资产定价中的高维预测问题

——A 股市场月度收益率的监督学习

摘要：本文是使用经典监督学习解决资产定价中的收益率预测问题的一次应用。针对金融数据中特有的低信噪比和高维预测等问题，本文引入了正则化方法进行特征选择；通过适当的损失函数和交叉验证设计，本文有针对性地解决了组合整体表现的预测目标问题以及样本时期先后的信息泄漏问题。在 2011 年 1 月至 2023 年 3 月的沪深主板 A 股的月度收益率预测中，本文逐步引入和改进 OLS、岭回归及 Lasso 模型，在样本内外均观察到预测性能的提升。

1. 背景介绍

随着数据、算力、模型的发展，预测标的不断增加，金融资产定价中的“高维预测”问题不断凸显。以多因子模型（多 Beta 模型）为例，经典的三因子和五因子模型（Fama & French, 1993; Fama & French, 2015）中孤立地组合少数稀疏预测指标的做法日渐被同时使用成百上千个大规模的预测特征的方式替代，这一趋势所带来的高维预测问题（Cochrane, 2011）使得正则化等方法的引入成为必要。不同于传统回归，正则化能通过添加惩罚项的方式进行变量选择，从而解决高维预测下的共线性等问题。然而，直接将统计学习方法引入资产定价问题可能带来问题，需要额外考虑。

将机器学习方法应用到资产定价所产生的问题来源于机器学习经典数据集和金融数据之间的差异。具体表现为：（1）数据信噪比低，预测难度大；（2）观测时期有限导致的目标观测数量远远少于用于预测的特征数量；（3）预测目标是投资组合表现最优，而非个体预测平均表现最优；（4）样本具有时间先后，因而交叉验证可能带来信息泄漏。由于以上特点，直接将经典的机器学习方法论应用于资产定价存在困难，需要根据实际场景进行调整。

本次研究希望通过一个最基础的资产定价场景，重新考察正则化回归估计在资产定价中可能遇到的问题，并且尝试给出改进方案。特别地，考虑资产定价中最核心的收益率预测问题：根据特征集 $\mathbf{x}_{i,t}$确定资产i在的持仓权重$f\left(\mathbf{x}_{i,t}\right)$，以使得未来的超额收益率最大（见下式）。

$$\max_{f} \mathrm{E}\left[r_{i,t+1} \mid \mathbf{x}_{i,t}\right] f\left(\mathbf{x}_{i,t}\right)$$

在过去的研究当中，主流的收益率预测模型采用了可解释性较好的线性表示，用一系列通过金融学理论支持的因子或者经由统计性质计算的因子来对未来的收益率进行预测。由于数据和算力的局限，这些经典模型的维度较低，表现为预测性指标数量远小于观测期数和资产数量。在这一设定下，模型的拟合在统计上是稳健的。

然而，随着时代发展，以上做法逐渐不再适用。一方面，金融研究不得不将数量庞杂的预测性指标纳入考虑，决定任一变量是否该被纳入模型所带来的模型种类将是呈指数级增长的。另一方面，机器学习的引入使得模型的非线性表示能力大大增加，线性预测能力一般（表现为当期指标取值与资产未来收益率的相关性系数较低）的指标也可以直接作为“特征”输入提升模型表现，因而，相比于“故事驱动”的因子研究，现代资产定价模型可以轻易构造海量具有潜在预测能力的指标。来自金融和机器学习两种学科的发展都使得收益率预测中的高维问题变得越发普遍。

在工业界，收益率预测的量化问题也长期受到关注。举例而言，World Quant 公司发表的 Alpha 101（Kakushadze, 2016）以及后续由其他机构发布的改进版本 Alpha 191 都清晰的呈现了“遗传规划”的特征，意味着潜在的特征构造数量级极为庞大；某些对冲基金已经将算力和大模型投入作为最核心的发展方向（暗涌 Waves, 2023）；在卖方研究方面，走在前列的“华泰人工智能系列”也经过多年测试认为“直接将多个滞后期的 K 线数据作为模型输入是深度学习训练的最好方式”。

以上应用实例均指向了高维特征下的收益率预测这一共性问题。本文尝试通过一个简单的线性预测模型，探讨引入机器学习所可能带来的问题以及所需进行的改进。囿于篇幅，本文仍然在线性模型范畴下进行探讨，通过引入 Ridge 和 Lasso 罚项来解决高维预测问题，并且设计不同的交叉验证（Cross Validation）依据来更好的提升模型在未来资产配置意义上的预测表现。

2. 模型与数据

2.1 模型

为了研究便利，本文考虑如下的预测问题：

$$r_{i,t+1} = \sum_{k=1}^{T} b_k \, r_{i,t-k} + \sum_{k=1}^{T} c_k \, r_{i,t-k}^2 + \sum_{k=1}^{T} d_k \, r_{i,t-k}^3 + e_{i,t+1}$$

其中 $r_{i,t+1}$ 代表需要预测的未来收益率。$r_{i,t-k}$ 为过去第 $t-k$ 期的收益。此处的收益率均值的是去除市场表现后的超额收益率，市场表现用当期全市场等权收益率计算。该模型仅仅利用了历史收益率以及其二阶、三阶矩作为预测。考虑这一相对简单的模型是因为：（1）引入大量滞后项和高阶矩，是产生海量预测性特征最常用的方式；（2）

用过去的收益率预测未来收益率，是一类被长期观测和实证确立的技术因子——“动量”和“反转”效应，其背后具有市场无效性和行为金融理论支持；（3）几乎收益率预测模型都不会遗漏历史已实现收益率，只使用收益率，可以避免讨论交互项带来的可解释性问题。

对于预测窗口，一般而言，收益率窗口期越长，可预测性越强（Pedersen, 2015），即月度收益率比周度、日度收益率更容易被预测。我们分别尝试了：（1）周度历史预测周度未来；（2）日度历史预测周度未来；（3）月度历史预测月度未来。最终发现在 A 股中周度收益率的可预测性一般（样本外 R2 均为负值），因此主要通过过去月度收益率的滞后项预测未来一个月的收益率。

2.2 模型改进

2.2.1 引入正则化

在以上模型中，由于特征的规模达到 360 个，相对于资产数量来说不可忽略，通过 OLS 最小化均方误差的方式有可能过多地对噪音进行了拟合，使得样本内 R^2 较大而在样本外较小。通过岭回归（Hoerl & Kennard, 1970）和 Lasso 方法（Tibshirani, 1996）引入惩罚项后有利于控制估计系数的规模，从而防止回归中一些特征的系数取值过高。上述两个模型的思路是通过正则化的方式来解决高维问题。OLS，Ridge 和 Lasso 的损失函数如以下三式所示：

$$\min_{\mathbf{g}} (\mathbf{y}-\mathbf{X}\mathbf{g})'(\mathbf{y}-\mathbf{X}\mathbf{g})$$

$$\min_{\mathbf{g}} \left[\frac{1}{N}(\mathbf{y}-\mathbf{X}\mathbf{g})'(\mathbf{y}-\mathbf{X}\mathbf{g}) + \gamma \mathbf{g}'\mathbf{g}\right]$$

$$\min_{\mathbf{g}} \left[\frac{1}{N}(\mathbf{y}-\mathbf{X}\mathbf{g})'(\mathbf{y}-\mathbf{X}\mathbf{g}) + \gamma \sum_{j=1}^{K} |g_j|\right]$$

2.2.2 自定义损失函数

模型预测的最终目的是构建投资组合，达到投资组合收益最大化的目的。因此在进行超参数选择时，仅仅最大化验证集 R^2 是平均意义上提升对单个资产未来收益率的准确预测，在不同的持仓规则下并不一定能最大化未来投资组合的期望收益。可以证明，在岭回归的 L_2 罚项中，只有当资产协方差为对角阵，即资产收益率之间存在相关性时，最大化 R^2 才和最大化投资者组合收益率等价。为了更好地实现资产定价需要的目标，本研究根据下式的权重构造多空投资组合并用投资组合的平均收益率和夏普率表现作为交叉验证和模型评估的损失函数。决定投资组合收益率的方式如下：

$$\hat{\omega}_{t-1} = \frac{1}{\sum_{i=1}^{N}|\hat{\mu}_{i,t-1}|}\hat{\boldsymbol{\mu}}_{t-1}$$

2.2.3 自定义交叉验证生成器

机器学习中一般的交叉验证方法直接将样本均分，这依赖于对样本独立分布的假设。实际的金融数据观测中，不同样本的特征和预测目标存在重叠，样本的实际产生时间也分时间先后。在训练神经网络等超参数复杂的模型时，需要特别注意验证集和训练集之间不出现信息重叠，即进行 Purging 和 Embargo（De Prado, 2018）。由于本研究仅仅设计了单一超参数的正则化线性模型，超参数调优中引入信息泄漏（information leakage）的可能性较小，因此直接根据时间先后采取“留一法”，将某一年的 12 个月留出作为验证集，其余年份的样本作为训练。最后根据训练集 10 年共 10 组交叉验证的平均表现选择参数。

2.2.4 引入贝叶斯先验的正则化缩放比例

上述正则化的惩罚对于所有特征一视同仁地进行同等程度的缩放。可以证明，由于已经对特征变量进行了标准化，投资组合收益率的方差并不受正则化参数放缩的影响（Nagel, 2021）。如果希望通过正则化来获得比 OLS 更好的投资组合表现，这些预测性特征的截面协方差必须不同。因此，如果想要让正则化真正对投资组合整体表现发挥作用，有必要根据先验知识将不同特征标准化到不同的截面方差。方差越大，收到同等尺度的惩罚后保留的影响越大。

对于本研究考虑的三组滞后项收益率，我们倾向于认为一阶矩、二阶矩、三阶矩对于预测未来收益率（截面差异）的贡献依次下降。因此，先验地将二阶矩的截面标准差缩放到 1/2，三阶矩的截面标准差缩放到 1/4；此外，也可以额外考虑预测能力随时间指数衰减，根据滞后期数进行缩放。后文中的 Ridge 3，Ridge 4 以及 Lasso 模型将采用新的缩放比例。

2.3 数据

本文采用的预测样本时间段为 2011 年 1 月至 2023 年 3 月的月度收益率。针对每一期预测样本，选用过去 120 个月的月度收益率作为特征。在投资标的上，选择所有具有 120 个月完整历史收益率（即上市 10 年以上）的沪深主板 A 股（股票编号以 00 或 60 开头）。数据来源为锐思（RESSET）数据库。我们汇总了数据库中各个月份可用于建模的个股数量如下图 1 所示。

本研究选用 2011 年后的数据，从而保证截面个股数量达到 600 以上。为了更方便地解读估计系数，将预测目标（未来超额收益）做截面中心化处理，所有 360 个特征则在截面上进行标准化处理，以便于线性模型的解释。在样本划分上，我们首先取 2021-2023 年的月度数据作为测试集，则一共有 35441 个观测；其余 2011 年-2020 年的共计 95862 条观测作为训练集。

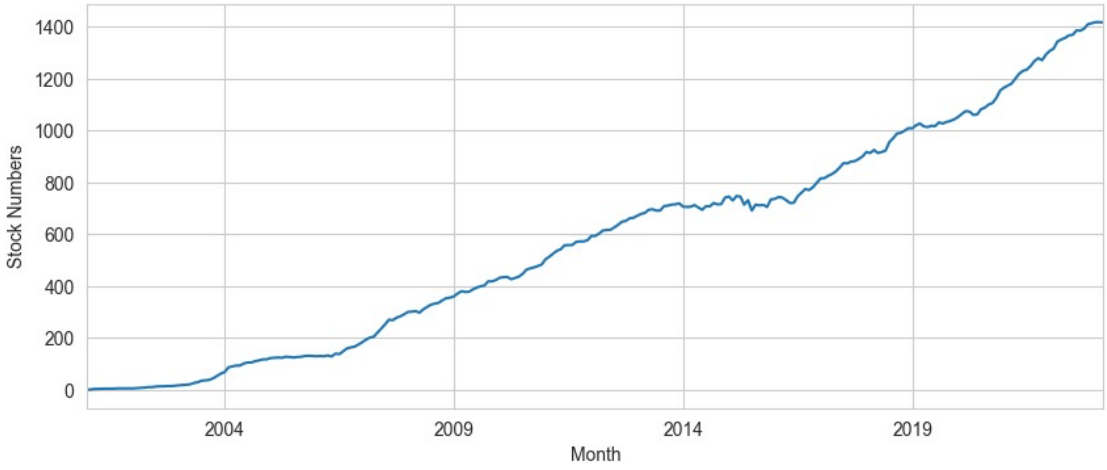

图 1: 各月份有效样本数量

3. 预测表现

综合以上讨论，本文进行了如下五组模型测试。结果如下表 1 和表 2 所示。模型设定为：（1）OLS 直接在训练集内训练，之后依次评估交叉验证平均表现以及验证集表现；（2）Ridge 1 模型用验证集拟合优度 R^2 进行超参数调优，选择“留一”法交叉验证后的最优（R2 最大的）模型；（3） Ridge 2 模型采用最大化交叉验证的投资组合验证集夏普率（SP）的方式进行超参数调优；（4）Ridge 3 对特征变量进行不一致缩放后根据交叉验证平均 R^2 选择最优参数；（5）Ridge 4 对特征变量进行不一致缩放后根据交叉验证投资组合夏普率选择最优参数；（6）Lasso 模型则延续 Ridge 4 的设定，采用 Lasso 惩罚。

表 1: 各模型设定、最优超参数、以及拟合优度R^2表现

				R^2		
模型	缩放比例	交叉验证依据	α	训练集	交叉验证	测试集
OLS	相同	-	-	1.11%	-0.83%	0.07%
Ridge 1	相同	R2	500000	0.28%	0.06%	0.15%
Ridge 2	相同	SP	100000	0.67%	0.00%	0.28%
Ridge 3	不同	R2	100000	0.51%	0.06%	0.31%
Ridge 4	不同	SP	500000	0.19%	0.06%	0.13%
Lasso	不同	SP	0.00005	1.00%	-0.47%	0.26%

表 2: 各模型投资组合表现

	交叉验证投资组合			测试集投资组合		
模型	均值	标准差	夏普比率	均值	标准差	夏普比率
OLS	3.87%	3.88%	1.10	9.28%	4.00%	2.32
Ridge 1	3.97%	4.22%	1.18	11.54%	5.16%	2.24
Ridge 2	4.08%	4.16%	1.18	11.03%	4.79%	2.30
Ridge 3	4.19%	4.28%	1.24	12.38%	4.99%	2.48
Ridge 4	4.15%	4.25%	1.24	12.50%	5.05%	2.48
Lasso	3.91%	4.19%	1.09	10.45%	4.44%	2.35

从交叉验证平均表现和测试集投资组合表现来看：（1）OLS 能取得样本内很高的 R^2，但是在验证集中的表现则为负，说明受到高维预测问题影响，直接用 OLS 模型是过拟合的；（2）通过引入投资组合夏普率作为损失函数，能够提高交叉验证投资组合的表现；（3）引入不同缩放比例后能够提升交叉验证的验证集投资组合表现，相应的，样本外测试集表现也更好；（4）Lasso 模型在调优后表现不如 Ridge，可能原因是经过惩罚后大部分特征变量被抑制，无法贡献预测能力。

以上测试结果表明，通过面对资产定价问题有针对性地对机器学习训练过程进行调整，确实能够达到提高样本外投资表现的目的。需要强调的是，本文计算的收益是在因变量截面去均值后的相对收益，也即在市场平均表现以上的超额收益中、超出所有沪深主板股票平均水平以上的收益。此外，本文考虑的投资组合未计入交易成本，且放宽了一定的做空约束，因此离真实的实盘表现还有一定距离，不构成投资建议。

3.1 系数规模比较

根据正则化理论，OLS 倾向于获得最高的估计系数，Ridge 倾向于抑制系数的规模，Lasso 则有变量筛选的作用，抑制大部分系数为 0 而仅仅保留少数非零系数。我们对线性模型的估计系数规模进行了可视化，以下图 2 中的三张子图分别呈现了滞后 120 期的收益率一次幂、二次幂、三次幂的系数规模。其中岭回归选择了 Ridge 4 组，以保证交叉验证方式和 Lasso 一致。

从系数规模中可以看出：（1）一次项的估计系数规模均较大，二次项估计系数显著不为 0 的则较少，三次项更少，符合本研究在惩罚项缩放中的先验设定；（2）一次幂项中，可以观测到最近一个月“反转”，2~12 月“短期动量”（Jegadeesh & Titman, 1993），滞后超过 12 月以前则“长期反转”（De Bondt & Thaler, 1985），符合投资界对于 A 股市场的一般印象；（3）一次幂项的表现有一定 12 个月的周期性，这也和文献中动量效

应的周期性一致（Heston & Sadka, 2008），具体地，每 12 个月大致出现一次“正”的估计系数；（4）二次幂项在最近 36 个月内更多表现出负的系数，表明过去收益率极端值较高，或者波动率交大的个股未来收益表现越差；（5）三次幂项的系数大部分为 0，少数滞后期则有非常大的系数，且方向不一，反映了极端收益具有的偏度对未来收益比较混乱的影响，并且可能是过拟合的来源。

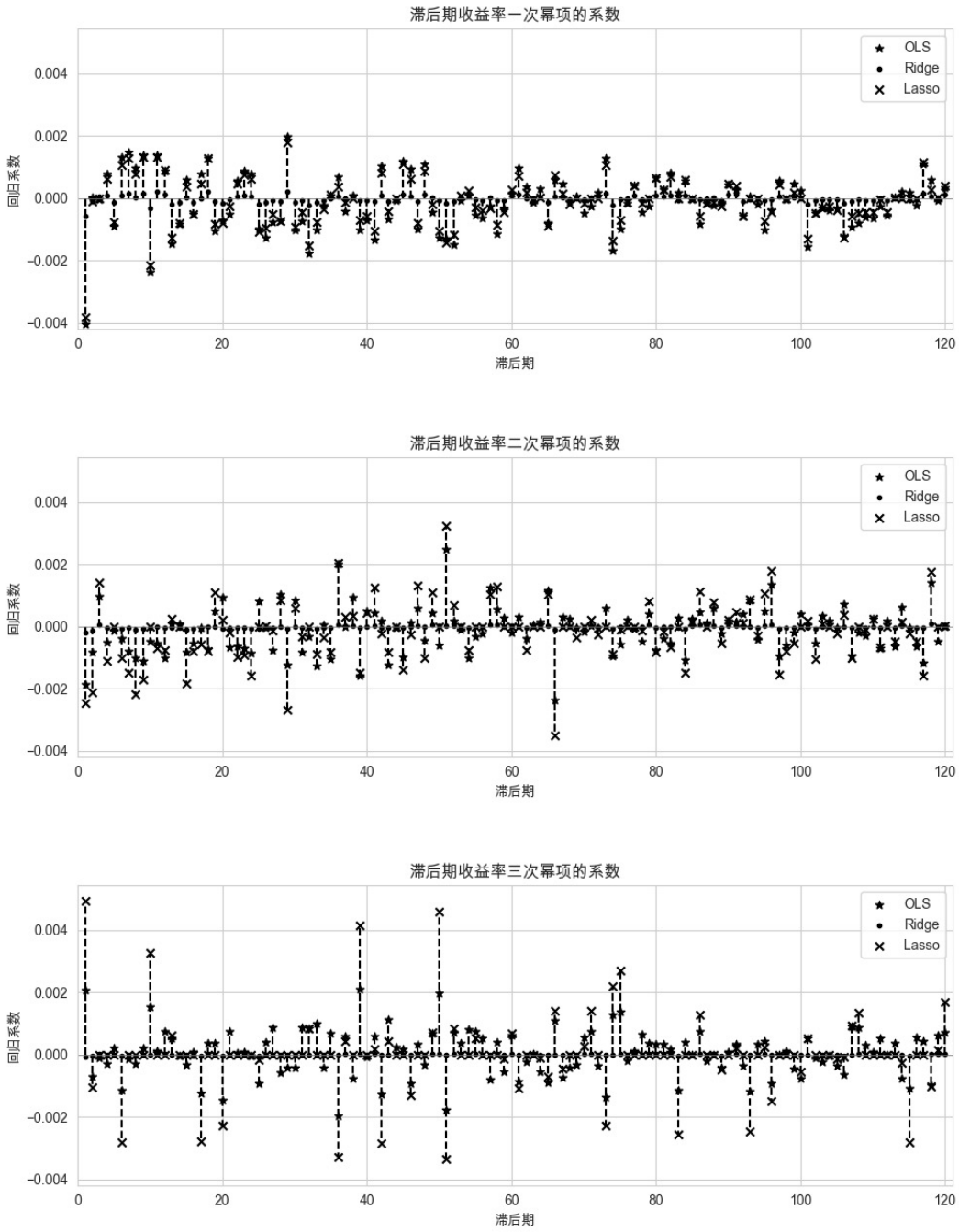

图 2: 各滞后期下不同模型估计的特征系数

对比不同模型设定下估计系数的规模，可以发现：（1）OLS 几乎会考虑所有项目的估计系数，即使在高阶幂上也只是估计系数规模较小，但系数之间差异不大；（2）岭回归（Ridge）中几乎所有特征的系数规模均大大被抑制，但都是接近 0 而非完全为 0，说明 Ridge 起到了控制高维特征过拟合的作用；（3）Lasso 中绝大部分系数为 0，说明 Lasso 起到了变量筛选的作用；（4）Lasso 中显著特征的系数规模和 OLS 重合度高，并且系数规模高于 OLS，说明 Lasso 在变量筛选之后会突出强调那些贡献较大的特征，可能带来的过拟合问题导致本文设定下的 Lasso 表现不如 Ridge 模型。

4. 总结与展望

本文是使用经典监督学习解决资产定价中的收益率预测问题的一次应用。针对金融数据中的信噪比问题、高维问题、投资组合预测目标问题以及样本时期先后信息泄漏问题，通过适当的损失函数和交叉验证设计，能够在机器学习框架下得到一定程度的解决。通过逐步改进 OLS，岭回归和 Lasso 模型，本文在样本内外均发现了以上改进下预测表现的提升。

本文的主要结论为：（1）通过引入收益率滞后项以及滞后项的高阶矩构建的线性模型能够对未来股票收率进行一定程度的预测，但是传统金融计量中的 OLS 对噪音过度拟合，样本外几乎没有预测能力；（2）正则化项的加入能缓解上述回归问题中的过拟合，保证模型在样本外也具有一定的泛化能力；（3）在交叉验证中选用投资组合整体的收益表现作为损失函数，能够确保超参数对于样本外的投资组合的整体表现是最优的；（4）可以根据研究者对于特征重要性的先验知识设置不同的正则化缩放尺度，从而使得正则化更有可能保留更具预测能力的特征；（5）Lasso 相对 Ridge 具有“变量筛选”的作用，表现为绝大部分特征系数取 0，而少数有重要特征的系数规模则较高。

本文作为 Nagel (2021) 一书在纽交所美股中实证结果的复显，从中国 A 股市场的角度再次发现了机器学习在金融应用过程中进行调整的必要性。类似的著述还包括 De Prado (2018), Creamer et al. (2021)，都是这一交叉领域的最新理论和实证经验。将机器学习方法论应用到金融市场是一门有关于“实践”的学问，公式推演仅仅表明一些理想化的理论性质，但囿于金融数据的低信噪比特征，脱开市场实际情况的理论建模往往很难为机器学习方法论的直接应用给出指导。本文利用可预测性较强的月度数据探索了线性模型中正则化方法的效果，不失为一种有益的尝试。

关于本文涉及的收益率预测问题，可以进一步尝试更加复杂的机器学习方法，囿于时间和篇幅限制，本文暂时未作尝试，有待后续探索。这些方法具体包括：树模型（Khaidem et al., 2016; Moritz & Zimmermann, 2016; Bryzgalova, Pelger, & Zhu, 2019），神经网络模型（Gu, Kelly, & Xiu, 2020; Chen, Pelger, & Zhu, 2023），也包括引入文本图

象数据后更加复杂的图神经网络模型。如何在应用这些模型时针对金融数据进行更合理的调整，是在尝试这些方法论之前都需要仔细斟酌的首要问题。

参考文献

暗涌 Waves. (2023, 五月 24). 疯狂的幻方：一家隐形AI巨头的大模型之路. *暗涌Waves*. https://mp.weixin.qq.com/s/Cajwfve7f-z2Blk9lnD0hA

Bryzgalova, S., Pelger, M., & Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Chen, L., Pelger, M., & Zhu, J. (2023). Deep Learning in Asset Pricing. *Management Science*.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance*, *66*(4), 1047–1108.

Creamer, G., Kazantsev, G., & Aste, T. (2021). *Machine Learning and AI in Finance*. Routledge.

De Bondt, W. F., & Thaler, R. (1985). Does the stock market overreact? *The Journal of finance*, *40*(3), 793–805.

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3–56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1–22.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Heston, S. L., & Sadka, R. (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics*, *87*(2), 418–445.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, *48*(1), 65–91.

Kakushadze, Z. (2016). 101 formulaic alphas. *Wilmott*, *2016*(84), 72–81.

Khaidem, L., Saha, S., & Dey, S. R. (2016). *Predicting the direction of stock market prices using random forest* (arXiv:1605.00003). arXiv. https://doi.org/10.48550/arXiv.1605.00003

Moritz, B., & Zimmermann, T. (2016). Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2740751

Nagel, S. (2021). *Machine learning in asset pricing* (卷 1). Princeton University Press.

Pedersen, L. H. (2019). *Efficiently inefficient: How smart money invests and market prices are determined*. Princeton University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.