非对称性和尾部 Alpha

aglv

2025 年 3 月 16 日

1 General

1.1 价格序列处理

尝试使用对数收益率来作为收益率值 $\rightarrow CVaR_{neg}$ 在 230101-240101 上表现不佳

1.2 标准化

对同一股票的前一段时间数据作为样本 or 某个时间截面针对所有股票计算均值标准差
按照 axis=1 求标准化，求得效果与原数据相同
取 rank？

1.3 离群值处理

1.4 缺失值处理

2 波动与收益率

《收益率的非对称分布与尾部蕴含的 Alpha》东方证券

2.1 Skewness

偏度 (skewness):z-score 后的三阶矩

$$Skewness = E[(\frac{X-\mu}{\sigma})^3]$$

remark:
偏度大小很受离群值的影响，三次方加大了偏差，实际上关系到了尾部分布

改进？

2.1.1 中性化后的特质偏度

$$Skew_{i,t} = \frac{1}{n}\sum_{d\in S(t)}\left(\frac{R_{i,d} - \bar{R}_i}{\hat{\sigma}_i}\right)^3$$

where $S(t)$ and n are the set of trading days and the number of tradin
$\bar{R}_i$ and $\hat{\sigma}_i$ are the sample mean and the sample standard deviation c

Idiosyncratic skewness is calculated following Boyer et al. (201
shown in Eq. (8) to daily stock return data observed during the pre
security i at the end of month t is calculated as follows:

$$IdioSkew_{i,t} = \frac{1}{n-2} \times \frac{\sum_{d\in S(t)} \varepsilon_{i,d}^3}{IdioVol_{i,t}^3}$$

where $IdioVol_{i,t}$ is calculated following Eq. (7).[11]

Finally, to calculate coskewness, we follow the approach discus
regression:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_{i,m}(R_{m,t} - R_{f,t}) + \varepsilon_{i,t}$$

where $R_{i,t}, R_{m,t}$, and $R_{f,t}$ are the monthly return of stock i, the mont
monthly risk-free rate at the end of month t, respectively. Then, f
for stock i in month t, denoted $CoSkew_{i,t}$, is estimated as follows:

$$CoSkew_{i,t} = \frac{E\left[\varepsilon_{i,t}(R_{m,t} - R_{f,t})^2\right]}{\sqrt{E\left[\varepsilon_{i,t}^2\right]E\left[(R_{m,t} - R_{f,t})^2\right]}}$$

where $\varepsilon_{i,t}$ is the regression residual calculated from Eq. (11) above.

特征偏度：ϵ_i 表示 Fama-French 三因子中性化之后的收益率

日内特质波动率：

$$iDVol_t = \sqrt{\sum_{i=1}^{N} \epsilon_{t,i}^2}$$

特质偏度:

$$iDSkew_t = \frac{\sqrt{N}\sum_{i=1}^{N} \epsilon_{t,i}^3}{iDVol^{3/2}}$$

2.1.2 ：由于偏度对离群值敏感，考虑一定程度上对样本去离群值后计算偏度

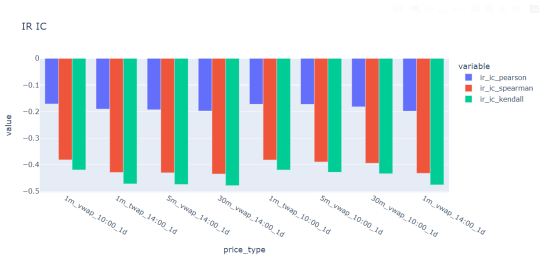
Method1: 用均值-方差筛选:

$$\text{新的样本空间：} \quad \{x|\mu - 2\sigma < x < \mu + 2\sigma\}(\text{正态时，包含 } 95\% \text{ 数据})$$

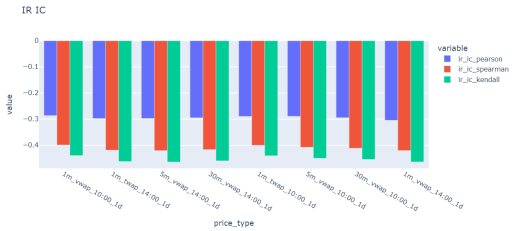
Method2: 直接简单去除极大值和极小值，用分布数筛选

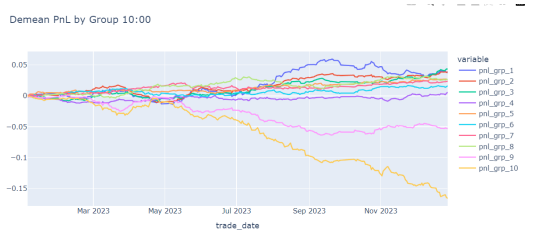
$$\text{新的样本空间:}\{x|X_{0.05} < x < X_{0.95}\}$$

基本形式的 skew, 数据范围 230101-240101:



简单去除一部分极大极小值的 skew:





2.1.3 考虑其他形式的类 Skewness

描述样本统计分布特征的统计量除了均值和标准差之外，还有中位数和四分位等，尝试把他们加到偏度计算中:

Groneveld and Meeden's coefficient:

$$skew(X) = \frac{\mu - v}{\mathbb{E}(|X - v|)}$$

其中 v 是中位数

考虑使用顺序统计量的 L-Momnets:

$$\lambda_r := \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} \mathbb{E}(X_{r-k:r})$$

其中 $X_{k:n}$ 表示 k 个 X 样本的第 K 顺序统计量

$$L - skewness := \lambda_3/\lambda_2$$

2.2 E_ϕ

E_ϕ: 左右尾概率作差表示分布的非对称性

$$E_\phi = P(x \geq k) - P(x \leq -k)$$

2.3 S_ϕ

S_ϕ: 基于熵构造原理 其中

以用于衡量一个分布是否对称。但由于单纯的距离无法体现分布非对称性的方向，即无法区分分布是左偏还是右偏，所以 Jiang 等人（2020）在距离前乘以E_φ因子的符号$Sign(E_\varphi)$，使得S_φ加入了非对称性方向的考量。

$$S_\varphi = Sign(E_\varphi) * \frac{1}{2}\left\{\int_{-\infty}^{-k}(f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}})^2 dx + \int_{k}^{+\infty}(f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}})^2 dx\right\}$$

S_φ定义中的x和k的含义与E_φ相同。由于$f_1(x)$、$f_2(x)$的真实分布未知，所以对二者的估计采用非参数估计中的核密度估计法，核密度函数选取高斯核函数。

$$\widehat{f(x)} = \frac{1}{nh}\sum_{i=1}^{n}\kappa(\frac{r_i - x}{h})$$

$$\kappa(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$$

$$h = 1.06\sigma n^{-\frac{1}{5}}$$

简化求 f:

$$\frac{\partial k(\frac{r_i - x}{h})}{\partial h} = -(\frac{r_i - x}{h})k(\frac{r_i - x}{h})$$

$$\frac{\partial^2 k(\frac{r_i - x}{h})}{\partial h^2} = (\frac{r_i - x}{h})^2 k(\frac{r_i - x}{h}) + \frac{r_i - x}{h}k(\frac{r_i - x}{h})$$

于是由于 rolling 时，会改变其中一个样本点引起的 h 的变化可以用二阶 Taylor 公式逼近，在 rolling 时计算 f 的速度提高(但是实际表明，其实这一步的时间成本并不算太大，计算复杂度主要由于积分的计算)

化简计算:

$$X_2 = -X + 2\mu_X$$

$$\begin{aligned} P(X_2 < x) &= P(-X + 2\mu_X < x) \\ &= P(X > 2\mu_X - x) \\ &= 1 - F(2\mu_X - x) \end{aligned}$$

$$f_2(x) = f(2\mu_X - x)$$

记 S_ϕ 的被积函数为 g

$$\begin{aligned} g &= (f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}})^2 \\ &= f(x) + f(2\mu_X - x) - 2[f(x)f(2\mu_X - x)]^{\frac{1}{2}} \end{aligned}$$

则

$$\begin{aligned} S_\phi &= \frac{1}{2} * sign * (\int_{-\infty}^{-k} g + \int_{k}^{\infty} g) \\ &= sign * (2 - (\int_{-\infty}^{-k}[f(x)f(2\mu_X - x)]^{\frac{1}{2}}dx + \int_{k}^{\infty}[f(x)f(2\mu_X - x)]^{\frac{1}{2}}))dx) \end{aligned}$$

$\Rightarrow$ 只需求一个密度函数，但是求积运算时间复杂度过大

计算复杂度优化:
考虑到计算积分的时间成本，将

$$sign(E_\phi) \to sign(skew)$$

由于被积分的差值函数 g 关是关于 y 轴对称的两个函数的差值平方，其值应该起伏变化较大，将求积分运算转化为关注极大值的均值

$$\text{被积差值函数的积分} \to E[i|i > series_k \quad for\ i\ in\ \text{diff-series}]$$

其中 $series_k$ 表示 series 序列的 k 分位数

2.4 $Asym_p$

$Asym_p$:在分布接近对称分布时表现更好

$$Asym_p = \begin{cases} -corr(f,F) & if \quad 0 < var(f) \\ 0 & if \quad var(f) = 0 \end{cases}$$

remark:
1、怎么计算两个函数之间的相关关系？(person/spearman):
→ 使用 f 和 F 作用于实际收益率的分点上的两个序列的相关系数
2、对于 F 的计算:
idea1: 在求积分端努力:
更换数值求积方法 (Newton-Cotes → Romberg),
减小求积区域:$(\infty, x] - > (\text{某个小值}, x)$
加大求积精度: 精度设置为 1e-2 时停止求职
idea2: 视作 iid 序列用中心极限定理做近似
idea3: 采用统计分布数来替代，定义经验分布函数:

$$F(x) = \frac{1}{n}\sum_{i=0}^{n} \mathbb{I}(X_i < x)$$

idea4
引入 Chebyshev 多项式:

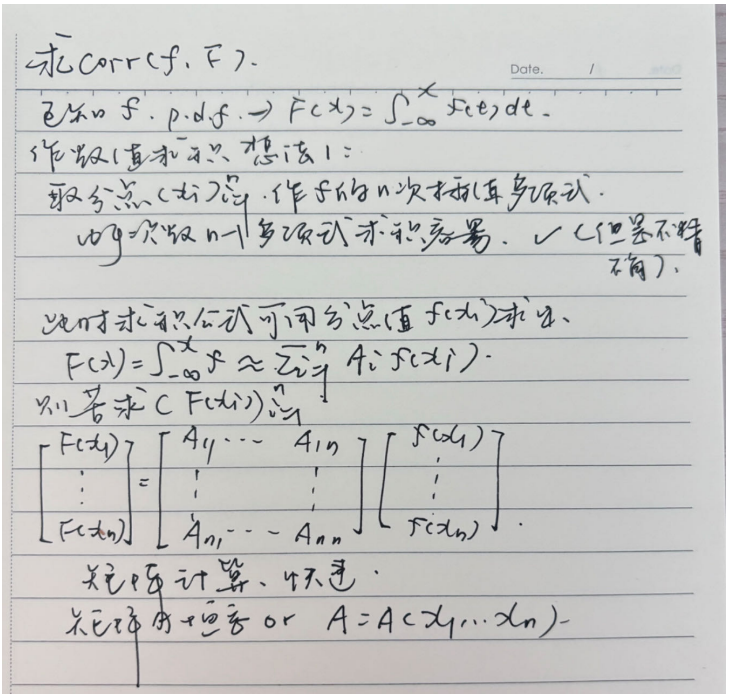
$$T_n(x) = \cos(n\, arccos x)$$

在 [-1,1] 上，取内积 $< f, g >= \int \frac{1}{\sqrt{1-x^2}} f(x)g(x)\, dx$，Chebyshev 多项式族在此内积下正交，由 Fourier 理论知 Chebyshev 多项式族给出了多项式函数关于任意实函数的最优二次逼近，于是用 Chebyshev 的分点做数值求积，得到的积分的精确度高

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx \approx \sum_{i=1}^{n} \omega_i f(x_i)$$

其中:

$$\omega_i = \frac{\pi}{n}; \quad x_i = \cos(\frac{2i-1}{2n})\pi$$



对于一般的积分区域 [a,b]:

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f(\frac{b-a}{2}t + \frac{b+a}{2})\,dt$$

由于 x_i 是收益率序列，分布在 0 附近，于是可以考虑收缩分布函数的积分区域从 (∞, x_i) 到 $(-10, 0)$ 此时权重 ω_i 恒定，上图的矩阵 A 固定，计算速度大大加快 or 不使用数据样本的 x_i 当做计算 corr 的数据，而是直接固定一组分点 $(yj)_{j=1}^m$, 此时只用计算 m 个权重矩阵 A_j

2.5 CVaR

$$CVaR_- = E[X|(X < VaR)\&(P(X < VaR) = 5\%)$$

$$CVaR_+ = E[X|(X > VaR)\&(P(X > VaR) = 5\%)$$

idea1：

是否能考虑将算术求和的线性期望改变成其他形式的期望来增加对极端情况的厌恶

e.g. 考虑扩大求期望范畴，即扩大上面限制条件 $P(X < VaR)$ 的值但是减少较为中心值的权重(考虑线性加权和指数加权)

idea2:

2.6 Tail Beta

：

$$\widehat{\beta_j^T} := \tau_j\widehat{(k/n)}^{1/\widehat{\alpha_m}} \frac{\widehat{VaR_j(k/n)}}{\widehat{VaR_m(k/n)}}$$

令$X_t^{(m)} = -R_{m,t}^e, t = 1,2\cdots n$，$n$为计算期间内交易日天数。然后将市场超额收益的相反数$X_t^{(m)}$从小到大排序，得到$X_{n,1}^{(m)} \le X_{n,2}^{(m)} \le \cdots \le X_{n,n}^{(m)}$。$X_t^{(j)}$、$X_{n,i}^{(j)}$定义方式同上。取$k$为$n$个交易日中$R_m^e$损失超出其 VaR 值的天数（$k \approx 0.05n$），则

$$\frac{1}{\widehat{\alpha_m}} = \frac{1}{k}\sum_{i=1}^{k} log\, X_{n,n-i+1}^{(m)} - log X_{n,n-k}^{(m)}$$

$$\tau_j\widehat{\left(k/n\right)} = \frac{1}{k}\sum_{t=1}^{n} 1_{\left\{X_t^{(j)} > X_{n,n-k}^{(j)}\ and\ X_t^{(m)} > X_{n,n-k}^{(m)}\right\}}$$

$\widehat{VaR_j(k/n)}$、$\widehat{VaR_m(k/n)}$分别表示股票j和市场第$k+1$大损失。尾部 beta 因子表示当市场出现极端下跌行情时，个股收益对市场收益的敏感性。若尾部 beta 越大，表示个股收益对市场的极端负收益越敏感，使得收益率低于尾部 beta 偏小的股票。

$$R_j^e = \beta_j^T R_m^e + \varepsilon_j, R_m^e < -VaR_m(\bar{p})$$

3 Further Topic

3.1 增加即时项

研报中的因子构建都是基于向前 rolling 一段时间 (记住 T) 的数据，每个数据在计算中的地位相当。因此磨灭了数据的时序特征，因子在 t 时间的因子值代表的是 [t-T+1,t] 这一段时间的分布的非对称性或者尾部特征，滞后性较强 (特别地，对于尾部特征的因子常常出现一段时间因子值不变的情况)，

在我们的假设下，[t-T+1,t] 的数据是随机变量 X_t 的 n 个样本，因此我们可以用这些数据来估计分布的性质，为了减弱计算的滞后性，我们通常需要减少窗口值 T，但是也使得估计的模糊性加大

当因子值 f_t 与 f_{t-1} 相同时，以尾部特征因子 CVaR 为例，说明样本 x_t 在样本 $x_{t-T} - x_{t-1}$ 中较为居中，尾部特征较差。我觉得应该视其为一个后验的信息，对原先计算的因子值做出调整

当 $f(t) < f(t-1)$ 时说明在 $t-T$ 时刻出现一个尾部值，或者 t 时刻的尾部值小于 T 时刻，而在经过了 T 时间后数据 t-T 后依然居于尾部，说明分布的尾部性较差，应该惩罚减小 f(t) 值，且惩罚系数性格大于相同的情形；当 $f(t) > f(t-1)$ 时，说明 t 时刻出现了一个尾部值，说明分布的尾部性较好，应该适当奖励增大 $f(t)$

于是我们需要找到一个激活函数 $s(x)$, 使得 $x \le 0$ 时 $s(x) < 0$, $x > 0$ 时 $s(x) > 0$ 考虑选取最简单的分段函数：

$$s(x) = \begin{cases} x & if \quad x > 0 \\ -0.5 & if \quad x = 0 \\ -1 + x & if \quad x < 0 \end{cases}$$

则：

$$f_t = f(t) + \lambda s(f(t) - f(t-1)) * f(t)$$

其中 λ 是惩罚系数，初步考虑设置在 0.05 左右
想法本质上来源于想增加样本 x_t 在计算因子值 f_t 的权重

3.2 重要性采样 Important Sampling

用重要性采样加速类似 Monte Carlo 方法从密度函数计算分布函数的收敛速度