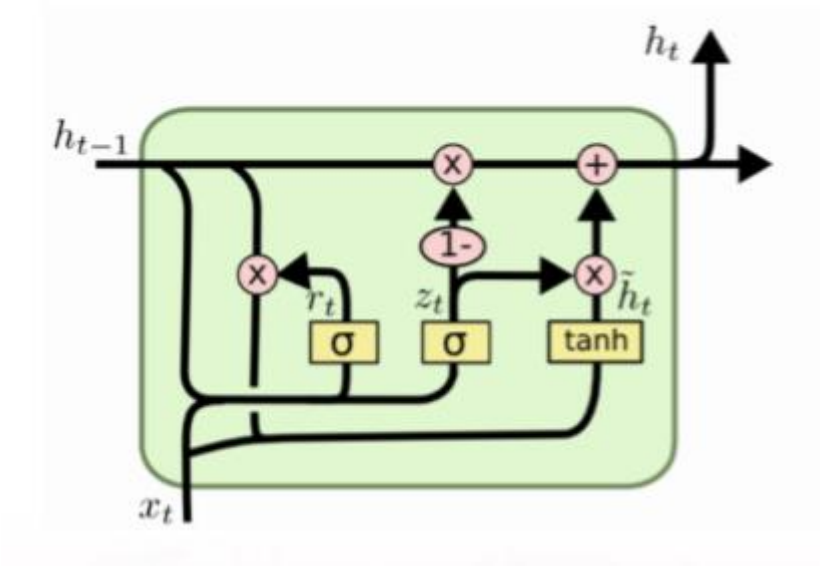


图 8-10 GRU 模型简图

$z_t = \sigma(W_z \times [h_{t-1}x_t])$是更新门，输出 0 到 1 之间的一个值，用于确定当前信息与上一个时期的信息的分配比例。

$r_t = \sigma(W_r \times [h_{t-1}x_t])$是重置门，用于衡量对上一个时期信息的保留程度。

$$\tilde{h}_t = tanh(W_r \times [r_t \times h_{t-1}, x_t])$$

$$h_t = (1 - z_t)h_{t-1} + z_t\tilde{h}_t$$

8.5.5 模型对比&案例

与连续预测模型一样，我们选用 2014 年至 2020 年上半年的沪深 300ETF 的涨跌幅作为模型的数据源。（见表 8-3）

在实例中的马尔可夫链模型，我们假定市场存在上行趋势和下行趋势两种状况s_1, s_2, s_3，这两种状况的会以$w_1, w_2, , w_3$的概率相互转化，下一日的涨跌可以通过上一日的涨跌来估计。

进一步的，我们把时间扩到 3 日的区间内来考虑，也就是当日的涨跌会由前三日的涨跌幅来估计，将上涨 0.5%及以上记为 1，下跌 0.5%以上记为-1，其余情况（平盘）记为 0，得到的马尔科夫过程可以表达为表 8-12。

		当前上涨	**当前震荡**	**当前下跌**
上一日	**上两日**	概率		
下跌	**下跌**	26.56%	34.38%	39.06%

		当前上涨	当前震荡	当前下跌
上一日	上两日	概率		
	震荡	15.49%	45.07%	39.44%
	上涨	26.09%	26.09%	47.83%
震荡	下跌	23.33%	48.33%	28.33%
	震荡	29.31%	44.83%	25.86%
	上涨	18.06%	48.61%	33.33%
上涨	下跌	38.24%	50.00%	11.76%
	震荡	50.82%	34.43%	14.75%
	上涨	30.65%	53.23%	16.13%

表格 8-12 马尔科夫过程

可以发现在上一日上涨而上两日平盘时，上涨概率最高，而在上一日下跌而上两日上涨时，下跌概率最高。

通过对于单纯价格序列的马尔可夫过程研究，我们可以掌握一些基本的技术面归因方法，探求前 n 日的波动对于当日波动的可解释性。同样的模型可以利用到基本面的研究上。

使用简单的 RNN 模型对沪深 300ETF 的涨跌幅进行预测，模型的输入时当前 etf60 日的历史 K 线数据和成交量，输出有两种模式，一种是当日的涨跌幅，一种是当日的涨跌情况，数据在时间序列上 Z 值化，以便模型更好的识别。RNN 模型使用 SGD 优化器，损失函数利用标准差衡量，对所有训练集数据训练 5 次。

当使用当日涨跌状况，利用 RNN 进行训练时有（表 8-13）：

使用模型	ACC	AUC
SimpleRNN	54.22%	56.15%

表格 8-13 RNN 预测结果

准确率达到 54%，说明模型有一定的择时能力，但并不明显。

同样的，使用 LSTM 进行预测，有（表 8-14）：

使用模型	ACC	AUC
LSTM	51.01%	25.60%

表格 8-14 lstm 预测结果

使用 LSTM 时发现 AUC 值极低，原因是在测试集中，LSTM 预测结果基本为下跌，导致上涨情况预测成功率极低。整体准确率一般，择时能力基本为 0。

最后使用 GRU 模型对沪深 300ETF 的涨跌幅进行预测。

GRU 模型使用 ADAM 优化器，损失函数利用标准差衡量，对所有训练集数据训练 5 次。

当使用当日涨跌状况进行训练时有表 8-15：

使用模型	ACC	AUC
GRU	0.5502008032128514	0.549213630406291

表格 8-15 GRU 预测结果

GRU 模型的预测成功率为 3 个时序是神经网络中最高，但 AUC 值略有不足。

此外-，使用当日涨跌幅作为预测目标得到的预测结果和实际值的标准差对比，如表 8-16

使用模型	标准差
SimpleRNN	1.1958%
LSTM	1.1467%
GRU	1.1649%

表格 8-16 模型对比

可以看到 LSTM 模型的标准偏差最小，GRU 次之，SimpleRNN 最大，按模型复杂度降序排列

8.6 本章小结

本章主要讲解了基于信息的收益率预测方式，要理解信息对于超额收益率的作用，一定要理解预测的基本公式

$$\Phi = \mathrm{Std}\{\mathrm{r}\} \cdot IC \cdot Scores$$

在基本公式的基础上，进一步的，我们针对单信息源和多信息源的收益率预测做出推算，单信息源的精炼预测为

$$\Phi = \mathrm{Std}\{\mathrm{r}\} \cdot Corr\{r, g\} \cdot z(T+1)$$

即对于单信息源 g，精炼预测的构成只需要三个变量来估计：收益率的标准差，IC 和 g 的标准分值。

多信息源的精炼预测为

$$\Phi = E\{\boldsymbol{r}|\boldsymbol{g}\} - E\{\boldsymbol{r}\} = \mathrm{Std}\{\mathrm{r}\} \cdot H \cdot Z = \mathrm{Std}\{\mathrm{r}\} \cdot IC_1^* \cdot Z_1 + \mathrm{Std}\{\mathrm{r}\} \cdot IC_2^* \cdot Z_2$$

其中$\mathrm{H} = [IC_1 \quad IC_2] \times \rho_g^{-1} = IC^T \times \rho_g^{-1} = [IC_1^* \quad IC_2^*]$

$$IC_1^* = \frac{IC_1 - \rho_{12} IC_2}{1 - \rho_{12}^2} \qquad IC_2^* = \frac{IC_2 - \rho_{12} IC_1}{1 - \rho_{12}^2}$$

此外，本章还有介绍了一些对于连续时序预测和离散时序预测的方法：连续时序模型主要是计量模型：ARIMA，GARCH，Prophet 模型。离散时序模型主要依靠成时间序列的模块参数，一般是一些特征处理的模型：RNN，LSTM，GRU，马尔科夫链等。

8.7 参考文献

[1]SciPy 官方文档
[2]华泰证券_华泰人工智能系列之九_人工智能选股之循环神经网络模型
[3]东北证券_经济指标周期及领先性确认的数理方法

[4]兴证期货_股指期货量化展期与择时对冲策略