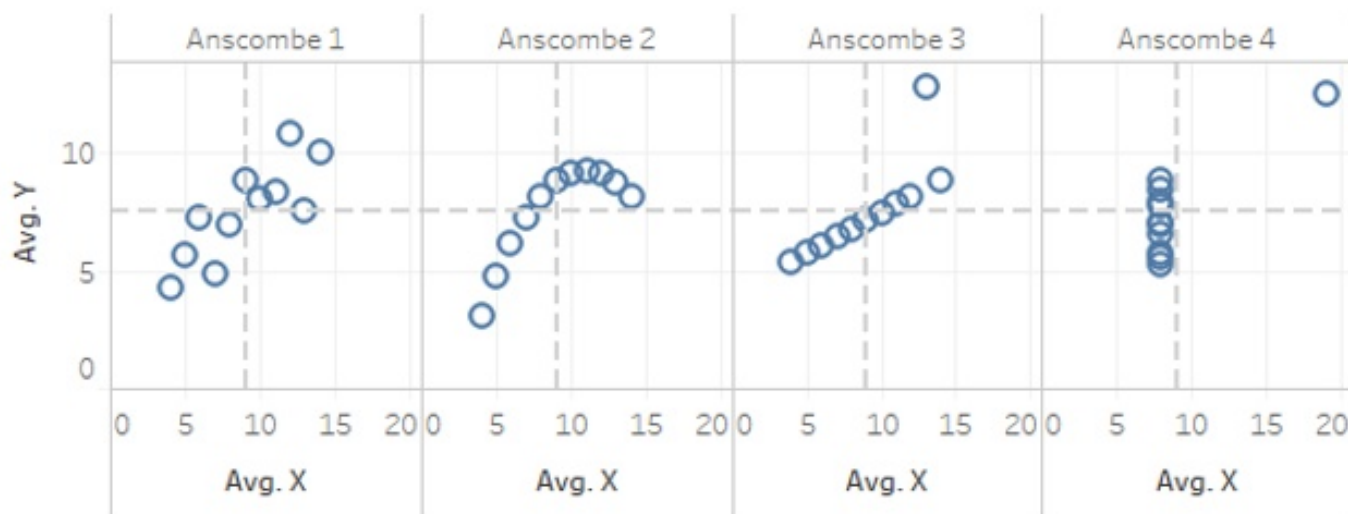


知乎

首发于
川流不息

Anscombe quartet



用 IC 评价因子效果靠谱吗？



石川

量化交易 话题的优秀回答者

已关注

淮浩、王坚等 85 人赞同了该文章

摘要

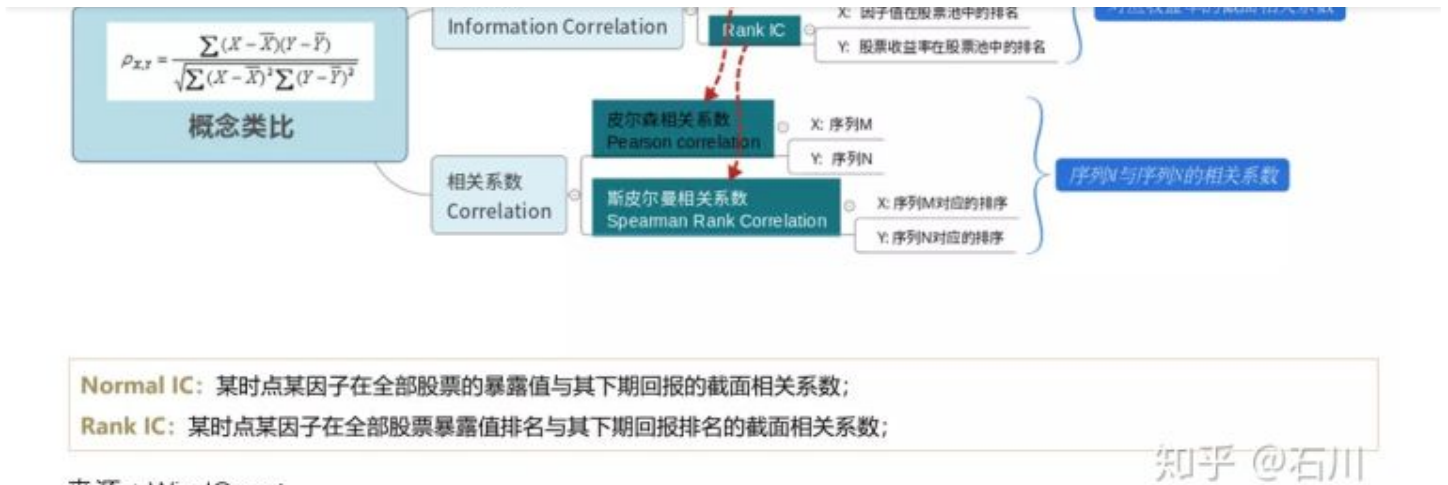
传统的 IC 或者 Rank IC 在评价因子选股效果时不够合理，有一些陷阱。基于 IC 进行因子配置不十分靠谱。本文提出对 IC 的一些改进，并建议使用加权 IC 来评判因子效果。

1 IC 和 Rank IC

在多因子选股实务中，人们热衷于动态评价因子在单期截面上的选股效果。为实现这个目标，通常的做法是用当期个股的因子取值（记为 x ）和下一期个股的收益率（记为 y ）在截面上计算信息系数（information correlation），简称 IC。IC 的计算方法通常有两种： x 和 y 的相关系数，以及 x 和 y 的秩相关系数（见下图）。第一种就是我们常说的 IC，第二种可以称作 Rank IC。



知乎

首发于
川流不息

这里简单介绍下秩相关系数。

秩相关系数 (rank correlation coefficient) 和相关系数类似，不同的是它考察的是两个随机变量之间的**单调相关性 (monotonic correlation)**。秩相关性对变量之间的线性或非线性的相关性不做假设。在计算秩相关系数时，使用的并不是观测值本身的数值，而是它们在各自样本中的排序。秩相关系数的取值在 -1 到 1 之间。

在统计学中，有多种计算秩相关系数的方法，其中最流行的要数 Spearman 秩相关系数，它以 Charles Spearman 命名。假设有两个随机变量 x 和 y 的 n 对儿观测值，Spearman 秩相关系数 r_s 的计算过程如下：

1. 首先将 x 和 y 的观测值转换成它们对应的排序 x_r 和 y_r 。
2. 对 x_r 和 y_r 采用传统的线性相关系数公式，则可得到 r_s ：

$$r_s = \frac{\text{cov}(x_r, y_r)}{\sigma_{x_r} \sigma_{y_r}}$$

下图是某因子在一段时间内的滚动 Rank IC 移动平均，从中我们能对常见选股因子 IC 的取值范围有个大概的了解。



知乎

首发于
川流不息

即 IC 的均值除以标准差) 的高低来动态进行因子的配置。

上面这些用法的核心前提是 IC 能够正确反映因子选股的能力。然而，真的是这样吗？**如果这个核心前提不成立，那么基于 IC 的各种因子择时、因子配置、因子打分恐怕难言靠谱。**

2 IC 中的陷阱

本节通过一个假想的例子说明 IC 和 Rank IC 计算中存在的陷阱。假设有十支股票，它们的因子取值从大到小如下表所示。此外，考虑这十支股票的两组假想的收益率序列。

因子取值	收益率序列一	收益率序列二
1.0	2.00 %	-0.80 %
0.8	1.60 %	-0.40 %
0.6	1.20 %	0.40 %
0.4	-2.00 %	0.80 %
0.2	-1.60 %	1.20 %
-0.2	-1.20 %	1.60 %
-0.4	-0.80 %	2.00 %
-0.6	-0.40 %	-1.20 %
-0.8	0.40 %	-1.60 %
-1.0	0.80 %	-2.00 %

很容易计算该因子和这两组收益率序列的相关系数均为 0.2909。**如果仅仅看 IC 这个单一指标的话，我们会认为该因子在当期的选股能力很不错。**但 IC 背后还有很多故事可讲。我们不妨把因子和这两组收益率序列画出来，并各自做一条线性回归线来看一看。

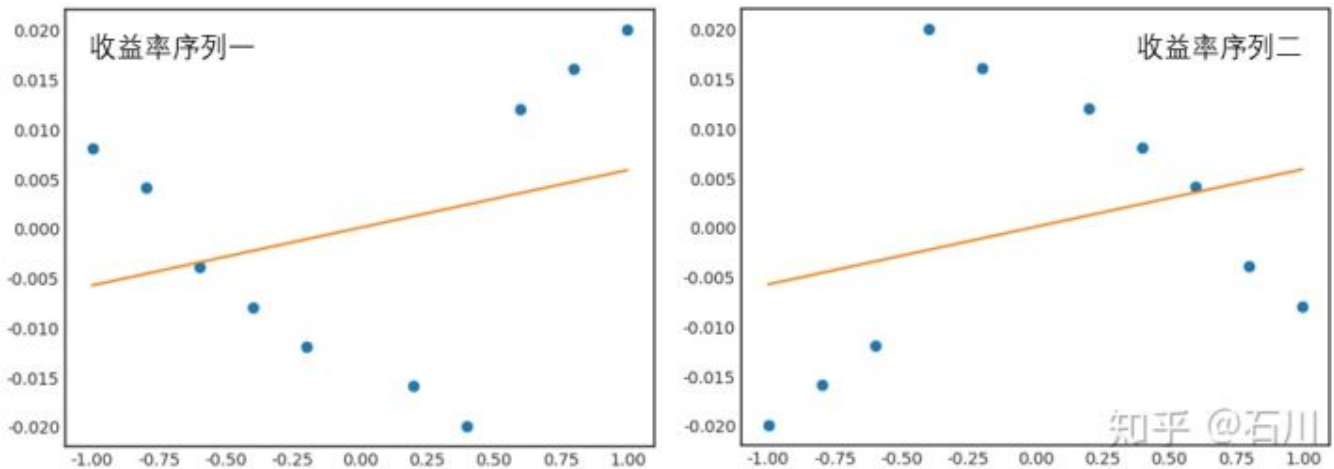
另 y 代表收益率， x 代表因子，则线性回归表达式为：

$$y = a + bx + \varepsilon$$

上式中斜率 b 和 x 与 y 的相关系数 ρ 满足如下关系：

$$b = \rho \frac{\sigma_y}{\sigma_x}$$





虽然 IC 一样，但是画出图来才看到这两组收益率序列和因子的关系大相径庭。假设从业务逻辑来说，个股的收益率和因子呈正相关，因此我们要选因子取值大的股票。但是，这个逻辑在上面两组收益率序列中会得到截然不同的结果：对于序列一，使用最大的因子取值可以选出收益率最高的股票；而对于序列二，使用最大的因子取值却选出了收益率相当差的股票。面对如此结果，IC 无辜吗？

如果使用 Rank IC 代替 IC，得到的也是同样的结论。这两组收益率和因子的秩相关系数均等于 0.3212。从这个数字背后解读不出任何超过这个数字本身的东西。

在量化投资中，我们喜欢并追寻能够精确计算出的数字。但这么做的前提是该数字有意义。在统计学家中流传着一个说法：

Numerical calculations are exact, but graphs are rough.

译：数值计算是精确的，而图形是粗糙的。

单一的统计量，比如上面的 IC 或者 Rank IC 却难以体现出图形反映出来的因子和收益率之间更多的关系。这说明如果我们仅仅看中 IC，可能会步入数据的陷阱。

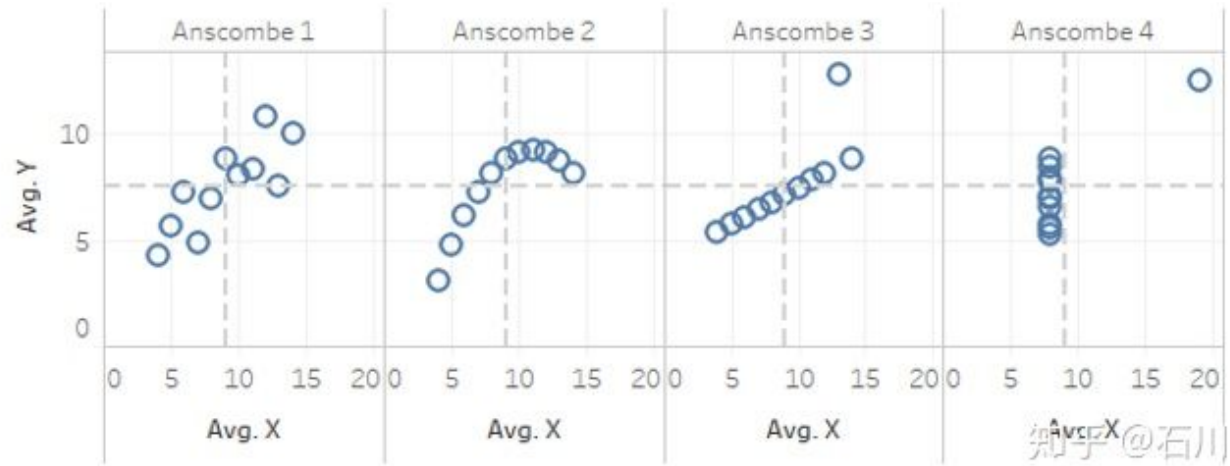
仅关注统计量而忽视图形信息本身最著名的例子当属**安斯库姆四重奏 (Anscombe's quartet)**。安斯库姆四重奏是四组基本的统计特性一致的数据，但由它们绘制出的图形则截然不同。每一组数据都包括了 11 个 (x_i, y_i) 点。这四组数据由统计学家弗朗西斯·安斯库姆 (Francis Anscombe) 于 1973 年构造，他的目的是用来说明在分析数据前先绘制图表的重要性，以及离群值对统计的影响之大。

下图就是这四组数据绘制出来的图形，可见它们截然不同：

1. 第一组描绘了 x 和 y 之间近似的线性关系；
2. 第二组中 x 和 y 表现出了明显的非线性关系；
3. 第三组中 x 和 y 之间存在线性关系，但由于一个明显的 outlier 的存在改变了数据的统计

关系”。

Anscombe quartet



这四组数据和它们的统计特征如下图所示。

第一组		第二组		第三组		第四组		性质	数值
x	y	x	y	x	y	x	y	x 的均值	9
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58	x 的方差	11
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76	y 的均值	7.50, 精确到小数点后两位
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71	y 的方差	4.125 ± 0.003
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84	x 与 y 之间的相关系数	0.816, 精确到小数点后三位
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47	线性回归线	y = 3.00+0.500x 分别精确到小数点后两位和三位
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04	线性回归 R-squared	0.67, 精确到小数点后两位
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25		
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50		
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56		
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91		
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89		

这个例子完美的诠释了统计量（比如本文的 IC）不能反映出数据的全部信息。更危险的是，一旦它们被错误解读和使用，将会导致完全错误的结果。

3 改进的 IC

上一节的例子是为了说明当使用个股的因子取值和下期收益率在截面上回归时，得到的 IC 或者 Rank IC 不能很好的反映出因子选股的效果。对于这种情况，可以考虑以下两种改进方法。

第一种方法是按照因子取值把个股分成 n 档（比如十档），然后将每一档视作一个投资组合，算投资组合收益率和投资组合因子在截面上的 IC 或 Rank IC。每一个投资组合中，可以按照等权或者市值加权来计算投资组合的收益率和因子取值。

因子取值之间的相关性。这就是使用因子构建投资组合、再计算 IC 的初衷。投资组合的收益率是一揽子股票的均值，也可以更好的消除收益率上的噪音。

第二种方法仍然从个股收益率和因子取值的 IC 出发，但是在计算时根据因子的业务逻辑（大到小、还是小到大的关系）来给 x 和 y 的取值赋权，从而得到 weighted IC。由于结合了从业务逻辑出发的权重，这个加权 IC 能更好的反映因子的选股能力。

下面以上一节的因子取值和两组收益率序列为例解释这一做法。假设从业务出发，因子取值越大越好。将十组 (x_i, y_i) 样本点按照因子值 x 从大到小排序，并假设它们的权重按指数衰减，系数为 0.9。这十组样本点的权重为：

权重	因子取值	收益率序列一	收益率序列二
0.1535	1.0	2.00 %	-0.80 %
0.1382	0.8	1.60 %	-0.40 %
0.1244	0.6	1.20 %	0.40 %
0.1119	0.4	-2.00 %	0.80 %
0.1007	0.2	-1.60 %	1.20 %
0.0907	-0.2	-1.20 %	1.60 %
0.0816	-0.4	-0.80 %	2.00 %
0.0734	-0.6	-0.40 %	-1.20 %
0.0661	-0.8	0.40 %	-1.60 %
0.0595	-1.0	0.80 %	-2.00 %

有了权重向量（记为 w ），就可以计算 x 和 y 之间的**加权均值、加权方差、加权协方差、以及加权相关系数（weighted correlation coefficient）**：

知乎

首发于
川流不息

$$\begin{aligned}\text{cov}(x, y, w) &= \sum_i w_i (x_i - E_w[x])(y_i - E_w[y]) \\ &= \sum_i w_i x_i y_i - (\sum_i w_i x_i) (\sum_i w_i y_i)\end{aligned}$$

$$\begin{aligned}\text{var}(x, w) &= \sum_i w_i x_i^2 - (\sum_i w_i x_i)^2 \\ \text{var}(y, w) &= \sum_i w_i y_i^2 - (\sum_i w_i y_i)^2\end{aligned}$$

$$\rho(x, y, w) = \frac{\sum_i w_i x_i y_i - (\sum_i w_i x_i) (\sum_i w_i y_i)}{\sqrt{\sum_i w_i x_i^2 - (\sum_i w_i x_i)^2} \sqrt{\sum_i w_i y_i^2 - (\sum_i w_i y_i)^2}}$$

根据上述定义，很容易计算出因子和这两组收益率序列的加权相关系数。它们分别为 0.4494（因子和第一组收益率序列），以及 0.0908（因子和第二组收益率序列）。从加权 IC 来看，显然**第一组的收益率序列比第二组收益率序列更能说明因子的选股能力**。

同样的，为了绘图说明加入权重的优势，对 x 和 y 进行 **weighted least squares 回归 (WLS)**：

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{10} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_{10} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{10} \end{bmatrix}$$

令 \mathbf{X} 代表系数矩阵（包括截距项系数 1 和 x ）， \mathbf{W} 表示由权重 w_i 作为第 i 个对角元素构成的对角矩阵，则带权重回归的解为：

$$\begin{bmatrix} a \\ b \end{bmatrix} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{W} \mathbf{y})$$

利用线性代数的运算法则，不难求出上式右侧的第一项逆矩阵为：

$$(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} = \begin{bmatrix} \frac{\sum w_i x_i^2}{\sum w_i x_i^2 - (\sum w_i x_i)^2}, & \frac{\sum w_i x_i}{(\sum w_i x_i)^2 - \sum w_i x_i^2} \\ \frac{\sum w_i x_i}{(\sum w_i x_i)^2 - \sum w_i x_i^2}, & \frac{1}{\sum w_i x_i^2 - (\sum w_i x_i)^2} \end{bmatrix}$$

回归式中右侧第二项为：

知乎

首发于
川流不息

$$\left[\sum w_i x_i y_i \right]$$

因此，加权回归的系数为（其中 a 为截距， b 为斜率）：

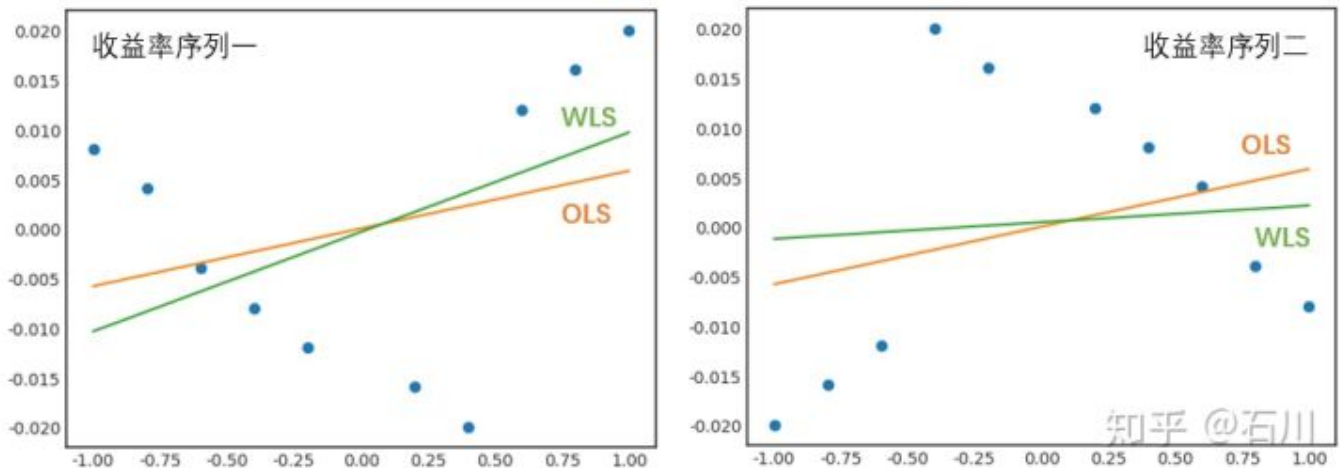
$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \frac{(\sum w_i x_i^2)(\sum w_i y_i) - (\sum w_i x_i)(\sum w_i x_i y_i)}{\sum w_i x_i^2 - (\sum w_i x_i)^2} \\ \frac{\sum w_i x_i y_i - (\sum w_i x_i)(\sum w_i y_i)}{\sum w_i x_i^2 - (\sum w_i x_i)^2} \end{bmatrix}$$

费了半天劲写出了 a 和 b 的表达式（其实从求解的角度，给出矩阵形式的求解足够了）只是想说明下面这件事儿。如果我们比较加权相关系数 $\rho(x, y, w)$ 以及加权方差（标准差）

$\text{var}(x, w)$ 和 $\text{var}(y, w)$ ，以及斜率 b ，则不难发现，和 OLS 一样，在加权回归中， ρ 和 b 仍然满足如下关系：

$$b = \rho(x, y, w) \frac{\sigma(y, w)}{\sigma(x, w)}$$

下面就来画图比较一下 WLS 回归和上一节 OLS 回归的结果。对于这两组收益率序列，OLS 回归的结果相同。但从选股的角度，我们知道如果因子对应的是第一组收益率，则该因子远比起其对应第二组收益率有效。但是 OLS 回归（和普通的 IC）无法体现这一点。而采用改进的 WLS（以及 weighted IC）来衡量的话，如果因子产生了第一组收益率序列，则它的 WLS 回归斜率为 0.01（大于 OLS 的斜率 0.0058）；如果因子产生了第二组收益率序列，则它的 WLS 回归斜率仅为 0.0017（小于 OLS 的斜率）。这说明通过使用基于因子业务规则的权重系数，WLS 比 OLS 更能判断因子和收益率之间的关系。



4 结语

知乎

首发于
川流不息

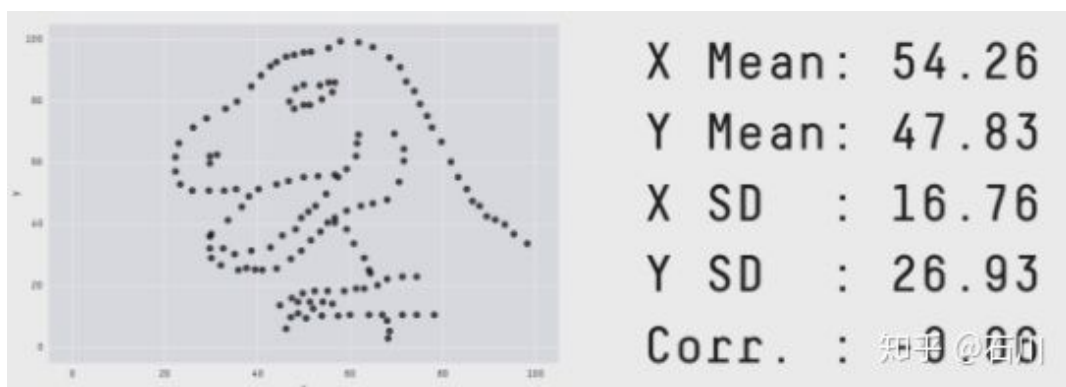
致错误的判断。金融数据已经信噪比极低了，我们当然不希望因为自己使用不当再加入不必要的噪音。

很多时候数据关系越复杂，统计量传递出来的信息可能越失真。

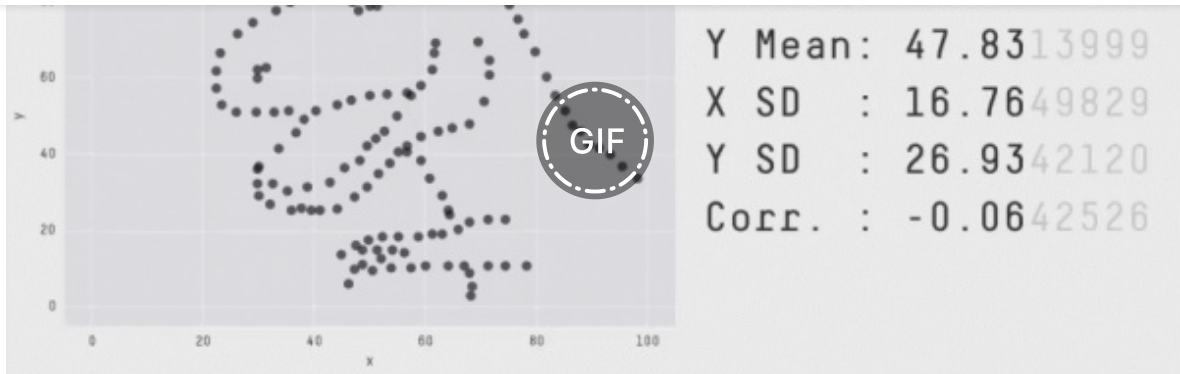
2017 年，来自 Autodesk Research 的 Matejka 和 Fitzmaurice 构建了当代版的“安斯库姆四重奏”（Matejka and Fitzmaurice 2017）。他们用计算机算法可以生成 x 均值、 y 均值、 x 标准差、 y 标准差、以及 x 和 y 相关系数相同的复杂数据集。比如下图中的 12 个完全不同的数据集就在上述五个统计量中取值完全一致—— x 均值 54.26， y 均值 47.83， x 标准差 16.76， y 标准差 26.93， x 和 y 相关系数 -0.06。



先别忙着惊讶，上述这些数据集都是由下面这张恐龙数据集（也有同样的统计量）构建来的（请点击 GIF 看动画）！



知乎

首发于
川流不息

有的朋友也许会说，IC 不够，再引入更多的统计量就行了。我们当然可以计算更高阶矩的统计量，但是因为数据的信噪比极低，这些样本数据计算出来的高阶统计量也存在大量误差。本文提出的改进方法属于从因子和收益率之间的内在逻辑出发——比如分档构建组合、或者给不同的权重。这些都是**以内在的逻辑为先验**，以期更好的判断因子的选股能力。

如果你在使用 IC 或者 Rank IC（以及 IR）来动态的评价、配置因子，那么本文希望能引发你的思考。在评价因子选股效果的道路上，我们也许还有很长的路要走。

参考文献

- Matejka, J. and G. Fitzmaurice (2017). *Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing*. CHI 2017 Conference proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems.

免责声明：文章内容不可视为投资意见。市场有风险，入市需谨慎。

原创不易，请保护版权。如需转载，请联系获得授权，并注明出处，谢谢。已委托“维权骑士”（[维权骑士_免费版权监测/版权保护/版权分发](#)）为进行维权行动。

编辑于 2019-07-03

多因子模型 量化交易 相关性分析

▲ 赞同 85 ▼ 6 条评论 ➤ 分享 ★ 收藏 ...

文章被以下专栏收录



推荐阅读

趁个车呀

完美并不美



《FoF的前世今生：从资产配置到因子投资》新版序

llang... 发表于因子投资那...




睡前消息替


马前卒

6 条评论


切换为时间排序


写下你的评论...



 韩秀一 1 年前


请问一下，如果是允许空的话，因子权重应该是两头高，中间低了吧？

 赞


 石川 (作者) 回复 韩秀一 1 年前

对的。对这种情况我也简单试了试，文中没有提。


 赞

 韩秀一 回复 石川 (作者) 1 年前

还有个问题想请教，因子分组的收益单调性不好的话，怎么来衡量这个因子的效应

 赞

展开其他 1 条回复

 Pandas 1 年前

我觉得配合着看因子分组的净值走势就能解决这个问题，如果top组收益一直显著高于其他组，那说明在每一期的高IC就是有序的，如果中间组收益率较高和top组不显著，那么高IC就

知乎



首发于
川流不息



石川 (作者) 回复 Pandas

1 年前

嗯 是的

👍 赞

