回测过拟合

1. 动机

在做 CTA 的参数优化时，遇到了一个问题，昨天我用截止昨天的数据做参数优化得到一组参数，今天我用截止今天的数据做参数优化得到一组参数，这样每天坐下来，参数的变动无规律，怎么解决该问题？为了解决该问题，我们阅读了一些研究，这些研究的主题是用历史数据做参数优化（即从很多的策略设置中筛选比较好的策略设置），选出来的策略设置很可能真正运行时表现差，一个主要原因是回测过拟合（什么是回测过拟合下面会详细介绍）。该主题虽然不是直接针对上面的问题，但是对解答上面的问题有帮助；更重要的是，这一主题让我们对参数优化或者回测有了更加清楚的认识，能够帮助我们构造出回测表现好且真正使用时表现也好的策略。 我们了解到的对该主题的研究有三个研究流派或研究组。分别是 Keith Fitschen 的《Building_Reliable_Trading_Systems》（下面简称 KF），Campell Harvey 和 Yan Liu 的研究（简称 HL），以及 Bailey 和 Marcos 的研究（下面简称 BM）。下面我们分别对这 3 个流派的观点或方法进行描述，然后进行比较。

首先，我们说一下过拟合的最直观的解释。过拟合就是用历史数据来对模型的设置进行调整，使得模型在这段历史数据上有很好的表现，但是如果模型的设置在调整的过程中针对于这段历史数据的特性也进行了相应调整，那么得到的模型在未来的表现就不会与在该历史数据上的表现相同。该直观解释对于后续理解比较重要。

2. KF

KF 对于过拟合（KF 全文用的词是 curve-fitting）的直觉是，给定历史数据长度，在众多策略设置（strategy configurations）中选择一个在该历史数据上表现最优的，如果在该最优策略设置下，在该历史数据上产生的交易数量较少，从而用这较少的交易数量计算的策略表现（比如每笔交易平均收益，年化收益率/年化最大回撤），其估计误差是较大的，从而用一个估计误差较大的最优表现值来推断该策略在以后表现会好的这种做法是不可靠的，KF 将其称作 curve fitting。下面以一个例子来说明过拟合。

具体为，假设历史数据为某个期货的价格的 10 年日数据（开盘价、收盘价、最高价、最低价），我们的策略是根据过去几日的收益率的正负来决定买入还是卖出该期货（开仓信号），持有几天后平仓，中间再设置一个止损参数，表示损失达到多少的时候平仓，再设置一个开仓过滤参数，比如今日收盘价比 20 前的收盘价高多少的时候才判断开仓信号。所以我们的策略由这 5 个参数决定：前 n 日收益率，持有日数 m，止损金额 x，开仓过滤值 y，假设 n，m，x，y 各自可以取 10 个值，那么我们的策略设置就会有 10^4 个。我们在每一个策略设置下在该历史数据上跑一遍，就会得到该策略设置下的好多笔交易，可以计算出每笔交易的收益，进一步计算出每笔交易平均收益，年化收益率，最大回撤，年化最大回撤等指标。最后我们根据某一指标选出最优的一个策略设置。KF 在操作的时候用的指标是年化收益率/年化最大回撤，但是在解释过拟合时用的是每笔交易平均收益，因为用每笔交易平均收益解释起来推导简单，我们也用每笔交易平均收益来解释。假设在某一最优策略设置下，得到的每笔交易的收益率为 $r_i, i=1,...,T$，从而每笔交易平均收益率为 $\bar{r}=\frac{1}{T}\sum_{i=1}^{T} r_i$，假设 r_i 与 r_j 是独立同分布，且均值为 μ，方差为 σ^2，则 $\bar{r} \sim N(\mu, \sigma^2/T)$，所以我们如果用 $\bar{r}$ 来推断该策略设置在

以后表现也应该是$\bar{r}$，其精确程度由σ^2/T决定。所以给定σ^2，如果我们用历史数据选出的最优策略设置对应的 T越小，用其对应的表现值（该例子为$\bar{r}$）来推断以后该策略的表现，这种推断的精确程度越低。什么情况会导致T越小呢，KF 经常举的例子就是在上面 4 个参数的基础上，再加一些过滤参数，这样筛出来的交易是这段历史数据上最赚钱的（或表现最好的），但是一般情况下，这样的交易笔数会较少，即T较小。

站在对过拟合最直观的角度，我对 KF 的理解是，在给定历史数据长度下，通过加各种过滤参数，使得产生的交易笔数变少，用交易笔数变少来表示该模型设置是适应了该历史数据的特性，然后用参数或者表现指标的估计值的误差较大来表示该模型设置以后的表现跟历史数据的表现会不一样。

我们在稍后会了解到另外 2 种（HL 和 BM）对过拟合的描述框架，站在他们的框架下，即使T较大，也存在过拟合。避免过拟合不能只是使得T较大，而是使得T相对于模型设置个数较大。

最后，KF 推荐了一个检验过拟合的容易操作的方法 BRAC（Build,Rebuild and Comparison）。具体操作步骤为：（1）用所有的历史数据来选择最优的策略设置，（2）将所有历史数据的最后一部分数据（比如一年）切下来，用剩下的历史数据来选择最优的策略设置，（3）比较（1）和（2）分别得到的最优策略设置在（2）切下来的数据上的表现，如果二者表现一致，则过拟合可能性小，否则，过拟合可能性大。KF 在其文中真正应用该方法的时候是比较了（1）和（2）两种情况下得到的最优策略设置的对应的参数是否一致。

怎样将 BRAC 方法与上面 KF 对过拟合的论述结合起来呢？KF 认为如果T足够大，那么$\bar{r}$的方差会很小，意味着用T比交易来估计得到的$\bar{r}$与用$T-s$笔（s为（2）中留出数据对应的交易笔数）交易来估计得到的$\bar{r}$相差很小，而我们假设$\bar{r}$是由策略设置对应的参数决定的，那么两种情况下得到的策略设置对应的参数差异应该很小。

3. HL

3.1 HL 思路

HL 没有直接提及过拟合这个词，但是其文章主题是用多个策略设置来做回测，得到的表现指标（如夏普率）往往高估，导致该策略设置真正使用时表现不好。其基于的框架是多重检验（multiple testing）。例如，我们有 N 个策略设置，每个策略设置对应用某个信号来预测股票收益率对应的策略设置，我们将 1 个信号作为 1 个策略设置来对待，每个策略设置下，用历史数据都能得到 1 串交易的收益率，从而得到收益率的平均值，以及该收益率平均值对应的 t 统计量。如果我们的 N=1，则用 t 统计量的绝对值大于 2 来判断该策略设置平均收益率是不是为 0 是可行的，因为这种情况下犯第一类错误（真实情况为平均收益率为 0，我们认为平均收益率不为 0）的概率是 5%；但是当 N 大于 2，尤其是当 N 较大的时候，用 t 统计量的绝对值大于 2 来判断该策略设置平均收益率是不是为 0，犯第一类错误的概率会很大。所以好多文章发现某一信号对股票收益率有预测作用，但是它没有他是从 100 个信号里面选出的 1 个，如果是按照按照 t 统计量绝对值大于 2 得到的结论，那么该发现很可能是假发现（false positive）。HL 针对该问题提出了一个解决方法，将普通方法得到的夏普率调低。下面详细介绍该方法。

3.2 N 个策略设置独立情况下的夏普率调整

给定一串历史收益率$(r_1, r_2, ..., r_T)$，$\hat{\mu}$和$\hat{\sigma}$分别表示用其算出的均值和标准差，T为交易笔数，y为交易年限，则检验零假设（平均收益率为 0）的 t 统计量为

$$t=\frac{\hat{\mu}}{\hat{\sigma}/\sqrt{T}},$$

而如果r_i表示每笔交易收益率，则夏普率为的估计为

$$SR=\frac{\hat{\mu}}{\hat{\sigma}}\cdot\sqrt{\frac{T}{y}},$$

则

$SR=t/\sqrt{y}$。

假设某个研究员研究了N个策略设置，每个策略设置之间是独立的，则在原假设：没有一个策略设置的平均收益率不为0下，至少有一个策略设置的t统计量大于观察到的t值的p值为

$$p^M=1-\prod_{i=1}^{N}\Pr(|z_i|\geq t)=1-(1-p^S)^N,\quad (3.1)$$

其中，z_i为服从标准正太分布的随机变量，p^S单个策略设置的t统计量大于观察到的t值的p值。可以看到，当$N=1,p^S=0.05$时，$p^M=0.05$；当$N=10,p^S=0.05$时，$p^M=0.401$。

通过让单次检验的p值等于多重检验的p值，我们就可以得到在单次检验下与多重检验下等价的t值（Ht），从而得到对应的夏普率，即调整后的夏普率（HSP）。即：

$$p^M=\Pr(|z|>Ht)=\Pr(|z|>HSR\cdot\sqrt{y})\quad (3.2)$$

3.3 N个策略设置不独立情况下的夏普率调整

将N个策略单独检验时的p值升序排列，$p_{(1)}\leq p_{(2)}\leq\ldots\leq p_{(N)}$，然后有3种调节p值的方法，分别为Bonferroni方法，Holm方法，和BHY方法。分别为：

Bonferroni方法：$p_i^{Bonferroni}=\min(Np_{(i)},1),i=1,\ldots,N$

Holm方法：$p_i^{Holm}=\min(\max_{j\leq i}\{(N-j+1)p_{(j)}\},1),i=1,\ldots,N$

BHY方法：$p_i^{BHY}=\begin{cases}p_{(N)}\ if\ i=N\\ \min(p_{i+1}^{BHY},\frac{N\times c(N)}{i}p_{(i)})\ if\ i<N\end{cases}$，其中$c(N)=\sum_{j=1}^{N}\frac{1}{j}$。

三种方法的特点：（1）Bonferroni方法和Holm方法对p值的调整，体现了N个策略设置里面至少有1个策略设置错判为有效的概率，而BHY方法体现的是错误率（N个策略设置里，错判为有效的策略占所有判定为有效的策略的比例）大于某一值的概率。前者控制的是绝对错误个数，后者控制的是错误比例。（2）Bonferroni方法只需要知道我们关注的某一策略设置的p值和总策略设置个数，但是后2种方法需要知道每一种策

略设置的 p 值，后 2 种的其他策略的 p 值该文章推荐从一个已经建立好的分布中取抽取。但是该分布只是关于多因子选股的。

通过上述 3 种方法计算完调整后的 p 值后，然后利用公式（3.2）计算出调整后的夏普率。调整后的 p 值就是（3.2）中的 p^M 。

4. BM

BM 在过拟合方面主要有 2 块内容，第 1 块内容是提供了一个严格的数学框架来定义过拟合和过拟合概率以及在其框架下某种特殊情况下的具体实施。第 2 块内容是以夏普率作为评价策略的指标，直接计算调整过拟合后的夏普率。第 1 块内容是非参数方法，第 2 块内容是参数方法。

4.1 过拟合概率（非参数方法）

该文章对理解回测过拟合很有帮助。

4.1.1 引言

回测是对一个算法策略如果在过去实施的话表现会怎样的历史模拟（A backtest is a historical simulation of how an algorithmic strategy would have performed in the past）。

什么样的金融实证发现是合格的？

金融发现通常涉及识别出一个低信噪比的现象，而且该信噪比由于竞争被降得更低。因为信噪比低，所以所以用来评估该现象是否存在的假设检验必须使用较大的样本。

当研究者在同一组数据上进行多个假设检验是，发现假正（false positives）的概率随着假设检验的个数的增加而增大。这会影响发现的合格性。

用历史数据去拟合参数使得策略在历史上表现最优，由于信噪比低，得到的参数很可能是从过去的噪音中获利而不是从未来的信号中获利。得到的结果就是回测过拟合（backtest overfit）。

以前的很多学术或实践文章几乎都没有说明他们的金融发现背后试验的次数（number of trials），所以他们的发现很可能是假正。该文章提供了一种专门应用在投资策略研究领域的方法，该方法通过计算的方式来控制由试验次数增加导致的增加的假正概率。

其他方法的缺点。

金融实践者应对过拟合的一个最常用的方法是 hold-out 方法，该方法优点是简单，但是缺点不少。（1）if the data is publicly available, it is quite likely that the researcher has used the “hold-out” as part of the IS（in sample） dataset。（2）Second, even if no “hold-out” data was used, any seasoned researcher knows well how financial variables performed over the time period covered by the OOS dataset, and that information may well be used in the strategy design, consciously or not。（3）hold-out 方法不适用与小样本。Weiss and Kulikowski 认为 hold-out 方法的样本数不应该小于 1000。（4）即使使用大样本，如果 OOS（out of sample）使用最后一段时间的数据，则我们得到的结果没有使用最新的信息，如果 OOS 使用最早一段时间的数据，OOS 的表现不具有代表性。（5）只要研究者试验了多个策略设置，过拟合总是存在。

另一个常用的处理过拟合的方法是，给一些金融变量建模，用这些金融变量的模型产生未来的随机情境（scenario），然后看策略在未来这些情境的表现。该方法的确定是用来产生未来情境的金融变量的模型有可能本身是过拟合的，也有可能对未来模拟得不够准确。

4.1.2 数学框架

考虑一个概率空间 $(\Gamma, F, Prob)$，其中，Γ 表示样本空间，其元素为 IS 样本和 OOS

样本组成的配对，F 近似理解为由样本空间的元素组成的集合的集合，即时间的集合，$Prob$ 是定义在 F 上的函数，自变量为事件，函数值为该时间的概率。我们的目标是估计下面的*策略选择过程*对应的过拟合概率：根据某个表现指标（如夏普率），使用回测，从 N 个策略里（这 N 个策略的标签分别为 $1,2,...,N$ ）选择一个最优的。给定表现指标，在 $(\Gamma, F, Prob)$ 上定义随机向量 $\mathbf{R}=(R_1,R_2,...,R_N)$ 和 $\bar{\mathbf{R}}=(\overline{R_1},\overline{R_2},...,\overline{R_N})$，分别表示 N 个策略设置在 IS 的表现和在 OOS 的表现。对于一个给定的样本 $c\in\Gamma$，它表示一个具体地由 IS 样本和 OOS 样本组成的配对，我们使用 $\mathbf{R}^c$ 和 $\bar{\mathbf{R}}^c$ 来表示对于该样本 c，这 N 个策略设置 IS 表现和 OOS 表现。

接下来的问题涉及这 N 个策略设置根据表现的 IS 排序和 OOS 排序。所以定义随机向量 $r=(r_1,r_2,...,r_N), r_i=1,...,N$ 和 $\bar{r}=(\overline{r_1},\overline{r_2},...,\overline{r_N}), \overline{r_i}=1,...,N$，分别表示这 N 个策略设置根据表现的 IS 排序和 OOS 排序。例如，表现指标为夏普率，$N=3$，对于某一样本 c，$\mathbf{R}^c=(0.5,1.1,0.7)$，$\bar{\mathbf{R}}^c=(0.6,0.7,1.3)$，$r^c=(1,3,2)$，$\bar{r}^c=(1,2,3)$。

另外，定义 $\Omega=\left\{(a_1,a_2,...,a_N)\middle| a_i=1,...,N \ for\ i=1,...,N \ and\ a_i\neq a_j \ for\ i\neq j\right\}$，可以看到，任意 r 或 $\bar{r}$ 都属于 Ω。定义 Ω 的子集 $\Omega_n^*=\left\{f\in\Omega\middle| f_n=N\right\}$，该集合下面的定义要用到。

回测过拟合定义（Backtest Overfitting）：我们说某个回测策略选择过程过拟合如果一个 IS 最优的策略设置在 OOS 的期望排序低于所有策略设置在 OOS 的中位数。即：

$$\sum_{n=1}^{N} E\left[\overline{r_n}\middle| r\in\Omega_n^*\right]Prob\left[r\in\Omega_n^*\right]\leq N/2 \quad (4.1)。$$

回测过拟合概率定义（Probability of Backtest Overfitting，PBO）：

$$PBO=\sum_{n=1}^{N} Prob\left[\overline{r_n}<N/2\middle| r\in\Omega_n^*\right]Prob\left[r\in\Omega_n^*\right] \quad (4.2)。$$

注意：上面的定义说的过拟合是针对最终的策略选择过程说的，不是针对策略下面的模型的模型拟合说的。例如，某个策略的信号使用过去移动窗口的数据做回归来确定的，那么上面的定义是针对移动窗口取多长说的，而不是针对回归方程式里的系数是多少说的。

关于上面的定义，我有个疑问：根据（4.1）的定义，过拟合不是一个 F 中的事件，所以过拟合不存在概率一说，但是（4.2）是另外一种意义上的过拟合概率的定义。该意义原文如下：

“Backtest overfitting is a deterministic fact (either the modelis overfit or it is not), hence it may seem unnatural to associate a probability to a non-random event. Given some empirical evidence and priors, we can infer the posterior probability that overfitting has taken place. Examples of this line of reasoning abound in information theory and machine learning treatises, e.g. [23]. It is in this Bayesian sense that we define and estimate PBO.”

关于这一段不太清楚什么意思。也许进一步阅读文献[23]才可能找到答案。

4.1.3 CSCV

上面的定义框架很一般，但是在具体地应用中计算 PBO 的时候，我们需要方法来估计上面抽象定义的概率，而估计概率很大程度上依赖于Γ中元素的选择。CSCV（combinatorially symmetric cross-validation）方法就是一个用来解决该问题的方法。具体如下。

第一步：构造一个$T\times N$的矩阵M。M的每一列表示对应策略设置在历史数据上回测得到的T个收益损失数据。对M的要求是：

（1）M是一个矩阵，M的同一行在时间上要同步；

（2）用来选择最优策略的表现指标必须能使用M的一列的子样本估计。

例如：如果表现指标是夏普率，我们希望独立同分布且正太的假设能在M的行与行之间成立。另外，如果不同的策略设置导致的交易频率不一样，那么观测应该被加总使得匹配一个共同的时间下标$t=1,...,T$。

第二步：把矩阵M在行的维度上切成S个子矩阵，S是偶数，每个子矩阵行数相同。所以每个子矩阵为T/S行，N列。

第三步：从S个子矩阵里，无放回抽取$S/2$子矩阵，总共有$C_s^{s/2}$种抽法，用集合C_S表示这$C_s^{s/2}$种抽法的集合。

第四部：对于集合C_S中的每个元素c，

（1）把抽出来的$S/2$个子矩阵按照原来的时间顺序合成$(S/2)(T/S)\times N=T/2\times N$的新矩阵$J$，称为训练集。

（2）用剩下的$S/2$个子矩阵按原来的顺序形成$T/2\times N$的新矩阵$\bar{J}$，称为测试集。（J和$\bar{J}$中子矩阵的顺序对某些表现指标无影响，如夏普率，对有些有影响，如最大回撤。）

（3）使用训练集得到 IS 表现向量$\mathbf{R}^c$，根据$\mathbf{R}^c$得到相应的 IS 表现排序向量r^c。

（4）使用测试集得到 OOS 表现向量$\bar{\mathbf{R}}^c$，根据$\bar{\mathbf{R}}^c$得到相应的 OOS 表现排序向量$\bar{r}^c$。

（5）决定n^*使得$r_{n^*}^c\in\Omega_{n^*}^*$。即$n^*$是 IS 表现最好的策略设置。

（6）定义$\bar{r}_{n^*}^c$的相对排序$\bar{\omega}_c:=\bar{r}_{n^*}^c/(N+1)\in(0,1)$，用来表示 IS 选取的最优策略设置在 OOS 的相对排序。

（7）定义 logit ： $\lambda_c=\ln\frac{\bar{\omega}_c}{1-\bar{\omega}_c}$。较高的 logit 表示 IS 和 OOS 表现一致，表明较低的回测过拟合。

第五步：收集所有的数据$\lambda_c, for\ c\in C_S$，计算 OOS 相对表现的概率分布。定义λ

发生的相对频率为

$$f(\lambda)=\frac{\sum_{c\in C_S}\chi_{\{\lambda\}}(\lambda_c)}{\#(C_S)}$$ （4.3），

其中 $\chi_{\{\lambda\}}(\lambda_c)=\begin{cases}1 if\ \lambda_c=\lambda\\0 if\ \lambda_c\neq\lambda\end{cases}$，$\#(C_S)$ 表示集合 C_S 中元素的个数。所以 $\int_{-\infty}^{+\infty}f(\lambda)d\lambda=1$。

4.1.4 过拟合统计量

通过 4.1.3，我们可以使用 4 种互补的分析来判断策略回测的可靠性。

（1）PBO：PBO 表示 IS 里最优的策略在 OOS 表现低于中位数的概率。利用 CSCV 方法，PBO 的估计为 $\phi=\int_{-\infty}^{0}f(\lambda)d\lambda$。PBO 至少有 3 种应用。（a）当 PBO 大于 0.05 时，拒绝模型。（b）PBO 可以在贝叶斯分析中用作先验概率，例如可以用它得到某个模型预测的后验概率。（c）(1-PBO)或 1/PBO 可以用作构建组合的权重。

（2）表现恶化（performance degradation）

对于所有的 $c\in C_S$，我们能得到 $(R_{n^*}^c,\bar{R}_{n^*}^c)$，回归 $\bar{R}_{n^*}^c=\alpha+\beta R_{n^*}^c+\varepsilon^c$，看 β 的显著性和正负号。大多数情况下 β 是负的。一个原因是，模型过度拟合到了过去的噪音上，以至于模型不适用于未来的信号。模型越是回测过拟合，记忆效应对未来的表现影响越大（记忆效应大概是说金融数据前一段时间越是偏离均值，接下来越容易出现反方向偏离，大概是均值回归）。

（3）损失概率（Probabilty of loss）:在 IS 最优的策略在 OOS 表现是负数的概率。即：$Prob(\bar{R}_{n^*}^c<0)$。在 ϕ=0 时，如果 $Prob(\bar{R}_{n^*}^c<0)$，那么 OOS 表现不好就不是过拟合的导致的，得找其他原因。

（4）随机占优（stochastic dominace）。用来评价 IS 的最优策略设置是不是比 IS 随机选择一个策略在 OOS 表现要好。

4.1.5 CSCV 方法的注意事项

（1）S 不能太大，太大会导致分出来的子矩阵不包含关键的时间结构，比如季节效应，月度效应；S 也不能太小，太小会导致 C_S 的元素个数太小，导致 logit 的频率分布估计不准。作者认为数据是 4~6 年，$S=16$ 是一个不错的选择。

（2）N 不能太小，太小会导致 logit 频率分布不连续，导致 PBO 估计误差大。推荐 $N>10$。

（3）T 是真正回测数据长度的 2 倍。

4.1.6 局限

4.1.2 的数学框架是具有一般性的，但是 4.1.3 的 CSCV 方法尽管设计的时候使用尽量少的假设和输入，但是 CSCV 毕竟是在 4.1.2 的基础上的具体实施，做了一些假设，所以还是有些局限性。

设计上的局限：CSCV 方法的构造训练集和测试集具有对称性，但是有的情况下，

使用 K-FCV（K 折交叉验证）更好。另外 CSCV 将样本表现分成了 S 个子矩阵，如果样本表现指标具有自相关性，那这样做会破坏该性质。最后，S 个子矩阵等权重，如果先验告诉我们不是等权重，那么应该使用不同的权重。

应用上的局限：（1）策略设置里应该包含所有理论上可行的策略设置。（2）该方法不评估回测的正确性。如果回测的时候输入的交易成本是错的，该方法就会基于错误的交易成本来评估过拟合概率。（3）该方法只考虑包含在 T 里的结构性变化。如果结构性变化发生在回测数据之外，我们的 PBO 不能将它考虑在内。所以这就要求回测数据要对未来具有代表性。（4）虽然高的 PBO 表明 N 个策略的选择存在过拟合，但是这 N 个策略里面可能包含有效的策略。例如：这 N 个策略都有效，且表现指标接近，那么 PBO 会很高。（5）CSCV 方法不能用来作为优化目标。把反过拟合的方法作为优化目标会产生新的过拟合。

4.2 缩减夏普率（Deflated Sharp Ration,DSR,参数方法）

该方法之所以称为参数方法，是因为该方法要利用到正太分布，且正态分布里的一些参数需要用数据估计。该方法思路是对夏普率做 2 项纠正，第 1 项纠正是如果收益率时间序列不是正太分布，那么常规的夏普率的估计误差受到偏度和峰度的影响，忽略该影响通常会使得夏普率的估计精度高估。第 2 项纠正是纠正过拟合的影响。

4.2.1 PSR（Probabilistic Sharp Ratio）-对非正太分布的调整

在比较宽松的条件下（是指收益率是 stationary and ergodic），夏普率的估计符合渐进正太分布，即：

$$(SR-SR)\xrightarrow{a}N(0,\frac{1+\frac{1}{2}SR^2-\gamma_3 SR+\frac{\gamma_4-3}{4}SR^2}{n}) \quad （4.4）$$

其中，SR 是夏普率的估计，SR 是真实夏普率，γ_3 是偏度，γ_4 是峰度，n 是观测个数。

为了体现偏度和峰度，作者定义了 PSR，其含义是 SR 大于某一基准夏普率 SR^* 的概率，具体为

$$PSR(SR^*)=Prob[SR>SR^*]=Z\left[\frac{(SR-SR^*)\sqrt{n-1}}{\sqrt{1-\gamma_3 SR+\frac{\gamma_4-1}{4}SR^2}}\right] \quad （4.5）$$

其中，$Z[\cdot]$ 表示标准正太分布的累积分布函数。对于很多对冲基金的策略，γ_3 为负（负的偏度），γ_4 大于 3（胖尾），所以在不考虑偏度和峰度时，（4.5）高估了夏普率大于基准夏普率的概率。需要注意的是，上面的夏普率不是年化夏普率，而是原始频率收益率的夏普率，引用作者原文，原因如下：

“It is not unusual to find strategies with irregular trading frequencies, such as weekly strategies that may not trade for a month. This poses a problem when computing an annualized Sharpe ratio, and there is no consensus as how skill should be measured in the context of irregular bets. Because PSR measures skill in probabilistic terms, it is invariant to calendar conventions. All calculations are done in the original frequency of the data, and

there is no annualization. This is another argument for preferring PSR to traditional annualized SR readings in the context of strategies with irregular frequencies”。

但是，具体应用的时候，只有基准夏普率是 0 的时候，年不年化无所谓；基准夏普率不为 0 的时候，必须年化才能比较。记平均每年的观测个数为q，则（4.5）的年化版本为：

$$PSR(SR^*) = Prob[SR > SR^*] = Z\left[\frac{(SR - SR^*)\sqrt{n-1}}{q\sqrt{1-\gamma_3 SR + \frac{\gamma_4 - 1}{4}SR^2}}\right]$$，其中SR和SR^*都是年化夏普率。

4.2.2 DSR（Deflated Sharp Ratio）

有N个策略设置，对每个策略设置回测，得到对应的夏普率估计$SR_n, n = 1,..., N$，假设这N个SR_n服从独立同分布，且服从正态分布，正太分布的均值为$E\left[\left\{SR_n\right\}\right]$，方差为$V\left[\left\{SR_n\right\}\right]$（这么假设的一个理由是这$N$个策略设置是同一大类策略下面的），则作者证明了，这N个SR_n的最大值的期望为：

$$E\left[\max\left\{SR_n\right\}\right] \approx E\left[\left\{SR_n\right\}\right] + \sqrt{V\left[\left\{SR_n\right\}\right]}\left((1\text{-}\gamma)Z^{-1}\left(1-\frac{1}{N}\right) + \gamma Z^{-1}\left(1-\frac{1}{N}e^{-1}\right)\right)$$

（4.6），其中$\gamma \approx 0.5772$。从该式可以看出，即使这N个策略设置真实无效，即$E\left[\left\{SR_n\right\}\right] = 0$，但是这$N$个策略设置的最大夏普率的期望不为 0，且随着$N$的增大而增大。

缩减夏普率（DSR）的定义就是应该（4.5），只不过其中的基准夏普率为（4.6）中的后半部分，即令

$SR^* = SR_0 = \sqrt{V\left[\left\{SR_n\right\}\right]}\left((1\text{-}\gamma)Z^{-1}\left(1-\frac{1}{N}\right) + \gamma Z^{-1}\left(1-\frac{1}{N}e^{-1}\right)\right)$，代入（4.5），得到缩减夏普率（DSR）的定义

$$DSR \equiv PSR(SR_0) = Z\left[\frac{(SR - SR_0)\sqrt{T-1}}{1-\gamma_3 SR + \frac{\gamma_4 - 1}{4}SR^2}\right] \quad （4.6）$$

其含义是，在扣除过拟合的影响后，同时考虑收益率的偏度和峰度，夏普率大于 0 的概率。个人观点：一个自然的理解是，如果不考虑偏度和峰度，只考虑过拟合，那么可以直接用$SR - SR_0$来表示调整过拟合后的夏普率。

上述推导，假设这N个SR_n服从独立同分布，如果不独立，那么上面的N应该是

其中独立的个数。在决定 N 个可能相关的随机变量里面有几个独立的变量时作者推荐的最好的方法是参考信息理论里面的熵（data compression, total correlation and multiinformation, e.g. Watabane [1960] and Studený and Vejnarová [1999]）。

4.2.3 什么是最优的试验（trials）数量

由于尝试的策略设置越多，即试验次数越多，过拟合的风险越大，所以作者推荐了一个最优的试验次数。作者文章里面没有给出证明，但是作者说对于该问题的解决最优停时理论（theory of optimal stopping, more concretely the so called “secretary problem”, or 1/e-law of optimal choice, see Bruss [1984]）提供了答案。具体如下：

从所有的理论上合理的策略设置里面随机抽取 $1/e$（约 37%）个策略设置，计算每个策略设置下的表现，然后从策略设置池里面剩下策略设置里逐个抽取策略设置，计算其表现，与其之前的所有策略设置的表现比较，直到它是最优的，则停止抽取。