

基于 AI+EI 的长期资产定价模型（附录）

附录 1 Wind-band 计算方法

Wind-band 在软件展示过程中引入了盈利数据，本文为了展示得更加简明，文中绘制的估值 band 均直接对齐相对估值。目前市场中较为常用的 Wind 的 PB-band 计算方法如下：

设定长度为 N 年（通常为三年）的滑动历史时间窗口，在每一个时间点：

1. 计算滑动时间窗口中数据的 50% 分位数（或平均值）作为中轨；
2. 计算滑动时间窗口中数据的最小值作为下轨，并计算中轨和下轨的差值；
3. 中轨加差值得到对应的上轨（生成 5 条分位数的 band 则需将差值除 2 作为 band 间距）。

附录 2 Transformer 及变种的讨论

时间序列预测领域中基于 Transformer 思想设计的 Informer（Zhou H et al., 2021）、Autoformer（Wu H et al., 2021）、FEDformer（Zhou T et al., 2022）、TFT（Lim B et al., 2021）等模型在不同领域展现出了极强的时序依赖捕捉能力。相比于 RNN/CNN 等传统深度学习网络结构，Transformer 结构天然适合对更长的时间序列依赖进行建模，但同时也存在一些问题，如：

1. 传统的自注意力机制只能体现单个时间点之间的数值关联而非时序趋势上的关联；
2. Attention 的计算复杂性为 $O(n^2)$ ，处理较长的时间序列计算效率过低；
3. 时序任务往往需要对趋势项和周期项进行拆分处理。

Autoformer 是基于 Transformer 提出的一种将时间序列分解嵌入深度学习模型的新型时序预测模型。该模型提出的自相关（Auto-correlation）机制首次实现了子序列级别相关性判别的 self-attention 机制，抓住了时间序列数据的主要特征，提升了信息利用程度和计算效率。Autoformer 在能源、交通、经济等领域大幅超越之前的同类模型。

本文认为 Autoformer 是一种更适配中枢问题的神经网络架构，其网络结构中使用自相关模块代替了 Attention 模块，通过快速傅里叶变换算法对子序列相关性计算进行加速，并使用时间序列分解模块对时间序列的趋势项和周期项进行渐进式拆解处理，弥补了传统 Transformer 结构的缺陷。

附录 3 经验分布算法

算法伪代码如下：

输入：时间序列 xs (pd.Series)

水库数据列表长度 R (int, default=100)

数据权重衰减周期 $decay$ (int, default=50)

输出 band 的分位数列表 qs (list, default=[0.1, 0.5, 0.9])

开始计算 band 的起始节点 min_periods (int, default=50)

输出: pb-band (pd.DataFrame)

1. 初始化空的水库数据列表 reservoir

2. 遍历时间序列

For i in range(len(xs)):

从输入时间序列中取出新的数据加入水库数据列表

if 水库数据列表数据量超过 R, 则:

生成 (0, 1) 之间的一个随机数

根据随机数和映射 F, 得到位于[0, R]之间的一个数组下标

删除水库数据列表中该下标对应元素

if i>min_periods:

根据 qs 中的值分别计算水库列表数据中对应分位数

存储到 pb-band 中

Return pb-band

附录图 1 经验分布算法

上述过程中从随机数到列表下标的映射方法如公式 1:

$$\text{key} = R - 1 - \text{Floor}\left[\frac{\ln(1+x*2^{\frac{R}{\text{decay}}})*\text{decay}}{\ln 2}\right] \quad (1)$$

其中, x 为 (0, 1) 区间内的随机数, R 为固定的水库数据列表长度, decay 为衰减周期, Floor 为向下取整运算。水库数据列表下标越小, 按照这种随机数到数组下标映射方式被剔除概率越大。不断加入新数据时, 就实现了“时间序列中的数据离当下时刻越远, 对计算分位数起到的潜在影响越小”的目标。此外, 在实际生产过程中, 由于计算时纳入了模型预测的未来数据, 所以需要在维护原有水库数据的基础上, 在计算分位数时并入未来数据。同时值得指出的是, 这种计算 band 的方式只需 O(n) 的计算复杂度, 所以可以在线计算。

附录 4 微分预期数据生产方式

二级市场中的交易产生了大量的高频市场数据, 但反映基本面变化的财报数据更新频率相对较低。此外, Wind 等数据商还提供了分析师群体对公司未来财务数据预测的平均值 (即一致预期数据)。为了更有效利用这些数据, 研究中需要对分析师预期数据进行预处理以满足后续建模需要。

分析师预期数据与财报披露数据的模式相同，以营收为例，一季度预期包含了 1~3 月的营收数据，而半年度预期则包含了 1~6 月的营收数据。针对这种数据统计，需要引入 TTM（Trailing Twelve Months）的方式对数据进行处理。此外需要注意的是，Wind 中给出的未来年度营收预测（FY1，同理 FY2 表示未来两年度的营收预测）的含义为分析师对该公司最近预测年度的预测营收的算术平均值，其中最近预测年度与个股年报是否披露有关，公布前该年度为上一年（新一年年初还未公布去年的财报），公布后则代表新一年。由此，针对这种数据切换模式，研究中需要对序列数据进行微分和积分操作，将分析师预期数据对应到 TTM 数据上。这种基本面财务数据构造方式基于简单假设，即每天的数据是通过将一年的数据值平均拆分得到的，这一过程可以称为微分。接下来需要得到更精确的高频营收预测数据更新，具体做法是根据观测时间点的不同对 FY1、FY2 进行按比例组合，这一过程称之为积分。按照此过程，对市场层面的交易数据进行降频得到周频级别数据。

本文用到微分预期数据包括，净资产收益率 ROE，盈利增速 NP-growth。相比较财务版本的 ROE、NP-growth 而言，微分预期版本的指标更高频，且由于包含了分析师预期而更加有前瞻性，因此适合用于数值建模。

附录 5 Autoformer_adapted 改进思路

正文第二章模型与方法中简要介绍了模型架构与改进方法，三处改进思路如下。

第一处改进是交换使用序列分解中的周期项和趋势项。我们先回顾 Autoformer 的核心机制，即在编码器和解码器中嵌入多个 Auto-correlation 以及 Series-Decomposition 单元。该机制的作用是在隐空间中对时间序列的趋势项和周期项先进行分离，周期项借助 Auto-correlation 模块基于自相关机制对相似子序列进行加权聚合，趋势项则采用渐进式分解-累加的方法进行提取。这种对趋势项、周期项分别进行处理的方法契合了时间序列数据的使用场景。我们发现，在估值中枢问题中，当我们把序列分解的周期设置到 1 年左右时，序列的周期项通常代表市场各类短期波动带来的扰动变化，而序列的趋势项才代表了长期走势。站在长期资产定价的角度来看，中长期趋势变化的重要程度要远高于短周期的扰动变化。因此我们直接将编码器、解码器中的 Auto-correlation 模块的输入由序列周期项切换为趋势项，使模型在渐进式分解的过程中更关注对趋势项相关性而非周期项相关性的挖掘。在实际金融市场中，由于市场环境等因素的变化，某些行业的估值会出现系统性上移或下移，即数据集分布漂移的情况。我们在实验环节发现，这一步的改进使模型得到的估值可以很好地应对数据漂移。

第二处改进在于数据的预处理上。由于金融价格数据属于典型的非平稳数据，如果直接预测估值，则不可避免地受到数据非平稳的影响。这里我们借用 N-Linear (Zeng A et al., 2023) 模型对数据集的预处理方法，对输入模型的行业估值数据进行预处理，即，输入数据减去序列的最后一个值，使模型可以将关注点从数据的相对变化转向变化值大小，由此减轻数据集

不平稳带来的潜在影响。值得指出的是，避免不平稳的另一种做法是差分，然而由于该问题的重点在于刻画中长期的趋势，所以差分方案相比较 N-linear 有内生性的劣势。

第三处改进在于使用适合金融数据的 loss 函数。金融数据分布通常是肥尾形态，有时还会出现偶发性的异常数据。我们使用 MAE-loss 代替 Auotoformer 中推荐的 MSE-loss 进行训练，从数学的角度讲，MAE 的导数比 MSE 的导数更稳健，从而避免数据集中部分年份的异常数据波动对训练效果产生过大影响。

附录 6 实验详细参数设定

参考 Autoformer 论文以及数据特点，本文在 PB 数据预测实验中采用了如下设定（所有数据均为周频级别）：利用过去四年的 PB、ROE(t1)、NPG(t1)、十年期国债收益率和行业所属板块共 9 维数据，预测未来一年的 PB，即输入 200 个周的 9 维特征，输出 50 个周的 PB 值。在数据集的特征构建上，实验中对代表行业身份信息的板块特征采用了 one-hot 静态编码的方式，针对 31 个申万一级行业，生成 5 维静态标签（即时间序列上全部为同一个值，分别对应了周期、消费、金融、成长和稳定五大板块），行业对应的所属板块赋值 1，其余均为 0，该编码方式相较于一维静态标签（直接按顺序编号 1~5）能够避免模型对编号的数值大小进行错误解读，更好地对相同板块的行业数据进行归类学习。

数据集的划分上，选取从 2007 年 1 月~2018 年 6 月的数据作为训练集，2018 年 6 月~2023 年 7 月的数据作为测试集。传统的 Transformer 模型、Informer 模型、未经修改的 Autoformer 模型采用相同实验参数进行了对比。

实验中的模型参数最终设定如附录表 1 所示。

附录表 1 模型训练参数设定情况

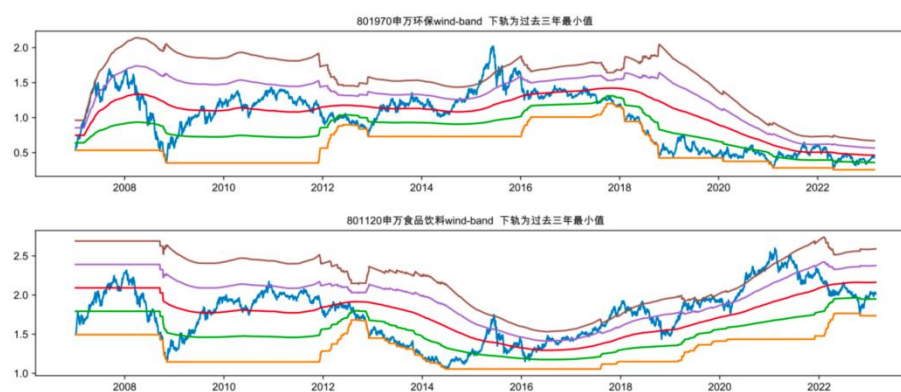
参数	中文解释	值
seq_len	输入序列长度	200
label_len	标签序列长度	100
pred_len	预测序列长度	50
e_layers	Encoder 层数	2
d_layers	Decoder 层数	1
d_model	隐藏层维度	512
moving_avg	MA 剥离窗口长度	13

factor	加权聚合因子	2
attn_head	注意力机制头数量	4

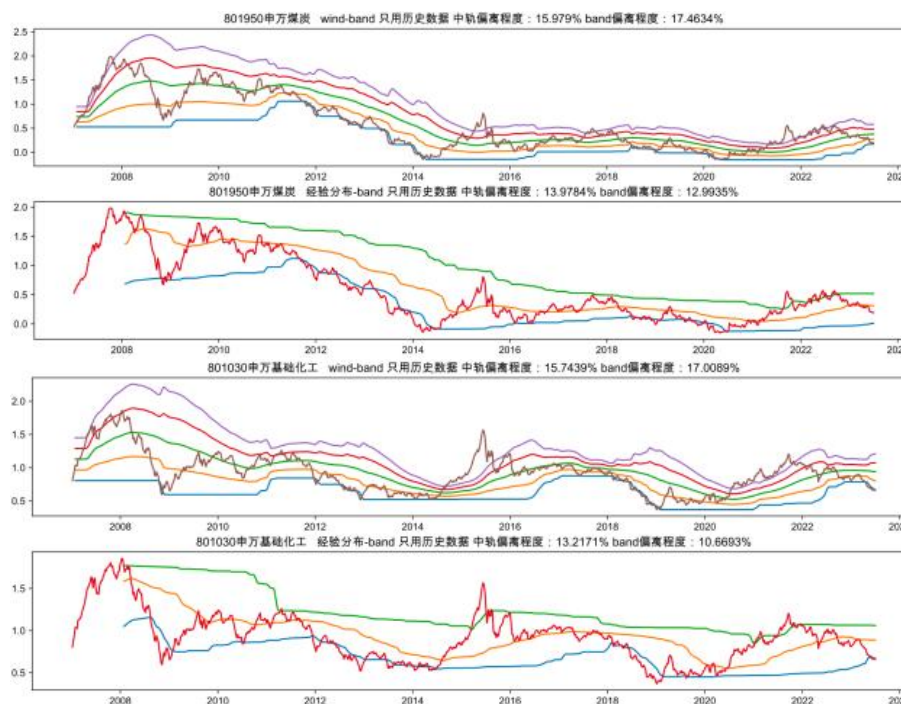
附录 7 击穿估值 band 的探讨

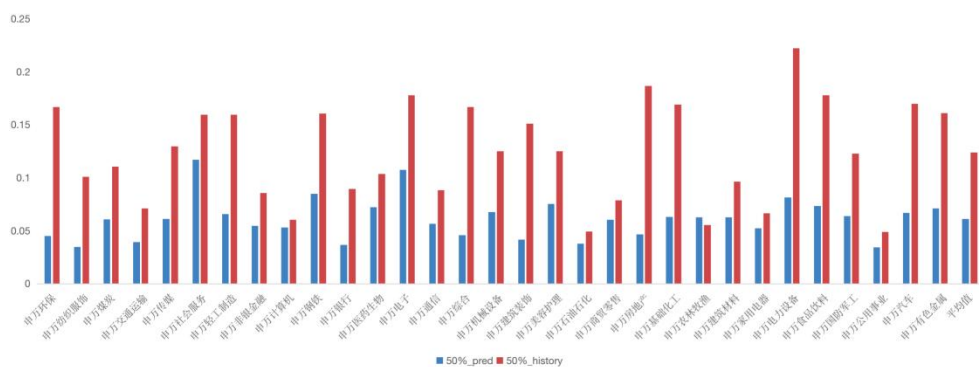
本文所提算法由于具备更强的领先性，故而相比 Wind 的算法并非避免了所有击穿的情况，而是尽可能减少由于基本面等因素变化带来的行业估值持续性漂移所带来的过早击穿 band 的情况（只用历史数据的算法在大幅涨跌的初期会过早地给出交易信号从而带来“假信号”）。经统计，Autoformer_adapted 给出的 band 在测试集区间 31 个行业上仍然有 18.8% 的样本击穿上下轨。因此，模型在尽量避免出现“假击穿”信号的同时，也仍然有相当比例会触发击穿上下轨的情况，进而说明该上下轨的宽度是合理的。

附图1 申万环保、申万食品饮料行业 Wind-band

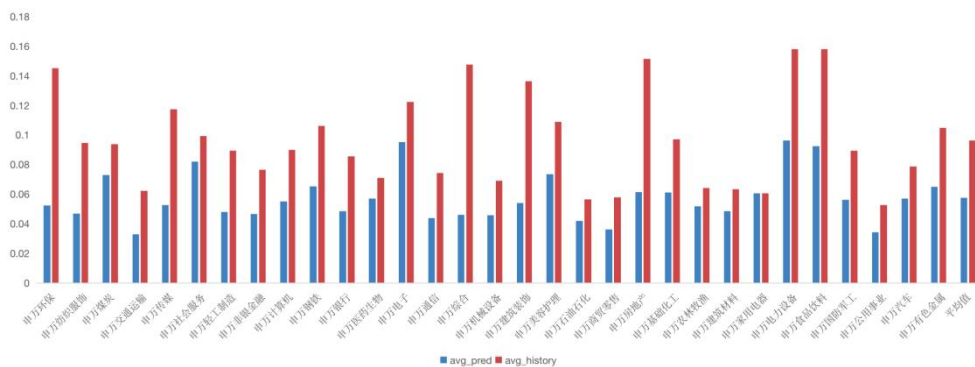


附图2 申万煤炭行业、基础化工行业 Wind-band、经验分布-band 对比
(只用历史数据)

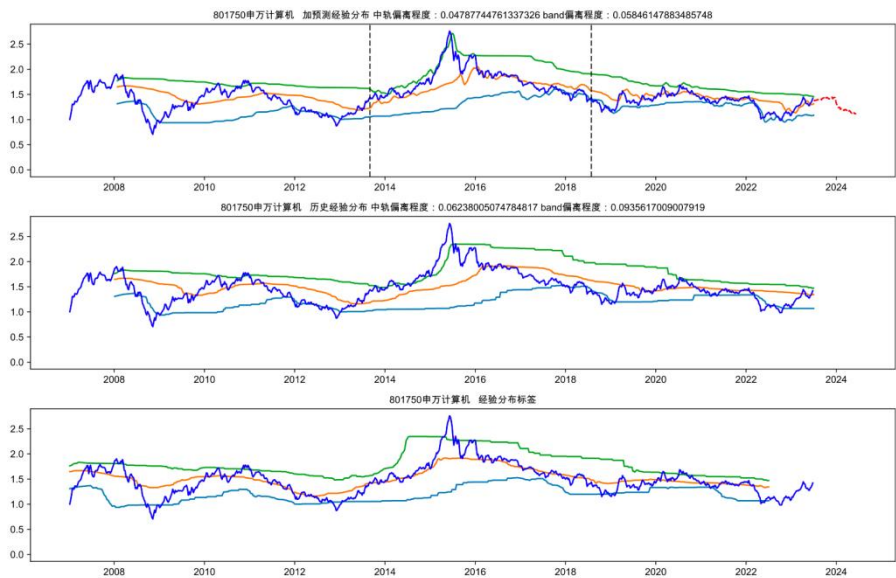




附图 6 申万一级行业（纳入预测数据）经验分布-band 整体偏离统计



附图 7 申万计算机纳入预测数据、只用历史数据、偷看未来数据计算的的经验分布-band 情况对比



附图 8 申万食品饮料纳入预测数据、只用历史数据、偷看未来数据计算的的经验分布-band 情况对比

