

降低波动率预测的低估风险

一种基于残差右尾的后处理方法

Yihua Yue Tongren Xiao

SYSU

July 2026

核心问题

给定一个波动率模型，能否在完全不重新设计其内部结构的情况下，同时实现两个目标：

- 降低预测低估的频率和幅度
- 提升预测精度

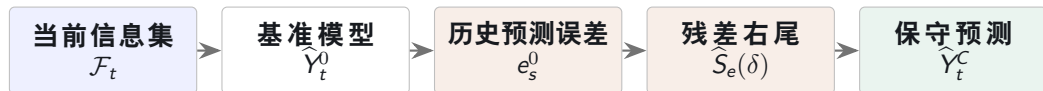
我们起码需要做下面几个事情：

- ① 定义预测目标；选取基准模型；
- ② 定义什么是低估；我们如何从历史预测误差中评估低估风险；
- ③ 如何规避低估

解决思路

先利用基准模型预测波动率，然后加上一个后处理层负责校准低估风险。

模型总览：基准预测加保守输出层



$$\hat{Y}_t^C(\lambda) = \max\{\varepsilon, \hat{Y}_t^0 + \delta_t(\lambda)\}$$

- 后处理层只改变预测输出，不改变基准模型的内部结构；
- $\delta_t(\lambda)$ 是我们基于历史预测偏差以及相应的分布得到的一个调整量，而不是由一个任意常数决定。

所以我们相当于在基准模型后面直接施加一个调整量，以此降低低估风险。令人惊讶的是，这个非常简单的操作，居然是的基准模型在 dow30 和 sp100 两个数据集上，预测精度也同时提升了。

第一步：定义预测目标

定义：未来波动率目标

对资产 i 、预测原点 t 和预测期限 h ，定义未来平均实现波动率为

$$Y_{i,t}^{(h)} = \frac{1}{h} \sum_{j=1}^h RV_{i,t+j}.$$

- $RV_{i,t+j}$: 资产 i 在未来第 j 天的实现波动率;
- h : 预测期限, 例如 1、5、21 或更长交易日;
- 在预测时点 t , $Y_{i,t}^{(h)}$ 尚未实现, 是我们要预测的未来对象。

例

当 $h = 5$ 时, 目标是未来五个交易日实现波动率的平均值, 而不是时点 t 当天的波动率。

第二步：定义信息集与基准预测

定义：时点 t 的信息集

$\mathcal{F}_t$ 表示在时点 t 可以获得的全部信息，例如历史收益率、历史实现波动率、市场状态变量或隐含波动率等等。

定义：基准预测

基准模型给出一个只依赖当前信息的点预测：

$$\hat{Y}_{i,t}^{0,(h)} = g_{i,h}(\mathcal{F}_t).$$

- $g_{i,h}$ 可以是 HAR、GHAR、HAR-IV、GHAR-IV 或其他预测模型；
- 本文方法不要求 $g_{i,h}$ 具有特定形式，只要求它能产生点预测。

第三步：定义残差与低估事件

定义：基准预测残差

$$e_{i,t}^{0,(h)} = Y_{i,t}^{(h)} - \hat{Y}_{i,t}^{0,(h)}.$$

低估： $e^0 > 0$

真实目标为 24，基准预测为 19：

$$e^0 = 24 - 19 = 5.$$

预测少报了 5 个单位。

高估或相等： $e^0 \leq 0$

真实目标为 17，基准预测为 19：

$$e^0 = 17 - 19 = -2.$$

预测高于真实值 2 个单位。

方向约定

残差定义为“真实值减预测值”，因此残差右尾正好对应波动率低估。

第四步：低估风险的两个评价指标

对一组已经实现的评价观测 $(Y_k, q_k)_{k=1}^n$ ，定义。其中， Y_k 是第 k 个评价时点已经实现的真实波动率， q_k 是模型在该时点事先给出的预测值。

$$UF(q) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}\{Y_k > q_k\}, \quad UPS(q) = \frac{1}{n} \sum_{k=1}^n (Y_k - q_k)^+,$$

其中 $x^+ = \max\{x, 0\}$ 。

UF：低估频率

$\mathbf{1}\{Y_k > q_k\}$ 是低估事件指示变量。UF 回答：低估发生得多不多？

UPS：平均短缺

$(Y_k - q_k)^+$ 只保留低估部分。UPS 回答：一次低估平均有多严重？

例

残差为 $(3, -1, 0, 5)$ 时， $UF = 2/4 = 50\%$ ， $UPS = (3 + 5)/4 = 2$ 。

第五步：预测时真实值未知，残差如何获得？



伪样本外残差

对过去每个历史时点 s ，严格模拟当时的信息约束：只用 $\mathcal{F}_s$ 产生 $\hat{Y}_s^0$ ，待对应的真实目标 Y_s 实现后，计算 $e_s^0 = Y_s - \hat{Y}_s^0$ 。

当前预测不使用未来真实值

当前时点只能使用已经形成的历史残差库估计右尾；未来真实值实现后，才会成为更晚时点的历史残差。

第六步：用残差右尾描述低估概率

定义：残差右尾生存函数

对残差随机变量 E ，定义

$$S_E(u) = \mathbb{P}(E > u).$$

当 u 越大，超过该阈值的残差概率通常越小。

样本中的经验右尾

给定历史伪样本外残差 $e_1^0, \dots, e_m^0$ ，经验右尾为

$$\hat{S}_e(u) = \frac{1}{m} \sum_{r=1}^m \mathbf{1}\{e_r^0 > u\}.$$

解释

$\hat{S}_e(2) = 0.20$ 表示历史残差中约 20% 大于 2；这刻画了“预测至少低估 2 个单位”的经验频率。

第七步：定义低估厌恶的目标函数

对于一个可实施预测 q_t ，定义

$$\mathcal{L}_t(q_t) = \mathbb{E}[(Y_t - q_t)^2 \mid \mathcal{F}_t] + \lambda \mathbb{E}[(Y_t - q_t)^+ \mid \mathcal{F}_t]$$

- 第一项是条件平方误差，惩罚所有方向的预测偏差；
- 第二项是条件平均低估短缺，只惩罚 $Y_t > q_t$ 的情形；
- $\lambda \geq 0$ 是低估惩罚权重，越大表示越重视避免低估。

模型目标

不是无条件地把预测抬高，而是在平方误差精度与低估风险之间选择一个有经济含义的平衡点。

目标函数的数值例子

设真实波动率为 $Y = 20$ ，低估惩罚权重 $\lambda = 2$ ：

预测 q	$(Y - q)^2$	$\lambda(Y - q)^+$	总损失
16	16	8	24
18	4	4	8
20	0	0	0
22	4	0	4

含义

当预测与真实值同样相差 2 个单位时，低估预测 $q = 18$ 的总损失为 8，高估预测 $q = 22$ 的总损失为 4。目标函数对方向进行区分。

第八步：把条件期望写成条件分布积分

记给定 $\mathcal{F}_t$ 的条件分布函数为

$$F_t(y) = \mathbb{P}(Y_t \leq y \mid \mathcal{F}_t), \quad \mu_t = \int_{-\infty}^{\infty} y \, dF_t(y).$$

若 $\mathbb{E}[Y_t^2 \mid \mathcal{F}_t] < \infty$, 则平方误差项为

$$\mathbb{E}[(Y_t - q)^2 \mid \mathcal{F}_t] = \int_{-\infty}^{\infty} (y - q)^2 \, dF_t(y).$$

对固定的 q 求导:

$$\begin{aligned} \frac{\partial}{\partial q} \int_{-\infty}^{\infty} (y - q)^2 \, dF_t(y) &= \int_{-\infty}^{\infty} -2(y - q) \, dF_t(y) \\ &= 2 \left(q - \int_{-\infty}^{\infty} y \, dF_t(y) \right) = 2(q - \mu_t). \end{aligned}$$

第八步：低估短缺项的积分形式与严格求导

由 $(Y_t - q)^+ = (Y_t - q)1\{Y_t > q\}$, 有

$$\mathbb{E}[(Y_t - q)^+ | \mathcal{F}_t] = \int_q^\infty (y - q) dF_t(y).$$

利用尾积分公式, 也可写成

$$\mathbb{E}[(Y_t - q)^+ | \mathcal{F}_t] = \int_q^\infty \mathbb{P}(Y_t > u | \mathcal{F}_t) du = \int_q^\infty [1 - F_t(u)] du.$$

若条件分布在 q 处无原子且尾部可积, Leibniz 公式给出

$$\frac{\partial}{\partial q} \int_q^\infty [1 - F_t(u)] du = -[1 - F_t(q)] = -\mathbb{P}(Y_t > q | \mathcal{F}_t).$$

为什么需要“无原子”条件？

若 $\mathbb{P}(Y_t = q | \mathcal{F}_t) > 0$, 短缺项在该点不可微, 只能使用次梯度; 连续条件分布时才可直接写成普通导数。

第八步：合并两项，得到内部最优条件

目标函数的积分表达为

$$\mathcal{L}_t(q) = \int_{-\infty}^{\infty} (y - q)^2 dF_t(y) + \lambda \int_q^{\infty} [1 - F_t(u)] du.$$

在 $\mathbb{E}[Y_t^2 \mid \mathcal{F}_t] < \infty$ 、尾部可积且 F_t 在 q 处无原子的条件下：

$$\mathcal{L}'_t(q) = 2(q - \mu_t) - \lambda \mathbb{P}(Y_t > q \mid \mathcal{F}_t).$$

内部最优解 q_t^* 满足 $\mathcal{L}'_t(q_t^*) = 0$ ，因此

$$2(q_t^* - \mu_t) = \lambda \mathbb{P}(Y_t > q_t^* \mid \mathcal{F}_t).$$

经济解释

左侧是提高预测带来的边际平方误差成本；右侧是提高预测所减少的边际低估短缺成本。最优点使二者相等。

第九步：从最优预测得到保守安全边际

把最优预测写成

$$q_t^* = \mu_t + \delta_t, \quad \delta_t \geq 0.$$

定义条件创新项

$$E_t = Y_t - \mu_t.$$

则

$$\mathbb{P}(Y_t > q_t^* \mid \mathcal{F}_t) = \mathbb{P}(E_t > \delta_t \mid \mathcal{F}_t),$$

从而得到核心方程

$$2\delta_t = \lambda \mathbb{P}(E_t > \delta_t \mid \mathcal{F}_t).$$

核心解释

安全边际越大，左侧的精度成本越大；但边际越大，残差超过它的概率越小。 δ_t 正是两个力量的平衡点。

理论创新项与实际基准残差的关系

理论对象

$$E_t = Y_t - \mu_t, \quad \mu_t = \mathbb{E}[Y_t | \mathcal{F}_t].$$

它围绕真实条件均值定义。

实际可计算对象

$$e_t^0 = Y_t - \hat{Y}_t^0.$$

它围绕基准模型预测定义，可以从历史伪样本外预测中获得。

后处理的识别逻辑

当基准预测 $\hat{Y}_t^0$ 能够较好地逼近条件均值 μ_t 时，基准残差 e_t^0 的右尾可以用来估计创新项 E_t 的右尾。后处理层因此只校准方向性风险，不重新承担全部波动率动态建模任务。

第十步：用经验残差右尾计算候选边际

理论方程中的条件概率由历史残差右尾替代：

$$2\delta_{i,t}^{(h)}(\lambda) = \lambda \widehat{S}_{e,i}^{(h)}\left(\delta_{i,t}^{(h)}(\lambda)\right).$$

其中

$$\widehat{S}_{e,i}^{(h)}(u) = \frac{1}{m_{i,h}} \sum_{r=1}^{m_{i,h}} \mathbf{1}\{e_{i,r}^{0,(h)} > u\}.$$

- 每个资产 i 可以有自已的残差右尾；
- 每个期限 h 可以有自已的残差右尾；
- 因而调整量是资产—期限特定的，而非全市场统一常数。

第十一步：最终后处理映射

定义保守预测为

$$\hat{Y}_{i,t}^{\mathcal{C},(h)}(\lambda) = \max \left\{ \varepsilon, \hat{Y}_{i,t}^{0,(h)} + \delta_{i,t}^{(h)}(\lambda) \right\}.$$

- $\hat{Y}_{i,t}^{0,(h)}$: 基准模型的原始点预测;
- $\delta_{i,t}^{(h)}(\lambda)$: 由历史残差右尾得到的保守边际;
- λ : 控制低估厌恶程度的模型参数;
- $\varepsilon > 0$: 保证预测保持正值, 适用于波动率预测和 QLIKE。

最重要的结构性质

这是一个加法后处理层: 基准模型产生什么预测, 后处理层就在其基础上增加一个由尾部风险决定的调整量。

一个完整的数值例子

设当前基准预测为 $\hat{Y}_t^0 = 18$, 历史残差近似服从 $E \sim \text{Unif}[-3, 3]$ 。对 $0 \leq \delta \leq 3$, 有

$$\mathbb{P}(E > \delta) = \frac{3 - \delta}{6}.$$

若 $\lambda = 6$, 核心方程给出

$$2\delta = 6 \frac{3 - \delta}{6} \implies \delta = 1.$$

因此最终预测为

$$\hat{Y}_t^C = 18 + 1 = 19.$$

真实值 $Y = 20.5$:

基准短缺 = 2.5,

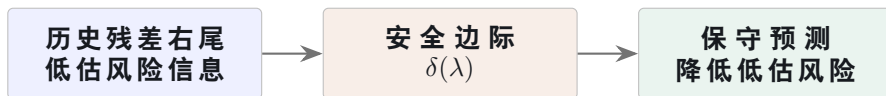
保守预测短缺 = 1.5。

真实值 $Y = 17.5$:

基准高估 = 0.5,

保守预测高估 = 1.5。

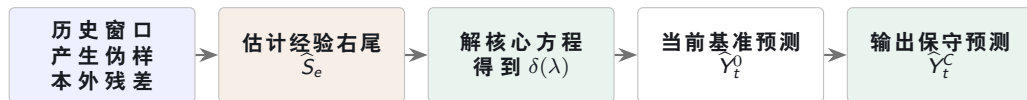
模型的经济含义：风险—精度权衡



核心权衡

提高 λ 通常会提高 δ ，从而降低低估频率和短缺；但过大的边际会增加高估和对称点损失。因此， λ 的作用不是“越大越好”，而是控制风险偏好的旋钮。

模型架构的严格时间顺序



当前汇报的边界

这里讲清楚模型对象和逻辑关系；具体训练窗口、验证集搜索和实证检验属于模型估计与实证部分，不改变上述架构。

最后总结：整个模型只有三层

第一层：基准预测

$$\hat{Y}_{i,t}^{0,(h)} = g_{i,h}(\mathcal{F}_t).$$

利用当前信息预测未来波动率目标。

第二层：残差右尾校准

$$2\delta_{i,t}^{(h)}(\lambda) = \lambda \hat{S}_{e,i}^{(h)}(\delta_{i,t}^{(h)}(\lambda)).$$

利用历史伪样本外残差估计低估风险并求解保守边际。

第三层：保守输出

$$\hat{Y}_{i,t}^{C,(h)}(\lambda) = \max\{\varepsilon, \hat{Y}_{i,t}^{0,(h)} + \delta_{i,t}^{(h)}(\lambda)\}.$$

谢谢！

模型主线

从基准预测出发，用历史残差右尾估计低估风险，通过非对称目标函数确定保守边际，最后将边际加回基准预测。

Questions and Discussion