幸运的因子

“琢璞”系列报告之四十一

深度报告

金融工程

本篇报告推荐 CR Harvey, Y Liu（2021）的《Lucky factors》发表在顶级期刊《Journal of Financial Economics》上，对于因子的选择文章提出了一种改良的方法，在挑选因子的过程中，可以更具有鲁棒性。他们的方法既可以应用于基于投资组合分类的资产定价测试，也可以应用于使用单个资产回报的测试，可以说应用性很强。

- 我们当下对因子的筛选，依然还是遵循主流的计量经济学方法，但是在时间维度上的单次回归有效性，可能来源于某种偶然，筛选出来的因子也未必是真正有效的因子，或者对原模型有增益效果的因子。
- CR Harvey, Y Liu（2021）的《Lucky factors》发表在顶级期刊《Journal of Financial Economics》上，提出了一种改良的方法，在挑选因子的过程中，可以更具有鲁棒性。简单来说，在挑选新因子的过程中，首先剥离掉原有因子的影响，而后在时间序列上将时间顺序打散，重复抽样，构建和计算概括统计量的分布特征，根据这一分布来检验新加入的因子是否具备增量的解释能力。他们的方法既可以应用于基于投资组合分类的资产定价测试，也可以应用于使用单个资产回报的测试，可以说应用性很强。
- 首先，文章描述了当前因子筛选存在的困境，由于因子的收益可能存在偶然性，基金经理的业绩表现也可能源于运气的成分，这在传统的框架下很难进行识别。
- 而后，他们详细地阐释了他们提出来的方法。使我们能够更好地识别运气成分。该方法基于一个自举采样方法（bootstrap，基于有限的数据使用可置换的重采样得到需要的样本数据），该方法能估计观测值的一般分布特征，重要的是，保留了数据中的横截面和时间序列相关性。我们的方法提供了具体的建议。例如，对于 5%的 p 值，我们的方法提供了一个临界检验统计量。在业绩评估中，这一临界检验统计值可以识别表现出色或逊色的基金。在资产定价中，它可以给我们提供一组特定的因子，即我们回答了有多少有效因子这个问题。
- 最后，文章做了实证模拟，对现有的 14 个被多篇文献所提及的风险因子根据他们提出的方法进行了重新挑选和排序，展示了该方法较强的普适性。
- 风险提示：本文内容和数据均依据美股的历史数据和已有的研究成果，在 A 股市场上的表现未得到验证。

相关报告

1、《“琢璞”系列报告之四十：因子敞口变动对共同基金业绩的影响》2021-08-18

2、《“琢璞”系列报告之三十九：结合隐含信息和历史信息的最优资产配置》2021-07-29

任瞳 S1090519080004
rentong@cmschina.com.cn
崔浩瀚 S1090519070004
cuihaohan@cmschina.com.cn

正文目录

图表目录

一、引言和推荐理由

从 2010 年以来，国内量化的发展始终绕不开对因子的研究，对于因子的组合，对因子的挖掘和对因子库的维护，一直是量化投资的“主流”方法。

我们当下对因子的筛选，依然还是遵循主流的计量经济学方法，但是在时间维度上的单次回归有效性，可能来源于某种偶然，筛选出来的因子也未必是真正有效的因子，或者对原模型有增益效果的因子。

CR Harvey, Y Liu（2021）的《Lucky factors》发表在顶级期刊《Journal of Financial Economics》上，提出了一种改良的方法，在挑选因子的过程中，可以更具有鲁棒性。简单来说，在挑选新因子的过程中，首先剥离掉原有因子的影响，而后在时间序列上将时间顺序打散，重复抽样，构建和计算概括统计量的分布特征，根据这一分布来检验新加入的因子是否具备增量的解释能力。他们的方法既可以应用于基于投资组合分类的资产定价测试，也可以应用于使用单个资产回报的测试，可以说应用性很强。

由于这篇文献篇幅很长，我们节选其中较为重要的内容进行了翻译，供投资者参详。

二、文献正文

摘要

我们提出了一种从一大组候选因子中进行选择的新方法，企图以此来解释期望收益的截面变化——候选中的许多因子可能是数据挖掘的结果。该方法对因子收益和资产收益的一般分布特征具有鲁棒性。我们考虑了时间序列和横截面相关性的可能性。该技术适用广泛的检验统计。我们的方法既可以应用于基于投资组合分类的资产定价测试，也可以应用于使用单个资产回报的测试。与最近的资产定价研究相反，我们对个股的研究发现，原始市场因素是迄今为止解释预期回报横截面的最重要因素。

关键词:因子，因子选择，变量选择，自举，数据挖掘，正交化，多重检验，预测回归，Fama-MacBeth，GRS。

1、研究背景

金融业中一些最重要的经济问题之间存在着共同的联系。例如，市场上有数千名基金经理，导致有些遵循随机策略的基金经理也可能表现出色，我们如何确定一名基金经理真正的“表现出色”？在众多其他变量中，我们如何评估股息率等变量是否能够预测股票回报？以最近发表的研究文章中的 300 多个变量作为候选因子，我们应该使用三因子模型还是五因子模型进行资产定价？常见的思路是多变量检验或数据挖掘。

本文提出了一种新的方法，使我们能够更好地识别运气成分。该方法基于一个自举采样方法（bootstrap，基于有限的数据使用可置换的重采样得到需要的样本数据），该方法能估计观测值的一般分布特征，重要的是，保留了数据中的横截面和时间序列相关性。我们的方法提供了具体的建议。例如，对于 5%的 p 值，我们的方法提供了一个临界检验统计量。在业绩评估中，这一临界检验统计值可以识别表现出色或逊色的基金。在资产定价中，它可以给我们提供一组特定的因子，即我们回答了有多少有效因子这个问题。

在预测回归中考虑下面的例子来说明我们面临的问题。假设我们有 100 个 X 变量来预测一个变量 Y。我们的第一个问题是，100 个 X 变量中是否有一个看起来是单独重要的。这并不像人们想象的那么简单，在传统层面上意义重大的东西可能是运气使然。我们还需要考虑 X 变量之间的相关性，因为如果 X 变量高度相关，则 t 统计量较大可能并非是这个变量自身造成的。假设这些问题已经得到解决，并且我们找到了一个重要的因子，那么我们如何寻找下一个呢？据推测，除了第一个变量有预测作用之外，第二个变量还需要用来预测 Y。如此可以尝试 99 个变量，这一额外的可预测性需要再次进行仔细检查。假设我们确定第二个变量是一个显著的预测因子，那什么时候停止继续检验？最后，假设我们不是预测回归，而是试图确定在横横截面回归中有多少因子是重要的，我们的方法又该如何改变？

我们提供了一个新的框架来回答上述问题。相比于现有的研究，我们的方法有如下特点：

首先，我们把数据挖掘（datamining）考虑在内。鉴于学术界和金融业在挖掘新因子方面的共同努力，这一点非常重要。数据挖掘对假设检验有很大的影响。在单个变量测试中，使用单变量 X 解释 Y 时，t 值为 2.0 足以在 95%的置信度拒绝原假设。但当有 100 个 X 变量且相互独立时，最大 t 值的 2.0 阈值对应于 99%的 p 值，甚至不接近可接受的 5%。我们的论文提出了适当的统计截止值来控制候选变量之间的关系。

提出横截面独立性假设是为了方便说明数据探索偏差，它是一个很强的假设。首先，对于我们的大多数应用中来说，这是不现实的，因为几乎所有的经济和金融变量都以复杂的方式内在地联系在一起。其次，脱离独立假设可能会对结果产生很大影响。例如，在我们前面的示例中，如果所有 100 个 X 变量完全相关，则无需进行多次测试调整，99%的 p 值会错误地高估原始值。最近关于共同基金业绩的研究表明，考虑横截面变量之间的依赖性可以极大地改变推断结果。

我们的论文提供了一个对变量之间的横截面依赖的形式和数量具有鲁棒性的框架，同时也保持了数据矩阵中的相关性信息，包括高阶矩和非线性相关性。此外，由于在有限样本中很难测量较高的矩相关性，并且这可能会使推断结果产生偏差，因此我们的方法考虑了不确定性采样（观察到的样本可能低估了从中提取样本的总体）并提供了不依赖于渐近近似的推断。

我们的使用一种叫自举采样的方法。当数据在时间上是独立的时，我们有放回的随机抽取时间段。重要的是，在抽取一个特定的时间段时，我们会绘制该时间点整个截面情况。以此来保留依赖于数据结构的同时期截面。此外，通过将重采样数据的大小与原始数据进行匹配，我们能够获得原始样本的采样不确定性。当数据随时间变化而变化时，我们使用分块采样以捕捉时间序列的变化，这与 White（2000）和 PolitisandRomano（1994）的想法类似。本质上，我们的方法重新构造了回归模型中的多假设检验问题，允许使用自举采样方法进行直观且无分布的推断。

从实证上看，我们展示了如何将我们的方法应用于预测回归和横截面回归模型，这两个研究领域的数据探索造成的偏差可能是最严重的。然而，我们的方法也适用于其他类型的回归模型。本质上，我们提供的是一种通用方法，当研究人员面临多个变量可供选择时，可以指定回归模型。

我们的文章补充了最近关于期望收益截面的多维性的文献。Harvey、Liu 和 Zhu（2016）通过研究学术界发现的 316 个因子，提供了一个调整数据挖掘的多重测试框架。Green、Hand 和 Zhang（2013）研究了 330 多个基于会计核算的收益预测信号，并通过适当组合这些信号展示出较大的多元化收益。McLean 和 Pontiff（2015）使用样本外方法研究发现异常的发表后偏差。这篇文献的总体发现是，许多发现的因子可能是错误的。但有多少因子是真正的因子呢？我们提供了一个新的测试框架，可以在回归模型的环境中同时处理多重测试、变量选择和测试相关性。

我们的方法受到许多有影响力的论文的启发并与之相关，特别是 Foster、Smith 和 Whaley（FSW，1997）以及 Fama 和 French（FF，2010）。在时间序列预测的应用中，FSW 在无预测性的原假设下模拟数据，以帮助识别真正的预测因子。对实际数据使用 bootstrap，可以应用到许多检验统计量，而不需要诉诸渐近逼近。更重要的是，我们的方法可以适用于研究横截面回归，其中风险负荷可能随时间改变的。在基金经理评估的应用中，FF(2010)(另见，Kosowskietal.，2006,Barrasetal.，2010,andFersonandChen,2015)采用了保留截面依赖性的自举采样方法。我们的方法与他们的不同之处在于，我们能够确定一个特定的临界值，通过这个对比临界值，我们得知一个经理的表现是否优异，或者一个因子是否有效的。

尽管已经发现了数百个因子，但在价值加权个股分析下夏普提出仍然是关键因子。我们也发现盈利因子的作用，但它的贡献很小。我们的等量加权分析提供了一些价值和规模因子的证据，但与价值加权分析一致，与市场因子相比，这些额外因素的贡献不大。引人注目的是，市场因子是公司金融实践中使用的主导因子(见，Graham 和 Harvey,2001)，但这一因子在资产定价研究中长期不受青睐。我们的论文组织如下。在第二部分中，我们将展示我们的测试框架。在第三部分，我们将我们的方法应用于风险因子的选择。我们提供基于传统投资组合分类的测试以及基于单个资产的原始测试的见解。最后一节给出了一些结束语。

2、提出的方法

我们的框架在预测回归的背景下得到了最好的说明。我们强调了该方法和目前的实践之间的差和该方法与研究的关联。

然后我们扩展该方法以适应横截面回归。

2.1 预测回归

假设我们需要使用 T*M 的包含时间序列 M 的自变量矩阵 X 预测 T*1 的收益向量 Y，即是矩阵 X 中的 i 列（Xi）表示变量 i 的时间序列。我们的目标是选择一个 M 回归量子集，以形成最佳预测回归模型。假设我们通过概括统计 ψ 来度量回归模型的拟合优度，我们的框架允许使用任意的性能度量 ψ，比如，R^2,t-统计量和 F-统计量。这一特性源于我们对 bootstrap 方法的使用，它不需要对概括统计进行任何分布假设来构建测试。相反，Foster,Smith 和 Whaley(FSW,1997)需要 R^2 上的有限样本分布来构建他们的测试。为了简化介绍，我们在描述方法时考虑通常的回归 R^2，但在必要时将指出不同之处.

我们基于自举采样的多重测试调整的增量因子选择流程包括三个主要步骤:

第一步、原假设下的正交化

假设我们已经选择了 k(0≤k<M)个变量，并想测试是否存在另一个显著的预测因子，如果存在，找到这个因子。在不丢失一般性的前提下，假设第前 k 个变量是预先选定的变量，我们在剩下$M-k$个变量中进行测试，即是$\{X_{k+j}, j=1,\dots,M-k\}$，我们的原假设是，这些候选变量都不能对 Y 提供额外的解释能力,见 White(2000)和 FSW(1997).这一步的目标是修改数据矩阵 X，使该原假设在样本内为真。

为此，我们首先将 Y 投影到预先选定的一组变量上，得到投影残差向量$Y^{e,k}$，这个残差向量包含了无法用预先选定的变量解释的部分信息。然后我们将$M-k$个候选变量与 $Y^{e,k}$ 进行正交化，使正交化的变量与整个样本的$Y^{e,k}$不相关。特别的，我们单独的将$X_{k+1}, X_{k+2}, \dots, X_M$投影到$Y^{e,k}$上并得到残差$X^e_{k+1}, X^e_{k+2}, \dots, X^e_M$,即是，

$$X_{k+j} = c_j + d_j Y^{e,k} + X^e_{k+j}, j = 1, \dots, M \quad (1)$$

其中 c_j 是截距，d_j 是斜率，X^e_{k+j}是剩余向量。通过构造，这些残差与$Y^{e,k}$的样本内相关性为零。因此，如果假设 X 和$Y^{e,k}$之间的联合正态性，它们似乎是独立于$Y^{e,k}$的。

这与 FSW(1997)的模拟方法类似，人工生成的独立回归量被用来量化多重检验的效果。我们的方法与 FSW 不同，因为我们使用的是真实数据。此外，我们使用 bootstrap 或分块 bootstrap 来近似检验统计量的经验分布。

我们实现了与 FSW 相同的目标，同时尽可能少地丢失回归量之间的相关信息。特别地，我们的正交化保证了$M-k$个正交化候选变量与样本内$Y^{e,k}$不相关。这类似于 FSW(1997)中模拟回归器与因变量之间的独立性要求。我们的方法是没有分布要求的，并在回归量中保持尽可能多的信息。我们简单地从每个候选变量中清除$Y^{e,k}$，从而保留所有与$Y^{e,k}$不线性相关的变量之间的分布信息。例如，所有变量(包括预选变量和候选变量)之间的尾部相关关系被保留。这是很重要的，因为在有限样本中，较高的矩相关可能会对检验统计量产生巨大的影响。

最近有关共同基金表现的文献也采用了类似的观点。特别是 Kosowski 等人(2006)和 FamaandFrench(2010)从基金收益中减去样本内拟合的 alpha，从而创造出“伪”基金，这些基金恰好产生样本内平均收益为零。类似地，我们正交化候选回归量，使它们与因变量中剩下的解释的相关性为零。

第二步、自举采样（bootstrap）

我们把预先选定的变量排列成$X^s = [X_1, X_2, \dots, X_k]$,把正交候选变量排列成$X^e = [X^e_{k+1}, X^e_{k+2}, \dots, X^e_M]$,注意，对于剩余向量$Y^{e,k}$和两个回归矩阵 X^s 和 X^e，行表示时间周期，列表示变量。我们对时间周期(即行)使用自举采样来生成不同回归模型概括统计量的经验分布。特别地，对于时间指标$t^b = [t^b_1, t^b_2, \dots, t^b_T]^T$的每一组截面数据，让相应的左右变量分别为 Y^{eb}、X^{sb} 和 X^{eb}.

下面的图表说明了我们如何自举。假设我们有五个周期，一个预先选择的变量 X^s，和一个候选变量 X^e，原始时间指标为$[t_1=1, t_2=2, t_3=3, t_4=4, t_5=5]^T$，通过可置换重采样，自举采样样本的时间索引的一种可能情况是$t^b = [t^b_1 = 3, t^b_2 =$

$2, t_3^b = 4, t_4^b = 3, t_5^b = 1]^T$,图中显示了我们如何将原始数据矩阵转换为基于新时间索引的引导数据矩阵。

$$[Y^{e,k}, X^s, X^e] = \begin{bmatrix} y_1^e & x_1^s & x_1^e \\ y_2^e & x_2^s & x_2^e \\ y_3^e & x_3^s & x_3^e \\ y_4^e & x_4^s & x_4^e \\ y_5^e & x_5^s & x_5^e \end{bmatrix} \begin{pmatrix} t_1 = 1 \\ t_2 = 2 \\ t_3 = 3 \\ t_4 = 4 \\ t_5 = 5 \end{pmatrix} \Rightarrow \begin{pmatrix} t_1^b = 3 \\ t_2^b = 2 \\ t_3^b = 4 \\ t_4^b = 3 \\ t_5^b = 1 \end{pmatrix} \begin{bmatrix} y_3^e & x_3^s & x_3^e \\ y_2^e & x_2^s & x_2^e \\ y_4^e & x_4^s & x_4^e \\ y_3^e & x_3^s & x_3^e \\ y_1^e & x_1^s & x_1^e \end{bmatrix} = [Y^{eb}, X^{sb}, X^{eb}]$$

返回具有 k 个预先选择的变量和 M-k 候选变量的一般情况，我们进行自举采样然后运行$M-k$个回归。每一个回归都涉及到 Y^{eb} 对数据矩阵 X^{eb} 中的一个候选变量的投影。让相关的概括统计量为$\Psi^{k+1,b}$，$\Psi^{k+2,b}$，…，$\Psi^{M,b}$，并让这些概括统计量中的最大值为Ψ_I^b，直观来说，Ψ_I^b测量了最佳拟合模型的性能，该模型使用正交候选变量列表中的一个变量来优化预先选定的回归模型。

最大统计量控制数据探索偏差。由于有$M-k$个因子可供选择，被选择的因子可能是因为随机选择而显得重要。与 White(2000)、Sullivan、TimmermannandWhite(1999)和 FSW(1997)做法类似，我们采用最大统计量作为控制多个假设检验的检验统计量。我们的自举采样方法允许我们获得在没有$M-k$变量为真的联合原假设下的最大统计量的经验分布。由于多次检验，这种分布与单个检验中检验统计量的原假设分布有很大的不同。通过将实现的(数据中的)最大统计量与该分布进行比较，我们的检验考虑了多重检验。

我们应该使用哪个统计数据来概况候选列表中一个变量的额外贡献？根据回归模型的不同，选择也不同。例如，在预测回归中，我们通常使用 R^2 或调整后的 R^2 作为概括统计量。在横截面回归中，我们使用 t 统计量来检验平均斜率是否显著。我们方法的一个特点是，它不需要一个显式的表达式来表示测试统计量的原分布。因此，它可以很容易地容纳不同类型的概括统计信息。相反，FSW(1997)只适用于 R^2。

对于我们方法的其余描述，我们假设度量候选列表中变量的增量贡献的统计量已经给出，对于第 b 个引导样本，我们一般将其表示为Ψ_I或Ψ_I^b。我们自举采样 B=10000 次得到结果$\{\Psi_I^b, b = 1,2,\dots,B\}$,表示为

$$(\Psi_I)^B = \{\Psi_I^b, b = 1,2,\dots,B\}. \quad (2)$$

这是Ψ_I的经验分布，它衡量了当一个正交化回归量被考虑时对回归模型的最大额外贡献。假如这些正交化回归量中没有一个是真正对总量预测有贡献的，$(\Psi_I)^B$给出了当原假设为真时这个最大额外贡献的分布，即$M-k$个候选变量都不为真。$(\Psi_I)^B$是 FSW(1997)中 R^2 最大值分布的自举采样模拟。与 White(2000)类似并优于 FSW(1997)，我们的自举采样方法本质上是无分布的，允许我们通过样本扰动获得检验统计量的确切分布。

我们的自举采样样本与原始数据具有相同的时间段数量。这使我们能够将原始数据的抽样不确定性与估计样本进行匹配。当数据中的时间相关性很小时，我们简单地将每个时间段作为抽样单位，并进行样本替换。当时间相关度较大时时，我们使用时间块作取样估计，详细解释在附录。在这两种情况下，我们只对时间段进行重新采样，保持截面的完整性，以保持变量之间的同期相关性。

第三步:假设检验和变量选择

对原始数据矩阵 X 进行处理，我们可以得到一个Ψ_I统计量，它衡量一个候选变量的最大额外贡献。我们用Ψ_I^d表示这个统计量。假设检验第(k+1)个显著因子是否存在，相当于将Ψ_I^d与Ψ_I在原假设下的分布进行比较，即$(\Psi_I)^B$。对于预先指定的显著性水平 α，例如 5%，如果Ψ_I^d超过$(\Psi_I)^B$的(1-α)百分位数，则拒绝原假设，即

$$\Psi_I^d > (\Psi_I)_{1-\alpha}^B, (3)$$

$(\Psi_I)_{1-\alpha}^B$是$(\Psi_I)^B$的(1-α)百分位数。

假设检验的结果告诉我们，在考虑多重检验后，在剩余的$M-k$个候选变量中是否存在显著的预测因子。如果结果是确定的，我们认为具有最大检验统计量的变量(即Ψ_I^d)为有效的，并将其纳入预先选定的变量列表中。在所有候选变量都加入或者排除之前，我们从第一步开始测试下一个预测量。否则，我们终止算法并得出最终结论，即预先选择的变量组是唯一重要的。

在 Fama-MacBeth 回归上的应用

我们的方法可以应用在 Fama-MacBeth 回归（FM 回归）上，Fama-MacBeth 回归用于计算因子载荷。将我们的方法应用于 FM 回归的一个障碍是横截面回归中的时变斜率。特别是，对每个时间段执行单独的横截面回归以获得横截面回归斜率的集合。我们通过查看时间维度上的平均横截面斜率系数来测试一个因子的显着性。因此，在 FM 框架中，原假设是总体的斜率为零。我们调整我们的方法，使这个条件完全适用于调整后的回归模型。

首先，我们需要正交化。假设我们在基础模型上运行 FM 回归并获得残差超额收益。特别地，在时间 t，让剩余超额收益的向量为Y_t。我们测试候选因子在解释预期回报横截面方面的增量贡献。假设候选因子的风险载荷向量（即β）为X_t。假设在时间 t 的横截面中有n_t个资产，因此Y_t和X_t的维度都是$n_t \times 1$。请注意，n_t可以是时间相关的，因为我们的方法可以直接处理不平衡面板。在典型的 FM 回归中，我们会将Y_t投影到X_t上。为了我们的正交化工作，我们逆转了这个过程，类似于我们在预测回归中所做的。举一个特殊情况的例子，我们将Y_t和X_t的集合堆叠到维度为$\sum_{t=1}^{T} n_t \times 1$的两个列向量中，并运行以下约束回归模型：

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} = \begin{bmatrix} \emptyset_1 \\ \emptyset_2 \\ \vdots \\ \emptyset_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \xi_{1\times 1} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1}$$

其中$\emptyset_T$是时间 t 的截距常数向量，$\xi_{1\times 1}$是标量，$[\varepsilon_1', \varepsilon_2', \cdots, \varepsilon_T']'$是将用于后续自举分析的预测回归量的向量。这是一个约束回归，因为我们在整个样本中有一个单一的回归斜率（即ξ）。如果允许不同的时间斜率，将得到普通的无约束回归模型，其中X_t逐个投影到Y_t上。具有单个斜率系数是我们实现 FM 模型样本内原假设的关键。

或者，我们可以将上述回归模型视为我们在预测回归中使用的正交化过程的变形。它将回报和因子载荷汇集在一起以估计单个斜率系数。然而，不同的是在不同的时间段使用不同的截距。这是很自然的，因为 FM 程序允许随时间变化的截距和斜率。为了从X_t中清除Y_t的影响，我们还需要考虑随时间变化的截距。在数学上，与时间相关的截距允许回归残差在每个周期内总和为零。这个性质证明非常重要，因为它允许我们在样本中形成 FM 零假设，我们将在后面提到。接下来，我们通过其平方和 ε′ε 归一化每个残差向量 ε，并生成正交化的回归向量：

$$X_t^e = \varepsilon_t / \varepsilon_t' \varepsilon_t, \quad t = 1,2,\cdots T$$

这些正交回归量是预测回归中正交回归量的 FM 对应物。它们满足横截面回归中的 FM 原假设。特别是，假设我们使用这些正交回归向量为每个周期运行横截面 OLS：

$$Y_t = \mu_t + \gamma_t X_t^e + \eta_t, \qquad t = 1,2,\cdots,T$$

其中μ_t是截距的$n_t \times 1$向量，γ_t是第 t 个周期的标量斜率，η_t是残差的$n_t \times 1$向量。我们在附录 A 中表明以下 FM 零假设在样本中成立：

$$\sum_{t=1}^{T} \gamma_t = 0$$

上述正交化是我们需要适应以将我们的方法应用于 FM 程序的唯一步骤。我们方法的其余部分遵循 FM 回归中的因子选择。特别是，通过预先选择的一组右侧变量，我们将其余的右侧变量正交化以形成联合零假设，即它们都不是真正的因素。然后我们自举来检验这个零假设。如果我们拒绝，我们将最重要的一个添加到预先选择的变量列表中，然后重新开始测试下一个变量。否则，我们将停止并以一组预先选择的变量结束。

实证部分：识别有效因子

风险因子的选择

原则上，我们的方法可以用来对所提出的风险因子进行排序。我们方法的一个吸引人的特点是它允许风险因子的数量

大于测试投资组合的数量，这在传统的多元回归模型中是不可行的。然而，我们在当前的论文中不追求这一点，而是专注于选定的一组有显著影响的风险因子。相比之下，个股避免了随意的投资组合构建。我们会将我们的方法，应用于流行的测试投资组合和个股。

下文对 Fama 和 French 等人提出的 14 个因子（包括市值、账面市值比、ROE、投资因子、毛利率、动量、波动率等）进行挑选和排序。

我们专注于因子在经济上合理以及统计学上的显著性。直观地说，在我们的上下文中，一个好的测试统计量应该能够说明在解释基线模型和向基线模型添加一个额外变量的增强模型之间的预期收益横截面时的差异。对于面板回归模型，让$\{a_i^b\}_{i=1}^N$和$\{a_i^g\}_{i=1}^N$分别是基线模型和增强模型的回归截距的横截面。令$\{s_i^b\}_{i=1}^N$为基线模型下回归截距的标准误差的横截面。我们的第一个检验统计量由下式给出

$$SI_{ew}^m \equiv \left(\frac{1}{N}\sum_{i=1}^{N}|a_i^g|/s_i^b - \frac{1}{N}\sum_{i=1}^{N}|a_i^b|/s_i^b\right)\Big/\frac{1}{N}\sum_{i=1}^{N}|a_i^b|/s_i^b$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}|a_i^g| - |a_i^b|/s_i^b}{\frac{1}{N}\sum_{i=1}^{N}|a_i^b|/s_i^b}$$

其中SI表示标准化截距，ew表示等权重，m表示均值。直观地，SI_{ew}^m测量绝对回归截距中的百分比差异，由基线模型下回归截距的标准误差缩放。如果增强模型改进了基础模型，SI_{ew}^m则为负。在原假设下生成的自举经验分布评估了改进的重要性，即增强模型中的附加变量在解释预期收益的横截面时的增量贡献为零。

虽然SI_{ew}^m计算了标准化后的平均绝对值的百分比差异，但它可能对横截面中的极端观察不稳健，尤其是当我们使用个股作为测试资产时。因此，我们还考虑了计算缩放中位数绝对截距百分比差异的稳健版本，即：

$$SI_{ew}^{med} = \left(median\left(\{|a_i^g|/s_i^b\}_{i=1}^N\right) - median\left(\{|a_i^b|/s_i^b\}_{i=1}^N\right)\right)/median\left(\{|a_i^b|/s_i^b\}_{i=1}^N\right)$$

其中med代表的就是中位数。

我们检验统计量有效性的一个关键假设是横截面平均$|a_i^g|$应小于横截面平均$|a_i^b|$如果增强模型中的附加因子是真有效的风险因子。在个人资产层面，$|a_i^g|$将小于$|a_i^b|$在人口中（即，我们有足够长的因子并返回时间序列），如果增强模型是真正的潜在因子模型。

案例：风险因子的挑选

将我们的模型用于经典的测试集合中，我们使用了 KenFrench 的在线数据库中提供的市值与账面市值比的 25 组顺序排列的投资组合。

然后我们在此基础上挑选 14 个风险因子。这 14 个风险因子(包含了市值、账面市值比、ROE、投资因子、毛利率、动量、波动率、质量等)。

表 1 1968 年至 2012 年数据的描述性统计

Table 1: **Summary Statistics, January 1968 - December 2012**

Summary statistics on portfolios and factors. We report the mean annual returns for Fama-French size and book-to-market sorted 25 portfolios and the five risk factors in Fama and French (2015a) (i.e., excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*)), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility (*civ*) in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). We also report the correlation matrix for factor returns. The sample period is from January 1968 to December 2012.

Panel A: Portfolio Returns

	Low	2	3	4	High
Small	0.009	0.078	0.085	0.106	0.120
2	0.039	0.074	0.095	0.101	0.108
3	0.047	0.082	0.082	0.093	0.119
4	0.062	0.061	0.077	0.087	0.090
Big	0.046	0.061	0.053	0.059	0.069

Panel B.1: Factor Returns

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
Mean	0.052	0.022	0.048	0.081	0.024	0.055	0.068	0.057	0.048	0.105	0.039	0.047	0.033	0.060
t-stat	[2.17]	[1.32]	[3.08]	[3.54]	[1.84]	[2.99]	[5.09]	[5.76]	[3.74]	[5.98]	[3.24]	[4.44]	[2.92]	[3.48]

Panel B.2: Factor Correlation Matrix

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
mkt	1.00													
smb	0.30	1.00												
hml	-0.32	-0.24	1.00											
mom	-0.14	-0.03	-0.15	1.00										
skew	-0.02	-0.05	0.23	0.03	1.00									
psl	-0.05	-0.04	0.03	-0.03	0.10	1.00								
roe	-0.19	-0.39	-0.11	0.51	0.19	-0.06	1.00							
ia	-0.39	-0.26	**0.69**	0.04	0.15	0.02	0.04	1.00						
qmj	-0.54	-0.54	0.02	0.26	0.13	0.03	**0.68**	0.15	1.00					
bab	-0.09	-0.07	0.40	0.18	0.24	0.06	0.25	0.35	0.19	1.00				
gp	0.08	0.06	-0.34	0.01	-0.01	-0.03	0.34	-0.26	0.45	-0.11	1.00			
cma	-0.41	-0.16	**0.71**	0.01	0.05	0.03	-0.10	**0.90**	0.07	0.32	-0.34	1.00		
rmw	-0.21	-0.42	0.11	0.10	0.27	0.03	**0.68**	0.05	**0.76**	0.26	0.49	-0.08	1.00	
civ	0.17	0.27	0.13	-0.18	0.04	0.05	-0.26	-0.00	-0.28	0.11	-0.00	0.04	-0.10	1.00

资料来源：Harvey C R, Liu Y. Lucky factors[J]. Journal of Financial Economics, 2021. 招商证券定量研究团队整理

表 2 将因子组合作为测试的资产

Table 2: **Portfolios as Test Assets**

Test results on 14 risk factors using Fama-French size and book-to-market sorted 25 portfolios. (See Table 1 for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., SI_{ew}^{m} and SI_{ew}^{med}), which measure the difference in equally weighted scaled mean/median absolute regression intercepts, are defined in Section 3.2. GRS reports the Gibbons, Ross and Shanken (1989) test statistic.

Panel A: Baseline = No factor							
		single test			single test		
Factor	SI_{ew}^{m}	5th-percentile	p-value	SI_{ew}^{med}	5th-percentile	p-value	GRS
mkt	**-0.607**	[-0.340]	(0.002)	**-0.672**	[-0.333]	(0.000)	4.290
smb	-0.209	[-0.243]	(0.072)	-0.108	[-0.257]	(0.215)	4.402
hml	0.189	[-0.100]	(0.999)	0.230	[-0.110]	(0.997)	4.050
mom	0.224	[-0.108]	(0.998)	0.256	[-0.120]	(0.998)	4.302
skew	-0.014	[-0.040]	(0.195)	0.007	[-0.053]	(0.731)	4.454
psl	0.043	[-0.038]	(0.946)	0.054	[-0.044]	(0.952)	4.286
roe	0.504	[-0.150]	(1.000)	0.470	[-0.144]	(0.999)	4.919
ia	0.607	[-0.157]	(1.000)	0.637	[-0.164]	(1.000)	4.553
qmj	0.820	[-0.275]	(0.990)	0.806	[-0.273]	(0.983)	5.594
bab	0.036	[-0.042]	(0.952)	0.030	[-0.055]	(0.908)	**3.718**
gp	-0.042	[-0.037]	(0.039)	0.026	[-0.049]	(0.892)	4.096
cma	0.450	[-0.143]	(1.000)	0.464	[-0.155]	(0.999)	4.238
rmw	0.268	[-0.126]	(0.991)	0.273	[-0.124]	(0.987)	4.325
civ	-0.281	[-0.140]	(0.000)	-0.283	[-0.141]	(0.002)	4.132
		multiple test			multiple test		
	min	**[-0.368]**	**(0.003)**		**[-0.373]**	**(0.000)**	

Panel B: Baseline = *mkt*					
	single test			single test	
SI_{ew}^{m}	5th-percentile	p-value	SI_{ew}^{med}	5th-percentile	p-value
-0.068	[-0.174]	(0.251)	-0.007	[-0.211]	(0.481)
-0.434	[-0.260]	(0.000)	-0.397	[-0.302]	(0.009)
0.218	[-0.071]	(0.999)	0.210	[-0.113]	(0.985)
-0.116	[-0.085]	(0.025)	-0.134	[-0.117]	(0.039)
-0.038	[-0.034]	(0.040)	-0.135	[-0.055]	(0.004)
0.375	[-0.106]	(1.000)	0.366	[-0.137]	(0.998)
-0.318	[-0.168]	(0.001)	-0.262	[-0.206]	(0.012)
0.560	[-0.134]	(1.000)	0.898	[-0.173]	(1.000)
-0.442	[-0.154]	(0.000)	-0.447	[-0.179]	(0.000)
0.202	[-0.087]	(1.000)	0.200	[-0.128]	(0.988)
-0.476	[-0.196]	(0.000)	**-0.500**	[-0.225]	(0.000)
0.055	[-0.056]	(0.991)	0.132	[-0.119]	(0.962)
-0.219	[-0.094]	(0.001)	-0.099	[-0.128]	(0.088)
	multiple test			multiple test	
min	**[-0.289]**	**(0.001)**		**[-0.342]**	**(0.000)**

Panel C: Baseline = *mkt* + *cma*					
	single test			single test	
Factor	SI_{ew}^{m}	5th-percentile	p-value	SI_{ew}^{med}	5th-percentile	p-value
mkt					
smb	-0.232	[-0.353]	(0.171)	**-0.295**	[-0.454]	(0.188)
hml	0.001	[-0.136]	(0.657)	0.013	[-0.230]	(0.615)
mom	0.091	[-0.067]	(0.981)	0.115	[-0.139]	(0.930)
skew	0.005	[-0.058]	(0.654)	0.093	[-0.134]	(0.896)
psl	-0.027	[-0.028]	(0.054)	0.222	[-0.069]	(0.992)
roe	0.911	[-0.128]	(1.000)	1.271	[-0.228]	(1.000)
ia	0.382	[-0.106]	(1.000)	0.631	[-0.181]	(1.000)
qmj	1.381	[-0.153]	(1.000)	1.857	[-0.242]	(1.000)
bab	0.101	[-0.069]	(0.991)	0.080	[-0.153]	(0.880)
gp	**-0.260**	[-0.073]	(0.061)	-0.084	[-0.104]	(0.081)
cma					
rmw	0.561	[-0.119]	(1.000)	0.644	[-0.188]	(1.000)
civ	-0.160	[-0.100]	(0.013)	-0.214	[-0.211]	(0.049)
	multiple test			multiple test	
min	**[-0.356]**	**(0.148)**		**[-0.464]**	**(0.253)**

资料来源：Harvey C R, Liu Y. Lucky factors[J]. Journal of Financial Economics, 2021. 招商证券定量研究团队整理

表 1 是对上述因子和数据进行了描述性统计，重点看表 2。

我们测试了这是 14 个因子，观察者 14 个因子中是否有能独立解释收益的因子。观察表 2 中的 Panel2，市值因子（mkt）是其中表现最好的，截距项减小了 61%。为了评估市值因子的重要性，我们按照我们的方法将 14 个因素正交化，这样它们对样本中预期回报的横截面的影响为零。我们引导以获得单个测试统计量的经验分布。然后我们根据这些经验评估已实现的测试统计数据。更准确地说，市值因子减少截距项 61%，SI_{ew}^{m}的经验分布下的 p 值为 0.002。

基于表 2 的一项有趣观察是，选择的最佳因子可能不是单个测试 p 值最低的因子。例如，在表 2 中的异质波动率因子（civ）单一测试 p 值较低，但市场因素是我们选择的第一个因子。因为市值因子对截距项的减小作用更加明显。

以此类推，根据 panel2 的情况，第二个入选的因子就是投资因子（cma），在已经选择了了市值因子的前提下，投资因子对截距项的减小作用最为明显。这样便实现了对因子的挑选和排序。

结论

我们提出了一种新方法，使研究人员能够应对金融经济学中多重测试的挑战。我们的方法基于自举方法（bootstrap），并考虑到一般分布特征、横截面以及时间序列相关性以及一系列测试统计数据。

我们将我们的方法应用于风险因子的识别。资产定价中提出了数百个因素来解释预期收益的横截面。有些可能只是偶然出现的重要风险因子。此外，实证资产定价研究长期以来一直怀疑：按某些特征排序的投资组合会影响新因素的发现。

最后，虽然我们已经将我们的方法应用于金融中的因子发现，但我们想强调的是，我们的方法可以应用于面临多重测试问题的金融或金融之外的任何回归模型。事实上，越来越需要新工具来导航大量的“大数据”。我们提供一个新的研究方向。

三、总结

从 2010 年以来，国内量化的发展始终绕不开对因子的研究，对于因子的组合，对因子的挖掘和对因子库的维护，一直是量化投资的“主流”方法。

我们当下对因子的筛选，依然还是遵循主流的计量经济学方法，但是在时间维度上的单次回归有效性，可能来源于某种偶然，筛选出来的因子也未必是真正有效的因子，或者对原模型有增益效果的因子。

CR Harvey, Y Liu（2021）提出的方法，最大程度上解决了因子收益存在偶然性的问题，使得模型结果更加的稳健和有说服力。

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

任瞳：首席分析师，定量研究团队负责人，管理学硕士，17 年证券研究经验，2010 年、2015 年、2016、2017、2020 年新财富最佳分析师。在量化选股择时、基金研究以及衍生品投资方面均有深入独到的见解。

崔浩瀚：量化分析师，浙江大学经济学硕士，5 年量化策略研究开发经验。研究方向是机器学习在金融领域的应用和多因子选股策略开发。

投资评级定义

公司短期评级

以报告日起 6 个月内，公司股价相对同期市场基准（沪深 300 指数）的表现为标准：

强烈推荐：公司股价涨幅超基准指数 20%以上

审慎推荐：公司股价涨幅超基准指数 5-20%之间

中性：公司股价变动幅度相对基准指数介于±5%之间

回避：公司股价表现弱于基准指数 5%以上

公司长期评级

A：公司长期竞争力高于行业平均水平

B：公司长期竞争力与行业平均水平一致

C：公司长期竞争力低于行业平均水平

行业投资评级

以报告日起 6 个月内，行业指数相对于同期市场基准（沪深 300 指数）的表现为标准：

推荐：行业基本面向好，行业指数将跑赢基准指数

中性：行业基本面稳定，行业指数跟随基准指数

回避：行业基本面向淡，行业指数将跑输基准指数

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。