基本面数据延时对量化模型影响研究

华泰期货研究所 量化组

陈维嘉

量化研究员

☎ 0755-23991517

✉ chenweijia@htfc.com

从业资格号：T236848

投资咨询号：TZ012046

基本面数据通常都会包含延时等情况。比如在当天交易开盘前，可能拿不到当天需要的数据，又或者即使拿到了一个数据，但这个数据值可能在日后出现改变。如果在量化模型回测时无法知道具体指标录入数据库的时间，而是根据指标的标注时间进行回测，则有可能在回测时引入未来数据，从而无法有效评估交易策略的收益与风险。为了考察基本面数据延时对量化策略收益的影响，这里主要考察钢联提供的 API 数据库，里面提供了数据的录入时间，从而可以确定数据的延时。对比数据录入时间和数据标注时间可以发现，黑色商品数据的延时主要发生在节假日前后，其他时间出现延时的情况较少。

另外，这篇报告使用 R-MIDAS 模型综合日频和周频的黑色基本面数据进行日度预测，并按照数据的标注时间和录入时间，把策略分为三组，第一组按照标注时间训练和预测，第二组按照录入时间训练和预测，第三组按照标注时间训练，但按照录入时间进行预测。三组测试分别在螺纹钢、热轧卷板、铁矿石、焦煤和焦炭中进行。其中第一组和第二组在各个品种中表现各异，而且包含未来数据的第一组在各个品种中并无明显优势。第三组的收益表现通常在第一组和第二组之间。研究结果表明，虽然基本面数据可能存在延时的情况，但如果按照标注时间训练模型，录入时间进行预测，则仍有可能获得较好收益。

相关研究

混频基本面量化模型在黑色商品板块中的应用 2018-12-28

基本面数据内容结构

这篇报告研究钢联 API 提供的基本面数据，这个数据库包含了螺纹钢、热轧卷板、铁矿石、焦煤、焦炭以及有色金属和农产品的基本面数据，但里面的因子是以黑色金属板块为主。黑色金属所涉及的基本面数据包括各个地区和大型钢厂的不同类型的钢材价格。钢材类型包括高线、普中板、线材、不锈钢和盘螺等。地区通常以区域划分例如华东、南方、北方等。大型钢厂包括重钢集团、宝钢集团和马钢集团等。除了现货价格外该数据库还包含了库存、钢厂产能、实际开供条数和开工率等指标。

这些基本面数据通常都会涉及到延时的情况，即某个指标的更新可能会晚于其标注时间。而在交易前从数据库下载往往只能得到一个空值，但在构建量化模型时却能根据标注时间获得一个非空值。这样的结果就是把未来数据引入到量化模型当中。钢联的这个数据库一个最大特点是提供了大部分数据的录入时间。如果把数据录入时间与交易时间对比，并谨慎地选择交易时间前所能获取的数据信息，就能把未来数据从回测中剔除掉。钢联的这版数据库里面每个指标的时间序列都包含了两个时间，一个是标注时间，这个时间是当前指标所描绘的时间点。例如“重钢集团：重庆市场价格：热轧板卷：Q235B：4.75*1500*C（日）”这个指标在 2018 年 9 月 20 日这一天对应了两个价格，一个是 10:26:10 录入的 4360，另一个是 14:35:56 录入的 4350。这个指标在 2018 年 10 月 17 日只对应了一个价格是 4240，但这个价格是 2018 年 10 月 18 日 10:37:38 才录入的，所以如果在 2018 年 10 月 18 日 10:37:38 前交易要参考这个指标的话是看不到 4240 这个价格的，只能看到 10 月 17 日之前的价格。另外值得注意的一个情况是录入时间有可能提前于标注时间，比较常见的是一些周频指标，标注时间是周五发布，但实际是周四就录入了。例如指标“53 家独立电弧炉企业产能利用率（周）”标注时间是 2017 年 2 月 3 日，但录入时间是 2017 年 2 月 2 日 17:18:00。这篇报告只考虑录入时间滞后于标注时间的数据对量化模型的影响，录入时间提前于标注时间的影响留待以后进一步研究。

根据最新数据原则，这篇报告对钢联的指标时间序列分别按照标注时间和录入时间进行重新篇排。这里的量化模型只用到日频和周频数据。这里对螺纹钢、热轧卷板、铁矿石、焦煤和焦炭这 5 个黑色品种分别从钢联数据库中寻找相关的指标。每个品种对应 20-100 个日频和周频指标。按标注时间编排时，非周末或节假日时指标的标注时间与交易日日期对齐，周末或节假日时，指标的标注时间比下一个交易日滞后一天。例如 2018 年 10 月 8 日为国庆假期后的第一交易日，这是与国庆前最后一个交易日，即 9 月 28 日对齐的指标日期可以到 10 月 7 日。因此，这么做很有可能引入了未来数据，因为 10 月 7 日非工作日，其对应指标有可能在 10 月 8 日更新。按录入时间编排时，假设当天有夜盘的话，指标的录入时间必须在 20:00 之前，如果没有夜盘则指标录入时间须在下个交易日 8:00 之前。根据此原则把日频基本面数据与期货市场数据每天日期对齐。周频基本面数据的录入时间通常在周五或周四，

但标注时间通常都是在周五，这里把所有周频数据统一在一周内最后一个交易日更新。所以如果周末有夜盘则把周频基本面数据和期货交易日对齐。周五无夜盘或周五非交易日则周频数据的录入时间可在下个交易日 8:00 之前。

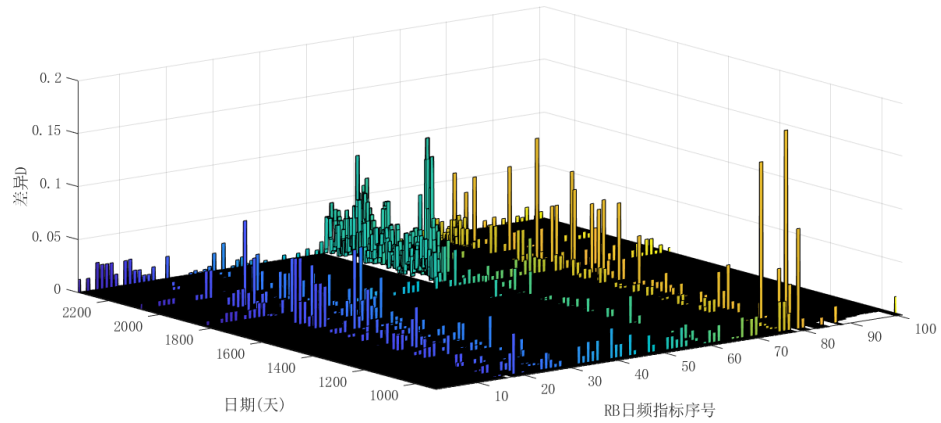
首先考察按照标注时间和录入时间分别构造的时间序列矩阵的差异。这里以涉及指标较多的螺纹钢和热轧卷板为例。这里的差异D可表示为

$$D = \left| log\left(\frac{V_c}{V_m}\right) \right| \tag{1}$$

其中V_c表示按照指标录入时间所提取的值，V_m表示按照指标标注时间所提取的值。

下图 1 至 4 分别标注作出了按照不同方法提取螺纹钢和热轧卷板指标值的差异D随时间变化的三维图。图中的时间左横轴使用数字代号表示，右横轴表示所使用的因子序号。图中的空白表示该指标数值在对应时间点缺失。图 1 中螺纹钢的日频指标差异主要集中在某些指标，每天都会出现一定的数据延时，但整体上即使出现延时相差的幅度也不大，大部分D值都在10%以下。少量指标的差异在某些时间点能达到 20%以上。

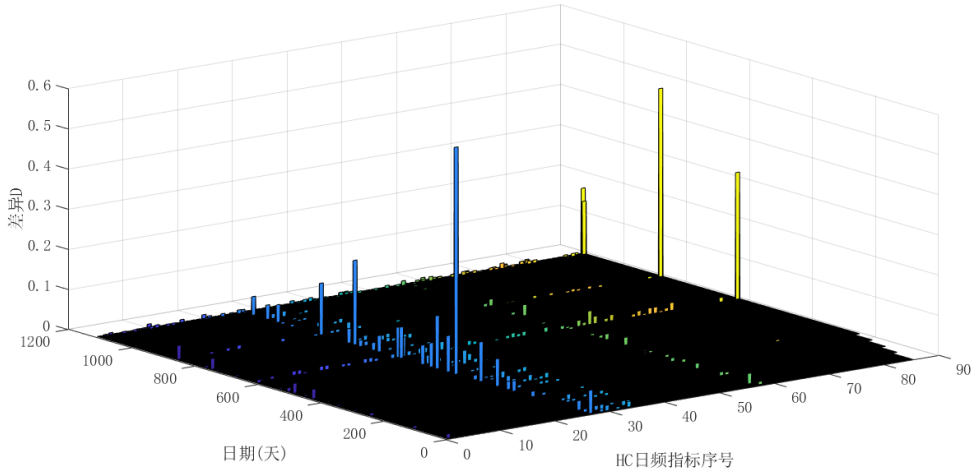
图 1： 螺纹钢日频数据差异(2013-1-4 至 2018-10-19)



数据来源：钢联

图 2 中热轧卷板数据的D值大部分都会比螺纹钢低些，出现差异的情况也会少些，主要集中在某几个特定时间段，但也有因子，例如 89 号因子“全国建筑钢材成交量：主流贸易商：合计（日）”会出现高达 46.9%的差异，这也可能是源于数据源使用了不同的统计方法引入的差异。

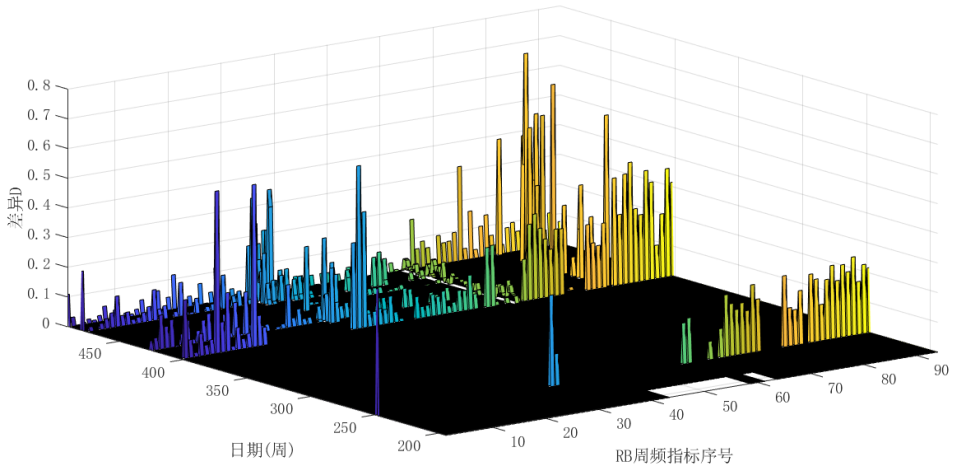
图 2： 热轧卷板日频数据差异(2014-3-24 至 2018-10-19)



数据来源：钢联

图 3 显示的的是螺纹钢周频因子差异，周频因子的差异幅度要比日频因子大，有些因子可以达到 70%以上，周频因子幅度较大的时间点主要集中在节假日前后，存在较大差异。在 200 至 400 上的时间段没有差异是因为这段时间缺乏入库时间记录。

图 3： 螺纹钢周频数据差异(2010-01-08 至 2018-10-19)

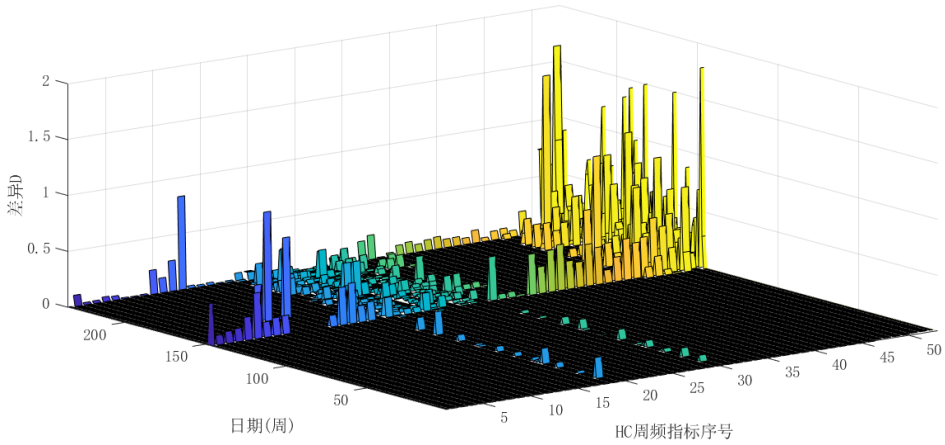


数据来源：钢联

图 4 显示的的是热轧卷板周频因子差异，周频因子的差异幅度也较日频因子大，某些因子幅度可以达到 150%以上。在某个时间点大量指标出现差异的情况通常发生在长假前后，例如 2017 年 1 月 27 日至 2017 年 2 月 2 日是春节长假，2017 年 2 月 3 日是正常交易日，如果按照标注时间在 2017 年 2 月 3 日 8:00 做预测,有些指标只能看到 2017 年 1 月 20 日的情况，因为 2017 年 1 月 26 日是长假前所以数据是缺失的。但是如果按照录入时间做预测，则可以看到 2017 年 2 月 3 日 8:00 前的数据，而 2017 年 2 月 3 日恰好是周五，有些数据被提前

录入，所以在 2017 年 2 月 3 日 8:00 做预测是可以看到，标注时间为 2 月 3 日但实际录入时间是 2 月 2 日的数据。这是造成两种排序方法相差较大的原因。

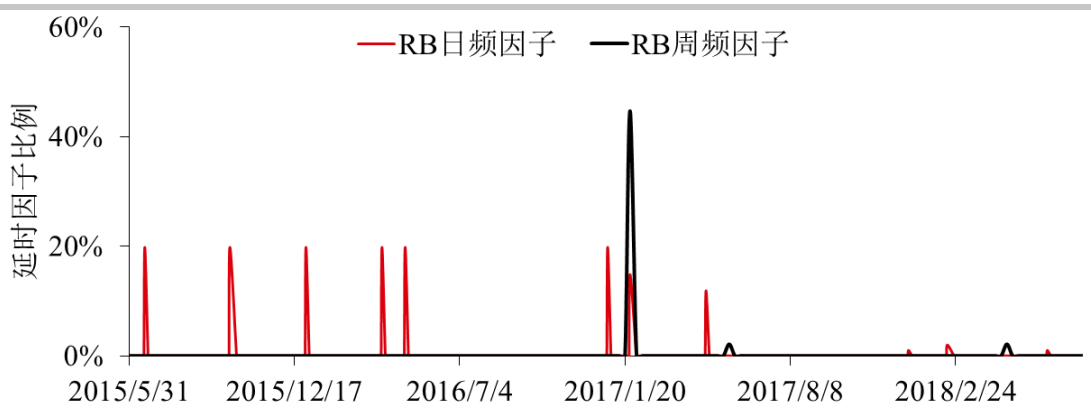
图 4： 热轧卷板周频数据差异(2014-3-28 至 2018-10-19)



数据来源：钢联

图 5 记录的是螺纹钢所使用的因子至少出现 1 天延时的比例，由图可见因子出现延时的比较随着时间不同会有所变化，日频因子出现延时的比例在 20%以下的水平，主要集中在五一、国庆和春节假期。周频因子存在录入时间的记录较短从 2017 年才开始，但是比例在 2017 年 1 月曾高达 40%。值得注意的是图 5 出现数据延时的比例要比图 1 和图 3 少，那是因为图 5 只考察了数据延时的情况，而图 1 和图 3 还包含了数据提前（录入时间领先于标注时间）的差异。

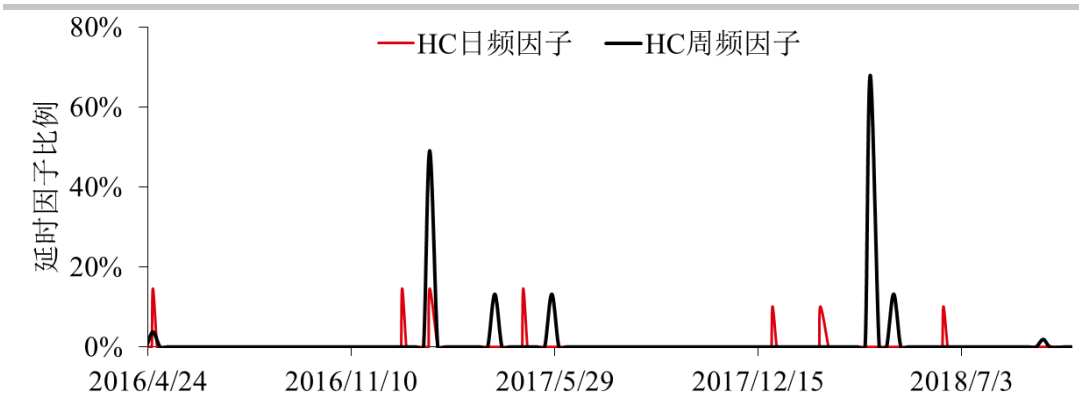
图 5： 螺纹钢至少出现 1 天延时的指标比例



数据来源：钢联

图 6 记录的是热轧卷板所使用的因子至少出现 1 天延时的比例，热轧卷板的日频因子出现延时的比例小于 20%的水平，主要集中在节假日。周频因子出现频率较低，但是数量较多。

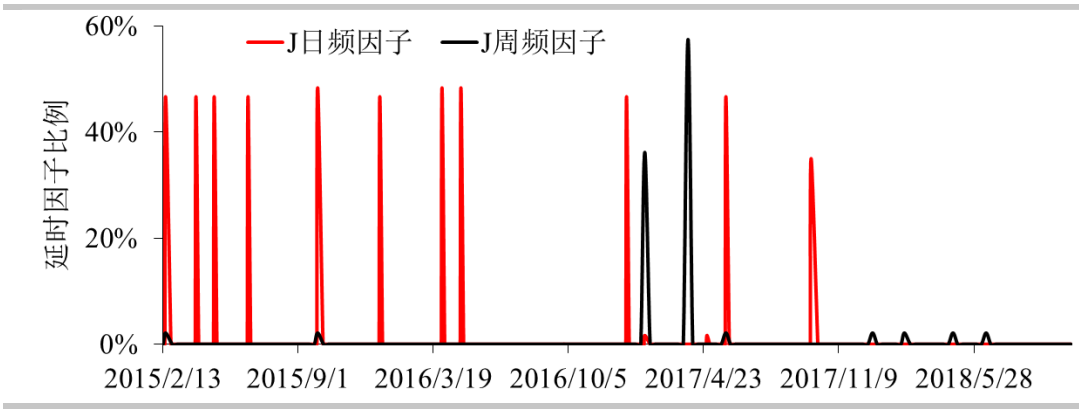
图 6： 热轧卷板至少出现 1 天延时的指标比例



数据来源：钢联

焦炭日频指标出现延时的频率较高，比例也大，比例通常高于 40%。周频因子除了在 2017 年春节和清明节出现较大比例延时外，其他时间出现延时的机会很少。

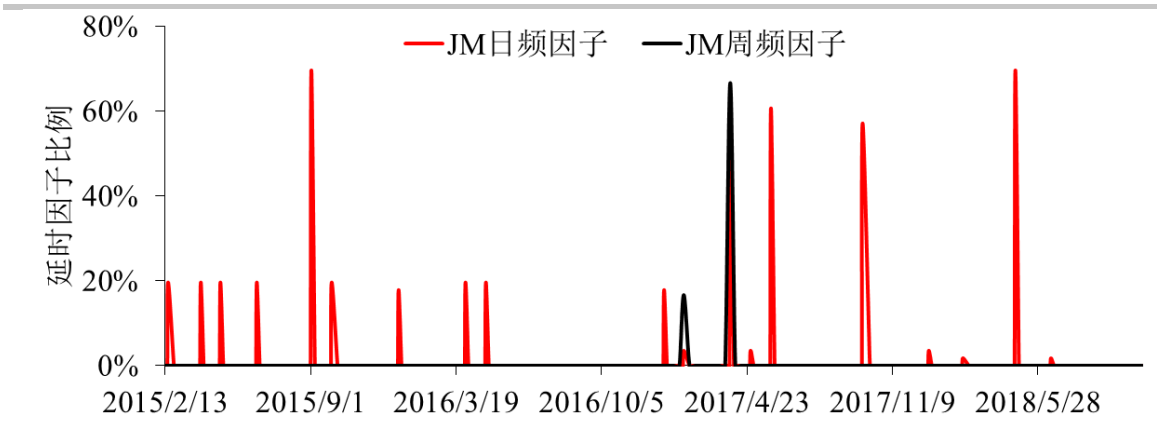
图 7： 焦炭至少出现 1 天延时的指标比例



数据来源：钢联

焦煤日频因子出现延时的频率也较高，比例超过 60%的是 2015 年 9 月 2 日，当天为抗战胜利日假期和 2018 年 4 月 27 日的劳动节假期。

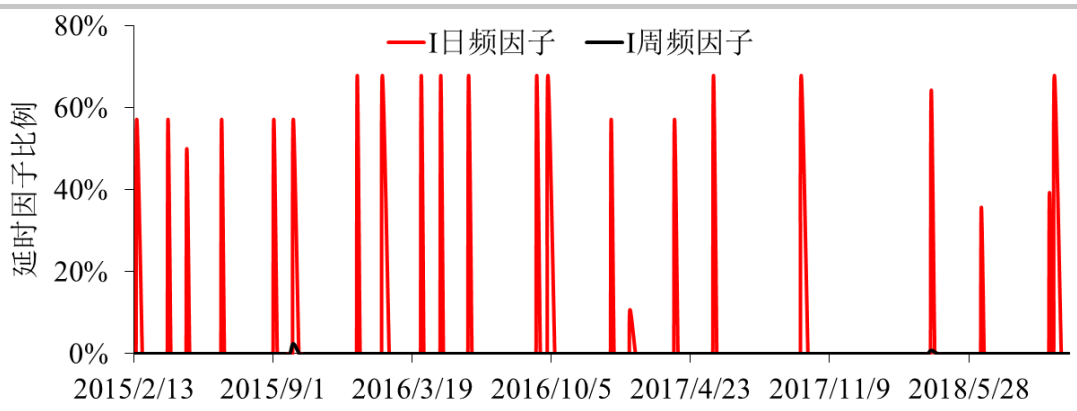
图 8： 焦煤至少出现 1 天延时的指标比例



数据来源：钢联

图 9 是铁矿石的情况，从图可见铁矿石日频因子出现延时的频率会更高一些。在一些短假期里也会延时，例如 2016 年 9 月的中秋节和 2018 年 6 月端午节等日子。周频因子出现延时的情况则较少。

图 9： 铁矿石至少出现 1 天延时的指标比例



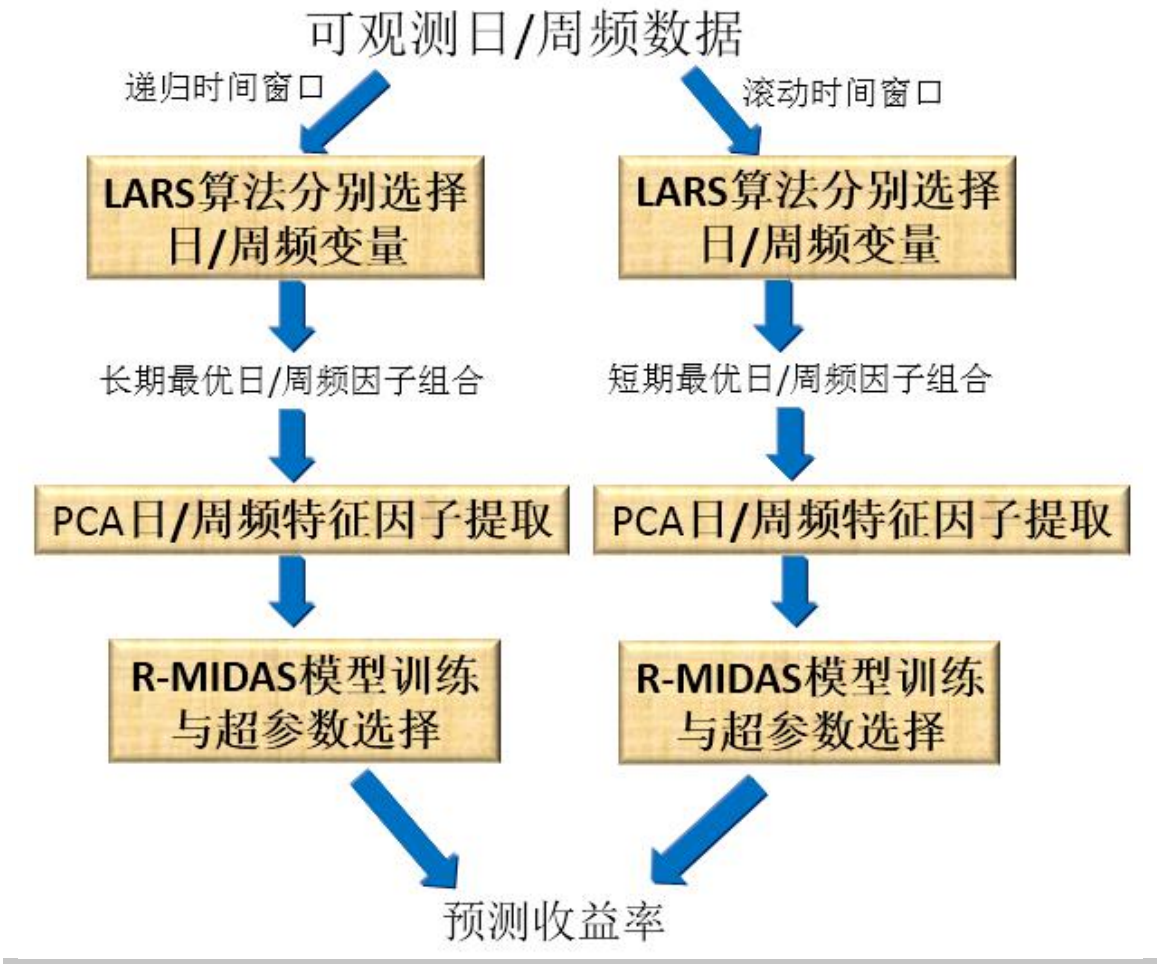
数据来源：钢联

混频基本面量化模型结构

从钢联的数据库可以了解到，黑色的基本面信息包含了多个频率的数据。这一系列相关的因素都会对整个黑色商品板块产生影响。目前使用的混频基本面量化模型只支持周频和日频数据，而且周频数据只能在每周最后一个交易日更新。

图 10 展示了混频基本面量化模型的结构，这个结构在回测时每周都会运行一遍，保证模型更新。首先从收集到的日频和周频数据开始，根据样本的时间标记把样本分成两份，一份包含所有历史数据，即递归时间窗口，而另一份则包含最新的历史数据，即滚动时间窗口。在这两个时间窗口下分别对数据日频和周频因子进行选择，运用最小角回归(LARS)寻找出最优的因子组合。在最优因子组合的基础上进行主成分分析(PCA)，把日频和周频信息进行进一步压缩得到特征因子。从这步开始模型就丧失了经济学意义，变得不可解释了。把得到的日频和周频特征放入 R-MIDAS 模型中训练，由于 R-MIDAS 包含了日度延时、周度延时以及特征维度等参数，这里使用交叉验证的方法对 R-MIDAS 模型进行超参数选择。最后在分别两个时间窗口下选出最优的模型，对其结果进行预测综合，得到最终的预期收益率。以上的方法每周五都会应用到因子选择、模型的更新和模型超参数选择上，这样模型使用的各种参数都是动态变化的，从而避免过度拟合和幸存者偏差等问题，但是经过这一系列处理，模型的可解释性就完全丧失了。

图 10： 混频基本面量化模型结构



数据来源：华泰期货研究院

R-MIDAS 模型原理

R-MIDAS 模型利用不同频率的特征主成份作为输入，进行信息综合，然后给出商品未来一天的收益率预测。Claudia Foroni 等人在 Using Low Frequency Information for Predicting High Frequency Variables 上对此模型做了较详细的描述。模型结构可以如下公式表示

$$\tilde{A}_i(L)X_t = b_i\left(L^{k-i}\right)Y_t + \xi_{it} \quad (2)$$

其中X_t为高频变量向量，Y_t为低频变量向量，$\tilde{A}_i(L)$为周期矩阵，$b_i\left(L^{k-i}\right)$为周期向量。$t = 0 + \frac{i}{k}, 1 + \frac{i}{k}, 2 + \frac{i}{k}, \dots; i = 0, \dots, k-1$。$k$为低频变量包含的高频变量周期数，例如当周频和日频混合时$k = 5$。$\tilde{a}_i(L)$和$b_i\left(L^{k-i}\right)$为延时线性多项式，例如$\tilde{a}_i(L) = \tilde{a}_{i1}L + \tilde{a}_{i2}L^2 + \cdots \tilde{a}_{ip}L^p$。因此$\tilde{a}_i(L)$和$b_i\left(L^{k-i}\right)$都包含周期结构，即周一至周五都有不同的多项式系数。L为高频延时算子，$L^j x_t = x_{t-j/k}$。值得注意的是在应用中$b_i\left(L^{k-i}\right)$的选择可以使得$b_i\left(L^{k-i}\right)y_t$只包含$y_t$的低频项。如果高频项的延时较多，则可能又引入了过多的可调参数，因此这里使用 Heiner Mikosch 等人在 Real-Time Forecasting with a MIDAS VAR 中提出的方法使用 Almon 多项式代

替高频延时项的系数，这个多项式的求解无需进行非线性迭代，求解非常方便，这就是 R-MIDAS 模型。该模型仍然是保持公式(2)所使用的架构，而只是对周期性矩阵$\tilde{A}_i(L)$的矩阵系数添加 Almon 多项式进行约束。Almon 多项式是一个线性系统，多项式在各延时项i上的权重w_i可用如下式子表示

$$w_i(\gamma_0,\dots,\gamma_P)=\sum_{p=0}^{P}\gamma_P i^p \tag{3}$$

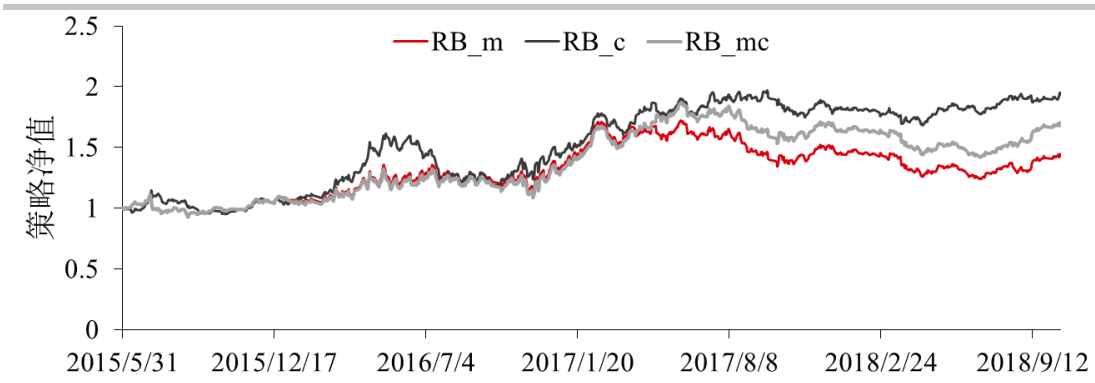
其中$\gamma_0,\dots,\gamma_P$是 Almon 多项式的待定参数，通常为了使w_i在各延时项上呈曲线分布$P\geq 3$。R-MIDAS 模型是结合了 Eric Ghysels，Claudia Foroni 和 Heiner Mikosch 等人的论文得出的，更详细的解释可以参考他们的原论文以及华泰期货 2018 年与 MIDAS 模型相关的量化报告。

策略回测效果

下面分析各个黑色品种策略的收益，由于各个商品品种的因子起始时间不同，期货合约上市时间也不同，所以这里使用各个黑色品种回测的起始时间也不相同，但截止时间都是 2018 年 10 月 19 日。开平仓信号按照模型给出的预测收益率正负来确定，预期收益率为正则做多，负则做空，不会有零持仓的情况。策略使用当天收盘价做预测，下交易日的开盘价做交易，由于这是日频策略，调仓较为频繁，交易手续费的影响也考虑在内。为了方便统一计算，交易手续费设置为单边 2%%。在回测的时候，策略都只使用 1 倍杠杆。起始时间是预留 2 年作为样本，之后开始滚动训练，例如螺纹钢策略的起始时间是 2015 年 5 月，那模型的训练时间就是从 2013 年 5 月开始的。为了对比因子延时对策略收益产生的影响，把数据分成三组第一组是按照标注时间(marked time)进行模型训练和预测，记为 m，这组策略在模型训练和预测时都可能使用到未来数据；第二组是按照录入时间(created time)进行模型训练和预测，记为 c，这组在模型训练和预测时都不会使用到未来数据；第三组是按照标注时间进行模型训练，然后按照录入时间进行预测，由于模型是每周训练，而延时时间通常会少于 1 周，但也可能多于 1 周，所以在模型训练时可能包含了未来数据，在使用模型预测时不会包含未来数据，而只会使用到录入时间领先于标注时间的提前数据。

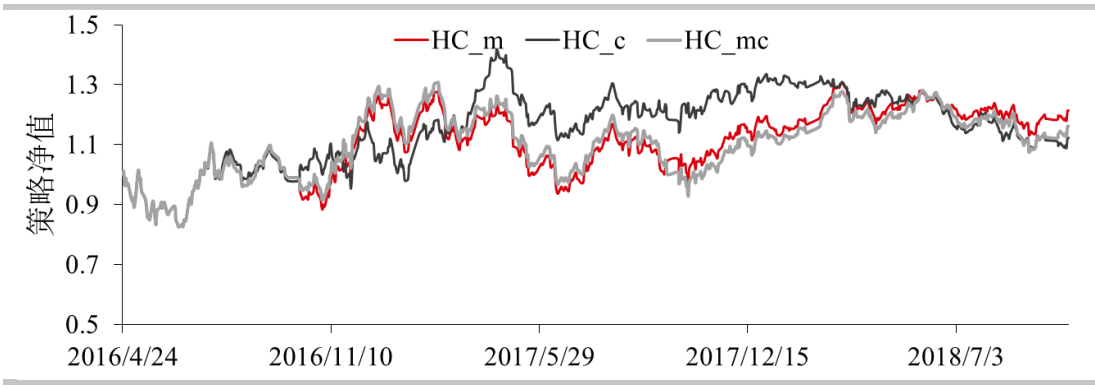
图 11 至 15 作出了 5 个黑色品种的策略净值曲线。螺纹钢和热轧卷板三组的净值曲线走势比较一致，长期来看差异并不显著，可能是因为这两个品种出现延时的因子比例较低。但是在大部分时间里，使用录入时间的策略在这两个品种中表现较好。

图 11：螺纹钢策略净值曲线



数据来源：钢联 华泰期货研究院

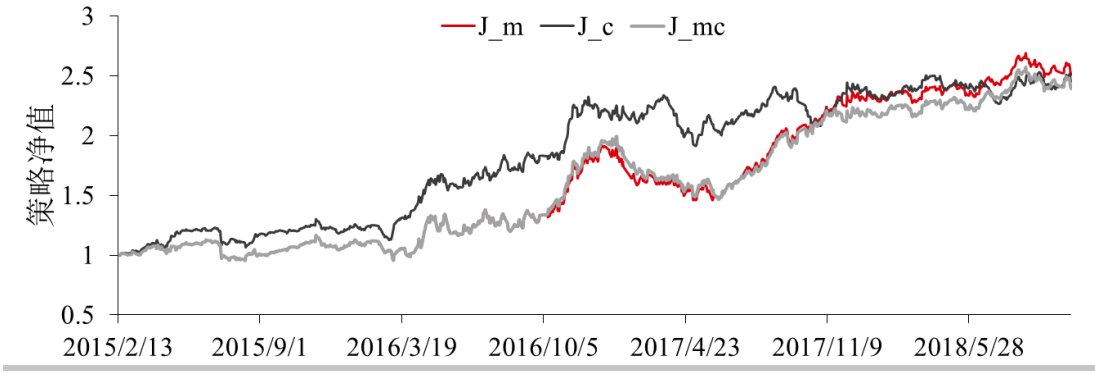
图 12：热轧卷板策略净值曲线



数据来源：钢联 华泰期货研究院

焦炭的情况跟螺纹钢和热轧卷板类似，在 2015 年 2 月至 2017 年 11 月，按照录入时间的策略表现较好，但在 2017 年 12 月以后三者的差别迅速减少。

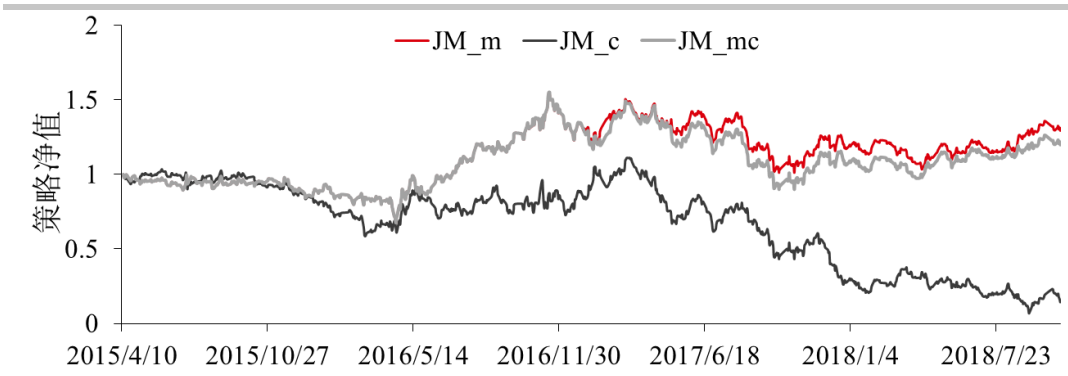
图 13：焦炭策略净值曲线



数据来源：钢联 华泰期货研究院

焦煤的情况跟之前都不一样，使用录入时间的 c 策略亏损严重，而使用标注时间的 m 策略以及 mc 策略则尚能盈利。

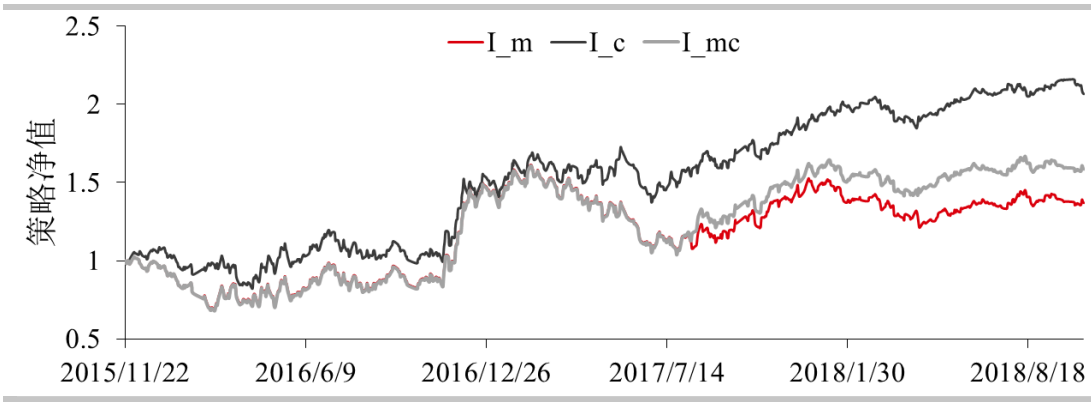
图 14：焦煤策略净值曲线



数据来源：钢联 华泰期货研究院

在铁矿石里使用录入时间的 c 策略长期领先于 m 策略和 mc 策略。使用标注时间的 m 策略收益反而是最低的。对照图 9 可以发现铁矿石的日频因子在节假日存在较频繁的延时，因此按照录入时间训练模型可能更能逼近真实情景。

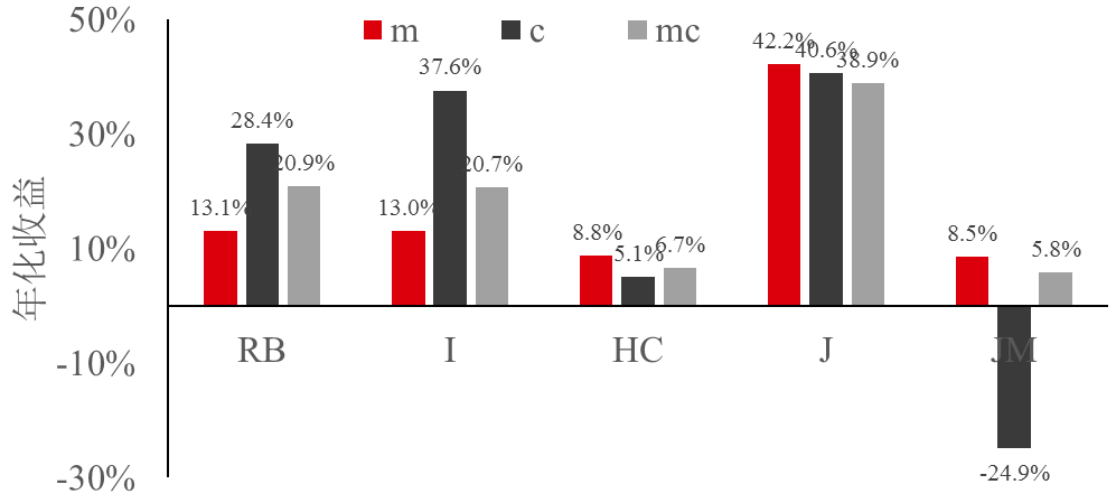
图 15：铁矿石策略净值曲线



数据来源：钢联 华泰期货研究院

图 16 对比了黑色各个品种按照不同时间训练和预测的收益，由图可见并没有哪种数据编排方式占有绝对优势。即使在训练和预测时都使用了未来数据的 m 组，也只在热轧卷板、焦炭和焦煤中表现好。而严格不使用未来数据的 c 组策略在螺纹钢和铁矿石中表现较好，其中的原因可能是因为螺纹钢交易量大，铁矿石较为国际化，这两个品种的市场关注度较高，所以基本面信息流通较快，按照录入时间训练模型和预测能够更加接近市场的真实情景。另外值得注意的是，m 组使用标注时间进行预测在实际交易中是无法实现的，而 c 组也只能在提供了录入时间的数据库中实现，mc 组更为贴近实际交易的情况，但 mc 组整体表现并不差。5 个黑色品种 m 组的平均收益为 17.13%，c 组平均收益为 17.35%，而 mc 组为 18.58%，虽然平均起来差别不大，但 mc 组的平均收益反而是最高的。那说明即使数据库没有提供录入时间，也有可能通过标注时间训练模型同时使用最新数据进行预测获取一定收益。

图 16： 各组策略收益对比



数据来源：钢联 华泰期货研究院

由于数据延时主要集中在长假前后，这里单独考察国庆和春节长假前一周和长假后一周 mc 策略的收益。下表统计了 2015 年春节至 2018 年国庆前后 5 个黑色品种的累计收益，虽然各个品种表现差异较大，但没有一个品种是在长假前后的平均收益是负的。考虑到基本面数据在长假前后延时比例较高，但对混频基本面策略的收益影响似乎有限。其中的原因可能是该策略的输入数据维度较高，部分品种的日频或周频因子就高达百个以上，单个因子对最终收益率的预测影响不大，同时该策略使用了两个时间窗口训练模型，而且对多个模型的输出进行了综合处理，所以单因子的微小波动对最终信号影响有限，这正说明了使用高维度数据预测的优势，也反映了该量化模型的稳定性。另外在长假前后商品波动通常会比平常大，但是单边高波动环境往往才是趋势策略产生收益的根源。从数据结果来看，只有 2015 年国庆和 2016 年春节国庆时，策略收益为负的，其他时间均为正值。因此从数据上似乎没有必要因为波动大而在长假前后停止该量化策略的交易，可是有一点不能忘记的是在构造量化模型的时候，都不可避免地进行过参数优化，长假前后模型表现好并不能排除是幸存者偏差带来的结果，所在长假前后交易还是谨慎为好。

表格 1 各品种 mc 组策略在国庆春节期间的收益

		HC	I	JM	J	RB	平均收益
春节前	20150216-20150217				1.37%		1.37%
春节后	20150225-20150227				-0.77%		-0.77%
中秋国庆前	20150921-20150930			-2.87%	1.30%	4.42%	0.95%
国庆后	20151009-20151008			-1.60%	0.00%	-2.36%	-1.32%
春节前	20160201-20160205		-4.08%	1.22%	1.32%	0.56%	-0.24%
春节后	20160215-20160219		-8.45%	0.23%	0.04%	-1.26%	-2.36%

国庆前	20160926-20160930	-2.95%	-6.06%	8.12%	4.35%	-3.38%	0.02%
国庆后	20161010-20161014	-4.23%	2.77%	0.50%	5.10%	-1.52%	0.52%
春节前	20170123-20170126	3.92%	7.98%	-2.03%	-4.23%	1.32%	1.39%
春节后	20170203-20170210	9.77%	-0.87%	15.10%	-9.56%	19.14%	6.72%
国庆前	20170925-20170929	-4.69%	4.18%	0.99%	8.24%	-5.10%	0.73%
国庆后	20171009-20171013	-1.43%	4.91%	4.28%	1.18%	2.88%	2.36%
春节前	20180205-20180214	-0.90%	1.74%	5.13%	5.05%	-0.38%	2.13%
春节后	20180222-20180302	6.78%	3.69%	-3.07%	0.14%	-2.57%	0.99%
国庆前	20180925-20180928	0.74%	-1.72%	2.28%	-1.26%	2.47%	0.50%
国庆后	20181008-20181012	1.66%	-1.10%	-6.13%	6.59%	0.31%	0.27%
	单品种平均收益	0.87%	0.25%	1.58%	1.18%	1.04%	0.83%

数据来源：华泰期货研究院

结果讨论

这篇报告介绍了钢联 API 数据库的内容和特点，由于这个数据库提供了各个因子的录入时间，这为考察基本面数据延时对量化策略的影响提供了条件。通过对比基本面数据的录入时间和标注时间可以发现，数据延时主要发生在非周六日的公众假期前后，尤其是在国庆和春节，基本面因子都有较大比例的延时，而且与前值通常差异较大，其他时间出现延时的情况较少。

在此基础上构建了三组基本面量化策略，第一组按照标注时间编排数据进行模型训练和预测，第二组按照录入时间进行模型训练和预测，第三组按照标注时间进行模型训练，但按照录入时间进行预测。前两组策略在五个黑色品种表现出较大差异，第三组策略收益在第一二组之间。由于模型使用了的输入因子较多，而且进行了模型预测综合处理，单因子延时影响有限。即使在长假前后该量化策略能在绝大多数情况下取得正收益。

免责声明

公司总部

地址：广东省广州市越秀区东风东路761号丽丰大厦20层、29层04单元

电话：400-6280-888

网址：www.htfc.com