

知乎

首发于
川流不息

尾部风险建模



石川

量化交易 话题的优秀回答者

已关注

黑猫Q形态等 64 人赞同了该文章

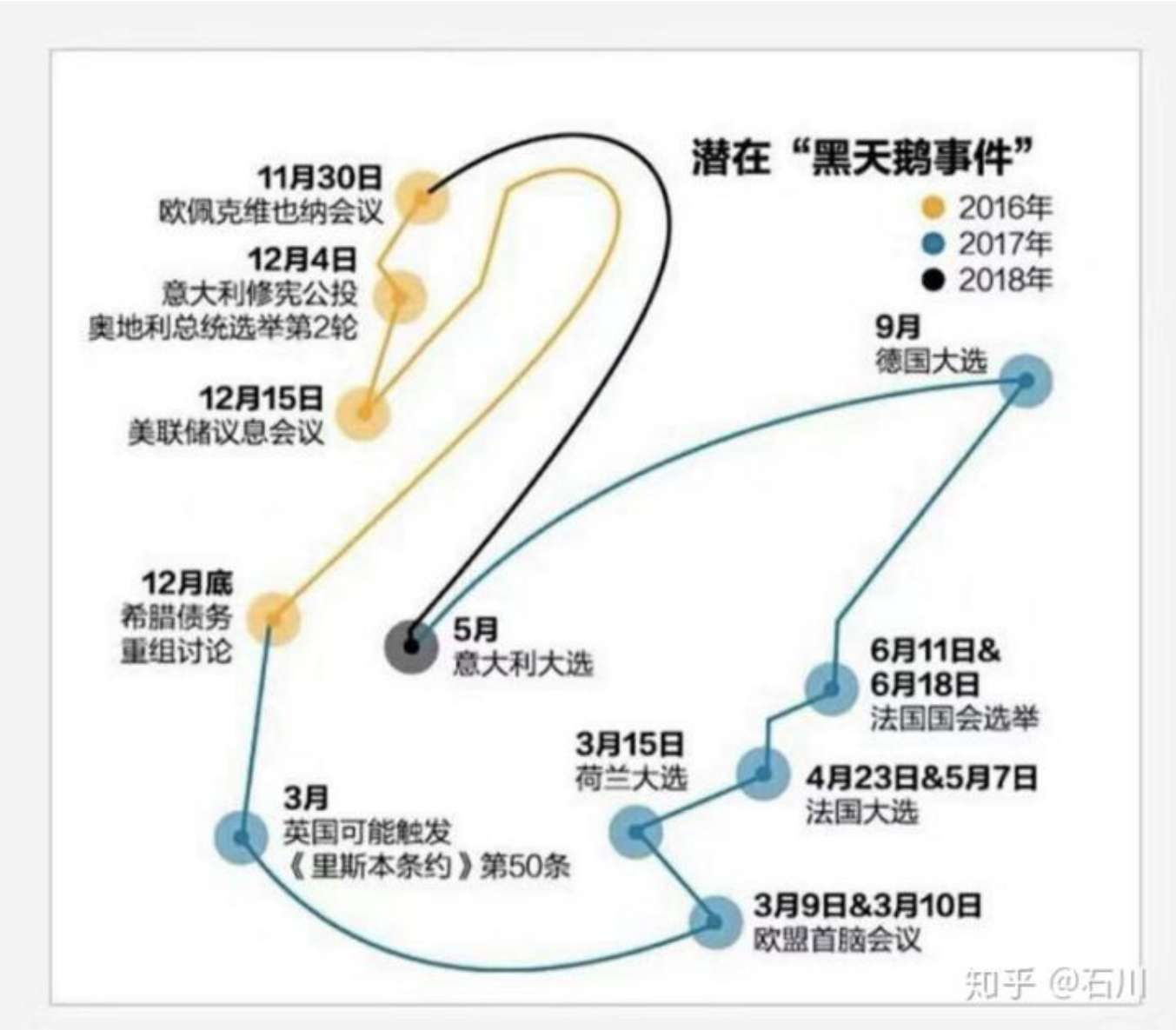
写在前面.....

这篇文章写在 2016 年年底。当时我做量化投资不久，希望从模型的角度学习尾部风险建模，因此**这篇文章算是一个学习笔记 + 一些简单的实证**。现如今，经过了近两年在 A 股和国内商品期货市场的摸爬滚打（赚了该赚的钱，也亏了该亏的钱），越来越感觉到对于**风险**而言，所有样本内的建模都是对样本外的过拟合，真正的风险是 unknown unknowns。

本文对尾部风险建模的目的是计算诸如风险价值 VaR（给定概率下亏损的阈值 —— **最小值**）、或 Expected Shortfall 这类度量风险的指标。关于 VaR 在刻画风险方面是否有用，知乎上有篇著名的讨论：《风险价值（VaR）是否是有史以来最蠢的衡量指标？》，我仔细学习了各位前辈的答案，收获和感慨颇多。对于我而言，对 VaR 的理解和认识远没有各位前辈深刻，但是这篇笔记仍然是一个很好的学习过程，所以我决定还是把它放在专栏里。它记录了学术界对尾部风险建模的一些努力，有助于我们更好的理解收益率的尖峰肥尾分布，以及市场参与者在试图理解风险时 —— 哪怕是一点点 —— 面临的困难。

1 风险控制和尾部建模

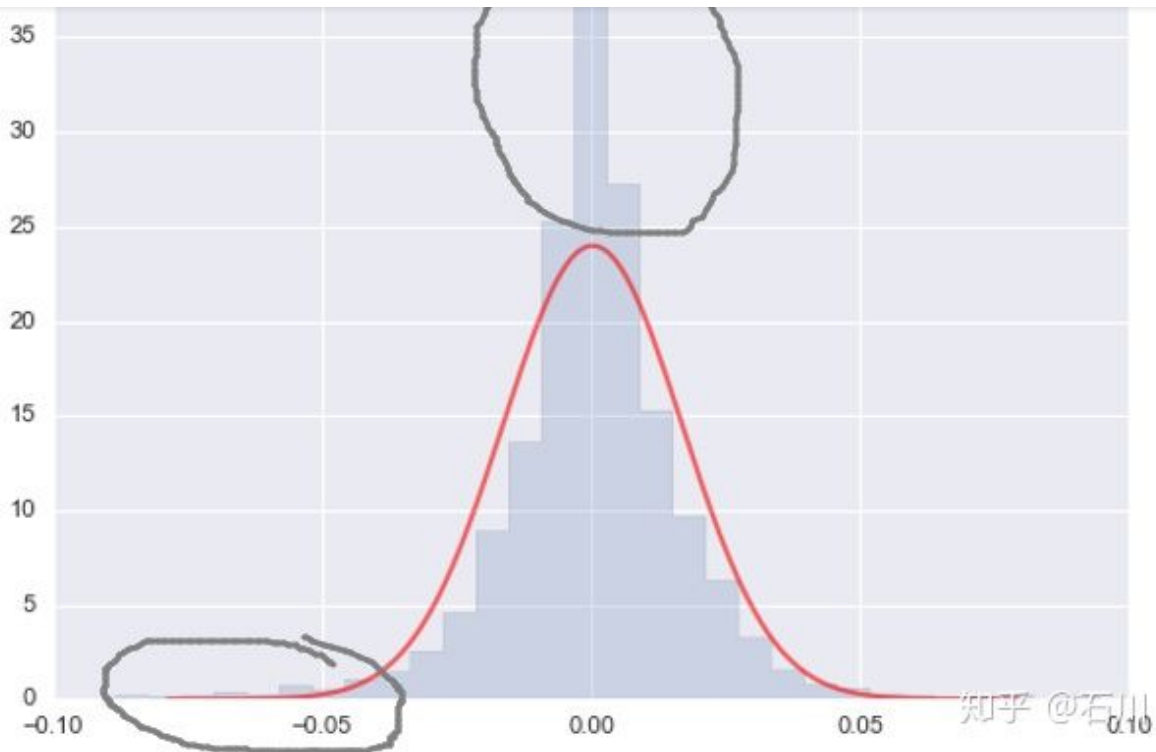
2016 年全球金融市场不太平，从英国脱欧到 Trump 当选美国总统再到意大利公投，“黑天鹅”事件频出，就连美联储也跟着添乱，嚷嚷了一年加息、故意扰乱市场对美国经济数据的解读。未来两年，潜在的黑天鹅更是一个接一个。



在这种背景下，风险控制再次回到人们的视线中。在金融领域，风险控制的目的是为了计算极端黑天鹅事件对金融资产造成的潜在损失（负收益率）的可能性以及冲击的大小。

先来看一个分布。下图为上证指数在过去 15 年内日收益率的分布。我们计算出日收益率的均值和标准差，便可以得到一个基于该均值和标准差的正态分布。下图比较了收益率的直方图和该正态分布。

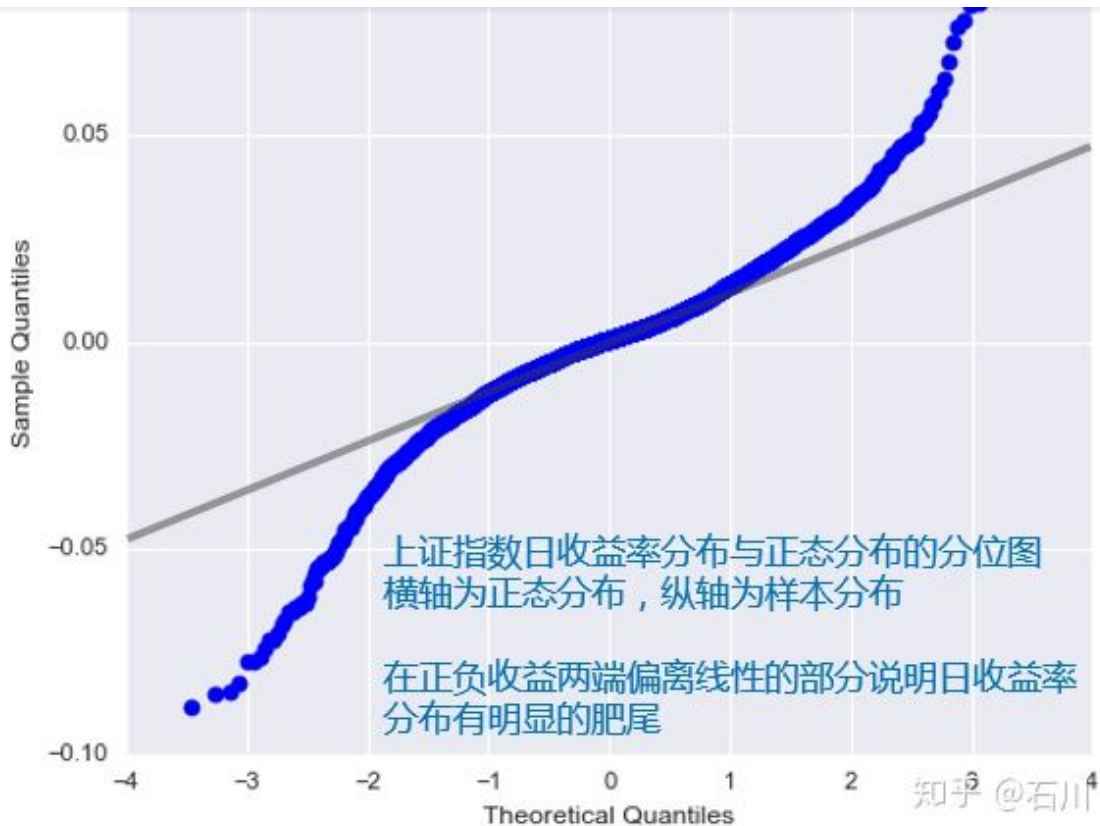
知乎

首发于
川流不息

不难看出，**上证指数日收益率的分布表现出明显的尖峰和肥尾特点，尤其是在负收益率部分**。比较日收益率分布和标准正态分布的分位图（下图），也可以清晰地验证这个结论。肥尾意味着上证指数实际发生极端收益率（从上图来看，尤其是极端跌幅）的概率要远远大于正态分布对应的概率。换句话说，**如果算出收益率的均值和标准差，然后构建一个正态分布来近似描述日收益率分布，这会造成很大的误差。**



知乎

首发于
川流不息

除了尖峰、肥尾的特点之外，另一个困扰“黑天鹅建模”的问题是，发生极端亏损（真正的黑天鹅）的历史样本太少了。比如说，我们想回答“上证指数每十年一遇的日收益率最大跌幅是多少”这个问题，回看上证指数过去 20 几年的历史，我们仅仅有可怜的 2 个样本点，无法根据它们构建有效的模型。

那么应该怎么办呢？在统计学上，广义极值分布（Generalized Extreme Value Distribution）可以用来对极端亏损建模。

2 极值建模

假设随机变量 X_i 代表某投资品的负收益率（亏损），它满足某未知分布

$F(x) = Pr(X_i \leq x)$ 。在下文中，我们用负收益率的绝对值代表亏损的大小（即 X_i 的取值为正数）。在这种描述下，当 X_i 的取值在其分布的右尾（right tail）时，便意味着该投资品发生了极端的亏损。

假设不同时间的亏损 X_i 是独立同分布的，并令 $M_n = \max(X_1, \dots, X_n)$ ，即 M_n 是 n 个样本中最坏的情况。广义极限分布理论解决的问题就是对 M_n 分布的建模。有了 M_n 的分布，我们就可以回答上面诸如“上证指数每十年一遇的日收益率最大跌幅是多少”的问题。

根据独立同分布的假设，我们可以写出 M_n 的 CDF 为：



由于分布 F 是未知的, F^n 自然也是未知的, 而经验分布函数对与 F^n 的估计也是非常差的。但是, 我们可以根据 Fisher-Tippet 理论 (Fisher and Tippet 1928) 来渐进逼近 F^n , 并以此得到 M_n 的分布。特别的, Fisher-Tippet 理论证明, 将 M_n 标准化后, 即

$Z_n = (M_n - \mu_n) / \sigma_n$, Z_n 的分布收敛于形式如下的广义极限分布:

$$H_\xi(z) = \begin{cases} \exp\{-(1 + \xi z)^{-1/\xi}\} & \xi \neq 0, \quad 1 + \xi z > 0 \\ \exp\{-\exp(-z)\} & \xi = 0, \quad -\infty \leq z \leq \infty \end{cases}$$

因此, 只要我们有足够多的原始负收益率样本数据 X_i , 我们可以用下式求出极端亏损 M_n 的分布:

$$\text{Prob}(M_n < x) \approx H_{\xi, \mu, \sigma} \left(\frac{x - \mu_n}{\sigma_n} \right) = H_{\xi, \mu_n, \sigma_n}(x)$$

在实际使用中, 广义极限分布 H 的参数 (ξ, μ, σ) 可以通过极大似然估计 (maximum likelihood estimation) 得到。为了估计这些参数, 我们必须有足够多个 M_n 的样本。为此, 我们可以将总长为 T 期的历史数据等分成单位长度为 n 的 m 个区间。每个区间中的最大亏损便是 M_n 的一个样本。这样我们就可以得到 m 个样本。这样, 便可以根据这些样本得到广义极限分布 H 的参数的估计。Embrechts et. al. (1997) 给出了详细的数学推导。

3 阈值外数据建模

在风险管理中, 风险价值 (Value at Risk, VaR) 是人们常说的一个概念。比如, 当我们说 1% 的日收益率的 $\text{VaR} = 6.8\%$, 它的意思是, 我们的目标投资品 (或者投资组合) 在当天有 1% 的概率可能产生超过 6.8% 的亏损 (因此 VaR 的数值是风险的最低值)。在给定的概率下, VaR 越大, 投资品的风险越大。

然而, 如果想计算 VaR 的大小, 上一节中对极值分布的模型并无法发挥作用。这是因为在计算 VaR 时, 我们必须对亏损分布的右尾进行建模, 而不单单是关注某一个极值 (注意, 在本文中我们用亏损的绝对值来描述亏损的大小, 因此亏损都是正数, 所以这里我们是对分布的右尾建模)。为此, 我们可以采用广义帕累托分布 (Generalized Pareto Distribution)。

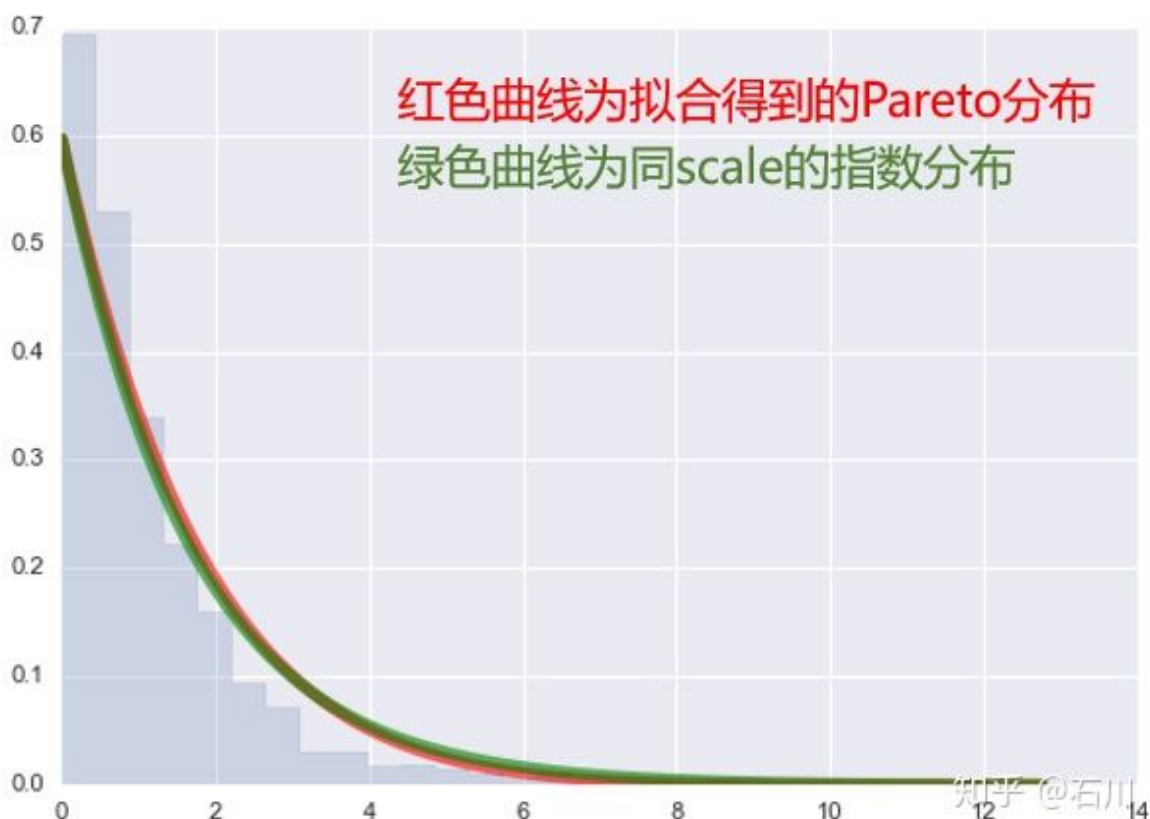
和上节一样, X_i 表示某投资品的一系列亏损, 并假设它们独立且满足某未知分布 F 。同样的, 定义 $M_n = \max(X_1, \dots, X_n)$ 。假设 u 为某一个给定的亏损阈值。在所有这些 X_i 中, 我们感兴趣的是那些大于 u 的样本, 即那些亏损超过阈值的样本点, 我们希望用它们来对 X_i 分布的右尾进行建模。超过给定阈值的亏损部分, 即 $X_i - u > 0$ 的部分, 可以由如下条件概率表示:

Embrechts et. al. (1997) 证明, 如果亏损 X_i 的极值 M_n 收敛于上节介绍的广义极限分布 H , 那么存在一个 u 的函数 $\beta(u)$, 使得 $X_i - u$ 满足如下形式的广义帕累托分布 G :

$$G_{\xi, \beta(u)}(y) = \begin{cases} 1 - (1 + \xi y / \beta(u))^{-1/\xi} & \text{for } \xi \neq 0 \\ 1 - \exp\{-y / \beta(u)\} & \text{for } \xi = 0 \end{cases}, \beta(u) > 0$$

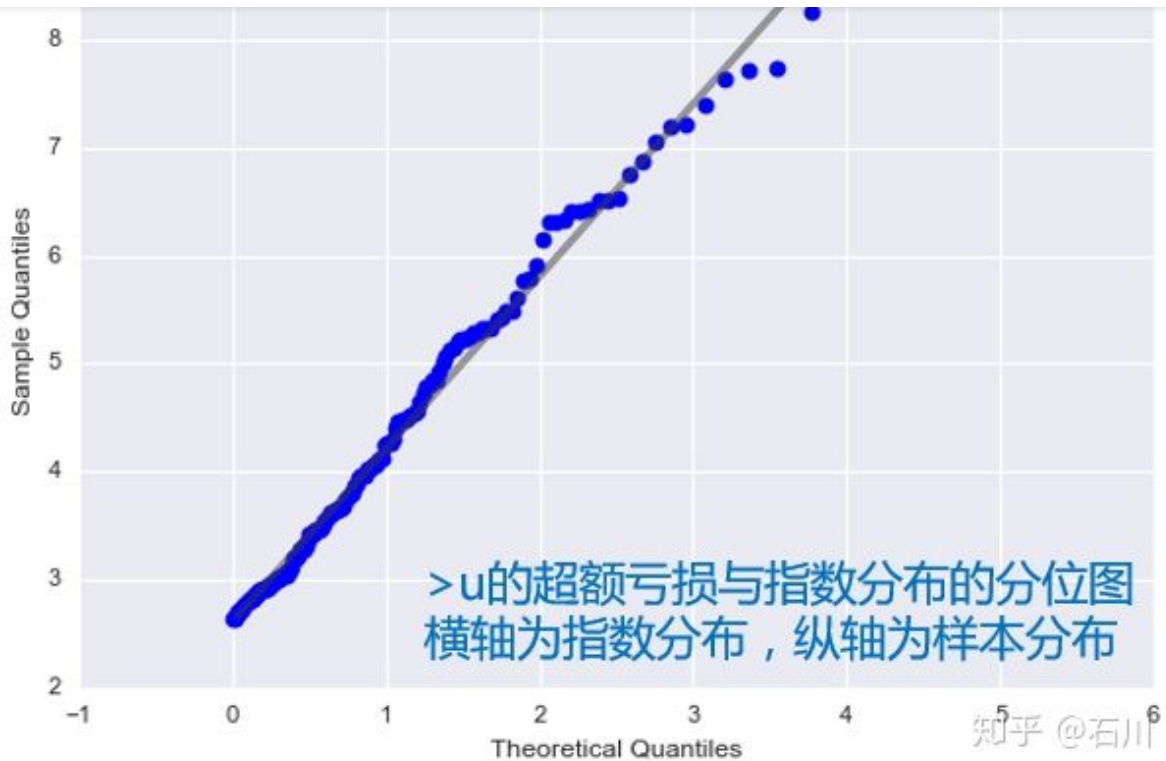
在实际应用中, 如果我们想对 X_i 的右尾建模, 只需确定阈值 u 。然后在 X_i 的所有样本中找出所有大于 u 的样本 (注: 我们用 X_i 的绝对值表示亏损的大小, 所以亏损在上述数学表达式中是正数), 将这些满足的样本各自减去 u 后得到超过 u 的部分, 然后用这些数据拟合广义帕累托分布 G , G 的参数由极大似然估计得到。

广义帕累托分布 G 的形状随着形状参数 ξ 的不同而不同。特别的, 当 $\xi = 0$ 时, G 就简化为指数分布。我们以过去 15 年上证指数日频的负收益率样本为例, 取阈值 $u = 2.65\%$ (即考察日收益率亏损超过 2.65% 的尾部分布), 得到了 G 的参数。其中形状参数的取值非常接近 0。下图为拟合得到帕累托分布和同比例的指数分布对比超额亏损的直方图的结果。可以看到红色的帕累托分布和绿色的指数分布非常接近。



此外, 我们也可以用超额亏损和标准的指数分布放在一起做分位图, 得到的结果如下。结果显示分位图近似的满足线性, 说明超额亏损的分布和指数分布十分接近。

知乎

首发于
川流不息

利用超额亏损对尾部分布建模后，我们便可以方便的求解风险价值。

4 风险价值

上一节曾经说过，风险价值描绘的是投资品在某一个指定的概率下亏损程度的阈值。在我们的定义下（即我们用正数来代表亏损的大小），风险价值就是某一给定概率下亏损 X_i 分布中右尾的某一个分位数。换句话说，只要根据给定的概率求出分位数，它的值就是这个概率对应的风险价值。

因此，通过广义帕累托分布 G ，我们便可以简单的推导出风险价值的公式。假设 $1-q$ 代表我们考虑的概率（比如我们想知道 5% 的概率对应的亏损，那么 $1-q = 0.05$ ），则其对应的风险价值为：

$$\widehat{\text{VaR}}_q = \mu + \frac{\hat{\beta}(u)}{\hat{\xi}} \left(\left(\frac{n}{k} (1-q) \right)^{-\hat{\xi}} - 1 \right)$$

其中， n 是亏损样本的总个数， u 是对应的阈值，可以由 $q = F(u)$ 求出。 k 是超过 u 的亏损样本的个数。在应用中， $(n-k)/n$ 可以作为对 $F(u)$ 的估计。因此，对于给定的概率 $1-q$ ，计算风险价值的步骤为：

1. 根据 q 和 $q = (n-k)/n$ 求出 k ；
2. 根据 k 求出 u ，即在所有亏损的样本中，找到对应的阈值 u ，使得满足 X_i 大于 u 的数为 k ；

由于风险价值关注的往往是 5% 甚至 1% 的亏损阈值，它们对应的是亏损分布中非常靠尾部的那些样本，因此只有当 n 足够大时，我们才可能得到足够多的超额亏损来建模。可惜的是，在这方面中国 A 股的年份太短了。

即便如此，我们仍然通过下面简单的实验来说明如何计算风险价值。这里我们考虑标普 500 指数（从 1930 年至今）和上证指数（从 2000 年至今）。此外，为了增加样本个数，我们考虑的风险价值对应的概率为 10%，而非极端的 5% 或者 1%。

对于标普 500，我们用每 15 年的数据来滚动建模，得到日收益率在 10% 概率下的风险价值。作为比较，我们用日收益率均值和标准差对应的正态分布同样求出 10% 概率下的风险价值。结果如下图所示。



上图说明以下几点：

1. 由于收益率存在明显的肥尾效应，**正态分布严重低估了风险价值（绿线持续的在红线之下）**；
2. 在 1929 年股灾之后的有一段时间，风险价值都在高位，这是因为计算的样本中有大量的高亏损样本；
3. 进入 21 世纪以来，风险价值有两次明显的跃升，分别对应着 2000 年的 .com 泡沫和 2008 年的次贷危机。

同样的，我们对上证指数建模。由于数据年份太短，我们用每 10 年的数据来滚动建模。结果如下图所示。同样的，正态分布建模严重低估了风险价值。此外，由于上证指数比标普 500 有更加明

知乎

首发于
川流不息

我们可以用更短的时间（即更少的样本）对上证指数进行滚动建模。**但是样本少一定会带来建模的误差。**下图为我们使用 5 年窗口进行滚动建模的结果。结果表明从 2008 年股灾开始后一直到 2014 年，上证指数的风险都非常大（注意，正态分布建模无法很好的描述风险价值的变化，且存在严重的低估）。在最近两年，随着 2015 年股灾和 2016 年 1 月份熔断引发的二次灾害，风险价值出现了两次迅速的蹿升。



知乎

首发于
川流不息

5 结语

做投资时，如何强调风险控制都不过分。然而，做好风控的前提就是能用正确的数学手段对其量化。为了控制风险，有人刻意限制仓位，有人“把鸡蛋放在不同的篮子里”。然而**分散投资不完全等价于分散风险**。“把鸡蛋放在不同的篮子里”不如“把鸡蛋放在一个篮子里，然后看好这个篮子”。从这个意义上说，对亏损的正确建模格外重要。

参考文献

- Embrechts, P. C. Klöppelberg, and T. Mikosch (1997). *Modelling Extremal Events*. Springer-Verlag, Berlin.
- Fisher, R. and L. Tippett (1928). *Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample*. Proceedings of the Cambridge Philosophical Society 24, 180-190.

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



▲ 赞同 64 ▼ 14 条评论 分享 ★ 收藏 ...

文章被以下专栏收录



川流不息

北京量信投资管理有限公司是一家在中国基金业协会备案登记的专业私募基金管理人...

关注专栏

推荐阅读



市场上最常见的7种仓位控制方法，你用了哪个？

汇商传媒 发表于外汇百科



跟踪沪深300指数的基金，到底选哪只？

码农投资



定投只会赚大钱的拼

菜导洪佳彪

14 条评论

切换为时间排序

写下你的评论...



黑猫Q形态

1 年前

你这个内容比人会说“黑天鹅是不可预测的！”然后转身就去“预测”收益

👍 2



石川 (作者) 回复 黑猫Q形态

1 年前

这篇文章最初的题目是黑天鹅预测的正确姿势。后来我觉得那题目太“自以为是”了，就改了，还加了一段前言。没想到这么“误导”啊（捂脸）。谢谢前辈留言。

👍 赞



黑猫Q形态 回复 石川 (作者)

1 年前

知乎

首发于
川流不息

👍 1



大柳

1 年前

半年前看了很多相关论文，你这是静态风险度量方法，还有很多动态度量方法，不过可用性不高

👍 1



石川 (作者) 回复 大柳

1 年前

👍。我还有很多需要学习的！

👍 赞



Syous

1 年前

额...这个正好是我的研究方向，我补充几句吧：1 EVT即极值理论建模只是一种边边缘分布建模方式，其实金融危机这十年来尤其是近年来，还有一种更常用的建模方式就是类GARCH建模，比如近两年文献经常用的GJR-GARCH过程。它的特点和优势大家可以自行查阅 2 其实边缘分布的建模还不是尾部风险的重点，因为实在找不到convinced的建模方式，还可以采用非参的办法。真正的重点在于，如何刻画“厚尾”，其中最流行的办法莫过于copula函数，即将一个联合分布函数分解为单独边缘分布的连乘和copula函数的乘积，从而大大简化计算和分别建模。顺便一提，08金融危机的技术性层面，就是因为李祥林的copula文献被各大投行滥用错用（选错了gaussian copula），从而导致对风险敞口估计错误。3 其实VaR的办法来估计风险阈值，在近年的文献中已经不常用。更常用的是求其条件VaR即CoVaR（conditional- VaR个体i损失超过阈值时，系统或者资产组合的阈值），对CoVaR求积分得到的尾部风险期望损失CoES(conditional expected shortfall)，以及反向的MES（marginal expected shortfall，系统或者组合低于阈值时，个体的期望损失如何---或者说，个体i的风险贡献度）。4 动态办法（包含regime switch这类）已经成为主流，比如copula函数的参数估计，目前以Patton的GAS动态过程为相对比较流行。5 如果想进一步深入研究，这几个作者的文献逃不掉：Acharya, Brownlees & Engle，以及Patton。

👍 3



石川 (作者) 回复 Syous

1 年前

谢谢！非常感谢如此认真、全面的回复！这方面我了解的还远不够，以后和您多学习。：)

👍 赞



Syous 回复 石川 (作者)

1 年前

哪里哪里，共同学习，共同学习，我也只是因为研究这一块儿所以有点心得。要论实践应用还得看大佬你的。

👍 赞 ↩ 回复 🗑 踩 🚩 举报

查看全部 7 条回复



知乎



首发于
川流不息

数和尺度参数也涉及到了GAS动态过程，还在学习中。

 赞

