$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}}$$

$$\text{其中} \quad d_{12} = \frac{1}{N}\sum_{i=1}^{N}\left((e_{it+1}^{(1)})^2 - (e_{it+1}^{(2)})^2\right) \tag{6.34}$$

上式中N为样本外的股票数量，d_{12}为两个模型在样本外的均方误差的差异，而$\bar{d}_{12}$和$\hat{\sigma}_{\bar{d}_{12}}$分别为$d_{12}$的均值和均值的标准误，$e_{it+1}^{(1)}$和$e_{it+1}^{(2)}$分别为两个模型对于股票$i$的预测误差。由定义可知，$DM_{12}$越大，表明模型1相对模型2表现越差；反之则表明模型1相对模型2表现更好。

基于从1957到2016年的长达60年的美股数据，Gu et al.（2020）仔细研究了不同模型的表现。他们考虑了94种公司特征和8个宏观变量及它们的交互项，并另有74个行业分类，得到总共94×（8+1）+74=920个特征。在此基础上，该文比较了13个预测模型，包括6个线性模型算法（即包含全部特征的OLS回归，只包含规模、账面市值比和动量的OLS回归、PLS、PCR、弹性网络，以及广义线性回归）、2个树模型（随机森林和GBDT）以及5个神经网络模型（分别包含1到5层隐藏层），且对于OLS、弹性网络、广义线性回归和GBDT，使用Huber稳健估计量。从全样本来看，OLS的表现非常糟糕（且对大盘股而言更是如此）。而只考虑三个特征的OLS，或者弹性网络等方法，通过添加额外的惩罚项，表现得到了显著的提升。此外，GBDT和随机森林表现也不错。然而在众多模型中，表现最好的非线性模型还要数神经网络模型，尤其是带3层隐藏层的神经网络模型。当采用样本外R^2_{OOS}为评价标准时也有类似的结果。除此之外，模型间的两两配对比较则有如下发现：所有带约束的线性模型的表现都显著优于普通OLS，而在降维方法（PLS/PCR）和惩罚性回归模型的表现则没有明显差异；树模型表现相比线性模型更好，但差异并不显著；神经网络表现显著优于线性模型，但相对树模型的改进则不够显著。

除了比较不同模型的表现外，Gu et al.（2020）的实证分析还有另外一个重要作用，即比较不同特征对于股票定价的重要性。该文将所有公司特征分为四大类：趋势类特征（例如各种动量和短期反转）、和流动性有关的特征、风险测度指标，以及基本面特征。他们发现线性模型普遍高度倾向趋势类特征，而非线性模型则会较为平均地关注多种公司特征。总体而言，趋势类特征的影响最为显著。

除上述针对美股的代表性研究外，近些年也有不少学者研究了机器学习算法在中国A股市场的表现，并有类似的发现。总体而言，带约束的线性模型表现优于OLS，非线性模型又优于线性模型。在非线性模型中又尤其以深度前馈神经网络（DFN）和XGBoost表现非常出色。此外，利用集成学习整合不同模型也可以进一步提升模型表现。而在特征重要性方面，已有研究表明，在A股市场中最为重要的因子是交易摩擦类（流动性）相关因子，这与美股市场有所不同。

6.8.4 主成分分析和因子选择

近年来，一些新的研究将无监督学习算法引入实证资产定价和因子投资，用

于改善基于线性回归的计量经济学估计方法的表现。因其简单有效，主成分分析（principal component analysis，即PCA）备受关注，而在资产定价领域的相关应用也主要集中在对PCA方法的应用和拓展上。这背后的原因与经典的因子收益率估计方法（如Fama–MacBeth回归）面临的问题有关。经典方法需要明确指定因子结构才能进行有效估计，且容易受到遗漏变量和测量误差的影响。另一方面，越来越多的研究指出，人们其实并不知道真实的定价因子是什么，反而更倾向于将真实因子视作是隐性的因子（latent factors），并利用降维的手段来同时估计因子暴露和因子溢价。PCA方法也由此进入实证资产定价的舞台。

隐性因子模型（latent factor model）是统计学中很常用的一个模型，在推荐系统等机器学习实践中有非常广泛的应用。对于资产定价问题，隐性因子模型的表达式为：

$$R_{it}^{e} = \beta_i' \boldsymbol{\lambda}_t + \varepsilon_{it} \tag{6.35}$$

其中，R_{it}^{e}是t期资产i的超额收益，λ_t是t期的因子溢价向量，β_i为资产i的因子暴露向量，而ε_{it}是随机扰动。乍看起来，模型（6.35）与一般的多因子模型并无差异，但它的特别之处是真实因子无从观测（隐性的含义），即人们并不知道λ_t的取值，因而因子暴露β_i也无从知晓。

此时PCA方法便派上了用场。它通过提取资产收益协方差矩阵的主成分来估计风险溢价和风险暴露。Giglio and Xiu（2019）在这方面做出了开创性的贡献，利用PCA构建了一种无须观测到全部的真实因子便可准确估计因子溢价的新方法。在隐性因子模型框架下，任一可观测因子的风险溢价等于它对隐性因子的暴露乘以隐性因子的溢价。计量经济学中的两个重要性质使得PCA在估计因子溢价时扮演了重要的角色。首先，利用线性多因子模型的旋转不变性，即便只能观察到隐性因子的某个满秩变换，也不妨碍估计可观测因子的溢价。其次，只要隐性因子足够强[9]，PCA总是可以复原对因子空间的某个旋转变换（Bai 2003）。通过结合这两个性质，Giglio and Xiu（2019）指出虽然真实因子不可观测，但利用PCA方法，仍可以准确估计因子溢价。

Giglio and Xiu（2019）基于美国市场的实证分析表明，相比经典方法，该利用PCA方法得到的估计量确实有显著的优势。一般来说，Fama–MacBeth回归结果高度依赖模型的控制变量。以动量因子为例，不控制其他因子和控制Fama–French三因子两种情况下，其因子溢价符号竟然相反，且都高度显著。而该PCA方法则能够获得令人满意的估计结果。对于可交易因子，其因子溢价与时序均值较为接近；对于加总的市场流动性、金融中介杠杆率等不可交易因子，其估计结果也与理论方向一致。

将PCA方法应用于资产定价的另一项研究来自Rapach and Zhou（2019）。该文首先通过稀疏PCA从120个宏观经济变量中提取了10个稀疏主成分，并指出这些主成分可大体对应债券的名义收益率水平、通胀率、产出率等经典指标，因而具有极好的可解释性；其次利用Giglio and Xiu（2019）的方法估计了这些稀疏宏观因子（sparse macro factors）的溢价，并发现债券的名义收益率水平、住宅和乐观情绪有显著的风险溢价。最终，他们用这三个因子和市场组合一起构建了一个稀疏宏观四因子模型，并发现该四因子模型具有同Hou–Xue–Zhang四因子模型和Fama–French五因子模型可比的解释力。

与Giglio and Xiu（2019）类似，Kelly et al.（2019）同样将真实因子视作不可观测的隐性因子并利用PCA方法同时估计因子溢价和资产的因子暴露。但他们同时指出经典的PCA只适用于估计静态模型，而对于动态条件资产定价模型则无能为力。为了解决这一问题，他们采用Kelly et al.（2017）提出的工具变量PCA方法（IPCA），引入大量公司特征作为股票因子暴露和超额收益的工具变量，构建了IPCA因子[10]。该方法受到以下两点的启发。第一，公司特征和因子暴露密切相关[11]，这使得用公司特征当作因子暴露的工具变量成为可能。其次，一家公司的各种特征会随着时间变化，这使得很难利用时间序列分析方法构建个股的条件预期收益率模型。以往研究者更多采用投资组合排序法，但该方法的局限在于只能处理较少的特征。一旦需要更多特征来充分刻画资产的预期收益率截面差异，它就会面临极大的挑战。而通过将因子暴露参数化为公司特征的函数，则可以较好地解决这一动态面板估计问题。

实证结果显示，IPCA方法的确具有较好的表现。相对经典的CAPM、Fama–French三因子模型等，有相同数量主成分因子的IPCA模型能够更好地刻画个股的风险，且经典因子相对于IPCA因子的增量信息非常有限。进一步地发现，随着因子数量增加，IPCA因子的样本外切线组合[12]的夏普比率也显著提升。当使用六个IPCA因子时，夏普比率高达惊人的4.05。相比之下，Fama–French五因子加上动量这六个因子的样本外切线组合的夏普比率仅有1.37。从均值—方差的角度来看，IPCA因子极高的夏普比率表明它们能在解释股票收益的共同运动的同时，通过因子暴露解释不同股票收益之间的差异。通过对比研究，Kelly et al.（2019）指出，真正起作用的是公司特征的动态变化，其对于理解因子暴露非常重要，这也呼应了IPCA想解决的核心问题，即如何在动态条件定价模型中得到因子溢价和暴露的估计。最后，通过分析每个因子对不同公司特征的暴露可以发现，IPCA因子有不错的可解释性。例如，第一主成分可近似理解为价值或杠杆率因子，第二主成分对应市场因子，第三和第四主成分则分别对应动量和短期反转因子。

上述这些利用PCA的研究虽然新颖，但它们仅仅利用了收益率的二阶矩信息，即协方差矩阵。Lettau and Pelger（2020）认为，这么做会丢失掉原始因子和收益率在截面上的关系，即一阶矩信息。为此，它们在经典PCA问题的目标函数中加入了代表一阶矩，提出了风险溢价PCA方法（risk premium PCA，PR-PCA）。实证分析表明，RP-PCA在绝大多数情况下都优于PCA和Fama–French五因子等经典模型，且统计检验表明，通过使用五个PR-PCA因子能够很好地反映股票的系统性风险，且同时能够解释它们收益率的截面差异。对因子构成进行进一步探索发现，这五个因子都有很好的经济学基础。

从上面相关最新研究的介绍可知，PCA及其拓展方法具备更好的解决资产定价问题的潜力。而在因子投资方面，也有两个思路可以考虑。一个思路是直接使用前述某种PCA方法提取主成分因子，并倒推出恰当的因子组合或公司特征权重，并据此构建股票组合。另一个思路则是利用各种PCA因子对公司特征以及经典因子的暴露，将它们映射为经典因子，然后以这些经典因子为基础进行资产配置。不过需要指出的是，由于投资者在现实中面临卖空约束，依据上述方法构建因子组合可能比经典因子更复杂、可投资性更低。因此，虽然它们提供了新的投资思路，但要真正用于实践，仍有不少细节需要进一步完善。

以PCA为代表的无监督学习应用于实证资产定价只是最近几年出现的新研究趋势，因此还不能将其视作高度成熟的方法。反之，它们的定位更多的是对经典

方法的有效改进。考虑到每种方法仍有其局限，未来仍有很多拓展工作值得进一步挖掘。此外，如何将通过PCA获取的信息成功地转化为投资实践也有待更多探索。但无论如何，这些方法都是极为有价值的探索，不仅标志着资产定价的大门对无监督学习方法敞开，也意味着因子投资领域多了一门令人充满期待的新武器。

6.8.5 机器学习的问题

机器学习算法固然强大，但人们在使用这些算法时也必须面对两个问题：（1）机器学习算法常常被视作黑箱，缺乏足够的可解释性；（2）机器学习算法也容易陷入过拟合。对于前者，以神经网络等算法为例，其内部往往非常复杂，黑箱性质使得人们难以真正理解其发现的特征与未来收益率之间的关系。幸运的是，这个问题并非完全无解。Dixon and Halperin（2019）指出可以通过计算特征对最终输出结果的影响程度来解释自变量和因变量之间的关系。另一方面，人们应在发现有显著预测能力的模型后，进一步考察有哪些特征是显著的，并梳理清楚特征和收益率之间可能的逻辑关联。一旦人们试图理解机器学习发现的规律时，它就变成了研究中的一块基石，帮助人们更好地理解数据背后的经济逻辑。

机器学习容易陷入过拟合有以下几个原因。首先，由于真实的资产价格路径只有一条，因而基于该路径反复训练模型本身就很容易过拟合。在这个过程中，也容易踏入Harvey et al.（2016）提出的p-hacking的陷阱。虽然相关研究往往采用了前向回测分析（walkforward backtesting），但该方法只能规避未来数据问题，并不能完全杜绝过拟合。事实上，这一问题对于因子研究而言可能尤为严重。典型的因子研究通常以股票月度收益为研究对象。即便以历史最为悠久的美股而言，较为完整的历史数据也只从1962年开始，即大约700个月的样本。对于机器学习算法而言，这一样本实在过小。另一个原因则与金融数据中的自相关性和异方差特征有关。在机器学习中一般使用交叉验证（cross validation）等方法来进行模型选择。但金融数据的序列的自相关性和异方差特征使得训练集中的信息会泄漏到测试集，从而导致交叉验证方法失效。最后，当给定资产的收益分布时，其预期最高夏普比率同波动率正相关。这造成波动率较大的资产，在一次历史回测中反而可能得到更高的夏普比率。因此，单纯的历史回测可能会高估因子的表现。

虽然有各种各样的问题，但也并不意味着人们就束手无策。例如，通过模拟生成多条（更长周期的）资产价格路径并分析不同场景下的表现，可以改善在历史价格路径上反复测试的问题。此外，通过确保训练集和测试集在时间区间上没有交集，可以改善前述交叉验证可能遇到的问题。而使用平减夏普比率（deflated Sharpe ratio）则可以部分解决夏普比率被高估的问题。

总体来说，机器学习在因子投资领域的应用仍处在早期阶段。但随着数据和算法日益成熟、研究者更加谨慎地应对上述问题，机器学习在未来注定能在因子投资研究中扮演更重要的角色。不过，机器学习是否可以取代经典的多因子模型成为最主流的方法，则有待进一步的考察。说到底，机器学习是一类数据模型方法，要在业务实践中发挥作用，仍有赖于对业务领域知识的理解。综合本节介绍的内容，本书作者认为有理由期待机器学习将扮演更重要的角色，但同时也相信机器学习在因子投资中的最佳路径在于和已有方法结合，而非取而代之。