arXiv:2601.03927v1 [q-fin.PM] 7 Jan 2026

金融指数跟踪问题不同建模方法的全面综述与分析

Vrinda Dhingra[1], Amita Sharma[1], Anubha Goel[2*]

通讯作者：电子邮箱： anubha.goel@tuni.fi;
贡献作者： vdmath.iitr@gmail.com； amita.sharma@nsut.ac.in;

摘要

指数跟踪（亦称被动投资）凭借其经济高效地复制特定市场指数表现的方式，在金融市场中日益受到青睐。本文综述全面介绍了指数跟踪领域发展出的各类建模方法与策略，并重点阐述每种方法的优势与局限。我们将指数跟踪模型归纳为三大框架：基于优化的模型、基于统计的模型，以及基于机器学习的数据驱动方法。基于标准普尔500指数数据集开展的全面实证研究表明，在基于优化的框架下，跟踪误差波动率模型能够实现最精准的指数跟踪；在基于统计的框架下，凸协整模型实现了最佳的收益—风险平衡；而在数据驱动框架下，固定噪声深度神经网络模型表现出较强的竞争力，同时换手率显著较低、计算效率较高。本文将对现有文献的批判性综述与比较性实证分析相结合，旨在洞察指数跟踪不断演变的格局及其对投资者和基金经理的实际意义。

关键词： 指数跟踪、跟踪误差最小化、跟踪投资组合、基数约束、机器学习、深度学习

1 引言

基金管理（或投资管理）是金融服务领域的一项核心职能，专业地管理多个资产类别中的资产，以代表投资者实现既定投资目标。投资者包括个人或机构，例如保险公司、养老基金和企业。基金管理涵盖资产选择、交易、监控、向利益相关方报告以及内部审计/治理等活动。其主要目标是在风险、成本和监管约束下，平衡中长期的资本增长与收入。

过去几十年来，在金融科技进步、市场日益复杂以及投资者偏好不断变化的推动下，投资格局发生了显著变化。这一演变的一个重要方面，是相对于标准普尔500指数或道琼斯工业平均指数（DJIA）等基准指数进行投资组合管理。总体而言，投资策略可分为两类：主动管理和被动 管理；本综述聚焦于这一分类体系中的被动指数跟踪。为界定研究范围，实证示例采用标准普尔500指数；本文综述的建模框架具有通用性，适用于各种资产类别。以下分别对两类策略进行说明：

(a) 主动管理 采用亲力亲为的方式，由基金经理主动决定买卖证券，以期跑赢特定的基准指数。其目标是实现

高于基准的收益率，借此发挥管理者的专业知识、研究成果和市场洞察。主动策略的范围从传统选股到更有组织的策略，例如增强型指数化；后者将基准复制与小幅战术性偏差相结合，旨在提升收益表现。

(b) 被动管理 旨在复制基准指数的表现，通常通过以相同权重持有相同的证券来实现；或者在完全复制成本较高时，持有能够高度近似该指数的代表性子集。其主要目标是以较低的费用、较低的换手率，并将低于基准的风险控制在有限范围内，实现与基准收益相匹配。常见的被动投资工具包括指数基金和交易所交易基金（交易所交易基金）。

表1：主动管理与被动管理策略比较

方面	**主动管理**	**被动管理**
方法	由研究与分析驱动	指数复制
Goal	跑赢市场	匹配市场表现
成本	主动交易和研究导致费用更高	由于交易和研究最低限度，费用较低
Risk	由于市场择时，风险可能较高	凭借更广泛的多元化，风险通常较低
风险敞口	公司和市场风险	市场风险
收益率	有望获得更高收益率	通常与市场回报相匹配
适合人群	寻求高于市场回报的投资者	寻求与市场回报相匹配的长期投资者

表1 总结了主动管理与被动管理在关键维度上的主要差异。多年来，凭借简单、透明、稳定且成本效益高等特点，被动投资在投资者和资产管理人中广受欢迎（见图1）。实证研究一再表明，长期来看，扣除费用后，许多主动管理型基金未能跑赢其基准，这凸显了被动投资替代方案的吸引力（Malkiel，2003年；ProndzinskiandMiller，2018年；Anaduet al.，2020年）。在此背景下，指数跟踪——构建一个能够复制基准风险收益特征，同时将跟踪误差和跟踪偏离度控制在较低水平的投资组合——已成为最重要的被动投资方法之一，为复制基准表现提供了一种系统且高效的方式。

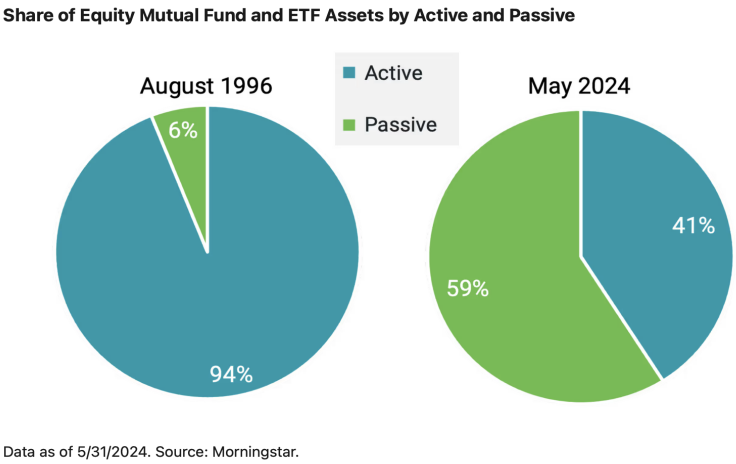


图1：三十年来主动与被动投资的占比（来源：Morningstar）

指数跟踪投资组合，例如指数基金（IFs）和交易所交易基金（ETFs），是应用最广泛的投资工具。指数基金是一种集合投资，通过完全复制或部分复制其成分股来跟踪基准的收益率，通常每天按资产净值（NAV）定价。相比之下，交易所交易基金的投资目标相似，但结构有所不同：它们在证券交易所全天交易，并采用

申购赎回机制，使市场价格维持在接近资产净值的水平。在市场上，投资者可以投资许多不同类型的交易所交易基金，例如行业型指数基金、宽基指数基金、市值加权指数基金、等权重指数基金、智能贝塔指数基金、国际指数基金和债券指数基金等。

指数跟踪的起源通常追溯至20世纪70年代中期，当时先锋集团的约翰·博格于1976年推出了首只指数基金（博格，2011）。该基金旨在复制标准普尔500指数的表现，为投资者提供了一种简单、低成本的方式，使其能够广泛分散地投资于美国股票市场。当时，这一理念显然打破了强调主动管理和股票选择的主流做法。此后，指数跟踪业务在全球范围内不断发展，产品提供商推出了跟踪国内外股票、固定收益及其他资产类别基准的产品。20世纪90年代交易所交易基金的出现进一步拓宽了产品种类，例如1993年的SPDR标普500信托基金（SPY）；此后，市场上还出现了区域性和单一国家的投资产品，例如在美国上市的iSharesMSCI德国ETF（EWG），以及莫蒂拉尔·奥斯瓦尔Nifty 50指数基金（印度）等共同基金产品。

被动投资，特别是指数跟踪的规模，已在全球范围内显著扩大。如图2所示，全球交易所交易基金的数量从2003年的276只增至2022年的8,754只，反映出市场结构正转向低成本、基于指数的投资工具。在美国，主动管理型基金的占比从2010年的81%降至2023年的52%，而指数共同基金和交易所交易基金的合计占比则增至股票基金市场的近60%。[1] 新兴经济体也呈现出类似的增长态势。例如，在印度，交易所交易基金交易量从2019—2020财年的约510亿印度卢比增至2024—2025财年的3.83万亿印度卢比，五年间增长逾七倍。[2] 这些趋势凸显了被动投资在全球范围内的普及：它在发达市场已发展成熟，并在发展中市场迅速扩张。

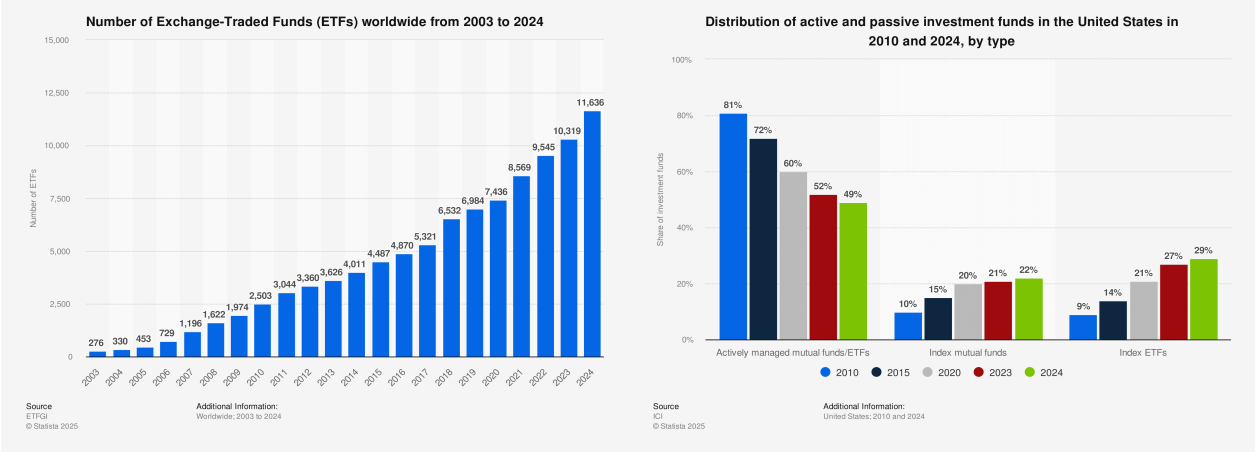


图2：多年来被动投资的增长。来源：*Statista 2025*

谁应投资于指数跟踪投资组合？ 指数跟踪投资组合旨在跟踪市场基准指数，因此其收益率应与标的指数的收益率相匹配，因费用和交易摩擦而产生小幅差异。这类投资组合因此对寻求可预测的广泛市场敞口、低成本和最少投资组合管理的投资者颇具吸引力。这些策略为所有投资者（无论经验丰富还是初学者）提供了一种易于参与且有益的投资选择。与通常交易更频繁、管理费更高且涉及自主决策的主动管理型基金不同，指数基金采用被动管理，并能实现大规模分散化。由于这类基金旨在匹配而非超越基准，因此是长期财富积累的高效工具，尤其适合注重成本、投资期限较长的投资者，或作为分散化股票投资组合中的核心配置。投资者仍应预期市场会经历完整的涨跌周期——这些指数跟踪器并不提供保本。

指数跟踪投资组合是如何构建的？ 指数跟踪作为一个研究领域的发展，其根源在于投资组合优化文献；在该领域中，这一问题最初被正式表述为最小化构建的投资组合与其基准指数之间的跟踪误差。在实践中，指数复制可以通过两种主要方式实现：完全复制和部分复制。完全复制是指按照与基准指数相同的比例持有其中的全部证券；对于成分股较少且流动性较高的窄基指数而言，这种方法是可行的，

[1]来源： https://www.americancentury.com/institutional-investors/insights/has-passive-investing-gotten-too-big/ [2]来源：
https://www.zerodhafundhouse.com/blog/a-comprehensive-guide-to-exchange-traded-funds-etfs-in-india

/

但对于包含众多成分股的宽基指数而言，由于交易、行政管理和流动性摩擦，其成本会变得高昂（参见Beasley 等人（2003年）；Canakgoz 和 Beasley（2009年）的相关讨论）。
因此，许多指数跟踪投资组合采用部分复制，选择具有代表性的证券子集来近似复制指数。此类选择通常受基数约束 支配，该约束限制投资组合中的资产数量，从而在与基准保持高度一致的同时，降低交易成本和运营复杂性。纳入此类约束会使问题具有组合性质且属于 NP难问题，因此需要采用先进的计算方法，并在再平衡过程中谨慎控制换手率。

最早且应用最广泛的建模方式，将跟踪误差定义为超额收益的均方根误差（RMSE），从而形成最小二乘二次规划（QP）框架（Beasley 等人，2003年）。随着时间推移，人们提出了启发式和元启发式技术——遗传算法、模拟退火和禁忌搜索——以高效处理这些复杂的建模问题（Beasley 等人，2003年；Guastaroba 和 Speranza，2012年；Derigs 和 Nickel，2003年；Krink 等人，2009年；Sant'Anna 等人，2017年）。关于基于启发式方法的全面综述，参见Silva 和 de Almeida Filho（2024年）。

优化框架还通过采用跟踪误差的其他定义得到了扩展，包括线性（Rudolf 等人，1999年）或绝对偏差（Guastaroba 和 Speranza，2012年），以及条件风险价值等基于风险的度量（Goel 等人，2018年）。这些进展旨在平衡可处理性、可解释性与稳健性，并降低对异常值和模型设定错误的敏感性。与此同时，一些统计框架开始利用指数与成分资产收益之间的关系，而不一定要求求解全局优化问题。基于回归的公式（Canakgoz 和 Beasley，2009年；Li，2020年）、正则化技术（Wu 等人，2014a；Fastrich 等人，2014年）、协整（Sant'Anna 等人，2017年，2020b）以及基于因子的方法（Corielli 和 Marcellino，2006年）进一步丰富了方法工具箱，往往能够构建适用于高维场景的稀疏且易于实施的投资组合。

近年来，人工智能技术显著拓宽了指数跟踪的方法体系。神经网络（Ouyang 等人，2019；Zheng 等人，2020a）、随机森林（Yuanyuan Cao 和 Yang，2022年）以及深度自编码器（Zhang 等人，2020年）已被用于捕捉资产收益率与指数收益率之间复杂的非线性依赖关系，从而实现更具适应性且可能更高效的投资组合构建。Dai 和 Li (2024年)设计了一种深度学习框架，可在基准约束下学习动态交易策略，并在标准普尔500指数数据上取得更优的样本外跟踪表现。Peng 等人(2023a)采用强化学习来构建这一问题，将指数跟踪视为一个考虑交易成本的无限期序贯决策任务；其深度强化学习（RL）智能体在长期跟踪精度和成本效率方面优于传统模型。Zheng 等人(2020b)提出了一种随机神经网络，通过重参数化直接纳入基数约束，从而能够构建稀疏且符合实际的指数复制投资组合，并实现最先进的跟踪精度。与此同时，研究人员也探索了可解释人工智能（XAI）：Zhang 和 De Smedt (2024年)将 SHAP（沙普利加性解释）与深度自编码器相结合，构建可解释的股票选择机制，并展示了性能和模型透明度的提升。研究者也开始关注稳健性，Bradrania 等人(2022年)提出了能够适应市场状态结构性转变的市场状态感知模型，增强了指数跟踪策略的韧性。这些进展展示了尖端人工智能方法——涵盖深度学习、强化学习、模型可解释性和自适应泛化——如何在现实约束下提升表现，并使指数跟踪与现代人工智能研究的核心挑战相契合，例如结构化学习、分布偏移下的学习以及约束感知推断。

本文为被动管理提供了什么？在这篇综述中，我们旨在全面分析过去三十年来发展起来的指数跟踪（指数基金/交易所交易基金）建模方法。从传统的基于优化的方法到近期的数据驱动技术，本文不仅梳理了指数跟踪的发展历程，还采用标准普尔500指数数据，依据常用评估指标（如跟踪误差和换手率）对这些方法进行了实证评估。Silva 和 de Almeida Filho的研究成果(2024年)对指数跟踪的解法进行了系统综述，通过对期刊和研究重点领域进行文献计量分析，涵盖了精确方法和启发式方法。与之不同的是，本文着重探讨建模框架本身，而非算法解法。此外，我们的研究通过在真实数据上对这些框架进行实证比较，补充了现有文献，并为不同方法的相对有效性提供了实践见解。本文的另一项贡献是开放获取代码，从而确保透明度、可复现性，并提供一个标准化基准，供未来研究者比较其策略。在实证比较中，我们报告了近十年的每日结果，时间跨度从2012年10月到2022年11月，并每三个月再平衡一次。所有复现代码和数据处理脚本均可

在以下网址获取：https://github.com/vrindadhingra/index-tracking-review-。因此，本文既是一篇学术综述，也是一项实用资源，服务于被动投资领域的学者和从业者，尤其是指数跟踪领域。

本文其余部分安排如下。第2 节概述综述方法和文章筛选标准。第3 节讨论指数跟踪方法的发展历程，并将各个里程碑归类为基于优化、基于统计和数据驱动的建模框架。第4 节介绍代表性模型的实证评估，第5 节总结主要发现及其影响，并指出未来研究方向。

2 综述方法论

本综述采用的方法包括三个主要阶段：(i) 在科学数据库中使用明确定义的关键词，系统检索相关研究论文；(ii) 通过结构化的筛选和审查过程，建立最终的研究文献集；以及(iii) 对从所综述文献中识别出的、最具代表性的指数跟踪建模框架进行全面的实证评估。本综述流程的整体设计借鉴了人工智能和金融优化文献中成熟的综述方法，尤其是Goodell 等人(2021年)和Silva 和 de Almeida Filho(2024年)采用的方法。

2.1 文献综述的文章筛选与过滤

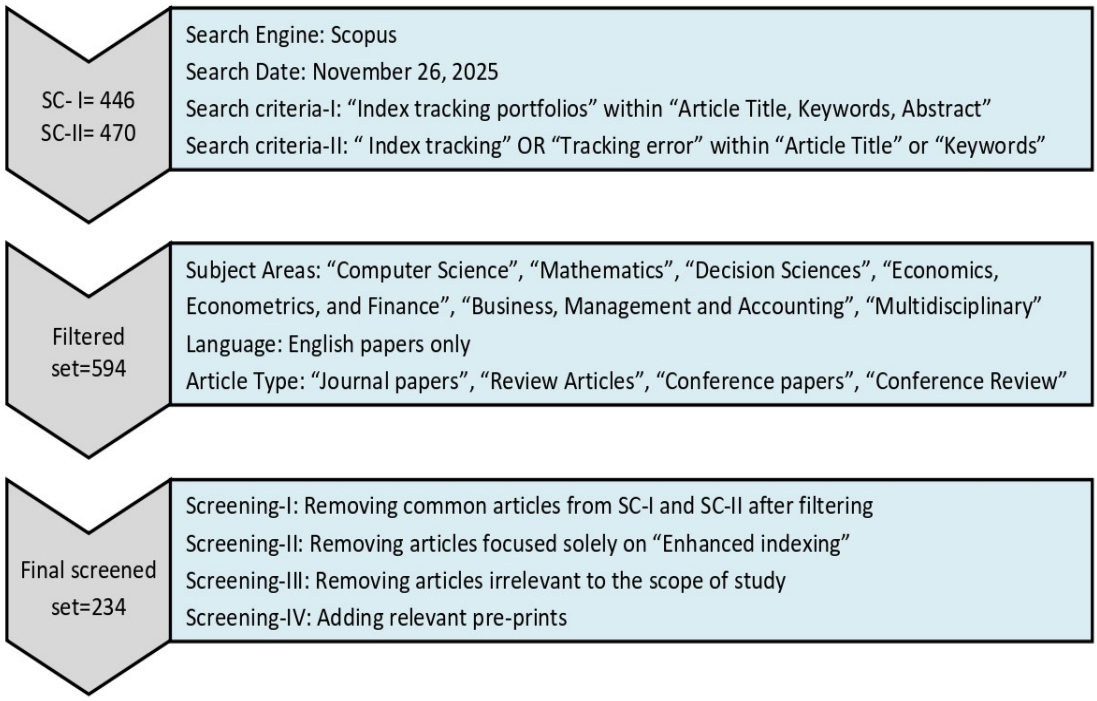


图3：文献综述文章的检索与筛选

构建全面综述的第一步，是识别并选取对指数跟踪领域做出重要贡献的相关研究论文。如图3所示，论文筛选流程分为三个主要阶段：**检索步骤**、**筛选步骤**和**审查步骤**。每个阶段都发挥着关键作用，旨在将文献范围逐步缩小，最终形成一组聚焦且高质量的研究成果，并确保其与本综述的目标直接契合。

1. **检索步骤**：第一阶段，我们使用*SCOPUS* 数据库开展了广泛的文献检索。选择该数据库是因为它广泛收录了金融、运筹学、人工智能和应用数学领域的同行评审期刊。为全面检索与指数跟踪相关的研究，我们采用了两项互补的检索标准。

- **搜索标准一（SC-I）：**我们使用关键词“指数跟踪投资组合”在文章标题、摘要和关键词中进行检索。此次初步检索共得到**446 篇文章**，涵盖了明确讨论指数跟踪投资组合构建的研究。
- **搜索标准二（SC-II）：**为进一步扩大检索范围，纳入使用其他相关术语的文章，我们使用关键词“指数跟踪”或“跟踪误差”进行二次检索。此次检索范围更为集中，专门针对文章标题和关键词部分，以确保检索结果与讨论这些核心概念的文章更加相关。二次检索共得到**470 篇文章**。

通过应用这两个标准，我们旨在确保研究纳入所有关于指数跟踪（IT）的重要且相关的文章。

2. **筛选步骤**：初步检索完成后，我们使用 SCOPUS 内置的筛选工具对收集到的文章进行筛选，剔除超出本综述范围的研究。具体而言：

- 仅保留以英语发表的文章，以确保解读和分析的一致性。
- 如图3所示，文章进一步根据学科领域和文章类型进行筛选，重点关注符合指数跟踪方法和金融分析范围的研究。

经过这一筛选过程，非学术出版物以及与指数跟踪没有直接关系的研究被排除。此阶段之后，文献集缩减至594篇文章，其中375篇来自SC-I，219篇来自SC-II。

3. **审查步骤**：这是筛选过程中的最后一个阶段，也是最关键的阶段，旨在确保纳入本综述的文章既互不重复又与主题相关。在此步骤中：

- 删除了同时出现在两种搜索标准或多个筛选类别中的重复文章。
- 对每篇剩余文章进行审查，以评估其与纯指数跟踪的相关性。排除了仅关注增强型指数化、无关的投资组合优化问题或更广泛的资产配置框架的研究。
- 为纳入最新进展，还审查了SCOPUS收录的预印本。在66篇可用预印本中，确定有**14** 篇文章与指数跟踪直接相关，且未与已发表的研究重复。纳入这些文章是为了确保综述反映最新的研究趋势。

系统地实施搜索、筛选和审查步骤后，最终**确定了233篇文章**，构成了关于指数跟踪的全面且最新的文献集。图4 展示了所选文章按年份的分布情况，凸显了随着时间推移，指数跟踪领域稳步增长且研究兴趣日益浓厚。此外，表2 列出了所选语料库中被引次数最多的文章，展现了塑造该领域的开创性贡献和有影响力的研究。

2.2 文章分类

基于审查过程得到精炼后的文章语料库，下一步是根据指数跟踪所采用的建模框架对文献进行分类和整理。依据以往研究所采用的方法论基础，我们将所选文章大致分为三大类：基于优化的框架、基于统计的框架和数据驱动框架。

- **基于优化的框架：**此类研究将指数跟踪问题表述为数学优化任务。这些方法的主要目标是在预算、基数和风险敞口等约束条件下，通过最小化指定的跟踪误差或偏差指标，构建能够紧密复制基准指数的跟踪投资组合。该领域的文献探讨了多种优化技术，包括线性规划、二次规划和组合优化。常用的跟踪误差指标包括均方误差、误差平方和、跟踪误差标准差和平均绝对偏差。根据误差指标和约束条件的选择，所得优化问题可能是线性或非线性的，往往会带来计算挑战，尤其是在大规模场景下。

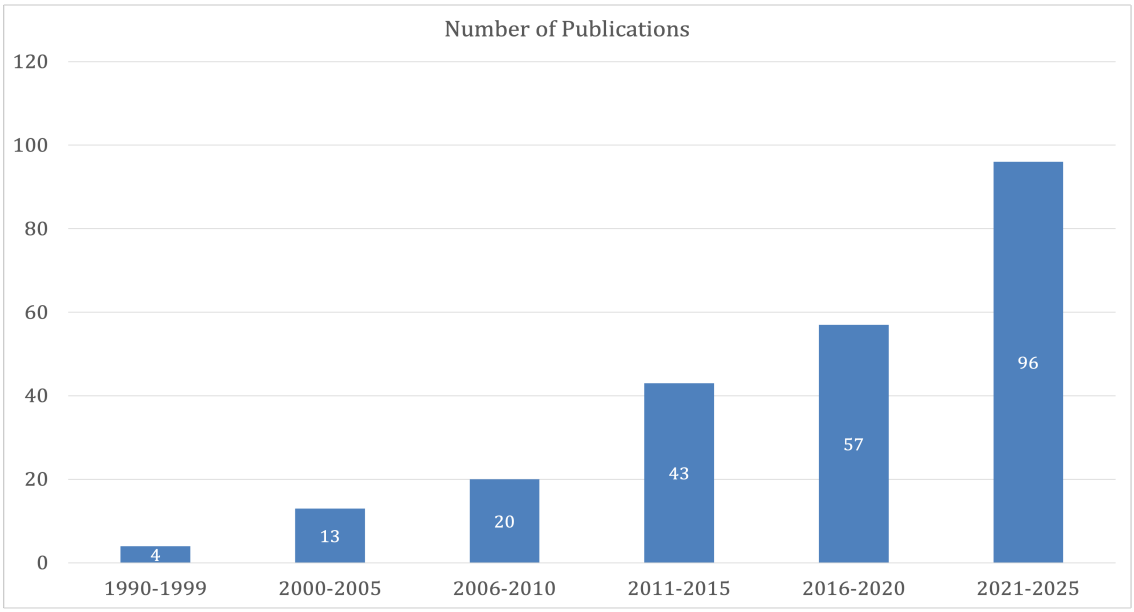


图4：逐年发表数量

表2：基于引用次数的高被引指数跟踪文章

作者	标题	TC
Beasley 等人 (2003年)	一种用于指数跟踪问题的进化启发式算法	266
Canakgoz 和 Beasley (1999年)	用于指数跟踪和增强型指数化	192
Rudolf 等人 (1999年)	用于跟踪误差最小化的线性模型	141
Guastaroba 和 Speranza (2012年)	核搜索：在指数跟踪问题中的应用	139
Dose and Cincotti (2005)	金融时间序列聚类及其在指数和增强型指数跟踪投资组合中的应用	120
Frino 和 Gallagher(2001)	跟踪标准普尔500指数基金	109
Gaivoronski 等(2005)	最优投资组合选择与动态基准跟踪	102
Ruiz-Torrubiano 和 Su´arez（2009年）	债券投资组合优化问题及其在指数跟踪中的应用：一种部分优化方法	97
Wu 等人(2014年b)	非负 Lasso 及其在指数跟踪中的应用	79
Konno 和 Watanabe (1996年)	一种用于指数跟踪的混合优化方法	77
Benidis 等 (2018年)	用于高维金融指数跟踪的稀疏投资组合	70
Corielli 和 Marcellino(2006年)	基于因子的指数跟踪	68
Chen 和 Kwon （2012年）	用于指数跟踪的稳健投资组合选择	67
Krink 等人 （2009年）	用于约束指数跟踪的差分进化与组合搜索	62
Strub 和 Baumann （2018年）	指数跟踪投资组合的最优构建与再平衡	59
Takeda 等人（2013年）	兼顾样本外表现与指数的稀疏性	58
Focardi and Fabozzi(2004年)	跟踪投资组合 一种基于时间序列聚类的指数跟踪方法论	53
Fastrich 等人（2014年）	指数跟踪中的基数与 q-范数约束	48
Sant'Anna 等人（2017年）	使用混合启发式方法进行受控资产数量的指数跟踪 结合遗传算法与非线性规划	39

• **基于统计的框架：**此类研究采用统计和计量经济学方法来设计和分析指数跟踪投资组合。这些方法侧重于使用回归分析、因子模型和协整方法等技术，对资产收益与基准指数之间的潜在关系进行建模。基于统计的框架强调可解释性和推断，有助于识别驱动指数变动的长期均衡关系、共同风险因子和解释变量。通过利用资产收益的统计结构，这些模型旨在实现有效跟踪，同时保持具有经济意义的投资组合构成。

• **数据驱动框架：**此类别涵盖了利用以数据为中心的机器学习方法进行指数跟踪的研究。数据驱动框架强调适应性和预测建模，能够动态纳入市场信息、历史数据和经济指标。该研究方向采用了多种机器学习技术，包括聚类算法、

结合回归的随机森林模型、支持向量回归、深度自编码器和神经网络。这些方法特别擅长捕捉资产收益与基准指数之间的非线性关系和复杂依赖关系，通常旨在提升市场状况动态多变时的跟踪表现。

这一分类用于梳理后续对指数跟踪方法的综述 es.

3 指数跟踪框架的演进

指数跟踪的理论基础源于现代投资组合理论和均衡资产定价。 马科维茨（1952）提出的均值—方差框架为构建有效投资组合奠定了数学基础，而Sharpe （1964）提出的资本资产定价模型（CAPM）则在均衡假设下，将市场投资组合确立为最优被动投资。法马Fama （1970）提出的有效市场假说进一步强化了这些理论成果。该假说认为，在信息有效的市场中，主动管理难以持续获得超额收益。 詹森（1968年）提供了支持这一观点的实证证据，记录显示，主动管理的共同基金经风险调整后，系统性地跑输市场基准。

这些理论和实证洞见共同为被动投资策略奠定了基础，并直接推动了指数基金的出现。这一范式转变最终促成先锋指数基金于1976年推出，其设计旨在复制标准普尔500指数的表现。该基金的成功标志着资产管理领域的转折点，确立了指数跟踪作为主动投资组合管理的一种可行且成本高效的替代方案。

在这些概念基础之上，Roll (1992年) 通过将跟踪误差波动率（TEV）定义为投资组合与基准收益率之差的方差，提出了指数跟踪（IT）的首个正式的基于优化的框架。Roll 通过最小化这一二次偏差，将指数跟踪重新表述为一个定义明确的均值—方差优化问题，从而为基准复制建立了严谨的数学表述。这一贡献标志着一个转折点，使指数跟踪在均值—方差范式下成为一个精确的数学优化问题。

后续研究通过提出替代性的跟踪误差度量，拓展了这一框架。尤其是，Rudolf 等人 (1999年) 提出了线性跟踪误差指标，包括跟踪投资组合与基准之间的平均绝对偏差、平均绝对下行偏差、最大偏差和最大下行偏差。这些表述增强了模型对异常值和非对称收益分布的稳健性。尽管如此，二次跟踪误差度量（例如 TEV）仍在学术研究和行业应用中占据主导地位，这得益于其良好的可微性、解析易处理性以及与凸优化技术的兼容性。

迈向实用化指数跟踪的一大步由 Beasley 等人（2003年）迈出，他们将交易成本和基数约束纳入均方根跟踪误差（RMSE）最小化框架。通过明确限制跟踪投资组合所持资产的数量，他们的模型考虑了基金经理面临的关键现实因素。然而，加入基数约束和换手率约束后，问题转变为混合整数非线性规划，因而成为NP难问题。为应对这一计算挑战，作者提出了基于种群的进化启发式算法，证明尽管问题具有组合性质，仍能高效获得近似最优的跟踪投资组合。

这一贡献推动了关于指数跟踪启发式和元启发式求解方法的研究不断发展。早期进展包括 Derigs 和 Nickel（2003年）提出的模拟退火方法，该方法采用线性多因子模型估计收益率和协方差；以及 Corielli 和 Marcellino（2006年）提出的因子驱动的构造启发式，该方法根据资产相对于基准的因子载荷，按顺序选择资产。随后，Zhu 等人（2010年）引入粒子群优化（PSO），以更有效地探索指数跟踪投资组合的高维搜索空间。

后续研究通过开发结合全局探索与局部改进的混合算法，拓展了早期的这些启发式方法，旨在同时提升收敛速度和解的质量。Scozzari 等人 (2013年) 提出了一种混合进化—局部搜索框架，用于最小化均方跟踪误差。研究表明，在进化算子中嵌入局部改进步骤，可显著提高跟踪精度。他们的研究结果凸显了在大规模指数跟踪问题中平衡分散化与开发的重要性。

基于这一思路，Sant'Anna 等人 (2017年) 提出了一系列专门化启发式方法，将基于回归的资产预筛选与自适应局部搜索机制相结合。通过在优化之前降低候选资产集的维度，他们的方法在多个基准指数上取得了具有竞争力的跟踪表现，同时大幅缩短了计算时间。近期的研究贡献延续了混合化与算法专门化的趋势。例如，Alvaro Rubio-García 等人 (2024年) 开发了一种针对基数约束指数跟踪问题的混合模拟退火框架，而 Ayón 等人(2024年) 提出了一种基于和声搜索的元启发式算法，采用问题专用搜索算子和双种群初始化方案，以加快在大型资产集合中的收敛。

总体而言，这些研究确立了启发式和元启发式优化作为实用指数跟踪的核心支柱，尤其适用于精确的混合整数形式在计算上难以求解的情形。它们还展现了研究方向的渐进转变：从通用元启发式算法转向具备领域知识的混合算法，明确利用金融结构、稀疏性和基准依赖性。

除启发式方法的发展外，另一项重要研究方向着眼于通过重新表述和凸化来提高指数跟踪模型的可处理性。具体而言，多项研究试图以线性或分段线性替代方案取代二次跟踪误差度量，从而将非线性混合整数规划转化为更易求解的线性或凸优化问题。该方向的一项重要贡献来自 Guastaroba 和 Speranza (2012年)，该研究在保持跟踪目标完整性的同时，实现了更高效的求解方法。

尽管跟踪误差波动率等二次跟踪误差度量仍占主导地位，研究者也探索了其他二次形式，以提升数值稳定性和计算效率。具体而言，Xu 等人 (2016年)提出误差平方和（SES）作为一种替代跟踪目标，其定义为投资组合收益与基准收益之间的偏差平方随时间累积后的总和。所得优化模型保留了二次结构，但其偏差汇总方式不同于 TEV。为高效求解相关问题，作者开发了一种非单调投影梯度算法，并展示了该算法在大规模指数跟踪实例上的有效性。尽管基于误差平方和的形式在计算上颇具吸引力，后续研究发现，当其与基数约束结合时，可能会导致投资组合高度集中，凸显了数值可处理性与分散化之间的权衡。

与基于优化的公式并行发展的是，大量文献逐渐构建了用于指数跟踪的统计框架。这些框架利用指数收益与成分资产收益之间的经验关系，通常无需明确求解全局跟踪误差最小化问题。这些方法强调推断、可解释性和结构依赖性，而不是直接优化。该方向最早的贡献之一是Corielli 和 Marcellino （2006年）提出的基于因子的指数跟踪框架。该框架将基准指数及其成分股建模为由少量潜在公共因子和特质成分驱动。通过识别因子载荷与指数因子载荷高度相似的资产，所构建的投资组合能够通过共享的系统性敞口，而非直接的收益率匹配，来复制指数动态。

基于回归的公式构成了另一类重要的统计型指数跟踪模型。这些方法试图将基准收益表示为所选成分资产收益的线性组合，从而使投资组合行为在期望意义上与指数保持一致。一个开创性贡献来自Canakgoz 和 Beasley （2009年），他们提出了一个混合整数回归框架，其中投资组合权重的选择方式使得投资组合收益对指数收益进行回归时，截距接近于零、斜率接近于一。

后续研究通过纳入稳健性和分布方面的考量，扩展了基于回归的指数跟踪。具体而言，分位数回归被用于应对重尾且非对称的收益分布下最小二乘估计的局限性。早期由 Mezali 和 Beasley（2013年）开展的研究表明，基于分位数的跟踪投资组合通过关注条件分位数而非均值行为，能够提供更好的下行保护。随后，Li（2020年）在高维场景下进一步发展了这一研究方向，提出了稀疏分位数回归模型，在保持对极端市场波动稳健性的同时，实现了有效的资产选择。较新的研究（如 Aguilar 等人 （2022年）探讨了多分位数和尾部概率跟踪目标，凸显了基于分位数的方法对于风险敏感型指数复制的重要性。

统计学指数跟踪研究中另一个颇具影响力的方向，建立在正则化和稀疏学习原理之上。这些方法对投资组合权重施加明确的惩罚项，以诱导稀疏性，从而在不引入二元决策变量的情况下实现资产选择。早期研究由 Wu 等人 (2014a提出，将非负

Lasso回归应用于指数跟踪，在确保投资组合具有经济可解释性的同时，构建仅做多投资组合，并自动控制投资组合规模。采用弹性网络惩罚项的扩展方法（Wu and Yang，2014年）通过结合 ℓ_1 和 ℓ_2 正则化，在稀疏性与稳定性之间取得平衡，缓解纯 Lasso解在存在高度相关资产时的不稳定性。更一般而言， q-范数正则化框架（Fastrich 等人，2014年）能够灵活控制稀疏程度和跟踪精度。从人工智能的视角来看，这些方法可视为金融领域监督式稀疏学习的早期形式，其中正则化在防止过拟合、提升高维资产空间中的泛化能力方面发挥着核心作用。

基于协整的方法是另一类重要的统计型指数跟踪模型，尤其适用于捕捉资产价格与基准指数之间的长期均衡关系。这些方法不着眼于匹配短期收益波动，而是专注于识别一组资产，使其价格过程与指数共享共同的随机趋势。Sant'Anna等人Sant'Anna 等人(2017年的早期研究表明，即使存在短期偏差，由协整资产构建的投资组合仍具有出色的长期跟踪特性。Sant'Anna 等人Sant'Anna 等人(2020b)进一步拓展了这一框架，提出凸协整建模方法，在保持与基准长期对齐的同时，提升了数值稳定性和可扩展性。对于持续存在共同变动的市场，基于协整的跟踪投资组合尤具吸引力，因为它们能够抵御暂时性冲击，并降低对短期噪声的敏感度。

早期的指数跟踪模型主要强调最小化跟踪误差，后续研究则凸显了明确控制下行风险和尾部行为的重要性，尤其是在市场压力时期。为此，一些研究将相干风险度量纳入跟踪框架。Wang 等人(2012年在平均绝对偏差模型中引入条件风险价值（CVaR）约束，结果表明，在跟踪精度未显著下降的情况下，可以有效限制下行风险。在此基础上，Goel 等人(2018年提出了双尾混合 CVaR（TMCVaR）模型，同时惩罚相对于基准的极端正偏差和负偏差，从而控制表现过优和表现不佳。该模型汇总多个置信水平下的 CVaR 度量，提升了稳健性，使跟踪行为更加稳定。较新的研究，例如Sant'Anna 等人(2022年)和Anis等人(2023年，进一步纳入动态或自适应风险度量，以在波动市场环境中共同提升跟踪表现和风险韧性。

与基于优化和统计的公式并行，近年来，越来越多的研究开始采用数据驱动和机器学习方法进行指数跟踪。这一转变源于高维金融数据日益丰富，以及人们需要对资产收益之间的非线性、时间性和结构性依赖关系进行建模，而这些关系难以通过经典参数框架捕捉。早期的数据驱动研究主要依靠聚类技术，识别出具有代表性的资产子集，使其整体表现能够近似基准指数。典型案例包括Focardi and Fabozzi (2004年) 和Dose and Cincotti (2005) 提出的层次聚类方法；这些方法利用欧几里得距离和基于相关性的距离度量，构建维度较低的稀疏跟踪投资组合。

后续研究通过引入更灵活的相似性度量和对资产动态更丰富的表征，拓展了这些思路。Tang 和 Li (2014年) 提出了软子空间聚类，使资产能够依据自适应特征权重而非统一的距离度量进行分组，从而提高了高维场景下的稳健性。 Zhang 等人 (2021年) 通过采用动态时间规整（DTW）距离的 k-中心点聚类，将时间结构纳入考量，从而更好地对齐不同时期的收益轨迹。最近，Goel 等人 (2024年) 将拓扑数据分析（TDA）引入指数跟踪，利用持续同调从资产收益路径中提取稳定的几何特征，并基于结构相似性而非逐点距离构建跟踪投资组合。

数据驱动研究的另一条互补脉络，聚焦于利用监督式和无监督式机器学习模型，学习成分资产收益与基准指数之间的非线性映射。基于自编码器的方法（Zhang 等人，2020年）学习低维潜在表示，以概括市场的共同结构；深度潜在表示模型（Kim and Kim，2020年）以及神经网络——包括前馈、随机和循环架构（Zheng 等人，2020a；Kwak 等人，2021年；Wang 等人，2024年）则能捕捉复杂的非线性和时间依赖关系。与此同时，研究者也探索了用于指数跟踪的基于树的方法和核方法：随机森林能够实现非线性交互建模和隐式特征选择（Cao 等人，2022年）；支持向量回归（SVR）在高维空间中提供基于间隔的泛化能力（Teng 等人，2017年）；强化学习框架则利用与跟踪表现相关的奖励信号，按序调整投资组合权重（Peng 等人，2023b）。

除了预测导向的机器学习模型外，越来越多的文献强调结构学习方法，利用资产收益的相似性、几何特征和谱性质。Zhang 等人 （2025）系统比较了多种聚类技术，包括K均值、K中心点、层次聚类和基于DTW的方法，并指出，尽管基于回归的模型通常能实现更高的跟踪精度，但基于相关性的层次聚类往往能构建风险更低的投资组合。基于TDA的跟踪研究持续发展，出现了采用持久同调指导正则化和资产选择的稀疏形式，无需大量交叉验证（Goel 等人，2026）。其他近期研究则纳入网络、谱和因子学习表示，例如将随机矩阵理论与网络中心性度量相结合，以筛选相关矩阵并识别有影响力资产（Grassetti，2025）；采用结构化主成分分析的特征值匹配方法，用于低换手跟踪（Cesarone 等人，2025）；以及嵌入自适应社区信息以提升风险调整后绩效的基于网络的模型（Xu 等人，2026）。与这些结构方法相辅相成的是稀疏学习方法，包括非凸 ℓ_p 正则化（Zapata Quimbayoand MorenoTrujillo，2025）和惩罚Huber损失回归（Li 等人，2025），它们在高维指数跟踪中持续提供计算高效且稳健的解。

指数跟踪模型的发展体现了一个渐进过程：从早期基于优化的公式，逐步转向更灵活的统计和数据驱动框架。这一过程既受到理论洞见的影响，也受到实际实现约束的塑造。如上所述，现有文献可归纳为三种宽泛且部分重叠的建模范式：基于优化的方法、基于统计的方法，以及数据驱动或基于机器学习的方法。每种范式都从不同的方法论视角处理指数跟踪问题，并在跟踪精度、可解释性、计算复杂度和可扩展性方面带来不同的权衡。

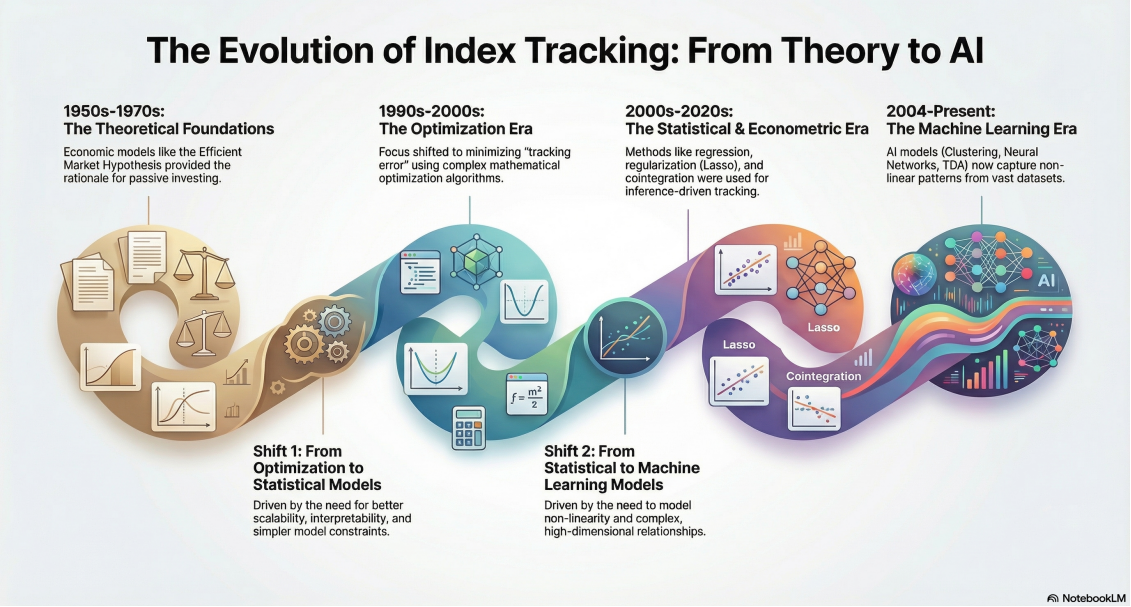


图 5：指数跟踪模型演进路线图：从理论发展到人工智能（使用 Notebook LM 制作）

在本节余下部分中，我们将使用统一的符号表示法，回顾这些类别中的主要建模框架。考虑一个包含 n项资产的资产集合，并观测 T 个时间段（或情景）。令$p_{i、t}$表示资产 i 在时间 t的历史价格，其中 $t = 1、2、\ldots\ldots、T$。资产 i 在时间 t 的收益定义为$r_{i、t} = p_{i、t} - p_{i、t-1}$。令 $R = [r_{i,t}] \in \mathbb{R}^{T\times n}$ 表示资产收益矩阵。指数跟踪投资组合由权重向量 $w = (w_1、w_2、\ldots\ldots、w_n)$ 表示，其中 w_i 表示投资于资产 i的资本比例。记基~~准指数在时间 t~~的收益为 I_t ，则指数跟踪投资组合n在时间 t 的收益为$R_{w、t} = \sum r_{i、t} w_{i\,i=1}$

3.1 基于优化的框架

在本节中，我们将讨论多年来发展起来的重要指数跟踪模型，这些模型采用基于优化的建模框架来确定指数跟踪投资组合。表3 根据不同的跟踪误差指标，列出了引用得分最高的文章。我们将结合其数学建模详细解释所有这些方法。通常，基于优化的指数跟踪问题旨在最小化投资组合收益与基准指数收益之间的跟踪误差。

表3：基于不同跟踪误差指标对指数跟踪问题中基于优化的方法所涉及的高被引文章进行分类

跟踪误差	作者	标题	TC
RMSE	Beasley 等人 (2003年)	一种用于指数跟踪问题的进化启发式算法	266
	Gaivoronski 等（2009年）	最优投资组合选择与动态基准跟踪	102
	Krink 等人（2009年）	用于约束	62
		指数跟踪	
MSE	Ruiz-Torrubiano 和 Su´arez （2009年）	指数跟踪的混合优化方法	77
	Sant'Anna 等人（2017年）	采用混合方法进行受控资产数量的指数跟踪	39
		结合遗传算法与非线性规划的启发式方法	
	Scozzari 等人 （2013年）	指数跟踪问题的精确方法与启发式方法	46
		带有 UCITS约束	
SES	Xu 等人 (2016年)	基数约束	42
		指数跟踪问题	
	Wang 等人 (2018年)	采用分层抽样和最优	6
		配置	
跟踪误差方差/跟踪误差标准差	Derigs 和 Nickel (2003年)	基于元启发式的投资组合优化决策支持	39
		以被动型投资中的跟踪误差最小化为案例研究	
		投资组合管理	
	Mutunge 和 Haugland（2018年）	最小化基数约束投资组合的跟踪误差	37
平均绝对偏差/绝对偏差	Rudolf 等人(1999年)	用于跟踪误差最小化的线性模型	141
	Guastaroba 和 Speranza (1999年)	核搜索：指数跟踪问题的一项应用	139
MADD	Rudolf 等人(1999年)	用于跟踪误差最小化的线性模型	141
最小-最大	Rudolf 等人(1999年)	用于跟踪误差最小化的线性模型	141
D最小-最大	Rudolf 等人 (1999年)	用于跟踪误差最小化的线性模型	141
MCVaR	Goel 等人(2018年)	使用混合	28
		风险价值	

标准的基于优化的指数跟踪问题（IT-OPT）对跟踪投资组合中的资产数量进行限制，是一个混合整数规划问题，定义如下：

$$\text{(IT-OPT)} \qquad \text{Min} \quad \text{TE}(R_{wt}, I_t) \tag{1}$$

满足以下约束

$$\epsilon_i z_i \leq w_i \leq \delta_i z_i, \quad i = 1, \ldots, n, \tag{2}$$

$$\sum_{i=1}^{n} w_i = 1, \tag{3}$$

$$\sum_{i=1}^{n} z_i \leq K, \tag{4}$$

$$z_i \in \{0, 1\}, \quad i = 1, \ldots, n. \tag{5}$$

这里，式（1）中的目标是最小化指数 R_{wt} 与 I_t 跟踪投资组合 I_t 的收益率之间的跟踪误差函数 TE(R_{wt}； $t = 1$、2、.....、T。式（4）中的约束是基数

约束，它将跟踪投资组合中的资产数量限制为 K，其中 $0 < K < n$。变量z_i 为二元变量；如果 $z_i = 1$，则资产 i 被纳入投资组合，而 $z_i = 0$ 则表示资产 i 未被纳入投资组合。约束（2）称为持仓约束或规模约束，它确保若资产 i 被纳入投资组合，则其权重 w_i 被限制在用户定义的参数 ϵ_i和 δ_i.

这些约束几乎存在于所有与指数跟踪问题相关的数学模型中。不同之处在于跟踪误差函数 TE(R_{wt}, I_t)，它可以呈现不同的结构形式，从线性到非凸。现在，我们介绍一些从文献中筛选出的常用跟踪误差函数，如表 3所示。

不同的跟踪误差函数及其对应的优化框架

1. **跟踪误差方差（TEV）**：基于方差的跟踪误差是指数跟踪文献中最早且应用最广泛的跟踪表现度量（Roll，1992年；Larsen 和 Resnick，1998年；Rudolf 等人，1999年）。它量化了投资组合收益与基准收益率之间差值随时间的波动性，反映了复制的一致性。跟踪误差方差（TEV）定义如下

$$\text{TEV} = \sigma^2(R_{wt} - I_t),$$

其平方根，即跟踪误差标准差（TESD），定义如下：

$$\text{TESD} = \sqrt{\sigma^2(R_{wt} - I_t)}.$$

用矩阵形式表示，若 $x = w - b \in \mathbb{R}^n$ 表示主动投资组合权重向量，其中 $b \in \mathbb{R}^n$ 为基准指数的权重向量，$\Sigma \in \mathbb{R}^{n\times n}$ 为资产收益的协方差矩阵，则 TEV 可写为

$$\text{TEV} = x^\top \Sigma x,$$

满足标准投资组合约束

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} b_i = 1, \quad 0 \le w_i \le 1$$

这可转化为

$$\sum_{i=1}^{n} x_i = 0, \text{and} - b_i \le x_i \le 1 - b_i \quad \text{for } i = 1, 2, \ldots, n.$$

T因此，使 TEV（或等价地，TESD）最小化的最终模型形式如下：

$$\begin{aligned} (\text{TEV}) \qquad & \text{Min} \quad x^\top \Sigma x \\ & \text{subject to} \\ & \sum_{i=1}^{n} x_i = 0, \\ & -b_i \le x_i \le 1 - b_i, \ i = 1, 2, \ldots, n \\ & \qquad (4)—(5) \end{aligned}$$

使 TEV（或等价地，TESD）最小化，可得到一个使其相对于基准指数的偏离波动率最小化的投资组合。尽管后来有人批评基于方差的度量忽略了系统性偏差，但由于其凸性、可解释性以及在实际指数跟踪应用中的经验稳健性，这类度量在学术研究和行业实践中仍占据核心地位。

2. **均方误差（MSE）：** 虽然 TEV 等基于方差的跟踪误差度量因简单且具有解析易处理性而早早广受欢迎，但后来它们遭到批评，被认为是“中心化”度量——也就是说，它们衡量的是投资组合超额收益围绕其均值的离散程度，而非围绕零的离散程度。因此，即使某个投资组合存在持续偏差（即其收益始终比指数高出或低于某个固定数值），且无法准确复制基准，也可能表现出较低的基于方差的跟踪误差。为解决这一局限，Beasley 等人(2003年) 提出了一种

广义跟踪误差函数，可直接衡量相对于基准收益的偏差，而不以均值为中心。这类度量可表示为

$$\mathrm{TE}_{\alpha} = \frac{1}{T}\left[\sum_{t=1}^{T} |I_t - R_{wt}|^{\alpha}\right]^{(1/\alpha)},$$

其中， α 决定对较大偏差的惩罚程度。一个常用的特殊情形是均方根误差（RMSE），对应于 $\alpha = 2$，其定义如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (I_t - R_{wt})^2}.$$

平方根的引入使 RMSE 函数变为非线性且非凸，从而形成混合整数非线性规划（MINLP）模型。因此，通常采用启发式或元启发式算法来获得近似最优解。例如，Beasley 等人 （2003年）提出了一种用于 RMSE 最小化的基于群体的启发式算法，后续研究也采用了类似的方法（Gaivoronski 等，2005；Krink 等人，2009年；Zhu 等人，2010年；Andriosopoulos 等人，2013年）。RMSE 的另一种变体是均方误差（MSE）跟踪误差函数，其表达式为

$$\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T} (I_t - R_{wt})^2 = \frac{1}{T}\sum_{t=1}^{T}\left(I_t - \sum_{j=1}^{n} r_{jt} w_j\right)^2.$$

该目标函数是凸二次函数，从而得到以下凸二次规划问题：

$$\text{(MSE)} \qquad \text{Min} \quad \frac{1}{T}\sum_{t=1}^{T}\left(I_t - \sum_{j=1}^{n} r_{jt} w_j\right)^2$$

$$\text{subject to (2)-(5).}$$

在本综述的实证研究中，我们考虑均方误差最小化问题。

3. **误差平方和（SES）：** SES 指标通过先汇总各期误差，再对总误差平方，来量化投资组合收益率与基准收益率之间的偏差。具体而言，将整个投资期限内各期的误差（定义为基准指数收益率与投资组合收益之差）相加，再将这一累计偏差平方，即得到 SES 跟踪误差函数。SES 函数在数学上表示为

$$\mathrm{SES} = \frac{1}{T}\left[\sum_{t=1}^{T} (I_t - R_{wt})\right]^2.$$

相应的跟踪模型为：

$$\text{(SES)} \qquad \text{Min} \quad \frac{1}{T}\left[\sum_{t=1}^{T} (I_t - R_{wt})\right]^2$$

约束条件为（2）-（5）。

与按期惩罚偏差的均方误差目标不同，SES 指标强调投资组合与基准之间整体方向的一致性。该函数仍为凸二次形式；然而，加入基数约束后，优化问题（SES）便成为一个 NP难问题。Xu 等人 (2016年)采用非单调投影梯度（NPG）算法应对这一挑战，并证明了该算法在最小化 SES 时识别近似最优跟踪投资组合的计算效率。

这些二次损失函数（TEV、均方误差和 SES）往往过度强调较大的偏差，可能无法准确契合投资者偏好。Rudolf 等人 (1999年)指出，由于基金经理的业绩费具有线性特征，相较于平方偏差，线性偏差能更准确地反映投资者的风险态度。基于这一理由，Rudolf 等人 (1999年)提出了四种用于指数跟踪的线性跟踪误差，具体如下。

4. **平均绝对偏差（MAD）**：Rudolf 等人提出的首个线性跟踪误差Rudolf 等人(1999年) 通过最小化跟踪投资组合收益与基准收益之间的平均绝对偏差（MAD）来实现，其定义如下

$$\mathbf{MAD} = \frac{1}{T}\sum_{t=1}^{T}|R_{wt} - I_t|.$$

由于绝对值函数不可微，因此引入两个非负辅助变量 y_t^+和 y_t^-，将其线性化；这两个变量分别表示跟踪投资组合收益相对于指数收益的正偏差和负偏差。据此，若

$$R_{wt} > I_t, \quad \text{then} \quad y_t^+ = R_{wt} - I_t \quad (\text{with } y_t^- = 0),$$

表示跟踪投资组合的表现优于指数，且若

$$R_{wt} < I_t, \quad \text{then} \quad y_t^- = I_t - R_{wt} \quad (\text{with } y_t^+ = 0),$$

表示跟踪投资组合的表现逊于指数。由于任一时刻 y_t^+或 y_t^-中只有一个为正 t，因此可将这些条件合并为如下单一约束：

$$R_{wt} - y_t^+ + y_t^- = I_t, \quad t = 1, 2, \ldots, T.$$

因此，以平均绝对偏差（MAD）为目标函数的指数跟踪问题可表述为以下线性规划问题：

$$\text{(MAD)} \qquad \text{Min} \quad \frac{1}{T}\sum_{t=1}^{T}(y_t^+ + y_t^-) \tag{6}$$

$$\text{subject to (2)-(5)}$$

$$R_{wt} - y_t^+ + y_t^- = I_t;\ t = 1, 2, \ldots, T \tag{7}$$

由于该表述是线性的，因此也是凸的，并且可通过标准线性规划技术求解，从而确保跟踪投资组合紧密复制指数收益。

5. **平均绝对下行偏差（MADD）**：虽然平均绝对偏差同时考虑投资组合收益与指数收益之间的正向和负向偏差，但在许多实际情景中，人们只关心指数的下行（或表现不佳）情况。在此类情况下，跟踪误差采用平均绝对下行偏差（MADD）衡量。Rudolf 等人 Rudolf 等人（1999年）提出了这种方法，它在平均绝对偏差模型中仅着重最小化负向偏差，其形式如下：

$$\text{(MADD)} \qquad \text{Min} \quad \sum_{t=1}^{T} y_t^- \tag{8}$$

约束条件为（2）-（5）

$$R_{wt} + y_t^- \geq I_t;\ t = 1, 2, \ldots, T \tag{9}$$

这种线性规划公式强调减少下行偏差，从而为必须避免相对于基准指数表现不佳的情景提供了一个稳健的指标。

6. **MinMax**：由Rudolf等人 （1999年）引入的第三种线性跟踪误差是MinMax。该跟踪误差旨在最小化投资组合与指数收益之间的最大（绝对值最大）偏差，代表一种稳健的“最坏情况”优化方法。从数学上讲，该跟踪误差指标定义为

$$\text{MinMax} = \min \max_t |R_{wt} - I_t|.$$

通过引入辅助变量 ξ 来界定绝对偏差，可以将该形式重新表述为一个线性规划（LP）问题，

$$\max_t |R_{wt} - I_t| = \xi \iff |R_{wt} - I_t| \leq \xi \quad \iff \quad R_{wt} - \xi \leq I_t, \quad R_{wt} + \xi \geq I_t, \quad \forall\, t = 1, 2, \ldots, T.$$

因此，MinMax 优化问题可表示为以下线性规划：

$$\begin{aligned} &\text{(MinMax)} \qquad \text{Min} \quad \xi \\ &\text{subject to (2)-(5)} \\ &R_{wt} - \xi \leq I_t;\ t = 1, 2, \ldots, T, \\ &R_{wt} + \xi \geq I_t;\ t = 1, 2, \ldots, T. \end{aligned}$$

该 LP 形式在整个期间内将最坏情况下的跟踪偏差降至最低，从而得到能够限制指数极端偏差的稳健跟踪投资组合，尤其适用于对单期大误差极度厌恶的保守型投资者。

7. **下行 MinMax（DMinMax）**：Rudolf 等人提出的第四种线性跟踪误差为Rudolf 等人(1999年)，即下行 MinMax。该方法将最大下行偏差降至最低，具体关注跟踪投资组合表现不佳于基准指数的情景。这种优化着重控制最坏情况下的表现不佳，提供了一种旨在将观测到的最大负偏差降至最低的保守策略。数学上，DMinMax 跟踪误差的形式为：

$$\text{DMinMax} = \min \max_t (I_t - R_{wt})^+.$$

通过引入表示最大下行偏差的辅助变量 ξ，并仅针对下行情景施加约束，即 $R_{wt} \leq I_t$，即可将其建模为线性规划问题：

$$\begin{aligned} &\text{(DMinMax)} \qquad \text{Min} \quad \xi \\ &\text{subject to (2)-(5)} \\ &R_{wt} + \xi \geq I_t;\ t = 1, 2, \ldots, T. \end{aligned}$$

这一线性形式直接限制了投资组合相对于基准指数的最坏情形表现不佳程度。DMinMax 模型尤其适合规避风险、寻求将相对于指数收益的显著下行偏差降至最低的投资者。

8. **基于混合 CVaR 的指数跟踪模型**：条件风险价值（CVaR）是一种一致的下行风险度量，用于衡量超过指定置信水平的预期损失 α (Rockafellar 和 Uryasev，2002年）。与对所有偏差一视同仁的基于方差的度量不同，CVaR 关注损失分布的尾部，因此能直接反映极端下行风险。在投资组合优化中，最小化 CVaR 可确保投资组合能够抵御罕见但严重的市场损失，因此自然适用于风险意识较强的指数跟踪。然而，正如Goel 等人(2018年所指出的，依赖单一置信水平 α 可能会忽略收益分布的其他部分。对此，混合 CVaR（MCVaR）框架汇总多个置信水平下的 CVaR 值，从而在保持凸性和一致性的同时，纳入更丰富的分布信息。

Goel 等人(2018年) 进一步扩展了该框架，提出了用于指数跟踪的**双尾混合 CVaR (TMCVaR)**模型。该方法对跟踪

投资组合相对于基准的下行偏差和上行偏差均予以惩罚，分别对应表现不佳和表现过佳，从而促进对基准收益动态的均衡复制。这两个组成部分通过权重参数 $\delta \in (0,\ 1)$ 进行组合，该参数用于控制对两侧尾部的相对重视程度：

$$\delta\,\mathrm{MCVaR}(-X_{wt}) + (1-\delta)\,\mathrm{MCVaR}(X_{wt}),$$

其中，$X_{wt} = I_t - R_{wt}$ 表示指数 I_t 相对于跟踪投资组合 R_{wt}的超额收益。
每个 MCVaR 分量都定义为在 m 不同置信水平下计算的条件风险价值度量的加权和，表示为：

$$\mathrm{MCVaR}(X_{wt}) = \sum_{k=1}^{m} \lambda_k^U \mathrm{CVaR}_{\alpha_k^U}(X_{wt}),$$
$$\mathrm{MCVaR}(-X_{wt}) = \sum_{k=1}^{m} \lambda_k^D \mathrm{CVaR}_{\alpha_k^D}(-X_{wt}),$$

其中 $\lambda^U = (\lambda_k^U,\ k=1,\ \ldots\ldots,\ m)$ 和 $\lambda^D = (\lambda_k^D,\ k=1,\ \ldots\ldots,\ m)$ 是满足以下条件的非负权重：$\sum_{k=1}^{m} \lambda_k^U = \sum_{k=1}^{m} \lambda_k^D = 1$。利用条件风险价值的线性化，Goel 等人 (2018年) 得到以下优化问题，该问题以投资组合权重为变量，最小化加权双尾 MCVaR 度量：

$$\textbf{(TMCVaR)} \quad \min_{\beta,w} \quad \delta\left(\sum_{k=1}^{m}\lambda_k^U\left[\beta_k^U + \frac{1}{(1-\alpha_k^U)T}\sum_{j=1}^{T}u_{jk}^U\right]\right) + (1-\delta)\left(\sum_{k=1}^{m}\lambda_k^D\left[\beta_k^D + \frac{1}{(1-\alpha_k^D)T}\sum_{j=1}^{T}u_{jk}^D\right]\right)$$

约束条件为

$$u_{jk}^U + \left(\sum_{i=1}^{n} r_{ij}w_i - I_j\right) + \beta_k^U \geq 0 \quad \text{for } k=1,\ldots,m,\ j=1,\ldots,T,$$

$$u_{jk}^D - \left(\sum_{i=1}^{n} r_{ij}w_i - I_j\right) + \beta_k^D \geq 0 \quad \text{for } k=1,\ldots,m,\ j=1,\ldots,T,$$

$$\sum_{i=1}^{n} w_i = 1,$$
$$w_i \geq 0 \quad \text{for } i=1,\ldots,n,$$
$$u_{jk}^U, u_{jk}^D \geq 0 \quad \text{for } k=1,\ldots,m,\ j=1,\ldots,T.$$

(4)—(5)

其中，变量 β_k^U 和 β_k^D 分别对应置信水平为 α_k^U 和 α_k^D时，上尾（上行）和下尾（下行）的风险价值（VaR）阈值。u_{jk}^U 和 u_{jk}^D 是非负辅助变量，分别表示在每个情景 j 和置信水平 k下，投资组合收益超出上尾和下尾相应 VaR 阈值的偏差。因此，u_{jk}^U 和 u_{jk}^D将分段条件风险价值分量线性化，而 β_k^U 和 β_k^D 则分别捕捉损失分布中与上行和下行风险相对应的条件分位数。

上述模型是一个线性规划，因此在捕捉偏差分布两端的同时，仍具有良好的计算可处理性。通过最小化双尾 MCVaR，该框架有效平衡了相对于基准的超额表现与表现不佳，从而在控制尾部风险的同时实现更紧密的复制。实证方面，Goel 等人(2018年) 报告称，与基于方差和平均绝对偏差的模型相比，所提出的基于双尾多置信水平条件风险价值的指数跟踪模型在多个全球指数上与基准的相关性更高，且跟踪误差更低。

表4：指数跟踪中不同基于统计的方法主题簇

主题	作者	标题	TC
最小二乘回归	Canakgoz 和 Beasley （2009年）	用于指数跟踪和 增强型指数化的混合整数规划方法	192
协整	Sant'Anna 等人(2017年)	使用协整进行指数跟踪和增强型指数化 以及采用内生投资组合选择的相关性	26
	Acosta-Gonz´alez 等人(2015)	关于指数跟踪与统计套利中的选择 通过协整选择股票：选股的作用	12
	Sant'Anna 等人(2020b)	基于凸方法求解指数跟踪问题 重构求解指数跟踪问题以实现协整	5
正则化	Wu 等人(2014a)	非负 Lasso 及其在指数跟踪中的应用	79
	Fastrich 等人(2014年)	指数跟踪中的基数与 q-范数约束	48
	Wu andYang(2014年)	非负弹性网络及其在指数跟踪中的应用	44
	Sant'Anna 等人(2020a)	基于 Lasso 的指数跟踪与统计套利多空 策略	23
基于因子	Corielli 和 Marcellino (2006年)	基于因子的指数跟踪	68
	Kwon 和 Wu （2017年）	基于因子的稳健指数跟踪	15
分位数回归	Mezali 和 Beasley （2013年）	用于指数跟踪和增强型指数化的分位数回归	33
	Li （2020年）	基于复合分位数的高效稀疏投资组合	5
	Aguilar 等人 (2022年)	用于高维指数跟踪的回归 利用分位数构建更优的跟踪投资组合	2

3.2 基于统计的指数跟踪优化模型

统计建模方法利用基于回归的技术以及金融数据固有的统计特性来构建指数跟踪投资组合。与传统优化模型不同，这些方法侧重于建立资产收益与基准指数之间的统计关系，从而能够稳健地捕捉复杂依赖关系。此类方法中的主要技术包括最小二乘回归、分位数回归、协整分析，以及 LASSO 和弹性网络等正则化回归模型；高被引文章列于表4。下文将简要介绍这些统计方法，并讨论它们在指数跟踪文献中的模型构建方式及应用。

1. 基于回归的模型： 基于回归的方法为刻画单项资产收益率与基准指数之间的关系提供了自然且直观的框架。通过将每项资产的收益表示为指数收益率的线性函数，回归能够捕捉资产与市场共同变动的紧密程度，并通过斜率和截距系数加以量化。在指数跟踪背景下，这种方法尤其有价值，因为它能够构建投资组合，使其综合收益动态模拟基准指数的收益动态。因此，这种基于回归的公式连接了统计估计与投资组合优化，同时兼具可解释性和可处理性。两种常用的基于回归的方法包括下文将讨论的最小二乘回归（LSR）和分位数回归（QR）。

（a）最小二乘回归方法

LSR方法由 Canakgoz 和 Beasley （2009年提出，将指数跟踪问题表述为一个基于回归的优化模型。首先，分别将每项资产的收益率与基准指数收益率进行回归，从而为每项资产得到一个线性回归方程，形式如下：

$$r_{it} = \alpha_i + \beta_i I_t + \epsilon_{it}, \quad i = 1, 2, \ldots, n; \quad t = 1, 2, \ldots, T,$$

其中，α_i 和 β_i 分别是资产 i的截距和斜率系数，而 ϵ_{it} 是回归残差。根据这些针对单个资产的回归，跟踪投资组合的整体截距（$\hat{\alpha}$）和斜率（β）定义为各个截距和斜率的加权和：

$$\hat{\alpha} = \sum_{i=1}^{n} w_i \alpha_i, \quad \hat{\beta} = \sum_{i=1}^{n} w_i \beta_i.$$

指数跟踪最小二乘回归模型的目标是找到由 K 项资产构成的最优子集，使其加权收益尽可能贴近基准指数，理想情况下，投资组合截距为零（$\hat{\alpha}=0$），斜率为一（$\beta=1$）。然而，对于现实数据，这未必能够实现。为此，作者提出了多种方法，其中采用的方法之一是两阶段方法：主要目标是使截距达到期望值零，次要目标是使斜率达到期望值一，即：先最小化 $|\hat{\alpha}-0|$，再最小化 $|\beta-1|$.。尽管模数目标是非线性的，但可以通过引入辅助变量 d 和 e将其线性化，使得

$$d \geq \hat{\alpha}, \qquad d \geq -\hat{\alpha}, \qquad e \geq \hat{\beta}-1, \qquad e \geq -\hat{\beta}+1, \qquad d \geq 0, \qquad e \geq 0$$

因此，该问题可归结为两个线性规划，并按以下两个阶段求解：
第一阶段（最小化截距）：在第一阶段，在满足基数和持仓约束的条件下，使截距 $\hat{\alpha}$ 最小化：

$$\begin{aligned} &\text{(LSR1)} \qquad \text{Min} \quad d \\ &\qquad\qquad \text{subject to (2)-(5)}, \\ &\qquad\qquad d \geq \sum_{i=1}^{n} w_i \alpha_i, \quad d \geq -\sum_{i=1}^{n} w_i \alpha_i. \end{aligned}$$

该阶段得到的最优截距值记为 α^{opt}。
第二阶段（最小化斜率偏差）：在第二阶段，在截距固定为其在第一阶段的最优值的条件下，使投资组合斜率 β 与 1 之间的偏差最小化：

$$\begin{aligned} &\text{(LSR2)} \qquad \text{Min} \quad e \\ &\qquad\qquad \text{subject to (2)-(5)}, \quad \hat{\alpha} = \alpha^{\text{opt}}, \\ &\qquad\qquad e \geq \sum_{i=1}^{n} w_i \beta_i - 1, \quad e \geq -\sum_{i=1}^{n} w_i \beta_i + 1. \end{aligned}$$

第二阶段得到的最终投资组合即为基于最小二乘回归的跟踪投资组合，该组合通过最小化截距和斜率的偏差，能够高度复制基准指数。

（b）分位数回归方法

分位数回归（QR）通过将响应变量的分位数建模为自变量的函数，扩展了标准最小二乘回归。这种方法能够捕捉条件均值以外的资产收益分布特征，因此在存在偏度或异方差性的情况下尤为有用。对于给定的分位数水平 $\tau \in (0,\ 1)$，资产 i 在时间 t 的收益的第 τ 条件分位数建模为

$$Q_\tau(r_{it} \mid I_t) = \alpha_{i\tau} + \beta_{i\tau} I_t, \quad i = 1, \ldots, n; \quad t = 1, \ldots, T,$$

其中，$\alpha_{i\tau}$ 表示截距，$\beta_{i\tau}$ 表示第 τ 条分位数回归线的斜率。对于每项资产 i，通过最小化分位数损失函数，得到系数 $\alpha_{i\tau}$ 和 $\beta_{i\tau}$：

$$\min_{\alpha_i, \beta_i} \left[\tau \sum_{t:u_t \geq 0} |u_t| + (1-\tau) \sum_{t:u_t < 0} |u_t| \right],$$

其中，$u_t = r_{it} - \alpha_{i\tau} - \beta_{i\tau} I_t$ 为分位数回归残差。这种非对称损失根据 τ的不同，以不同方式惩罚低估和高估。该问题等价地可表示为

通过引入非负变量u_t^+和u_t^-表示残差的正部分和负部分的线性规划，使得$u_t = u_t^+ - u_t^-$：

$$\min_{\alpha_{i\tau},\beta_{i\tau},u_t^+,u_t^-} \left[\tau \sum_{t=1}^{T} u_t^+ + (1-\tau) \sum_{t=1}^{T} u_t^- \right]$$

$$\begin{aligned} \text{s.t.} \quad & u_t^+ \geq r_{it} - \alpha_{i\tau} - \beta_{i\tau} I_t, \\ & u_t^- \geq -(r_{it} - \alpha_{i\tau} - \beta_{i\tau} I_t), \\ & u_t^+,\ u_t^- \geq 0. \end{aligned}$$

求解该线性问题可得资产特定的最优系数 $\alpha_{i\tau}$ 和 $\beta_{i\tau}$。在投资组合层面，τ 分位数处的聚合分位数回归线可近似表示为资产特定估计值的加权组合：

$$\hat{\alpha}(\tau) = \sum_{i=1}^{n} w_i \alpha_{i\tau}, \quad \hat{\beta}(\tau) = \sum_{i=1}^{n} w_i \beta_{i\tau},$$

这些聚合值为投资组合收益的 τ 分位数提供了线性近似，使我们能够构建用于指数跟踪（$\tau = 0.5$）或增强型指数化（$\tau < 0.5$）的投资组合（Mezali 和 Beasley，2013年）。

目标是构建一个跟踪投资组合，使其截距尽可能接近零，并且在选定的分位数水平 τ 处的斜率接近 1。按照类似于最小二乘法的两阶段优化（Canakgoz 和 Beasley，2009年），该问题采用两阶段形式：首先最小化 $|\hat{\alpha}(\tau) - 0|$ 然后最小化 $|\beta(\tau) - 1|$ 二者均为线性规划。

第一阶段（最小化分位数 τ 处的截距）： 第一阶段最小化绝对截距偏差 $|\hat{\alpha}(\tau) - 0|$，以找到分位数回归截距尽可能接近零的投资组合。与 LSR 类似，通过引入辅助变量 $d \geq 0$ 将问题线性化，并求解：

$$\begin{aligned} \text{(QR1)} \qquad & \text{Min} \quad d \\ & \text{subject to (2)-(5)}, \\ & \sum_{i=1}^{n} w_i \alpha_{i\tau} - d \leq 0, \quad \sum_{i=1}^{n} w_i \alpha_{i\tau} + d \geq 0. \end{aligned}$$

令 d^* 表示第一阶段得到的最优截距偏差。

^第二阶段（最小化分位数 τ 处的斜率偏差）： 第二阶段通过最小化 $|\beta(\tau) - 1|$，寻找一个分位数回归斜率尽可能接近 1 的投资组合，同时保留第一阶段得到的最优截距偏差 d^*。同样，通过引入变量 $e \geq 0$ 将问题线性化，并求解：

$$\begin{aligned} \text{(QR2)} \qquad & \text{Min} \quad e \\ & \text{subject to (2)-(5)}, \quad \textstyle\sum_{i=1}^{n} w_i \alpha_{i\tau} = d^*, \\ & \sum_{i=1}^{n} w_i \beta_{i\tau} - e \leq 1, \quad \sum_{i=1}^{n} w_i \beta_{i\tau} + e \geq 1. \end{aligned}$$

这一两阶段分位数回归框架构建了一个稳健的跟踪投资组合，旨在复制特定分位数处的基准指数收益率，使投资者能够更好地管理分布风险。该框架还可纳入交易成本及其他现实约束，体现出其在实际投资组合管理应用中的灵活性。

2. **采用正则化模型的回归：** 回归模型凭借高效处理高维

数据集并防止过拟合，在指数跟踪文献中备受关注。正则化会在目标函数中引入惩罚项，促使产生稀疏解，从而构建更简单、更易解释的投资组合。投资组合选择中常用的正则化方法包括 LASSO（最小绝对收缩和选择算子）和弹性网络，它们分别通过惩罚投资组合权重的 ℓ_1 范数和组合 ℓ_1-ℓ_2 范数来实现稀疏性。这些方法在大型市场指数中构建稀疏投资组合时尤其有益：它们通过选取有限数量的代表性资产，同时保持跟踪精度。下文中，$\|\cdot\|_1$ 和 $\|\cdot\|_2$ 分别表示 ℓ_1 范数和 ℓ_2 范数，其中

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \quad \text{and} \quad \|x\|_2 = \sqrt{\sum_i x_i^2}.$$

(a) 非负 LASSO

非负 LASSO 由Wu 等人(2014a)提出，是指数跟踪中用于构建稀疏投资组合的有效两阶段方法。其核心思路是先从大量资产中选出一组 K代表性资产（其中 K 表示所需的基数），然后确定能够有效跟踪基准指数的最优权重。

在第一阶　　段中，使用以下非负 LASSO 回归模型进行资产选择：

$$\text{(NNL)} \qquad \min_{w \geq 0} \frac{1}{T}\|I - R_w\|_2^2 + \lambda\|w\|_1, \tag{10}$$

参数 $\lambda \geq 0$ 控制投资组合稀疏性（资产选择）与指数收益率跟踪准确性之间的权衡。该模型在施加稀疏性惩罚项 ℓ_1 的同时，最小化投资组合收益率与指数收益之间的均方跟踪误差。由于目标函数由凸二次损失项和线性正则化项构成，并受非负约束限制，因此这一阶段构成一个凸二次规划问题。由于 λ 的选择会显著影响资产选择，因此通过二分搜索算法确定。该迭代方法从一个区间 $[\lambda_{\min}, \lambda_{\max}]$开始，通过反复求解中点值 λ_{mid}处的非负 LASSO 问题，逐步缩小该区间，直至获得具有所需基数的投资组合。

从第一阶段选出候选资产集合后，第二阶段通过求解以下非负最小二乘回归问题来进一步优化投资组合权重：

$$\text{(NNLS)} \qquad \min_{w_{\mathcal{S}} \geq 0} \frac{1}{T}\|I - R^{\mathcal{S}}\|_2^2, \tag{11}$$

其中，$R_{\mathcal{S}}$ 表示与第一阶段选定资产相对应的投资组合收益，$w_{\mathcal{S}}$ 为这些资产的比例权重。该公式通过最小化选定子集的投资组合收益与指数收益之间的均方跟踪误差来进行优化。该问题仍属于凸二次规划，可使用标准数值求解器高效求解。在我们的实现中，我们使用 R 中 nnls 包的 nnls() 函数，通过主动集算法计算非负最小二乘解。随后对系数进行归一化，使投资组合权重之和为一。第二阶段的优化通过确定预选资产之间的最优分配，提高了跟踪精度。

因此，两阶段非负 LASSO 方法能够构建稀疏、易于解释且实际高效的指数跟踪投资组合，尤其适用于标准普尔500指数和罗素2000指数等高维指数。

(b) 非负弹性网络

虽然非负 LASSO 在选择稀疏资产集合方面效果显著，但在预测变量高度相关的情景下可能面临挑战。为了解决这一局限，Wu and Yang (2014年)提出了非负弹性网络（NN-EN），它分别结合了 LASSO 和岭回归的 ℓ_1 和 ℓ_2 惩罚项。非负弹性网络的一般形式如下：

$$\text{(NN-EN)} \qquad \min_{w \geq 0} \|I - R_w\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2, \tag{12}$$

其中，λ_1，$\lambda_2 \geq 0$ 分别是控制稀疏性和收缩的正则化参数。ℓ_1 范数（$\|w\|_1$）有助于提高稀疏性，而 ℓ_2 范数（$\|w\|_2^2$）则通过对相关预测变量进行分组来稳定解，从而增强资产选择的稳健性。

与非负 LASSO 类似，NN-EN 估计量通常采用两阶段优化策略来进行指数跟踪：

- **第一阶段（资产选择）：** 通过求解公式（12）中的非负弹性网络问题，确定候选资产的稀疏子集。通过网格与二分搜索优化调优参数 λ_1 和 λ_2：对于预定义网格中的每个 λ_2，在 λ_1 上执行二分搜索，以确保选定资产的数量不超过基数约束 K。这种自适应搜索机制确保最终模型的非零权重最多为 K，兼顾稀疏性与准确性。
- **第二阶段（最优权重估计）：** 选定资产子集后，通过求解约束非负最小二乘（NNLS）问题（即公式（11））计算最优权重。此步骤（求解 NNLS 问题）仅使用第一阶段选定的资产，确定最能复现指数收益的最优投资组合权重。

这种混合式非负弹性网络 + NNLS 流程利用非负弹性网络的优势，选择收益序列稳定且相关的资产，同时通过非负性和稀疏性保留可解释性。Wu and YangWu and Yang（2014年）的实证结果表明，与非负 LASSO 相比，非负弹性网络方法能够实现更优的跟踪表现，尤其是在资产收益呈现强交叉相关性的市场中。

3. **基于协整的模型：** 协整是一个统计概念，最初由 Granger（1981年）提出，后由 Engle 和 Granger（1987年）进一步发展。该概念建立在这样一种认识之上：两个或多个非平稳时间序列的线性组合可能是平稳的，这表明它们具有共同的长期趋势。在金融市场中，股票价格通常表现出非平稳行为，协整有助于确定不同股票之间的共同趋势，从而有利于预测未来走势。简单来说，如果两只股票存在协整关系，就意味着它们会随时间共同变动。在指数跟踪的背景下，可以应用协整来识别其价格相对于市场指数走势接近的股票。对于时间序列 X_t 和 Y_t，如果线性组合 $\beta_1 X_t + \beta_2 Y_t$ 是平稳的，那么 X_t和 Y_t 就被称为协整。我们研究以下两种利用协整生成跟踪投资组合的方法，这些方法由 Sant'Anna 等人（2017年）和 Sant'Anna 等人（2020b）提出。

(a) 基于模拟的协整方法

由 Engle 和 Granger (1987年) 提出的协整理论基础表明，尽管单个时间序列可能是非平稳的，但它们的某些线性组合可以是平稳的。这意味着，由此类资产构成的投资组合能够与基准指数保持长期均衡。基于这一原理，Sant'Anna 等人(2017年) 提出了用于指数跟踪的基于模拟的协整框架。该方法假设指数和成分资产价格序列均为一阶单整，即 $I(1)$，且它们的对数可以通过协整向量建立线性关系。形式上，长期关系建模如下：

$$\log(J_t) = \beta_0 + \sum_{i=1}^{n} \beta_i \log(p_{it}) + \varepsilon_t,$$

其中，J_t 表示时间 t的指数价格水平，$p_{i,\,t}$ 是资产 i的价格，β_i 为协整系数，ε_t 表示偏离长期均衡的程度。如果残差序列 ε_t 是平稳的，那么指数与所选资产就被认为是协整的。

由 Sant'Anna 等人 (2017年) 提出的方法论是反复选取资产随机子集，并检验相应的线性组合与指数是否满足协整条件。具体而言，该方法在每个训练窗口内执行以下步骤：

(1) **随机资产选择：** 从包含 n项资产的完整资产集合中随机选取一个包含 K 项资产的子集 S（其中 K 为所需基数）。(2) **协整**
方法：**回归：** 使用普通最小二乘法估计所选子集 S 的长期关系

$$\log(J_t) = \beta_0 + \log(P_t^S)\beta + \varepsilon_t,$$

其中，$\log(P_t^S)$表示 S 中选定资产的对数价格，ε_t 表示指数与选定资产之间长期均衡的偏差。

(3) **平稳性检验：** 使用增广迪基–富勒（ADF）检验来检验残差 ε_t 的平稳性。在 5% 显著性水平下拒绝单位根零假设（$p < 0.05$）的子集将被保留为候选协整投资组合。

(4) 基于模拟的投资组合选择：对上述过程进行足够多次迭代（随机子集）。在所有满足 ADF平稳性条件的子集中，选择残差平方和（SSR）最小的子集作为最优投资组合，即 $\min \sum_t \varepsilon_t^2$。随后将最优回归的回归系数 β_j 归一化，以确定投资组合权重，如下所示：

$$w_j = \frac{\beta_j}{\sum_{i=1}^{K} \beta_i}, \quad j = 1, \ldots, K.$$

归一化权重 w_j 被分配给相应的所选资产子集，而所有未选资产的权重均为零。

基于模拟的搜索能够高效探索可能的资产组合，适用于资产规模庞大的市场；在此类市场中，穷举式协整检验在计算上往往不可行。事实证明，由此得到的协整投资组合能够提供稳定的跟踪表现和更优的样本外表现，尤其是在传统的基于方差的方法无法捕捉长期依赖关系的非平稳市场中。

（b）凸协整模型

为克服传统基于协整的指数跟踪模型需要事先进行股票选择的局限性，Sant'Anna 等人（2020b）提出了一种凸混合整数非线性规划（MINLP）形式，该形式能够内生地确定跟踪投资组合的构成。该模型通过将基数约束直接嵌入优化问题中，扩展了Sant'Anna 等人（2017年）的框架，从而能够同时估计协整向量以及投资组合中所包含的资产子集。

从前文介绍的协整回归框架出发，

$$\log(J_t) = \beta_0 + \sum_{i=1}^{n} \beta_i \log(p_{it}) + \varepsilon_t,$$

残差项 ε_t 表示与长期均衡的偏差。优化目标是最小化所有时间段内的残差平方和，确保选定资产与基准指数共同构成平稳组合。由此得到以下凸混合整数非线性规划（MINLP）形式

$$\text{(Cvx-CoInt)} \qquad \min \sum_{t=1}^{T} \left[\log(J_t) - \beta_0 - \sum_{i=1}^{n} \beta_i \log(p_{it}) \right]^2$$

$$\text{subject to}$$

$$\sum_{i=1}^{n} \mu_i = K, \quad \mu_i \in \{0,1\}, \quad \beta_i \geq 0, \quad \beta_i \leq \mu_i, \quad i = 1, \ldots, n,$$

其中，μ_i 是一个二元决策变量，用于指示资产 i 是否纳入指数跟踪投资组合；K 指定所需的资产数量。对 β_i 施加的非负约束构成卖空限制，确保模型在实际应用中的可行性。

T通过将目标函数表示为矩阵形式，可以证明其凸性 as

$$y = Ax + b,$$

其中，$y = \log(J)$、$A = \log(P)$是 n 资产的对数价格矩阵，$x = \beta$ 是系数向量。将最小二乘目标函数定义为

$$f(x) = \|y - Ax\|_2^2 = (y - Ax)^\top (y - Ax),$$

关于 x 的梯度和海森矩阵由下式给出

$$\nabla f(x) = -2A^\top (y - Ax), \qquad \nabla^2 f(x) = 2A^\top A.$$

二阶条件证实，对于任意 A ，该问题都是凸的，因为 $A^\top A$ 是半正定的（Boyd 和 Vandenberghe，2004年）。由此得到的凸混合整数非线性规划重构省去了早期基于模拟的方法所采用的随机选择步骤，生成的跟踪投资组合既稳定，计算效率也高。Sant'Anna 等人（2020b）提供的实证证据表明，与传统的协整投资组合相比，这种基于凸优化的表述在样本内和样本外都能实现更低的换手率和更小的跟踪误差。

4. **基于因子的模型：** 基于因子的模型由 Corielli 和 Marcellino （2006年）提出，为构建低维指数跟踪投资组合提供了强大的统计框架。这些模型利用了这样一种思想：少数潜在公共因子驱动资产价格的共同变动，从而能够构建简约的跟踪投资组合，有效复制基准指数的表现。通过识别与这些因子最为密切相关的资产，可以构建有效复制基准指数的投资组合。不同于直接作用于收益率的传统基于优化的模型，这种方法首先对数价格的底层结构进行建模，以识别代表性资产，然后使用简单收益率执行最终的跟踪优化。

形式化地，设 X 为表示在 T 个时间段内观测到的 n 个候选资产对数价格的（$T \times n$）矩阵。应用主成分分析（PCA）提取影响资产价格的主要公共因子：

$$X = FL^T + E,$$

其中， F 为公共因子（主成分）的（$T \times k$）矩阵， L 为因子载荷的（$n \times k$）矩阵， E 表示特质误差的 $T \times n$ 矩阵。提取的因子数量 k 通常按照能够解释资产价格较大部分变异的原则进行选择。基于因子的指数跟踪流程包括以下详细步骤：

(a) **因子提取：** 首先对对数价格矩阵 X应用主成分分析（PCA），得到因子矩阵 F。这些提取出的因子能够捕捉系统性市场行为，并作为资产选择的依据。

(b) **通过线性回归进行资产选择：** 进行资产选择时，使用普通最小二乘法（OLS）回归，将每项资产的对数价格 X_i 对提取出的因子序列 F 进行回归

$$X_i = F\beta_i + \varepsilon_i, \quad i = 1, \ldots, N,$$

其中， β_i 表示各资产对公共因子的特定载荷， ε_i 表示特异性成分，通常假设其服从零均值噪声过程。每次回归得到的决定系数 R_i^2衡量因子对第 i项资产价格变动的解释程度，$i = 1$，……， N。根据资产的 R_i^2值进行排序，并选取排名前 K项资产（其中 K为所需基数），组成子集

$$\mathcal{S} = \{i_1, \ldots, i_K\}.$$

这确　　保留的仅是与潜在因子结构最为契合的资产。

(c) **确定最终最优权重：** 确定代表性资产后，使用其简单收益率来确定最终投资组合权重。最终投资组合权重 w 通过求解均方误差最小化问题得出：

$$\min_w \ \|R_t^{\mathcal{S}} w - I_t\|_2^2, \quad \text{subject to} \quad w_i \geq 0, \quad \sum_{i \in \mathcal{S}} w_i = 1,$$

其中， $R^{\mathcal{S}}$ 是选定资产的收益率矩阵。

基于因子的方法能够构建稳健且易于解释的跟踪投资组合，在高维情境下尤其有效，Corielli 和 Marcellino (2006年的实证研究也证实了这一点。该方法有效结合了降维、数据驱动的资产选择和约束优化，从而提升了样本内解释力和样本外跟踪表现。接下来，我们将介绍用于指数跟踪的数据驱动方法。

表5：基于不同统计方法和机器学习方法的指数跟踪主题聚类

主题	作者	标题	TC
聚类	Dose and Cincotti (2005)	金融时间序列聚类及其在指数和增强型指数跟踪投资组合中的应用	120
	Focardiand Fabozzi (2004年)	基于时间序列聚类的指数跟踪方法论	53
	Wu 等人（2017年）	用于跟踪的基于聚类的约束方法标准普尔500指数	22
深度学习	Kimand Kim (2020年)	通过深度潜在表征学习进行指数跟踪	33
	Ouyang 等人(2019)	基于深度神经网络的指数跟踪	29
神经网络	Zheng 等人 (2020a)	具有基数约束的指数跟踪：一种随机神经网络方法	21
	Kwak 等人 (2021年)	用于指数跟踪投资组合优化的固定噪声神经网络优化	24
随机森林	Yuanyuan Cao 和 Yang (2022年)	结合随机森林与多重共线性用于指数跟踪的建模	2
拓扑学习	Goel 等人 (2025)	基于拓扑数据的稀疏投资组合选择分析聚类	
	Goel 等人 (2026)	风险降低型稀疏指数跟踪投资组合：一种拓扑数据分析方法	

3.3 指数跟踪的数据驱动模型

本节将探讨指数跟踪文献中涌现的各种数据驱动方法。具体而言，我们重点介绍若干入选的先进方法，涵盖深度学习技术、基于因子的模型、聚类技术、随机森林和支持向量回归。表5根据引用次数列出了用于指数跟踪的顶尖数据驱动模型。现在，我们将详细介绍本综述分析中筛选出的几种技术。

1. **基于聚类的方法**
聚类分析又称数据分割，是一种根据相似性将资产归入同质簇的技术，确保同一簇中的资产比不同簇中的资产表现出更高的相似性。早期研究中，Focardi 和 Fabozzi (2004年) 以及 Dose and Cincotti (2005) 将聚类技术引入指数跟踪和增强型指数编制框架。具体而言，

   Dose and Cincotti (2005) 对 S&P500 数据集应用层次聚类，使用以下两种关键相似性度量：

   - **收益率的相关距离：**该度量使用以下指标评估资产收益率序列的相似性

   $$d(X,Y) = \sqrt{2\,(1 - c_{XY})},$$

   其中，c_{XY} 是资产 $X = (x_t)$ 与资产 $Y = (y_t)$ 收益率之间的相关系数 t =1，2，.....，T。

   - **百分比** **价格百分比差异：**这种替代度量方式通过比较资产价格水平
   百分比 差异来计算距离。距离定义为

   $$d(X,Y) = \min\{d_1(X,Y), d_2(X,Y)\},$$

   with

   $$d_1(X,Y) = \min_{a\in\mathbb{R}} \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t - a y_t}{x_t}\right)^2,$$

   and

   $$d_2(X,Y) = \min_{a\in\mathbb{R}} \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t - a y_t}{a y_t}\right)^2.$$

   层次聚类从每项资产各自构成一个簇开始，然后反复合并最相似的簇对，直到所有资产都归入同一个簇。此过程的结果

会以一种称为树状图的图表呈现。根据聚类结果，资产选择可采用以下两种方法之一：

- **Clust1：** 将资产划分为 K 个簇，其中 K 为期望的基数，随后从每个簇中随机选择一项资产，使每个簇都得到同等代表。
- **Clust2：** 将资产划分为少于 K 个簇，并按照簇的大小比例从中进行选择。

这两种基于聚类的资产选择方法都采用如下定义的相关距离度量 s

$$d_{ij} = \sqrt{2\,(1-\rho_{ij})},$$

其中， ρ_{ij} 表示资产 i 和 j收益序列之间的皮尔逊相关系数。该距离能够捕捉收益共同变动的相似性，且与尺度无关。采用完全连接法进行层次聚类，以形成紧凑且彼此分离良好的簇。随后，对基于聚类的选择结果进行优化，使选定资产的收益与指数收益之间的均方跟踪误差最小。将选定资产集合的索引记为 S，最终投资组合权重 w 通过求解以下均方误差最小化问题得到：

$$\min_{w} \ \|R_t^{\mathcal{S}} w - I_t\|_2^2, \quad \text{subject to} \quad \sum_{i\in\mathcal{S}} w_i = 1 \ \ w_i \geq 0, ; i \in \mathcal{S}$$

其中， $R^{\mathcal{S}}$ 为选定资产的收益率矩阵。因此，该方法将数据驱动的划分与数学规划相结合，以构建稳健的跟踪投资组合。

2. **稀疏支持向量回归**
支持向量回归（SVR）是面向回归任务的支持向量机（SVM）扩展，基于结构风险最小化原理构建，即使在数据有限的情况下，也以能够实现较高的泛化性能而闻名。由于能够灵活地对非线性关系建模，SVR 已广泛应用于文本分类、图像处理和生物医学数据分析等领域。在指数跟踪领域，Takeda 等人（2010年）将 SVR 模型应用于指数跟踪，而De Leone （2011）则制定了相应的 0–1混合整数版本，以确保满足基数限制。然而，这两种模型都存在可扩展性问题。为解决这一问题，Teng 等人（2017年）提出了两种稀疏 SVR 模型，将显式基数约束纳入其中，以构建简约且计算效率高的跟踪投资组合。

(a) **稀疏 ϵ-SVR 模型**:
稀疏 ϵ-SVR 模型旨在最小化跟踪误差，同时通过ℓ_0-范数约束实现稀疏性。该模型的形式如下：

$$\begin{aligned}
\text{(SVR-IT)} \quad & \min_{w,\,\xi} \ \frac{1}{2}\|w\|_2^2 + C_1 \sum_{t=1}^{T} \big(c(\xi_t^u) + c(\xi_t^l)\big) \\
& \text{subject to} \\
& R_{wt} w - I_t \leq \epsilon + \xi_t^u, \\
& I_t - R_{wt} w \leq \epsilon + \xi_t^l, \\
& \xi_t^l, \xi_t^u \geq 0, \quad t = 1, \ldots, T, \\
& e^T w = 1, \\
& \|w\|_0 \leq K, \\
& 0 \leq w_j \leq u, \quad j = 1, \ldots, n.
\end{aligned}$$

其中， ξ_t^u 和 ξ_t^l 是松弛变量，用于表示相对于允许误差容限的偏差。 $\epsilon \geq 0$表示最大偏差容忍度，其中参数 $C_1 > 0$ 是超过给定 ϵ 的偏差的权重； $c(\xi)$ 是偏差 $\xi \in \mathbb{R}_+$的损失函数，而 u 表示资产权重的上限（通常用于防止过度集中）。术语 $\|w\|_0$ 表示 w中非零元素的数量，通过限制投资组合中的资产数量来实现稀疏性。

为方便起见，损失函数 $c(\xi)$采用高斯形式 $c(\xi)=\frac{1}{2}\xi^2$，从而形成平滑的二次优化景观，并得到以下优化模型

$$(\epsilon\text{-SVR}) \quad \min_{w,\xi} \quad \frac{1}{2}\|w\|^2 + C_1 \sum_{t=1}^{T} \left[(R_{wt}w - I_t - \epsilon)_+^2 + (I_t - R_{wt}w - \epsilon)_+^2\right]$$

$$\begin{aligned}&\text{subject to}\\ &e^T w = 1,\\ &\|w\|_0 \le K,\\ &0 \le w_j \le u, \quad j = 1,\dots,n.\end{aligned}$$

在 ϵ-支持向量回归模型中，最大偏差容忍度 ϵ 是一个先验参数。如果无法对 ϵ做出良好估计，则可将该参数视为变量，从而得到一个新的模型，即下文所述的 ν-支持向量回归模型。

(b) **稀疏 ν-支持向量回归模型**:
第二种变体通过将容忍度水平 ϵ作为决策变量，引入了灵活性，并通过额外参数 C_2权衡其取值与跟踪精度。由此得到以下 ν-支持向量回归模型的稀疏版本:

$$(\nu\text{-SVR}) \quad \min_{w,\epsilon,\xi} \quad \frac{1}{2}\|w\|^2 + C_1 \sum_{t=1}^{T} \left[(R_{wt}w - I_t - \epsilon)_+^2 + (I_t - R_{wt}w - \epsilon)_+^2\right] + C_2\epsilon$$

$$\begin{aligned}&\text{subject to}\\ &e^T w = 1,\\ &\|w\|_0 \le K,\\ &\epsilon \ge 0,\\ &0 \le w_j \le u, \quad j = 1,\dots,n.\end{aligned}$$

在第二种模型中，参数 C_2 通过调节允许的容忍度 ϵ，明确控制跟踪精度与模型复杂度之间的权衡。

这两种稀疏支持向量回归模型由于包含 ℓ_0-范数，本质上都是非凸的。为高效求解这两个模型，Teng 等人(2017年)采用了惩罚近端交替线性最小化 (PALM)算法，并结合稀疏投影算子。该迭代过程交替求解 w 中的凸子问题并进行阈值化，以保持稀疏性，并收敛至驻点。这种稀疏支持向量回归方法为指数跟踪问题提供了有效且计算上实用的解，尤其适用于需要精确控制投资组合基数和跟踪精度的情景。

3. **随机森林**
将机器学习与统计建模技术相结合，为构建稀疏且高效的投资组合提供了强大的工具，尤其适用于高维指数跟踪问题。支持向量回归（SVR）是生成稀疏跟踪投资组合的一种方法，而另一种颇具前景的机器学习框架则基于随机森林（RF）。得益于其集成特性，随机森林模型能够有效捕捉非线性依赖关系、处理多重共线性，并内置变量重要性度量，因此非常适合从庞大的金融资产范围中识别少量有影响力资产。

在此背景下，Cao等人（2022年）提出了两种融合方法的两阶段方法，将随机森林的预测能力与岭回归的稳定性相结合。第一种方法将基于随机森林的聚类与岭回归相结合，而第二种方法则采用随机森林回归，随后进行岭回归。在这两种策略中，随机森林模型作为机器学习组件，用于执行资产选择或分组；岭回归则构成统计建模阶段，用于估计最优投资组合权重。这些混合框架展示了集成学习与正则化回归的融合如何生成稀疏、稳健且计算高效的指数跟踪投资组合。

符号*Notation*

令 $\{h(R,\theta_k)\ k=1,\ 2,\ \ldots\ldots,\ K\}$ 表示随机森林中的 K 棵决策树集合，其中 $R\in\mathbb{R}^{T\times n}$ 表示跨越 T 个时间段的资产收益矩阵， θ_k 表示第 k 棵树的参数。每棵树 $h(R,\theta_k)$ 都基于从原始数据中有放回抽取的自助抽样样本进行训练。在决策树的每个节点处，选取 m 个特征的随机子集（$m\ll n$），并使用能够最大化不纯度降低准则的特征进行分裂，例如信息增益、基尼指数或均方误差（Ayll´on-Gavil´an 等人，2024年）。
观测值 R_t 的聚合随机森林预测通过对所有 K 棵树的输出取平均值获得：

$$H(R_t)=\frac{1}{K}\sum_{k=1}^{K}h(R_t,\theta_k).$$

随机森林模型通过评估每个特征对总体预测准确度的贡献，提供变量（特征）重要性度量。该特征重要性排名为后续基于混合随机森林的指数跟踪模型中的资产选择奠定了基础。

随机森林聚类与岭回归相结合（RF 聚类+岭回归）

随机森林聚类 + 岭回归模型由 Cao 等人 (2022年)提出，将集成学习与惩罚回归相结合，以识别具有影响力的资产并确定稳定的投资组合权重。随机森林部分捕捉非线性依赖关系，并通过聚类和变量重要性度量执行特征选择；岭回归则作为统计优化步骤，用于估计平滑且稳健的投资组合权重。

步骤 ***1***。用于资产选择的随机森林分类
训练随机森林分类器，利用涵盖所有 n资产的历史收益率，预测指数收益率的方向 （即判断其为正还是负）。每个输入实例由某一天的资产收益向量构成；相应的输出标签定义为：若指数收益率为正，则为 1，否则为 0。

该分类器是由 L 决策树 $\{h(R_t,\theta_l)\ l=1,\ 2,\ \ldots,\ L\}$组成的集成模型，其中 θ_l 表示第 l棵树的参数。每棵树都使用从原始数据中有放回抽取的自助抽样样本进行训练。在每个树节点处，随机选取一部分 $m\ll n$ 特征，并选择使不纯度降低准则（例如基尼指数或分类误差）最大化的特征进行分裂。对于给定的观测值 R_t，集成模型通过对所有树的预测进行多数投票，得出最终预测：

$$H(R_t)=\arg\max_{y\in\{0,1\}}\sum_{l=1}^{L}I\big(h(R_t,\theta_l)=y\big),$$

其中， $I(\cdot)$为指示函数。这种集成构建方式降低了方差并避免过拟合，同时为特征重要性估计奠定了基础。

步骤 ***2.*** 特征重要性评估
训练好随机森林分类器后，使用准确率平均下降值（MDA）量化每项资产的重要性。该指标衡量随机置换某个特征（即资产收益率序列）的值、从而打破其与目标变量之间的关系后，预测准确率的下降幅度。具体而言，资产 i 的准确率平均下降值分数计算如下：

$$\mathrm{MDA}_i=\mathrm{Acc}_{\mathrm{base}}-\mathrm{Acc}^{(i)}_{\mathrm{perm}},$$

其中，$\mathrm{Acc}_{\mathrm{base}}$ 是随机森林在验证数据上的基准准确性，$\mathrm{Acc}^{(i)}_{\mathrm{perm}}$ 是在样本中随机置换第 i 个特征后得到的准确性。较高的 MDA_i 表明资产 i 对预测指数收益率方向具有更高的相关性。

步骤 ***3.*** 资产子集选择
根据 MDA_i 分数按降序排列资产，并选取排名前 K 个资产（其中 K 为所需基数）组成缩减候选集 $\mathcal{S}\subseteq\{1,2,\ldots,n\}$，其中 $|\mathcal{S}|=K$。该子集代表与指数表现关联最为密切的资产。

步骤 ***4***：用于投资组合优化的岭回归
使用选定的资产子集 $\mathcal{S}$，通过岭回归估计最优跟踪投资组合权重：在最小化均方跟踪误差的同时施加 ℓ_2 惩罚，以控制多重共线性。优化问题表述如下

$$\hat{w}_{\mathcal{S}} = \arg\min_{w_{\mathcal{S}} \geq 0} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(I_t - \sum_{i \in \mathcal{S}} r_{it} w_i \right)^2 + \lambda \sum_{i \in \mathcal{S}} w_i^2 \right\},$$

其中， $\lambda > 0$ 是控制收缩程度的正则化参数。
通过最小化样本外均方误差，采用 k 折交叉验证选取正则化参数 λ 。随后，将得到的权重估计值 $\hat{w}_{\mathcal{S}}$ 归一化，以满足投资组合约束

$$\sum_{i \in \mathcal{S}} \hat{w}_i = 1, \qquad \hat{w}_i \geq 0,$$

从而确保投资组合完全投资且仅做多。
这种集成的随机森林聚类 + 岭回归框架利用随机森林的特征选择能力识别信息量最大的资产，同时使用岭回归生成稳定且易于解释的投资组合权重。这种组合能够有效捕捉资产收益中的非线性依赖关系，并生成稀疏而稳健的跟踪投资组合，因此特别适用于高维指数跟踪问题。

随机森林回归与岭回归相结合（随机森林回归+岭回归）

与基于聚类的方法类似，Cao 等人(2022年提出的随机森林回归与岭回归相结合的方法采用两阶段策略，将随机森林的非线性预测能力与岭回归的稳健性结合起来。与基于分类的策略不同，该模型直接进行回归，以预测指数收益的幅度，从而捕捉资产收益与基准指数表现之间的线性和非线性依赖关系。

步骤 ***1***：用于资产选择的随机森林回归
训练随机森林回归模型，以$\{I_t\}_{t=1}^{T}$ 历史资产收益率$\{r_{it}\}$ 为预测变量，预测指数收益序列。令 $\{h(R, \theta_l)$ $l = 1, 2, \ldots, L\}$ 表示由 L 棵回归树组成的集成，每棵树均基于 $R \in \mathbb{R}^{T \times n}$ （资产收益矩阵）的一个自助抽样样本进行拟合。对于给定观测值 R_t，预测的指数收益率通过对所有 L 棵树的输出取平均值获得

$$\hat{I}_t = \frac{1}{L} \sum_{l=1}^{L} h(R_t, \theta_l).$$

这种集成方法使模型能够学习资产收益之间复杂的非线性关系和相互作用，同时通过自助聚合和随机特征选择保持泛化能力。与基于分类的方法一样，训练后的模型还会提供特征重要性得分，用于为模型的下一阶段确定一个稀疏资产子集。

步骤 ***2.*** 特征重要性评估：
通过均方误差增加百分比（置换重要性度量，PIM）量化每项资产对指数预测的贡献，该指标评估随机置换第 i 个特征时预测准确度的下降程度。形式上，对于资产 i，PIM_i 得分定义为

$$\text{PIM}_i = 100 \times \frac{1}{L} \sum_{l=1}^{L} \frac{\text{MSE}_l^{(i,\text{perm})} - \text{MSE}_l}{\text{MSE}_l},$$

其中，MSE_l 表示第 l 棵回归树的均方误差，$\text{MSE}_l^{(i,\text{perm})}$ 表示置换第 i 个特征后的误差。PIM_i 值越高，表明资产 i 对指数收益的预测影响力越强； $i = 1$，……， n。

步骤 ***3***：资产子集选择
根据资产的 PIM$_i$ 值进行排序，并选取排名前 K 个资产，构成子集$\mathcal{S} \subseteq \{1, 2, \ldots, n\}$，其中包含 $|\mathcal{S}| = K$。这种数据驱动的选择方法能够识别出最能有效复制基准指数的资产。

步骤 ***4***：用于投资组合优化的岭回归
投资组合权重采用与 RF 聚类中的 + 岭回归框架所述相同的、经过岭正则化的跟踪误差最小化程序进行估计。

该混合模型将随机森林回归的非线性预测能力与岭回归的正则化强度相结合，可生成稀疏且高度稳定的指数跟踪投资组合。它能高效识别重要资产、缓解多重共线性并提升样本外跟踪表现，因此尤其适用于大规模、高维的指数跟踪应用。

4. **深度自编码器**
自编码器是一种旨在学习数据有效表示的神经网络。其实现方式是将高维输入数据压缩至低维潜在空间，再从这种压缩形式中重建原始数据。训练过程旨在最大限度地减少信息损失，可通过多种损失函数进行衡量。常见的选择是输入向量与输出向量之间的 ℓ_2 -范数差异，尽管也可以使用其他损失函数（Ouyang 等人 (2019);Kim and Kim (2020);Zhang 等人 (2020)).

在指数跟踪的背景下，自编码器使用构成市场指数的股票收益数据进行训练。对于给定期间 t，收益数据表示为向量 $\mathbf{r}_t = [r_{1t}, r_{2t}, \ldots, r_{nt}]$ 其中r_{it} 表示 t时第 i个资产的收益。

总体 用于训练自编码器的损失函数定义为平方差之和
（ℓ_2-范数）所有时间段内原始收益向量与重构收益向量之间的平方差之和 ds:

$$L = \sum_{t=1}^{T} \|\mathbf{r}_t - \hat{\mathbf{r}}_t\|_2^2$$

其中 t是 t 训练样本的指数，$\mathbf{r}_t$ 是时间段 t 的原始收益向量，而 $\hat{\mathbf{r}}_t$ 是重构 时间段 t 的收益向量。
为评估每只股票对重构的贡献，需计算每只股票的单独重构损失。这有助于识别哪些股票包含与基准指数相关的最多共同信息。股票 i 的单独损失定义如下：

$$L_i = \sum_{t=1}^{T} \|r_{it} - \hat{r}_{it}\|_2^2$$

其中 $\hat{r}_{it}$ 是资产 i 在时间 t 的重构收益。 L_i 值越小，表示第 i个资产损失的信息越少，说明其与整体指数越相似。利用这些信息，可以根据资产与基准指数之间的共同信息对资产进行排序。选取共同信息最高的前 K 个资产，即可构建一个紧密跟踪指数的跟踪投资组合。在 Zhang 等人 (2020年）中，作者针对资产选择考虑了六种自编码器架构，具体如下：

(a) **单隐藏层自编码器（SH-AE）**：该基准模型由编码器—解码器结构组成，包含一个神经元数量少于输入层的单一隐藏（瓶颈）层。编码器将标准化资产收益数据压缩为低维潜在表示，而解码器则尝试根据这一压缩编码重构原始输入。重构误差以每种资产输入与重构输出之间的均方误差衡量，并作为排序标准——重构误差最低的资产被视为最能代表潜在市场结构。

(b) **稀疏自编码器（SP-AE）**：该模型保留了单隐藏层自编码器的相同结构，但在隐藏层激活值上引入 ℓ_1 惩罚项以实现稀疏性。这会促使模型对任何给定输入仅激活少数几个神经元，从而增强特征选择性并减少冗余。通过突出资产收益中最显著的特征，稀疏自编码器提高了资产排名的稳健性和可解释性。

(c) **收缩自编码器（CON-AE）：** 收缩自编码器将基于雅可比矩阵的惩罚项纳入损失函数，使学习到的表示对输入中的微小扰动具有不变性。具体而言，它通过添加一个与编码器雅可比矩阵的弗罗贝尼乌斯范数成正比的收缩项，来惩罚编码器对输入变化的敏感性：

$$\left\| \frac{\partial h(\mathbf{r}_t)}{\partial \mathbf{r}_t} \right\|_F^2,$$

其中， $h(\cdot)$表示编码器映射。在实际应用中，该雅可比范数通常用编码器权重矩阵的平方弗罗贝尼乌斯范数近似，作为局部敏感性的高效替代指标。这种正则化能够提升潜在空间的稳定性和平滑性，使模型尤其擅长识别其收益动态呈现一致长期结构的资产。

(d) **堆叠自编码器（STCK-AE）：** 堆叠自编码器通过在编码器和解码器组件中采用多个非线性隐藏层，扩展了基础自编码器架构。这种分层深度结构——编码器通常由规模依次为 64、32 和 16 个神经元的层构成——使模型能够提取资产收益之间复杂的高阶相关性。通过多个非线性变换逐步压缩输入数据，模型能够捕捉收益动态中的全局和局部结构，形成更丰富的潜在表示，从而更有效地识别对市场内在结构贡献最大的资产。

(e) **去噪自编码器（DEN-AE）：** 去噪自编码器通过刻意向输入数据添加随机噪声，并训练网络重建原始的未受污染版本，从而增强稳健性。将受污染的输入表示为

$$\tilde{\mathbf{r}}_t = \mathbf{r}_t + \eta_t,$$

其中， $\eta_t \sim \mathcal{N}(0,\ \sigma^2)$表示添加到原始收益向量$\mathbf{r}_t$上的高斯噪声。网络通过最小化 $\tilde{\mathbf{r}}_t$ 与$\mathbf{r}_t$之间的重构误差进行训练，从而学习资产收益的抗噪稳定表示。这一过程提升了模型的泛化能力，并增强了其应对金融时间序列中常见随机波动的韧性。

(f) **变分自编码器 (VAR-AE)：** 变分自编码器通过学习一个将输入映射到潜变量分布而非单一编码的编码器，引入概率潜在表示。给定标准化的收益向量$\mathbf{r}_t \in \mathbb{R}^n$，编码器输出高斯潜在分布的均值和对数方差，

$$(\boldsymbol{\mu}_t, \log \boldsymbol{\sigma}_t^2) = f_\phi(\mathbf{r}_t),$$

并通过重参数化技巧获得一个潜在样本，

$$\mathbf{z}_t = \boldsymbol{\mu}_t + \exp\!\left(\tfrac{1}{2}\log \boldsymbol{\sigma}_t^2\right) \odot \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, I).$$

这里， I 是单位矩阵。解码器 g_θ 将$\mathbf{z}_t$ 映射回输入空间，以重构$\hat{}\mathbf{r}_t = g_\theta(\mathbf{z}_t)$。训练时最小化重构项与库尔贝克–莱布勒散度之和，以将潜在分布正则化为标准正态分布：

$$\mathcal{L}_{\mathrm{VAE}}(\theta,\phi) = \sum_{t=1}^{T} \underbrace{\left\|\mathbf{r}_t - \hat{\mathbf{r}}_t\right\|_2^2}_{\text{reconstruction}} + \lambda_{\mathrm{KL}} \underbrace{\sum_{j=1}^{d} \tfrac{1}{2}\left(\mu_{t,j}^2 + \sigma_{t,j}^2 - \log \sigma_{t,j}^2 - 1\right)}_{\mathrm{KL}\left(\mathcal{N}(\boldsymbol{\mu}_t, \mathrm{diag}(\boldsymbol{\sigma}_t^2)) \,\|\, \mathcal{N}(\mathbf{0}, I)\right)},$$

其中， d 是潜在维度， $\lambda_{\mathrm{KL}} > 0$ 用于控制 KL 正则化的强度（代码中为：lambda kl）。这种概率瓶颈有助于形成平滑且彼此解耦的潜在因子，在防止过拟合的同时捕捉共同的市场结构。

在实现中，我们使用双层编码器生成（$\boldsymbol{\mu}_t$，log $\boldsymbol{\sigma}_t^2$），应用重参数化$\mathbf{z}_t$，并使用双层解码器重构 $\hat{}\mathbf{r}_t$。

所有自编码器变体均遵循相同的训练与选择流程。模型使用标准化资产收益数据进行训练，通过最小化原始收益率与

重构收益率之间的均方重构误差。针对每项资产，计算其平均重构损失，并据此按照资产与整体市场结构的相似程度进行排序。为兼顾代表性与多样性，采用混合选择规则——选取重构误差最小的 70% 资产（共性资产）和误差最大的 30% 资产（特异性资产）。随后，将最终选定的资产送入投资组合优化阶段进行权重估计。

选定一组资产后（将该集合记为 $\mathcal{S}$），通过优化模型确定跟踪投资组合，使基准指数收益率与选定资产所产生的收益率之间的均方误差最小（参见第3节的均方误差最小化）。

5. **深度学习** 近年来，深度学习技术，尤其是深度神经网络（DNN），已成为指数跟踪中进行资产选择和投资组合构建的强大工具。在此背景下，Ouyang 等人 (2019) 提出了一种融合深度自编码器与深度神经网络的混合两阶段方法。该方法利用自编码器的表征能力进行资产选择，并借助神经网络的预测能力进行投资组合优化。该框架包含以下顺序步骤：

深度神经网络框架

Ouyang 等人 (2019) 提出了一种基于深度学习的两阶段指数跟踪方法，将通过深度自编码器进行资产选择与通过深度神经网络（DNN）进行投资组合优化结合起来。该方法论包括以下步骤：

(a) **使用深度自编码器进行资产选择**
通过重构历史资产收益率并选择重构误差最小的资产，深度自编码器可用于识别代表性资产。具体而言，给定历史资产收益率$R \in \mathbb{R}^{T\times n}$，自编码器学习参数 θ，以最小化重构损失函数：

$$\min_{\theta} \|R - \hat{R}\|_F^2,$$

其中 $\hat{R} = g_\theta(R)$ 表示重构后的收益率，其参数由权重和偏置 θ确定。之后训练计算每项资产 i 的重构误差，如下所示：

$$d_i = \frac{1}{T}\sum_{t=1}^{T}(r_{it} - \hat{r}_{it})^2.$$

排名最前的 h 个重构误差最低的资产 d_i 被选中，用于构成缩减后的候选集合 $\mathcal{S}$。

(b) **使用深度神经网络进行投资组合权重优化**
使用选定资产 $\mathcal{S}$训练前馈式（深度神经网络），根据相应的资产收益预测基准指数收益。通过最小化均方误差损失来优化网络参数 ϕ ：

$$\min_{\phi} \frac{1}{T}\sum_{t=1}^{T}(I_t - f_\phi(R_{t,\mathcal{S}}))^2,$$

其中$R_{t,\mathcal{S}}$ 表示选定资产在时间 t的收益率， $f_\phi(\cdot)$表示由 ϕ参数化的深度神经网络预测函数。此步骤使模型能够捕捉选定资产与基准指数之间复杂的非线性依赖关系。

(c) **通过敏感性分析确定投资组合权重**
随后，通过计算预测指数收益率相对于每个输入特征的基于梯度的敏感性，利用训练好的网络获得投资组合权重。资产i 的平均绝对梯度计算如下：

$$s_i = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\partial f_\phi(R_{t,\mathcal{S}})}{\partial r_{it}}\right|, \quad i \in \mathcal{S}.$$

这些敏感性得分经过归一化处理，以获得满足条件的投资组合权重

$$w_i = \frac{s_i}{\sum_{j\in\mathcal{S}} s_j}, \qquad \sum_{i\in\mathcal{S}} w_i = 1, \quad w_i \geq 0.$$

该方法有效利用深度学习动态选择资产并分配权重，与传统方法相比，实证表现出色，同时显著降低了跟踪误差和投资组合复杂度。

固定噪声深度神经网络（Deep NNF）

Kwak 等人提出的固定噪声深度神经网络（Deep-NNF）方法（2021年）为投资组合选择和指数跟踪提供了一种创新方法。与将资产收益直接作为输入的传统前馈神经网络不同，Deep-NNF 模型通过一个独立于观测市场数据的潜在噪声向量生成投资组合权重。该框架独特之处在于，摒弃了将显式资产收益作为输入，而仅使用固定的随机噪声向量作为深度神经网络的输入。这种由噪声驱动的方法天然地强调投资组合多元化，并避免对历史数据过拟合。

Deep-NNF 方法论包括以下两个主要阶段：
第一阶段：通过固定噪声神经网络进行资产选择。
首先，构建一个深度前馈神经网络 $f(\cdot)$，并以固定生成的随机噪声向量 $\xi \in \mathbb{R}^n$ 作为其唯一输入。网络架构由多个带有丢弃率和 ReLU 激活的全连接层组成，表示如下：

$$\mathbf{w} = f(\xi; \theta),$$

其中，θ 表示网络参数（权重和偏置）。输出层采用 softmax 激活函数，以生成有效的投资组合权重：

$$w_i = \frac{e^{u_i}}{\sum_{j=1}^{N} e^{u_j}}, \quad i = 1, \ldots, N,$$

其中，u_i 表示资产 i在最后一层的激活前输出。通过最小化实际投资组合收益与基准指数收益之间的均方误差来学习网络参数 θ，即

$$\mathcal{L}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \left(I_t - R_{wt}\right)^2.$$

训练完成后，根据相应投资组合权重的大小选出排名前 h 的资产。

第二阶段：部分复制与投资组合优化。
第二阶段中，仅使用选定的 h 项资产的收益率重新训练神经网络。具体而言，使用与所选资产对应的资产收益矩阵$\mathbf{r}_t^{(h)} \in \mathbb{R}^h$，并采用新的固定噪声向量 $\xi^{(h)} \in \mathbb{R}^h$。模型再次最小化均方跟踪误差：

$$\min_{\theta_h} \frac{1}{T} \sum_{t=1}^{T} \left(I_t - \mathbf{w}_h^\top \mathbf{r}_t^{(h)}\right)^2,$$

以获得优化后的最优投资组合权重 $\mathbf{w}_h$。最后，将这些优化后的权重扩展回完整的资产空间，并将未选定资产的权重设为零，从而完成部分复制策略。这种两阶段 Deep-NNF 框架利用固定噪声输入中固有的随机性以及深度神经网络的灵活性，构建稳健且多元化的投资组合；与传统指数跟踪方法相比，该框架展现出更优的跟踪精度和泛化能力（Kwak 等人，2021年）。

4 数值研究

为说明所评述指数跟踪（IT）模型的实际应用意义，我们对标准普尔500指数开展了一项比较数值研究。样本包括该指数及其成分股在十年期间的每日收盘价，时间范围为2012年12月10日至2022年8月11日，涵盖 $T = 2524$ 个交易日。对于资产 $i \in \{1, \ldots, n\}$ 和交易日 $t \in \{1, \ldots, T\}$，令$p_{i,t}$表示收盘价，其简单日收益率为

$$r_{i,t} = \frac{p_{i,t} - p_{i,t-1}}{p_{i,t-1}}. \tag{13}$$

保留在该期间具有完整价格历史的成分股，共得到 $n = 462$ 项资产。选择这一期间是为了使具有可用数据的成分股数量最大化。

4.1 方法论

我们评估了涵盖三种建模范式的 29 种代表性指数跟踪模型：(i) 基于优化的方法（8 种模型）、(ii) 基于统计的方法（7 种模型）和 (iii) 数据驱动方法（14 种模型）。每种模型均采用滚动窗口设计进行评估：先使用两年的样本内窗口（504 个交易日），再使用三个月的样本外测试窗口（63 个交易日）。窗口每次向前滚动 63 天，共形成 32 个评估窗口。为确保可比性，我们统一施加基数约束 $K = 45$ （约占资产总数的 10%）。

我们采用下文详述的多种指标评估样本外表现。此外，我们对模型的跟踪误差进行成对统计检验，以评估观察到的表现差异是否显著。最后，我们开展两阶段比较：(i) 在各建模范式内进行评估，以确定每个类别中的领先模型；随后 (ii) 对各范式中表现最佳的代表模型进行跨范式评估。这种方法不仅揭示了绝对表现，也凸显了基于优化、基于统计和数据驱动框架在指数跟踪方面各自的优势与局限。

4.2 绩效指标

跟踪投资组合的表现通过以下五类互补指标进行评估：(i) 跟踪表现、(ii) 收益表现、(iii)风险表现、(iv)风险调整后表现，以及 (v) 资产权重特征。设 $r_{p,\ t}$ 和$r_{b,\ t}$ 分别表示在长度为 H 的样本外评估期内，时点 t 的投资组合收益和基准收益。平均投资组合收益表示为 $\bar{r}_p = \frac{1}{H}\sum_{t=1}^{H} r_{p,\ t}$。

—

1. **跟踪表现：** 为评估投资组合对标准普尔500指数的复制程度，我们采用两个互补指标：(i) 跟踪误差，用于衡量偏差幅度；(ii) 相关性，用于反映方向上的共同变动。

(a) **跟踪误差（TE）：** TE 衡量主动收益序列（$r_{p,\ t} - r_{b,\ t}$）的变异性，即投资组合平均偏离基准的程度。其计算公式为

$$\mathrm{TE} = \sqrt{\frac{1}{H-1}\sum_{t=1}^{H}\Big((r_{p,t} - r_{b,t}) - \overline{(r_p - r_b)}\Big)^2},$$

其中 $\overline{(r_p - r_b)} = \frac{1}{H}\sum_{t=1}^{H}(r_{p,\ t} - r_{b,\ t})$。跟踪误差值越小，表示对基准的复制越精确。（b）**与指数的相关性：**相关性用于评估投资组合与基准之间线性协动的强度。其计算公式为

$$\rho(r_p, r_b) = \frac{\mathrm{Cov}(r_p, r_b)}{\sigma_p \sigma_b},$$

其中，σ_p 和 σ_b 分别为$r_{p,\ t}$ 和$r_{b,\ t}$的标准差。

2. **收益表现：**为评估跟踪投资组合的增长潜力，我们报告三个简单指标：评估期内的平均收益、观测到的最低收益和最佳观测收益。

（a）**平均收益：**平均收益率 $\bar{r}_p$ （已在上文定义）可用于衡量跟踪投资组合的典型表现。（b）**最低收益：**观测到的单期最低收益，

$$r_{\min} = \min_{1 \le t \le H} r_{p,t},$$

用于捕捉下行极值。

(c) **最高收益：** 观察到的最高单期收益率，

$$r_{\max} = \max_{1 \le t \le H} r_{p,t},$$

表示实现的最佳结果。

3. **风险表现：** 为评估跟踪投资组合的稳定性，我们考虑两个指标：波动率（收益率的总体变异程度）和平均回撤（相对于峰值的典型损失）。

（a）**标准差：** 投资组合收益率波动率的度量记为 σ_p，其计算公式为

$$\sigma_p = \sqrt{\frac{1}{H-1}\sum_{t=1}^{H}(r_{p,t} - \bar{r}_p)^2}.$$

数值越低，表明投资组合表现的稳定性越高。

（b）**平均回撤：** 设累积财富过程为 $W_t = \prod_{s=1}^{t}(1 + r_{p,s})$。时刻 t 的回撤定义为

$$DD_t = \frac{W_t - \max_{1\le s\le t} W_s}{\max_{1\le s\le t} W_s} \le 0.$$

平均回撤则为

$$\overline{DD} = \frac{1}{H}\sum_{t=1}^{H}|DD_t|,$$

该指标反映了整个评估期内相对于峰值水平的平均比例损失。

4. **风险调整后绩效：** 如果不考虑风险敞口，仅看绝对收益可能会产生误导。为了进行更公平的比较，我们评估将超额收益与不同风险度量相关联的绩效比率。只有当投资组合实现正的超额收益时，这些比率才具有明确的解释意义。

(a) **夏普比率：** 表示平均超额收益与标准差的比值，

$$\text{Sharpe Ratio} = \frac{\bar{r}_p - r_f}{\sigma_p}, \quad \bar{r}_p - r_f > 0,$$

其中，σ_p 是投资组合的标准差。数值越高，整体风险调整后收益越好。

(b) **索提诺比率：** 通过以负向偏差替代总波动率，聚焦于下行风险 ，

$$\text{Sortino Ratio} = \frac{\bar{r}_p - r_f}{\sigma_d}, \quad \bar{r}_p - r_f > 0,$$

其中，σ_d 是负收益的标准差。数值越高，表明对损失的防护越强。

(c) **特雷诺比率：** 衡量单位系统性风险所对应的超额收益，

$$\text{Treynor Ratio} = \frac{\bar{r}_p - r_f}{\beta_p}, \quad \bar{r}_p - r_f > 0,$$

其中，β_p 是相对于基准的投资组合贝塔，体现了相对于市场敞口的表现。

(d) **信息比率：** 用于评估主动收益相对于跟踪误差的表现，

$$\text{Information Ratio} = \frac{\bar{r}_p - \bar{r}_b}{\text{TE}}, \quad \bar{r}_p - \bar{r}_b > 0,$$

其中，$\bar{r}_b$ 是基准收益的平均值。数值越高，表示跑赢基准的效率越高。

5. **换手率与时间复杂度：** 为评估所构建投资组合的成本效率和稀疏性，我们考虑三项指标，其中两项基于投资组合权重，另一项基于所用时间。

(a) **换手率：** 用于衡量各再平衡窗口内的平均交易活动。换手率越高，意味着交易成本越高。其定义为

$$\text{TR} = \frac{1}{N-1}\sum_{k=1}^{N-1}\sum_{i=1}^{n}\left|w_{i,k+1} - w_{i,k}\right|,$$

其中 $w_{i,\,k}$ 是第 k 个窗口中资产 i 的权重，N 是滚动窗口的总数 s.

(b) **资产数量：**跟踪投资组合中具有非零权重的资产的平均数量。这反映了投资组合稀疏性，数值越小表示投资组合越集中。(c) **求解器时间：** 求解器在所有滚动窗口中获得最优或近似最优解所耗费的 CPU 时间总和。该指标反映了各建模框架的计算可处理性和可扩展性，从而为模型复杂度与现实适用性之间的权衡提供了实践视角。

4.3 统计检验

为评估不同模型的跟踪误差差异是否具有统计显著性，我们进行成对单侧配对 t检验。对于每一对模型（i, j），令 $\mathrm{TE}_{i,\,k}$ 和 $\mathrm{TE}_{j,\,k}$ 分别表示窗口 k 中的样本外跟踪误差，其中 $k = 1, \ldots\ldots, N$。零假设和备择假设定义如下

$$H_0 : \bar{\mathrm{TE}}_i = \bar{\mathrm{TE}}_j \quad \text{vs.} \quad H_A : \bar{\mathrm{TE}}_i < \bar{\mathrm{TE}}_j,$$

其中，$\bar{\mathrm{TE}}_i = \frac{1}{N}\sum_{k=1}^{N} \mathrm{TE}_{i,\,k}$ 表示模型 i 的平均跟踪误差。

配对设定考虑了时间依赖性，因为两个模型均在相同的滚动窗口上进行评估。由此得到的 p 值汇总于一个矩阵中，该矩阵展示了竞争模型之间具有统计显著性的支配关系。

4.4 标准普尔500指数特征

我们使用三项描述性统计量——平均收益率、标准差（波动率）以及与指数的相关性——来描述研究期间（2012年10月12日至2022年11月8日）的标准普尔500指数及其成分股。

1. **日均收益率分布：** 图 6 展示了所有成分股平均日收益率的横截面分布。大多数资产集中在适度的正收益水平附近，由于少数高收益离群值的存在，该分布呈右偏。垂直虚线分别表示成分股平均收益率的截面均值（红色）和横截面中位数（蓝色）。中位数低于均值，反映出少数极端离群值将平均值向上拉高；这表明，推动收益分布右尾的股票相对较少，而大多数股票的表现较为温和。

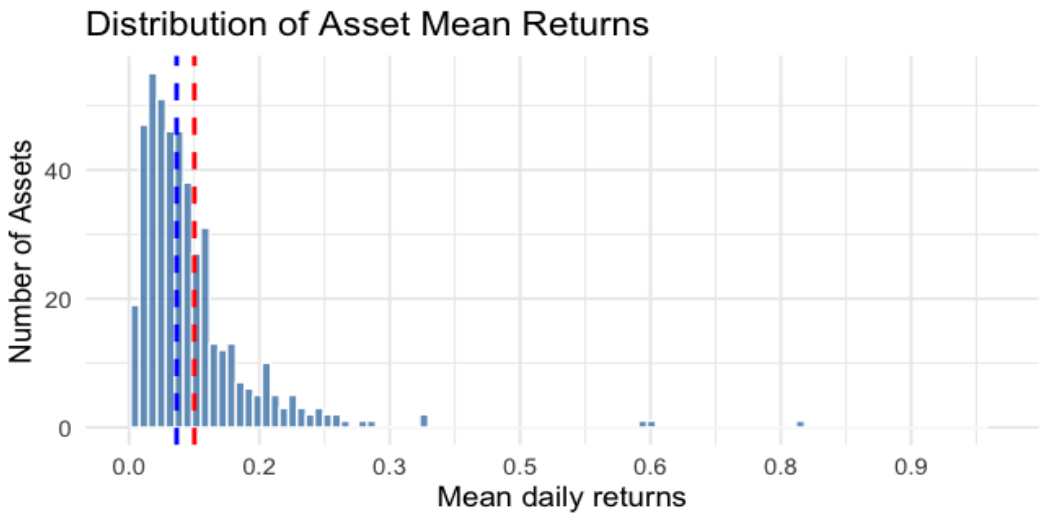


图6：2012年12月10日至2022年8月11日样本期内标准普尔500指数成分股平均日收益率的分布。红色虚线表示成分股平均收益率的截面均值，蓝色虚线表示横截面中位数。

2. **日波动率分布：** 图 7展示了日收益率标准差的分布，该指标用于衡量资产波动率。大多数资产的波动率介于 0.2 至 0.5 之间，而该分布呈右偏，少数波动率极高的股票标准差超过 1。垂直虚线分别表示成分股波动率的截面均值（红色）和中位数（蓝色）。均值位于中位数右侧，凸显了少数极端离群值抬高平均值的影响；相比之下，中位数更能代表典型股票的波动率水平。

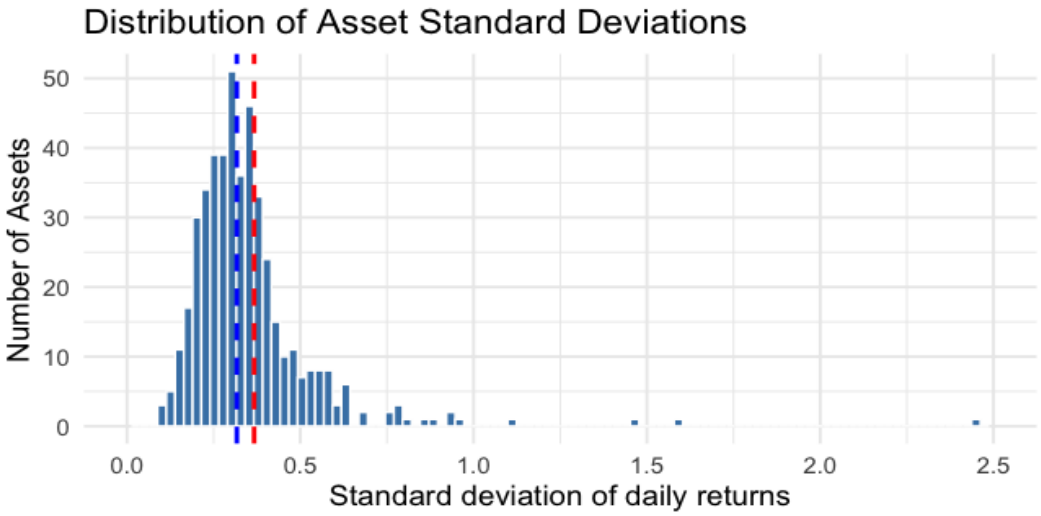


图7：样本期（2012年12月10日至2022年8月11日）标准普尔500指数成分股日收益波动率（标准差）的分布。红色虚线表示成分股波动率的截面均值，蓝色虚线表示成分股波动率的横截面中位数。

3. **资产相关性的分布：** 图 8 展示了各资产收益率与标准普尔500指数收益率之间相关性的分布。大多数成分股的相关性较高，处于0.6至0.9之间，表明它们与市场基准共同变动的程度较强。该分布呈中度左偏，其中有相当数量的资产相关性较弱，甚至为负。垂直虚线分别表示资产相关性的截面均值（红色）和中位数（蓝色）。中位数略低于均值，凸显了部分高度相关资产对均值的拉升作用。

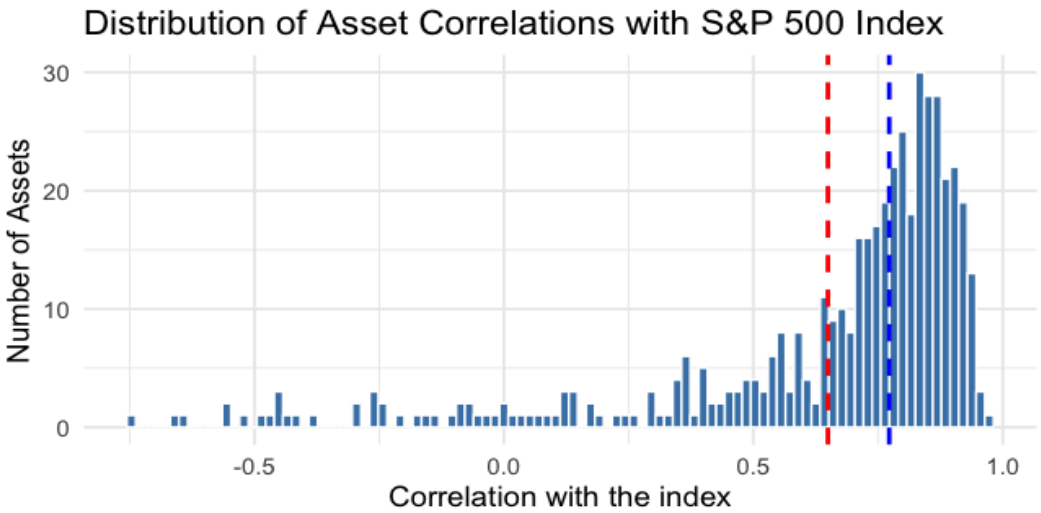


图8：2012年12月10日至2022年8月11日这10年期间，资产与标准普尔500指数的日相关性分布。红色虚线表示平均相关性，蓝色虚线表示资产间相关性的中位数。

4.5 样本外分析

现在，我们转向对比评估指数跟踪模型。分析重点在于样本外表现，并在前文所述的滚动窗口内进行评估。为便于理解，表 6完整列出了所考察的模型，并将其分为三大类：基于优化、统计和数据驱动方法；表 7则列出了模型规格，包括实证研究中考察的参数值及其他设定。

表6：研究中评估的指数跟踪模型分为三类：基于优化、统计和数据驱动模型。该分类体系为比较性样本外分析提供了框架。

优化模型	统计模型	数据驱动模型
1. 均方误差（MSE）	1. 最小二乘回归（LSR）	1. 聚类模型1（Clust 1）
2. 误差平方和（SES）	2. 分位数回归（QR）	2. 聚类模型2（Clust 2）
3. 平均绝对偏差（MAD）	3. 回归 +非负 LASSO（NNL）	3. 稀疏 ϵ-支持向量回归模型（ϵ-SVR）
4. 平均绝对下行偏差（MADD）	4. 回归 + 弹性网络（NNEN）	4. 稀疏 ν-支持向量回归模型（ν-支持向量回归）
5. MinMax	5. 协整 + 模拟（Coint-Sim）	5. RF 聚类 +岭回归（RF-Clust）
6. 下行 MinMax（DMinMax）	6. 凸协整（Cvx-CoInt）	6. 随机森林回归 + 岭回归（RF-Reg）
7. 跟踪误差方差（TEV）	7. 因子模型（FBM）	7. 深度自编码器（六种变体：SH-AE、Con-AE、DEN-AE、SP-AE、STCK-AE、VAR-AE）
8. 尾部混合条件风险价值（TMCVaR）		8. 深度神经网络 (DNN)
		9. 固定噪声深度神经网络 (DNNF)

表7：本研究评估的指数跟踪模型及其规格

模型名称	模型规格
MSE	基数约束（资产数量），$K = 45$
SES	基数约束（资产数量），$K = 45$
MAD	基数约束（资产数量），$K = 45$
MADD	基数约束（资产数量），$K = 45$
MinMax	基数约束（资产数量），$K = 45$
DMinMax	基数约束（资产数量），$K = 45$
TEV	$K = 45$，基准指数权重通过最小二乘回归确定后进行归一化
TMCVaR	$K = 45$, $\alpha =$ (0.9,0.75,0.5,0.1,0.01), $\delta = 0.5$
LSR	$K = 45$
QR	$K = 45$, $\tau = 0.5$
NNL	基数 $K = 45$，使用二分法得到的正则化参数 λ，初始范围为 0 到 100
NNEN	基数 $K = 45$，
FBM	因子数量，$k = 5$，基数，$K = 45$ 资产
CoInt-Sim	$K = 45$，迭代次数$= 10000$，ADF 检验的显著性水平 $= 0.05$，ADF 检验的最大滞后阶数$= 1$，ADF 检验的自动滞后准则='赤池信息准则'
Cvx-CoInt	基数，$K = 45$；使用 Gurobi 求解凸混合整数二次规划
Clust 1	聚类数 ，$K = 45$
Clust 2	聚类数 ，$K = 45$
ϵ-支持向量回归	$K = 45, C_1 = \{0.1, 1, 10, 50\}, \epsilon = \{0.001, 0.005, 0.01, 0.05\}$
ν-支持向量回归	$K = 45, C_1 = \{0.1, 1, 10, 50\}, \epsilon = \{0.001, 0.005, 0.01, 0.05\}, C_2 = 1$
RF-Clust	树的数量=100，基数 $K = 45$, $\alpha = [10^{-4}, 10^{-2}]$，交叉验证折数=5
RF-Reg	树的数量=100，基数 $K = 45$, $\alpha = [10^{-4}, 10^{-2}]$，交叉验证折数=5
SH-AE	隐藏层：16 个神经元（ReLU），优化器：Adam，损失函数：均方误差，训练轮数：50，批次大小：32, $K = 45$
SP-AE	隐藏层：16 个神经元（ReLU），稀疏性惩罚 $\lambda = 10^{-4}$ (L1)，训练轮数：50，批次大小：32, $K = 45$
CON-AE	隐藏层：16 个神经元（ReLU），收缩惩罚 $\lambda = 10^{-4}$，训练轮数：50，批次大小：32，$K = 45$
STCK-AE	编码器：$[64, 32, 16]$, 解码器：$[32, 64]$, 激活函数：ReLU，训练轮数：50，批次大小：32，$K = 45$
DEN-AE	隐藏层：16 个神经元（ReLU），高斯噪声（0.1），训练轮数：50，批次大小：32，$K = 45$
VAR-AE	潜在维度：16，KL 正则化 $\lambda_{KL} = 10^{-4}$，训练轮数：50，批次大小：32，$K = 45$
DNN	通过堆叠自编码器进行资产选择（编码器：64,32,16；解码器：32,64；激活函数：ReLU）；优化器：Adam；损失函数：均方误差；训练轮数：100；批次大小：32；深度神经网络（层数：64,32,1；激活函数：ReLU）；优化器：Adam（学习率 $= 0.01$）；损失函数：均方误差；训练轮数：100；批次大小：16
DNNF	$K = 45$，隐藏层=6，每层神经元数=64，激活函数=ReLU，丢弃率=0.5，优化器=Adam（学习率=0.01），训练轮数=100，批次大小=16，输出=Softmax归一化

需要注意的是，为确保比较公平，每种机器学习模型均采用其基于经验确定的最优配置进行训练，以确保训练损失函数收敛。由于各模型的学习目标和收敛特性因架构而异，因此各模型的训练轮数和批次大小无需完全一致。不过，按照第4.2节所述，所有模型均在相同的滚动窗口上进行评估，采用相同的基数约束和一致的评估指标。对于所有优化模型，$\epsilon_i 0$ 和 δ_i 的值分别对所有资产设为 0 和 1 i。

表8：基于优化的指数跟踪投资组合的样本外表现。面板I报告各模型的表现指标；面板II报告针对跟踪误差进行单侧配对 t检验所得的成对 p值。最佳值以**粗体**显示，最差值以斜体显示。

面板I：样本外表现									
模型	MSE	SES	MAD	MADD	MinMax	DMinMax	TEV	TMCVaR	标准普尔500指数
跟踪表现									
跟踪误差	0.00146	0.01490	0.00146	0.00150	0.00145	0.00148	**0.00142**	*0.00159*	0.00000
相关性	99.19%	*69.21%*	99.19%	99.18%	99.20%	99.18%	**99.25%**	99.09%	100%
收益表现									
平均收益	0.00050	**0.00124**	0.00045	0.00047	0.00048	0.00046	*0.00044*	*0.00044*	0.00042
最低收益	-0.11905	*-0.17794*	-0.12037	-0.12355	-0.12187	**-0.11693**	-0.12619	-0.12235	-0.11984
最高收益	0.10101	**0.18139**	0.09855	0.10533	0.09841	*0.09811*	0.10090	0.10920	0.09383
风险表现									
标准差	0.01147	*0.02032*	**0.01143**	0.01168	0.01154	0.01154	0.01162	0.01172	0.01142
平均回撤	**0.03496**	*0.08524*	0.03618	0.03586	0.03584	0.03562	0.03675	0.03735	0.03835
风险调整后绩效									
夏普比率	0.04316	**0.06084**	0.03944	0.04054	0.04175	0.04004	0.03816	*0.03791*	0.03650
索提诺比率	0.05872	**0.08670**	0.05349	0.05525	0.05689	0.05435	*0.05179*	0.05191	0.04931
特雷诺比率	0.00050	**0.00100**	0.00045	0.00047	0.00048	0.00046	*0.00044*	*0.00044*	0.00042
信息比率	0.05365	**0.05499**	0.02348	0.03779	0.04484	0.03064	0.01875	*0.01742*	–
换手率和时间复杂度									
换手率（TR）	1.15647	*1.91312*	1.18348	**1.13976**	1.21212	1.23969	1.18920	1.31755	–
资产数量	45	2	45	45	45	45	45	45	–
求解器时间	17.52小时	1.81秒	17.58小时	17.52小时	17.76小时	18.04小时	18.01小时	18.22小时	–

面板II：跟踪误差 t 检验的成对 p值								
p值	MSE	SES	MAD	MADD	MinMax	DMinMax	TEV	TMCVaR
MSE	******	1.00000	0.55610	0.95268	0.68441	0.67954	0.06967	0.89609
SES	1.4E-16	******	2.3E-16	1.8E-16	1.6E-16	1.4E-16	1.3E-16	1.4E-16
MAD	0.44390	1.00000	******	0.92645	0.62114	0.64187	0.06082	0.89603
MADD	0.04732	1.00000	0.07355	******	0.09033	0.12427	0.00278	0.61012
MinMax	0.31559	1.00000	0.37886	0.90967	******	0.55783	0.06581	0.85050
DMinMax	0.32046	1.00000	0.35813	0.87573	0.44217	******	0.06129	0.83036
TEV	0.93033	1.00000	0.93918	0.99722	0.93419	0.93871	******	0.98849
TMCVaR	0.10391	1.00000	0.10397	0.38988	0.14950	0.16964	0.01151	******

4.5.1 基于优化的模型的绩效评估

表 8 报告了八种基于优化的指数跟踪投资组合的样本外表现。为清晰起见，下文按前文提及的绩效标准归类并重点讨论关键结果。

1. **跟踪表现：**

   • 除 SES 外，所有模型的跟踪误差均较小，处于 0.0014–0.0016 范围内。TEV模型的跟踪误差最低（0.00142），其次是 MinMax（0.00145）。相比之下，SES表现较差，跟踪误差为 0.01490。

   •与跟踪误差结果一致，大多数模型与标准普尔500指数表现出非常高的相关性（高于 99%）。TEV再次表现最佳（99.25%），MinMax紧随其后。SES明显偏离，其相关性降至 69.21%。

2. **收益表现：**

   • 由于投资组合规模非常小，SES 模型在所有优化模型中实现了最高的平均收益（0.00124）和最高收益（0.18139）。然而，其最低收益也最低（-0.17794），对应着最大的最大损失。与其他模型相比，这种组合反映出其收益特征高度波动且不稳定。

   • 对于其余模型，平均收益紧密聚集在0.00042的指数平均收益附近，取值介于0.00044和0.00050之间。MSE投资组合的平均收益率最高（0.00050），其次是MinMax（0.00048）。同样，最高收益仅在0.098至0.109之间小幅变化，最低收益则介于 -0.116 和 -0.126之间，与基准值（0.0938和 -0.1198）高度接近。这表明，除SES外，基于优化的投资组合仅存在小幅偏差，便能再现基准的收益特征。

- 值得注意的是，TEV投资组合实现了最低的跟踪误差和最高的相关性，其平均收益（0.00044）也非常接近指数均值，进一步印证了其在跟踪和收益两个维度上的一致性。

3. **风险表现：**
   - SES 模型再次显得与众不同，其波动率最高（0.02032），几乎是指数及其他优化投资组合（0.0114–0.0117）的两倍。它的回撤也最大（0.08524）。这些结果与其极为稀疏的资产选择有关（尽管存在基数约束，仍仅选择了两个资产），这限制了分散化并增加了不稳定性。
   - 相比之下，所有其他投资组合的波动率均与指数值 0.01142 相近。MSE 投资组合的波动率最低（0.01147），而 MADD 的波动率最高（0.01168）。平均回撤也始终低于指数基准 0.03835，范围从 0.03496（MSE）到 0.03735（TMCVaR）不等。这体现了基于优化的选择所带来的益处：它能构建分散化投资组合，其下行风险低于持有完整指数的下行风险。
4. **风险调整后绩效：**
   - SES 投资组合在所有风险调整比率上均取得最高值（Sharpe 0.0608、索提诺比率 0.0867、特雷诺比率 0.0010、信息比率 0.0550），这主要得益于其异常高的平均收益率。尽管如此，SES 并非理想的跟踪器：其持仓极度集中（仅两项），导致风险显著偏高，无论是波动率（0.0203）还是回撤（0.0852）均如此，反映出其分散化效果不佳。
   - 在其余较稳定的跟踪器中，MSE 的整体风险调整比率表现最强（Sharpe 0.0432、索提诺比率 0.0587、特雷诺比率 0.00050、信息比率 0.0537），其次是 MinMax；TMCVaR 的表现往往最弱（例如，信息比率为 0.0174）。排除 SES 后，所有优化投资组合的 Sharpe、索提诺比率和特雷诺比率均高于指数（其范围分别为 0.0379–0.0432、0.0518–0.0587 和 0.00044–0.00050，而指数分别为 0.0365、0.0493 和 0.00042），这支持了基于优化的选择在样本外的有效性。在这一组中，TEV 和 TMCVaR 的比率与指数最为接近。

5. **换手率和资产数量：** 除 SES 仅选择两项资产外，所有基于优化的投资组合均满足45项资产的基数约束。这种极端集中度解释了 SES 尽管具有有利的风险调整比率，表现却不稳定。SES 的换手率也最高（1.913），表明其频繁进行再平衡；相比之下，其他模型的换手率维持在约1.14–1.32这一较窄范围内。

6. **计算效率：** 我们根据32个滚动窗口的总运行时间评估基于优化的模型的计算效率。所有优化模型均使用 R 编写，并通过 R 库接口调用 Gurobi 优化器求解，每个滚动窗口统一设置1800秒的时间限制。除 SES 模型外，所有优化模型在每个窗口中均持续运行至达到该时间限制，平均每个窗口的运行时间约为1800秒。这意味着在32个窗口中累计计算耗时约16小时。实际上，由于求解器开销和系统资源分配，观察到的实际运行时间略高，为17–18小时。因此，这些模型在当前资源设置下计算成本较高[3] 。值得注意的是，由于所有底层公式均为凸规划，因此在收敛时，所报告的解均为全局最优解。如果能够使用更先进的计算基础设施，运行时间负担将大幅降低。

7. **统计显著性（面板II）：** 面板II报告了 p值，这些值来自第 t节所述的成对单侧4.3检验，用于评估某个投资组合的跟踪误差是否低于另一个投资组合。若某模型所在列中有许多显著条目，则说明该模型在跟踪精度方面优于其他投资组合。TEV列包含多个显著条目：在5%的显著性水平下，其跟踪误差低于SES、MADD和TMCVaR；在10%的显著性水平下，则低于MSE、平均绝对偏差、MinMax和DMinMax。同时，TEV行没有显示出任何显著的拒绝结果，证实没有其他模型优于TEV。综合来看，这些结果表明，在常用显著性水平下，TEV在基于优化的公式中实现了最稳健的跟踪精度。

综合这些发现，八种基于优化的投资组合之间呈现出明显差异。SES 模型是表现最弱的跟踪器：它仅持有两项资产，分散化程度有限，与指数的相关性仅为

[3]对于基数约束 TMCVaR 模型，每个滚动窗口均在1800秒的时间限制内求解。尽管大多数窗口都得到了基数为45项资产的可行投资组合，但有三个窗口（8、10和28）产生了退化解，所含资产显著较少。单独重新运行这些窗口后，均得到了完整基数的解，这表明批量运行可能因收敛时间受限而提前终止。这凸显了混合整数投资组合模型对求解器资源约束的敏感性，以及某些窗口的数值条件性。

69.21%。其异常高的平均收益率被较高的波动率和回撤所抵消，因此稳定性和可靠性不足。

在其余投资组合中，MSE 在平均收益率和风险调整比率方面总体表现最佳，MinMax 紧随其后。然而，它们在跟踪误差方面的统计优势有限，仅显著优于 SES 和 MADD。

相比之下，TEV 始终占据优势。它实现了最低的跟踪误差、最高的相关性（99.25%），并且在跟踪精度方面的统计表现优于所有其他模型。同时，其收益和风险特征与指数高度一致，使 TEV 成为本次比较中最有效、最可靠的基于优化的指数跟踪模型。

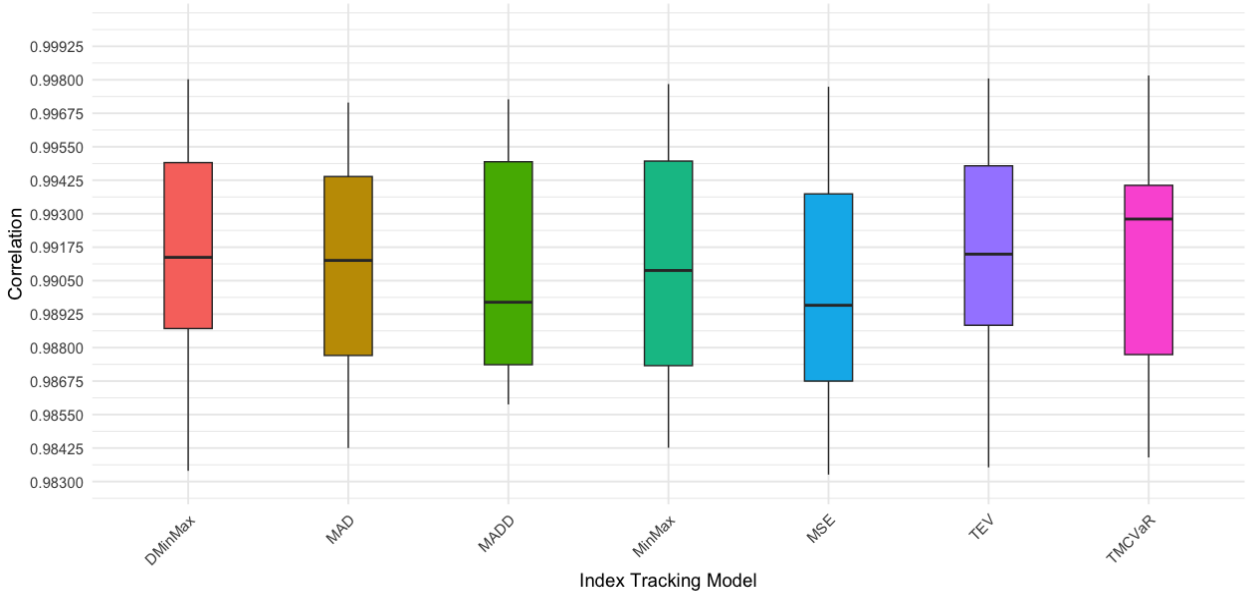


图9：标准普尔500指数与基于优化的跟踪投资组合在32个滚动样本外窗口中的相关性箱线图。

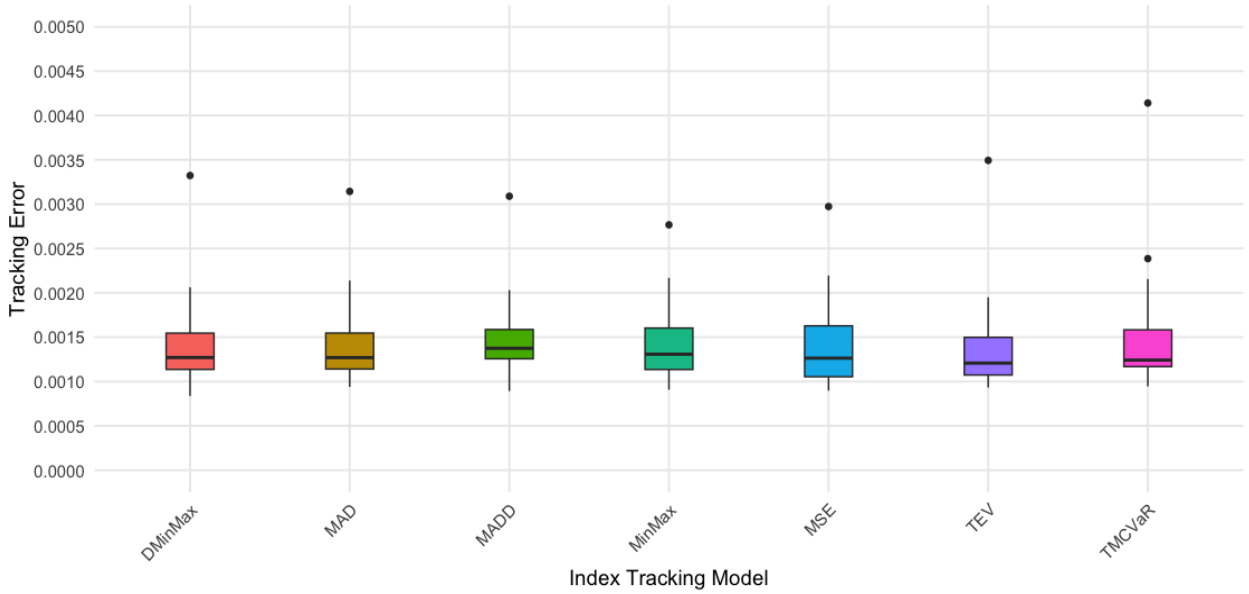


图10：基于优化的指数跟踪投资组合在32个滚动样本外窗口中的跟踪误差箱线图。

表9：基于统计方法的指数跟踪投资组合的样本外表现。面板I报告表现指标，面板II展示基于跟踪误差的单侧配对 t检验所得到的两两 p值。在每个区块中，Lasso 和 Elastic Net 类惩罚项中的最佳值以**粗体**显示，总体最佳值以斜体显示。

面板I：样本外表现

模型	LSR	QR	NNL	NNEN	因子	CoInt-Sim	Cvx-CoInt	指数
跟踪表现								
跟踪误差	*0.01322*	0.01307	0.00303	0.00206	0.00312	0.00246	**0.00148**	0.00000
相关性	*63.00%*	64.70%	97.77%	98.73%	96.75%	98.01%	**99.17%**	100%
收益表现								
平均收益	**0.00069**	0.00105	0.00055	0.00050	0.00053	*0.00047*	0.00049	0.00042
最低收益	*-0.13449*	**-0.09626**	-0.12917	-0.12493	-0.11511	-0.12531	-0.11753	-0.11984
最高收益	**0.18836**	0.16664	0.12265	0.11057	0.10422	0.11434	*0.09713*	0.09383
风险表现								
波动率	0.01700	*0.01713*	0.01301	0.01225	0.01223	0.01214	**0.01148**	0.01142
平均回撤	*0.21365*	0.08094	0.04244	0.03771	0.04687	0.04270	**0.03626**	0.03835
风险调整后绩效								
夏普比率	0.04077	**0.06146**	0.04217	0.04056	0.04297	*0.03836*	0.04226	0.03650
索提诺比率	0.05964	**0.08938**	0.05775	0.05556	0.05808	*0.05205*	0.05740	0.04931
特雷诺比率	0.00074	**0.00108**	0.00049	0.00047	0.00051	*0.00045*	0.00049	0.00042
信息比率	0.02091	**0.04869**	0.04357	0.03889	0.03493	*0.01998*	0.04629	–
换手率与复杂度								
换手率（TR）	1.49575	1.17703	**0.56941**	0.58561	1.04997	*2.02480*	1.12265	–
资产数量	3	3	32	38	38	43	45	–
求解器时间	3.98秒	6.02秒	1.90秒	7.46秒	7.62秒	19.49分钟	18.10小时	–

面板II：跟踪误差 t-检验的两两 p-值

p-值	LSR	QR	NNL	NNEN	因子	CoInt-Sim	Cvx-CoInt
LSR	******	0.64100	1.4E-12	6.3E-13	2.4E-12	3.5E-13	1.2E-13
QR	0.35900	******	4.8E-16	2.0E-16	4.4E-16	8.5E-17	6.0E-17
NNL	1.00000	1.00000	******	4.5E-07	0.81668	5.6E-09	1.6E-10
NNEN	1.00000	1.00000	1.00000	******	1.00000	0.99830	2.4E-10
因子	1.00000	1.00000	0.18332	3.4E-10	******	1.3E-05	5.7E-14
CoInt-Sim	1.00000	1.00000	0.99994	0.00169	0.99998	******	2.8E-10
Cvx-CoInt	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	******

图 9 展示了各基于优化的模型在滚动窗口中与指数相关性的分布情况。由于 SES 的相关性分散在更广的范围内，会影响可视化效果，因此将其排除。对于其余模型，相关性均紧密聚集在 0.99 附近，表明它们具有始终较强的跟踪精度。其中，TMCVaR 的中位数相关性最高，而 TEV 的四分位距最窄，反映出其在各窗口间具有更高的稳定性。TEV 与 TMCVaR 的相关性范围存在重叠；TEV 更稳定，而 TMCVaR 的中位数值略高。相比之下，DMinMax、MAD 和 MSE 等模型的分布范围略宽，且最小值更低，表明它们在某些窗口中的表现较弱。总体而言，尽管所有模型（SES 除外）都取得了稳健的结果，但 TEV 仍是最可靠的跟踪器，TMCVaR 位居第二。

图10 展示了基于优化的模型在滚动窗口中的跟踪误差分布；由于 SES 的误差范围极大，因此未将其纳入。其余所有投资组合的误差均非常低，通常集中在 0.001 至 0.0015 之间，凸显了它们复制基准指数的有效性。其中，TEV 和 MinMax 的中位数误差最低，且分布范围相对较窄，表明其表现出色且稳定。TEV 更为突出，其整个取值范围均低于 MinMax，因此在最小化误差方面是最准确的跟踪器。相比之下，MSE 和 TMCVaR 的波动性较大，偶尔出现较高的异常值，说明它们在不同窗口中的表现不够稳定。MAD、MADD 和 DMinMax 则处于中间水平，跟踪误差较为稳定，但并非表现最佳。总体而言，除 SES 外，所有优化模型都展现出可靠的跟踪精度，其中 TEV 最为可靠，其次是 MinMax，而 MSE 和 TMCVaR 的误差分布则相对更分散。

在考　　察了基于优化的公式后，接下来我们评估基于统计的模型　　s.

4.5.2 基于统计的指数跟踪优化模型的绩效评估

表9 展示了表6中介绍的七种基于统计方法的指数跟踪模型的样本外表现。结果呈现出几个明显的规律。

1. **跟踪表现：**

   • 跟踪误差：凸协整模型（Cvx CoInt）的跟踪误差最低（0.00148），其次是 NNEN 模型（0.00206）。模拟协整的误差也相对较低（0.00246）。相比之下，LSR 和 QR 等基于回归的基准模型的误差明显更高，这反映出它们的分散化程度有限，且复制指数变动的能力较弱。

   • 与指数的相关性：与基于优化的模型不同，后者与基准指数的相关性始终超过 99%；统计组中只有 Cvx CoInt 接近这一水平（99.17%）。NNEN 和 CoInt Sim 紧随其后，相关性分别为 98.73% 和 98.01%。相比之下，LSR 和 QR 的表现仍然不佳，相关性仅为 63–65%，凸显了它们无法可靠地捕捉指数动态。NNL 和因子模型的表现居中，相关性达到 96–98%。

2. **收益表现：**

   • 基于回归的模型（LSR 和 QR）平均收益异常突出，其中 QR 的平均收益在所有模型中最高（0.00105）。然而，这是以分散化不足为代价的，因为这两种模型在各个滚动窗口中都只选择了三种资产。因此，它们呈现出极端结果：LSR 的最高收益最大（0.18836），但最低收益也最低（-0.13449）。

   • 相比之下，其余五种模型的平均收益处于较窄的范围内（≈ 0.00047–0.00055），略高于指数的平均收益（0.00042）。其中，NNL的平均收益最高，因子模型紧随其后。在这个多元化的组别中，NNL的最高收益也最高（0.12265），但同时承受了最严重的损失，表明其结果波动性更大。

   • 值得注意的是，Cvx CoInt不仅展现出最佳的跟踪精度，在收益分布方面也与指数最为接近。其平均收益（0.00049）略高于基准（0.00042），且其最高收益和最低收益均与指数相差无几。这表明，Cvx CoInt投资组合不仅将跟踪误差降至最低，其对基准收益特征的复制也比其他模型更为忠实。

3. **风险表现：**

   • 受分散化程度有限的制约，基于回归的模型（LSR和QR）表现出最高的波动率和回撤。它们集中的资产配置使投资组合价值面临大幅波动，凸显了其作为指数跟踪器的脆弱性。

   • 在分散化投资组合中，NNL 的风险最高，其标准差最高（0.01301），平均回撤也位居第二（0.04244）。

   • 相比之下，基于协整的方法（Cvx CoInt 和 CoInt Sim）以及 NNEN 的风险值最低且最稳定。特别是，Cvx CoInt 不仅表现优于其他方法，也与基准指数最为接近，进一步彰显了其稳定性和可靠性。

4. **风险调整后绩效：**

   • 得益于极高的平均收益，QR 的 Sharpe、索提诺比率和特雷诺比率均为最高。然而，这些高比率具有误导性，因为它们源于极端收益表现，同时伴随着较差的分散化程度和较高的波动率。另一方面，LSR 未能取得有竞争力的比率：相较于较高的风险，其平均收益偏低，因此无法实现有意义的风险调整后收益。

   • 在充分分散化的组别中，因子模型和 Cvx CoInt 在所有风险调整指标上表现最强，NNL 紧随其后。值得注意的是，所有五种分散化跟踪器的 Sharpe、索提诺比率和特雷诺比率均超过基准。这表明，统计框架在设定得当时，能够提供优于直接指数复制的风险调整后绩效。

5. **换手率和资产数量：**尽管目标基数设为45项资产，基于回归的投资组合（LSR 和 QR）却始终只选择2–3项资产。这种有限的分散化推高了它们的换手率，反映出频繁且集中的重新配置。相比之下，其余五种模型分配了32至45项资产，因此交易成本显著较低。其中，NNL

和 NNEN 的换手率数值最为理想。值得注意的是，Cvx CoInt 在协整模型系列中表现突出，不仅换手率相对较低，还能稳定地选择全部45项资产，因此最严格地遵守了基数约束。

6. **计算效率：** 基于回归的模型（LSR 和 QR）以及惩罚回归（NNL 和 NNEN）速度极快，在所有滚动窗口中均能在几秒内得出解。相比之下，协整模型的计算开销要大得多。CoInt Sim 在 32 个窗口中平均需要约 20 分钟才能求得解，而 Cvx CoInt 则经常在每个窗口达到 30 分钟的求解器时间上限。这一差异凸显了一个重要的权衡：协整模型虽然具有更优的跟踪精度和稳定性，但代价是计算成本大幅提高。

7. **统计显著性（面板II）：** p值矩阵显示，Cvx CoInt 的跟踪误差在统计上显著低于几乎所有竞争模型；与基于回归的基准模型（LSR、QR）及其他统计跟踪模型（NNL、NNEN、因子）相比，差异高度显著（$p < 0.05$，且通常为 $p < 0.001$）。至关重要的是，Cvx CoInt 对应的行中没有显著拒绝结果，这证实没有其他模型能够持续优于它。NNEN 是表现第二出色的模型，与大多数跟踪器相比，其跟踪误差（TE）显著改善，但与 Cvx CoInt 相比则没有。最后，CoInt Sim 表现不俗，其误差显著低于若干同类模型，但不及 NNEN 及其凸对应模型。

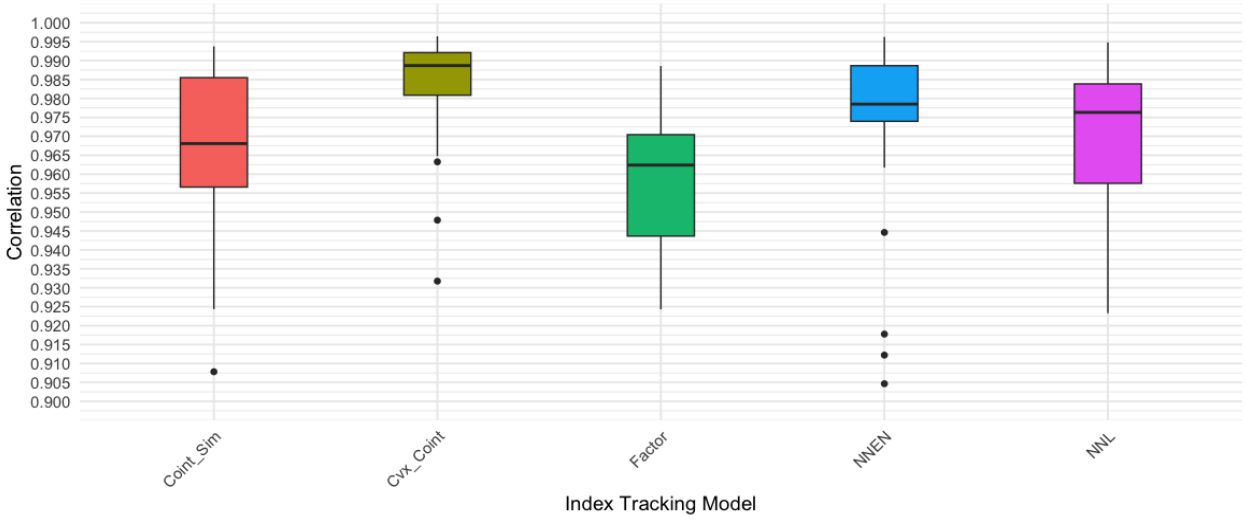


图 11：基于统计的指数跟踪模型与指数相关性的箱线图分布

总体而言，对基于统计的指数跟踪模型的分析凸显了其表现上的明显差异。基于回归的方法（LSR 和 QR）无法构建分散化投资组合，仅配置少数几种资产，因此与基准的相关性仅为 63–65%。在五种分散化程度较高的模型中，基于协整的投资组合成为最有效的跟踪器，其中 Cvx CoInt 的表现始终优于其他模型。NNL 模型虽然实现了最高的平均收益，但代价是波动率和回撤较高，因此作为实用策略的吸引力有所降低。相比之下，Cvx CoInt 和 CoInt Sim 的风险值都较低，其中 Cvx CoInt 在收益和风险度量方面与指数的吻合度最高。统计显著性检验进一步印证了这一优势：Cvx CoInt 的跟踪误差显著低于所有竞争者，没有模型被发现优于它。综合来看，这些发现表明，在统计建模范式下，Cvx CoInt 是指数跟踪基金的最佳候选模型。从投资角度看，这些结果表明，基于协整的策略，尤其是 Cvx CoInt，能够最可靠地复制指数。尽管这类策略需要更多计算资源，但其跟踪优势经过统计验证，是基于回归的方法无法比拟的。

图 11 展示了基于统计的模型与指数之间相关性的分布，其中排除了 LSR 和 QR，以避免它们显著较低的相关性造成失真。在其余跟踪器中，协整模型（CoInt Sim 和 Cvx CoInt）表现出较高的相关性，但稳定性有所不同。Cvx CoInt 展现出最稳健的表现，相关性紧密聚集在 0.99 附近，且与 CoInt Sim 相比具有更窄的四分位距，并且

NNEN 和 NNL 也取得了接近 0.98 的较高中位数相关性，但其波动性略大，并存在少数向下的异常值。相比之下，因子模型呈现出更宽的分布和更低的中位数相关性（约为 0.96），凸显了其较弱且稳定性较差的跟踪能力。总体而言，基于协整的模型在这一类别中提供了最可靠的相关性表现，其中 Cvx CoInt 明显树立了基准。

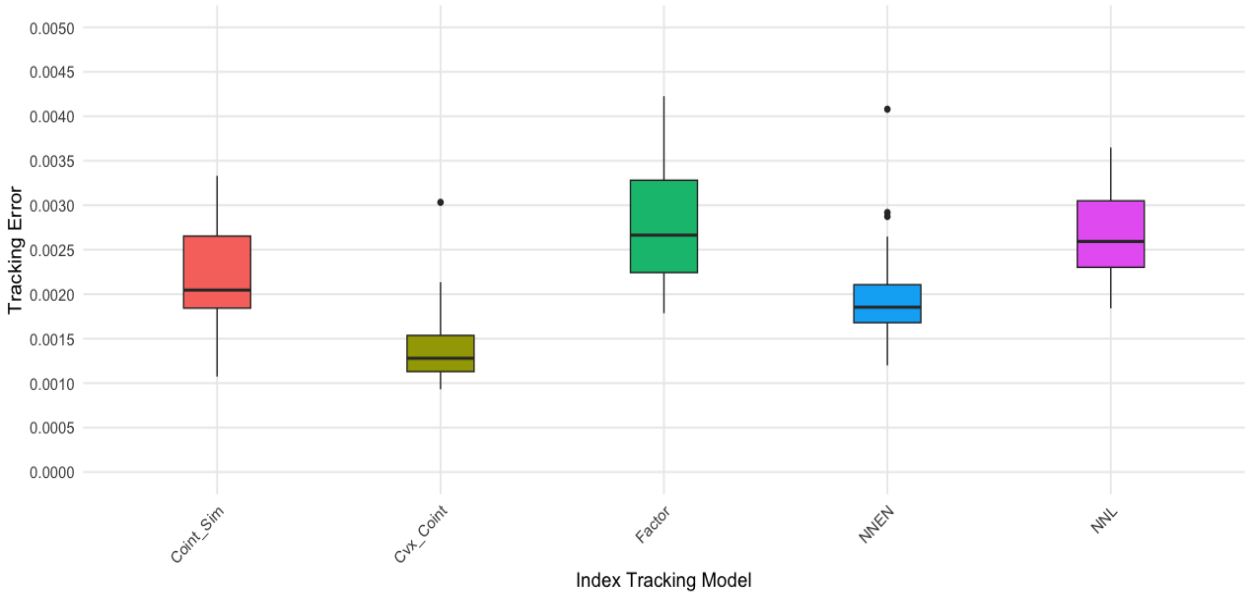


图12：基于统计的指数跟踪模型的跟踪误差箱线图分布

接下来，图12展示了基于统计的模型的跟踪误差分布；同样，为避免其误差范围大得多而造成失真，图中排除了 LSR 和 QR。在其余方法中，Cvx CoInt 的跟踪误差中位数最低，四分位距最窄，凸显了其作为最有效统计模型的可靠性。CoInt Sim 的表现也不错，尽管其变异性略高，但仍能保持较低且稳定的误差水平。相比之下，Factor、NNEN 和 NNL 的中位数更高、离散程度更大，且异常值出现得更频繁，这表明与协整模型相比，它们的准确性较低、稳定性较差。

综合来看，图示证据证实，尽管 CoInt Sim 是颇具竞争力的跟踪器，Cvx CoInt 在所有基于统计的方法中显然实现了最准确、最稳定的跟踪表现。下一小节将转向数据驱动类别，探讨机器学习算法为指数跟踪提供的替代方法。

4.5.3 数据驱动的指数跟踪模型绩效评估

表 10 报告了14种数据驱动的指数跟踪投资组合的样本外表现，涵盖五类方法：聚类（Clust 1、Clust 2）、支持向量回归（ϵ-SVR、 ν-SVR）、随机森林（RF-Clust、RF-Reg）、深度自编码器（SH-AE、Con-AE、DEN-AE、SP-AE、STCK-AE、VAR-AE）和深度学习（DNN、DNNF）。与上一节中的基于回归的模型不同，所有数据驱动方法都能构建分散化程度较高的投资组合，每个时间窗口通常配置40–45项资产。详细的跟踪表现总结如下。

1. **跟踪表现：**在这些模型中，DNNF表现最佳，实现了最低的跟踪误差（0.00153）和最高的与指数的相关性（99.11%），因此成为表现最出色的数据驱动跟踪器。RF-Reg位居第二，其跟踪误差为0.00271，与指数的相关性为98.10%，展现了基于树的集成方法在捕捉非线性关系方面的有效性。

   相比之下，基准深度神经网络（DNN）的表现不及其增强版本，跟踪误差最高（0.00498），相关性最低（90.14%）。基于支持向量回归的模型表现也不佳，跟踪误差相对较高（ϵ-SVR：0.00475； ν-SVR：0.00405），相关性也较弱（分别为 91.60% 和 93.51%）。

Table 10: Out-of-sample performance of data-driven index-tracking portfolios. Panel I reports performance measures; Panel II shows pairwise p-values from one-sided paired t-tests on tracking error. Best values are shown in **bold** and worst in *italics*.

Panel I: Out-of-sample performance

Models	SH-AE	Con-AE	DEN-AE	SP-AE	STCK-AE	VAR-AE	ϵ-SVR	ν-SVR	RF-Clust	RF-Reg	DeepNN	DeepNNF	Clust1	Clust2	Index
Tracking performance															
Tracking error	0.00402	0.00403	0.00406	0.00415	0.00388	0.00380	0.00475	0.00405	0.00288	0.00271	*0.00498*	**0.00153**	0.00271	0.00262	0.00000
Correlation (%)	94.10	93.99	94.12	93.71	94.48	94.70	91.60	93.51	97.22	98.10	*90.14*	**99.11**	97.23	97.39	100.00
Return performance															
Average return	0.00053	0.00053	0.00049	0.00062	0.00057	0.00046	0.00046	0.00047	0.00056	0.00049	0.00040	0.00047	0.00042	0.00045	0.00042
Minimum return	-0.13007	-0.12409	-0.13379	-0.12429	-0.12524	-0.12589	-0.11366	**-0.09710**	*-0.13389*	-0.12335	-0.12952	-0.11643	-0.11955	-0.11149	-0.11984
Maximum return	0.12554	0.12444	**0.12906**	0.12884	0.12376	0.11690	0.12005	*0.09370*	0.10342	0.12204	0.10793	0.10149	0.11475	0.09608	0.09383
Risk performance															
Volatility	0.01186	0.01175	0.01202	0.01187	0.01182	0.01182	0.01171	**0.01057**	0.01216	*0.01277*	0.01086	0.01144	0.01156	0.01152	0.01142
Avg drawdown	0.04079	0.03891	*0.04438*	**0.03478**	0.03780	0.04365	0.04230	0.03528	0.03885	0.04385	0.03782	0.03650	0.03859	0.03870	0.03835
Risk-adjusted performance															
Sharpe ratio	0.04443	0.04522	0.04113	**0.05234**	0.04846	0.03891	0.03960	0.04405	0.04586	0.03838	0.03725	0.04065	*0.03635*	0.03892	0.03650
Sortino ratio	0.06085	0.06192	0.05635	**0.07198**	0.06668	0.05325	0.05429	0.05919	0.06151	0.05257	0.05099	0.05537	*0.04938*	0.05308	0.04931
Treynor ratio	0.00054	0.00055	0.00050	**0.00064**	0.00059	0.00047	0.00049	0.00054	0.00054	0.00045	0.00047	0.00047	*0.00043*	0.00046	0.00042
Information ratio	0.02742	0.02843	0.01910	**0.04925**	0.04022	0.01141	0.00989	0.01215	0.04907	0.02702	*-0.00247*	0.03167	0.00133	0.01214	–
Turnover and complexity															
Turnover (TR)	0.82699	0.85497	0.82795	0.81242	0.85659	0.87958	1.72386	1.68549	1.26397	1.11341	**0.59538**	0.59882	1.77349	*1.85803*	–
No. of assets	45	45	45	45	45	45	45	45	45	45	45	45	42	40	–
Solver time	32.05s	31.79s	32.01s	33.08s	46.67s	43.14s	27.84s	48.91s	34.47s	37.46s	2.20m	51.63m	4.99s	5.02s	–

Panel II: Pairwise p-values of tracking error t-tests

p-values	SH-AE	Con-AE	DEN-AE	SP-AE	STCK-AE	VAR-AE	ϵ-SVR	ν-SVR	RF-Clust	RF-Reg	DeepNN	DeepNNF	Clust1	Clust2
SH-AE	******	0.45387	0.34268	0.84042	0.06520	0.19645	0.94554	0.71912	2.52E-05	1.95E-07	0.99999	4.27E-11	3.90E-07	1.80E-06
Con-AE	0.54613	******	0.37443	0.88885	0.10054	0.21707	0.92788	0.74956	1.29E-04	8.34E-07	0.99999	8.68E-11	8.73E-07	3.03E-06
DEN-AE	0.65732	0.62557	******	0.93842	0.19752	0.30417	0.96628	0.77674	2.14E-04	3.05E-06	0.99999	8.03E-10	3.82E-06	2.37E-05
SP-AE	0.15958	0.11115	0.06158	******	0.02025	0.06822	0.90975	0.55734	5.20E-05	1.09E-06	0.99997	1.64E-10	7.19E-07	3.33E-06
STCK-AE	0.93480	0.89946	0.80248	0.97975	******	0.60735	0.98350	0.92454	3.38E-04	6.74E-07	1.00000	1.59E-10	2.81E-06	1.35E-05
VAR-AE	0.80355	0.78293	0.69583	0.93178	0.39265	******	0.95216	0.89787	2.01E-05	6.30E-07	0.99999	1.70E-13	2.92E-08	1.29E-08
Eps-SVR	0.05446	0.07212	0.03372	0.09025	0.01650	0.04784	******	0.15963	1.16E-04	1.22E-05	0.98727	6.56E-08	2.32E-05	1.06E-04
Nu-SVR	0.28088	0.25044	0.22326	0.44266	0.07546	0.10213	0.84037	******	1.40E-06	2.58E-07	0.99991	7.58E-13	1.15E-09	4.42E-08
RF-Clust	0.99997	0.99987	0.99979	0.99995	0.99966	0.99998	0.99988	0.99999	******	0.04684	1.00000	1.06E-12	0.05163	0.05130
RF-Reg	1.00000	1.00000	0.99999	0.99999	0.99999	0.99999	0.99999	1.00000	0.95316	******	1.00000	1.78E-09	0.63888	0.53653
DeepNN	7.73E-06	5.61E-06	4.58E-06	3.19E-05	5.75E-07	3.38E-06	0.01273	8.53E-05	1.21E-09	5.94E-12	******	7.30E-14	4.14E-12	4.21E-10
DeepNNF	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	******	1.00000	1.00000
Clust1	1.00000	1.00000	0.99999	0.99999	0.99999	1.00000	0.99998	1.00000	0.94837	0.36112	1.00000	6.77E-12	******	0.38878
Clust2	1.00000	1.00000	0.99998	0.99999	0.99999	1.00000	0.99989	1.00000	0.94870	0.46347	1.00000	9.91E-15	0.61122	******

在自编码器系列中，VAR-AE 是表现最佳的变体，其 TE 最低（0.00380），相关性最高（94.70%）；紧随其后的是 STCK-AE（0.00388）。其余自编码器变体的表现相近，相关性集中在 94–95% 区间，较支持向量回归有所改善，但不及 RF-Reg 和 DNNF。

2.**收益表现：**与基于优化和统计方法的投资组合相似，几乎所有数据驱动模型（DNN 除外）的平均收益都高于指数。表现最强的是 SP-AE，其平均收益率最高，为 0.00062；其次是 STCK-AE，为 0.00057。就极端结果而言，基于支持向量回归的模型下行风险相对较低，其中 ϵ-SVR 和 ν-SVR 的最低收益最不严重（分别为 -0.1137 和 -0.0971）。相比之下，RF-Clust 的回撤最深，最低收益为 -0.1339。在上行方面，DEN-AE 的最高收益最高（0.1291），SP-AE 紧随其后（0.1288），这体现了自编码器架构捕捉强劲收益阶段的潜力。

值得注意的是，DNNF——在 TE 和相关性方面表现最佳的跟踪器——与基准收益特征高度吻合，平均收益为 0.00047，最大值为 0.1015，最小值为 -0.1164。Clust 2 是基于聚类的方法之一，在所有收益指标上也与指数高度一致，因此在这一类别中表现稳定，但不够突出。

3. **风险表现：**在数据驱动模型中，ν-SVR 的波动率最低，标准差为 0.01057，而 SP-AE 的平均回撤最小，为 0.03478。深度学习投资组合在风险度量方面也表现良好：DNN 和 DNNF 的标准差分别为 0.01086 和 0.01144，整体排名第二和第三。这两种模型也跻身回撤较低的投资组合之列，其中 DNNF 尤其接近基准的波动率（指数：0.01142）。

相比之下，随机森林模型的波动性更大，其中 RF-Reg 的标准差最高（0.01277），因此是这一组中风险最大的投资组合。基于自编码器的跟踪模型处于中等范围（0.01175–0.0120），其中 Con-AE 的数值较低，DEN-AE 的数值较高。值得注意的是，DEN-AE 不仅在自编码器中表现出最高的标准差（0.01202），回撤也最大，反映出其面临更大的下行风险。

4. **风险调整后绩效：** 总体而言，所有数据驱动模型在各项风险调整指标（Sharpe、索提诺比率、特雷诺比率和信息比率）上均优于指数，凸显了它们创造相对于风险而言更高收益的能力。SP-AE 表现尤为突出，在所有比率上均取得最高值（Sharpe $= 0.0523$、索提诺比率 $= 0.0720$、特雷诺比率 $= 0.00064$、信息比率 $= 0.0493$）。其出色表现得益于较高的平均收益率与相对适中的风险水平相结合；从投资角度看，这使其成为该类别中最具吸引力的模型。

STCK-AE 是最接近的竞争者，在各项比率中始终排名第二。事实上，大多数基于自编码器的模型（VAR-AE 除外）都保持着高于平均水平的比率值，这主要归因于其较高的平均收益率。

在另一端，Clust 1 的表现最弱，Sharpe、索提诺比率和特雷诺比率均为最低。基准深度神经网络的表现也不佳，其平均收益低于指数，导致其信息比率无法定义。相比之下，DNNF 的表现更加均衡：尽管其 Sharpe（0.0407）、索提诺比率（0.0554）和特雷诺比率（0.00047）排名居中，但其信息比率（0.0317）排名第四，证实了其稳定且高度多元化的特征。

5. **换手率与资产数量：** 基于自编码器的投资组合保持相对适中的换手率（≈ 0.81–0.88），反映出其在各窗口中的配置动态较为稳定。相比之下，基于聚类的模型产生了最高的换手成本（Clust $1 =$1.77、Clust $2 =$ 1.86），基于支持向量回归的模型的换手率也偏高（≈ 1.68–1.72）。

深度学习模型（DNN 和 DNNF）的换手率数值最低（≈ 0.60），在全部 14 个投资组合中是成本最高效的策略。值得注意的是，除聚类方法外，所有模型均满足 45 项资产的基数约束；聚类方法配置的股票略少（Clust 1 为 42 只，Clust2 为 40 只）。分散化程度有限且换手率偏高，凸显了基于聚类的跟踪器在实际投资环境中的相对低效。

6. **计算效率：**14 种数据驱动方法的求解器运行时间差异显著。聚类模型速度最快（每个窗口 ≈ 5s），而大多数自编码器、支持向量回归和随机森林模型的求解时间在实际可接受的 27–49s 范围内。相比之下，深度学习模型的计算需求明显更高：DNN 需要 2.2 分钟，而 DNNF 在全部 32 个窗口上的运行时间最长可达 51.6 分钟。这些结果凸显了训练深度架构时的计算权衡：与较浅的机器学习模型或统计学习器相比，深度架构能够提升跟踪精度，但耗时也显著增加。

7. **统计显著性（面板II）：** p值矩阵表明，DNNF 在这一类别中明显占据主导地位：与所有竞争者相比，其列中的数值绝大多数都很小（$p < 10^{-10}$），这证实了

其在跟踪误差方面具有显著优势。在自编码器系列中，SP-AE 相较于若干同类模型取得了统计显著的改进（在多重比较中为 $p < 0.05$ ），进一步凸显了其相对强劲的表现。相比之下，基准深度神经网络在统计上未能胜过任何模型，这与其较弱的实证结果一致。RF-Reg 的表现则喜忧参半：虽然与部分自编码器相比具有竞争力，但仍逊于 DNNF。这表明，尽管集成方法稳健，却无法匹敌采用噪声正则化的深度学习方法的精确性。

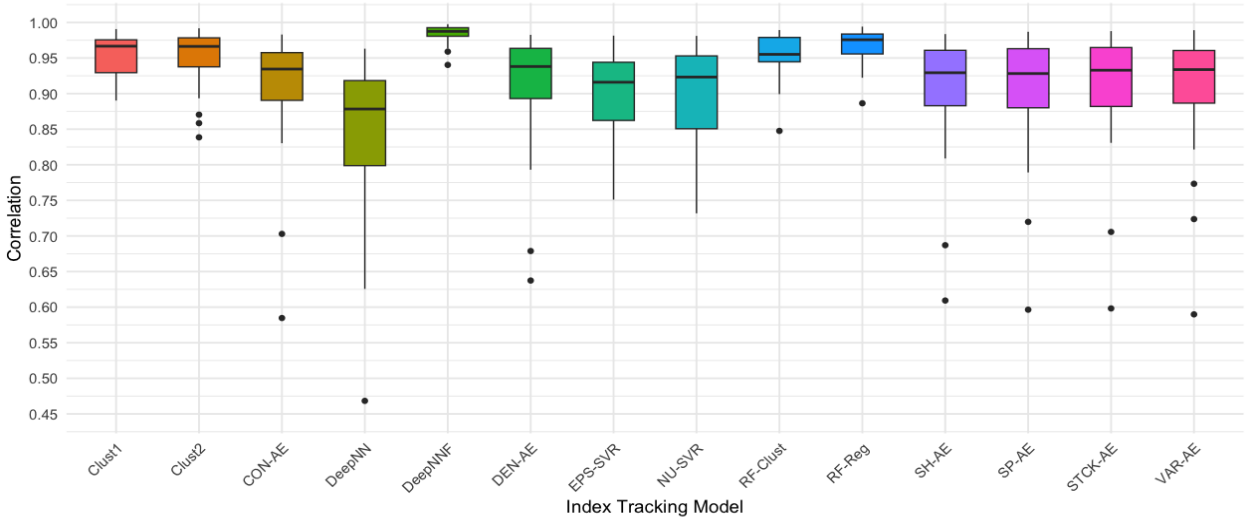


图 13：数据驱动指数跟踪模型与指数相关性的箱线图分布 ls

总体而言，DNNF 是最有效的数据驱动指数跟踪模型，兼具最低的跟踪误差、最高的相关性、较强的统计显著性以及最低的换手成本。SP-AE 在自编码器系列中表现突出，在保持较高跟踪精度的同时，取得了更优的 Sharpe 和索提诺比率，且得到统计验证的支持。 ν- 支持向量回归和 RF-Reg 是颇具竞争力的替代方案，不过二者的换手成本都较高。相比之下，基准深度神经网络在各项指标上始终表现不佳，而且是唯一一个平均收益低于指数的模型。

图 13 展示了 14 种数据驱动模型与指数相关性的分布。DNNF 明显占据优势，其相关性紧密聚集在 0.98 以上，表明其在所有评估窗口中都具有出色的可靠性。基于聚类的方法（Clust 1 和 Clust2）表现同样出色，相关性稳定保持在 0.93–0.97 区间。同样，RF-Reg 和 RF-Clust 的中位数相关性约为 0.95，与聚类方法相当。

自编码器变体（Con-AE、SH-AE、SP-AE、STCK-AE、VAR-AE、DEN-AE）和基于支持向量回归的模型（ϵ-SVR 和 ν-SVR）呈现中等偏高的相关性（0.90-0.95），但其分布比 DNNF 更分散、波动也更大。相比之下，基准深度神经网络的表现最弱，行为极不稳定，偶尔出现低于 0.5 的异常值。综合来看，这些结果表明，DNNF 是最稳健、最可靠的数据驱动跟踪器，其次是基于聚类和随机森林的方法；而较简单的自编码器和支持向量回归则处于中游水平。

图14 展示了数据驱动模型跟踪误差的分布。DNNF 的表现显然优于其他所有模型，其跟踪误差最低且最稳定，数值紧密集中在极小范围内，凸显了它在各评估窗口中的稳健性。基于聚类的模型（Clust 1 和 Clust 2）表现也颇具竞争力，中位数较低，离散程度适中。相比之下，深度神经网络和基于支持向量回归的方法（ϵ-SVR 和 ν-SVR）的中位数误差较高，波动范围大且异常值众多，表明其跟踪精度较弱、可靠性较低。基于自编码器的变体（SH-AE、SP-AE、STCK-AE、VAR-AE、Con-AE、DEN-AE）表现居中，其误差分布尚可接受，但明显逊于 DNNF 和基于聚类的方法。

综合来看，这些结果进一步表明，DNNF 是最准确、最稳定的数据驱动跟踪器，基于聚类的方法紧随其后，而支持向量回归和基准深度神经网络则是这一类别中效果最差的。

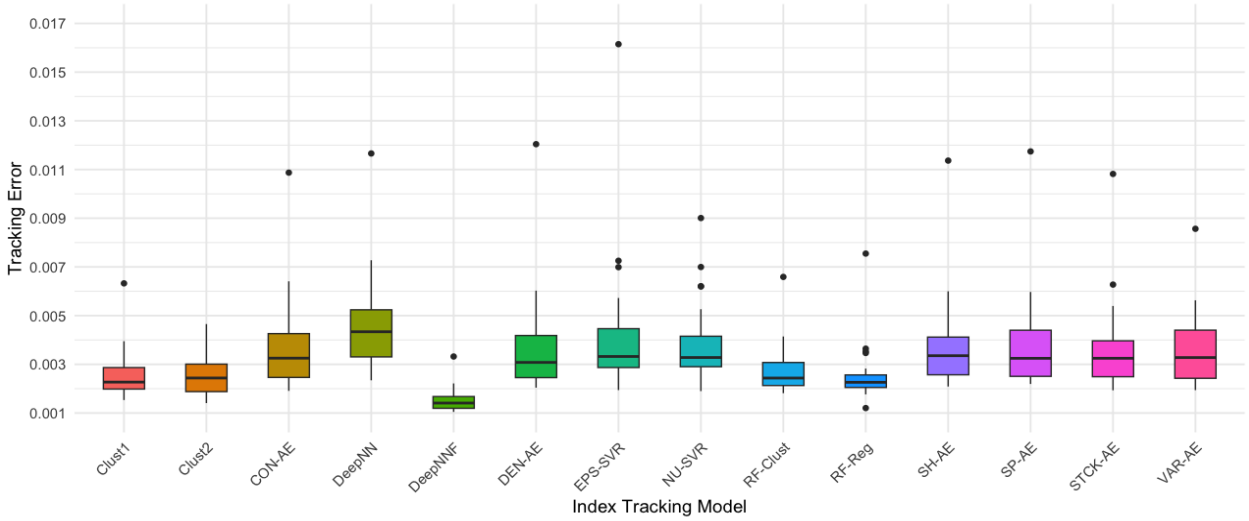


图14</style】：数据驱动的指数跟踪模型样本外跟踪误差的箱线图分布

哪种建模框架通常更适合跟踪标准普尔500指数? 现在，我们从比较的角度考察三种建模范式——基于优化的模型（8 个变体）、基于统计的模型（7 个变体）和数据驱动模型（14 个变体），以评估它们在复制标准普尔500指数方面的相对有效性。

• 跟踪误差及与指数的相关性：除多元化不足（仅配置了 2–3 项资产）的 SES 外，其余基于优化的投资组合均实现了极低的跟踪误差，范围为 0.00142–0.00159（中位数 $= 0.00146$），且相关性极高，为 99.09%–99.25%（中位数 $= 99.09\%$）。基于统计的模型（不包括多元化程度较差的 LSR 和 QR）表现稍弱，跟踪误差值介于 0.00148–0.00303 之间（中位数 $= 0.00206$），相关性为 96.75%–99.17%（中位数 $= 98.64\%$）。数据驱动模型的表现差异最大：虽然表现最佳的 DNNF 达到了 $= 0.00153$ 的跟踪误差和 $= 99.11\%$ 的相关性，但该类别整体的跟踪误差范围更广，为 0.00153–0.00498（中位数 $= 0.00388$），相关性范围为 90.14%–99.11%（中位数 $= 94.3\%$）。这些结果表明，作为一个整体，优化形式能够实现最准确、最稳定的指数复制；统计方法排名第二，而数据驱动方法的表现差异更大，尽管其中也有一些表现突出的模型。

• 信息比率：就风险调整后超额收益而言，优化模型的信息比率介于 0.0174–0.0550 之间（中位数为 $= 0.0342$），而统计模型的范围为 0.0209–0.0545（中位数为$= 0.0436$）。相比之下，数据驱动模型的信息比率介于 0.0099–0.0493 之间（中位数为 $= 0.0270$）。虽然最高的信息比率由一个基于优化的模型取得，但统计框架的整体表现最为出色，其中位数最高，且数值分布更为集中。这表明，当目标不仅是紧密复制指数，还包括改善风险调整后收益时，统计形式尤为有效。

• 换手率：作为交易成本的替代指标，换手率进一步凸显了不同范式之间的差异。基于优化的投资组合换手率介于 1.14–1.91 之间（中位数为 $= 1.20$），这意味着需要频繁再平衡，且交易成本较高。基于统计的模型报告的中位数相近，但略低，为 1.18。相比之下，数据驱动模型的成本效率显著更高，换手率介于 0.60–1.86 之间，中位数为 0.87。这些结果凸显了一个根本性的权衡：优化模型实现了最紧密的跟踪，但代价是较高的换手率；而数据驱动模型则在跟踪精度与成本效率之间实现了更好的平衡。

• 计算效率：求解器运行时间呈现出最后一项差异。除 SES 外，所有优化模型的计算开销都很大，由于基数约束形式较为复杂，每个评估窗口平均需要 17–18 小时。在统计模型中，只有 Cvx CoInt 的计算开销相当（18 小时），其他模型均可在几秒内求解。数据驱动模型总体上效率最高，其中聚类、随机森林、支持向量回归和自编码器方法都能在几秒内完成；不过，深度学习模型（DNN 和 DNNF）的计算开销仍然较大，需要数分钟而非数秒。从实现角度来看，统计方法和大多数机器学习方法实用得多

适用于大规模或高频应用，而如果无法获得充足的计算资源，基于优化的公式可能难以采用。

综合来看，比较证据揭示了三种范式之间明显的权衡：优化模型能够实现最精确的指数复制，但换手率和计算成本较高；统计模型在风险调整后表现方面尤为突出，尤其是信息比率；数据驱动方法则能以较低成本提供可扩展的解决方案，跟踪精度适中。因此，“最佳”框架取决于投资者的优先考量：严格复制、改善收益—风险权衡，或提高效率与可扩展性。下一节将通过对各类别中表现最佳模型的直接比较，进一步凸显这些差异。

4.5.4 三种范式中最佳模型的比较分析

本节根据各建模范式中表现最出色的模型的样本外跟踪能力、收益—风险特征和统计稳健性，对优化、统计和数据驱动这三种范式进行比较。确定的三种模型为：(1) 基于优化的公式中的 **TEV** ，(2) 统计方法中的 **Cvx CoInt**，以及 (3) 数据驱动方法中的 **DNNF** 。

表11：不同建模范式下最佳指数跟踪模型（优化：TEV；统计：Cvx-CoInt；数据驱动：DNNF）与标准普尔500指数的比较。面板I报告绩效指标；面板II展示基于单侧配对 t检验对跟踪误差进行检验所得的成对 p值。模型中的最佳值以**粗体**显示；最差值以斜体显示。

面板I：样本外表现				
模型	TEV	Cvx CoInt	DNNF	标准普尔500指数
跟踪表现				
跟踪误差	**0.00142**	0.00148	*0.00153*	–
相关性（%）	**99.25**	99.17	*99.11*	–
收益表现				
平均收益	0.00044	**0.00049**	0.00047	*0.00042*
最低收益	*-0.12619*	-0.11753	**-0.11643**	-0.11984
最高收益	0.10090	0.09713	**0.10149**	*0.09383*
风险表现				
波动率	*0.01162*	0.01148	0.01144	**0.01142**
平均回撤	0.03675	**0.03626**	0.03650	*0.03835*
风险调整后绩效				
夏普比率	0.03816	**0.04226**	0.04065	*0.03650*
索提诺比率	0.05179	**0.05740**	0.05537	*0.04931*
特雷诺比率	0.00044	**0.00049**	0.00047	*0.00042*
信息比率	0.01875	**0.04629**	0.03167	–
换手率与复杂度				
换手率（TR）	*1.18920*	1.20571	**0.59882**	–
资产数量	45	45	45	–
求解器时间	18.01小时	*18.10*小时	**51.63分钟**	–
面板II：跟踪误差检验的两两 p值 t				
p值	TEV	Cvx CoInt	DNNF	
TEV	******	0.95202	0.99344	
Cvx CoInt	0.04798	******	0.81170	
DNNF	0.00656	0.18829	******	

表11 报告了它们的比较表现。三者均能非常精准地复制标准普尔500指数，相关性超过99%，跟踪误差极低。其中，优化模型 TEV 的基准复制最为准确，跟踪误差最低（0.00142），相关性最高（99.25%）。统计模型 Cvx CoInt 紧随其后，但在收益与风险之间实现了最佳平衡，其 Sharpe、索提诺比率和信息比率均为最高。数据驱动的 DNNF 模型虽然在纯粹的跟踪精度方面稍逊一筹，却最具成本效益且计算效率最高，兼具较低的换手率，求解器运行时间以分钟计而非小时。表11 的下半部分列出了跟踪误差检验的成对 p值。在 5% 的显著性水平下，TEV 在统计上优于 Cvx CoInt（$p = 0.04798$）和 DNNF（$p = 0.00656$）。差异

Cvx CoInt 与 DNNF 之间的差异在统计上并不显著（$p = 0.18829$），这表明在实践中两者的跟踪误差相当。综合来看，TEV 是精度最高的复制模型，Cvx CoInt 的风险调整后表现最强，而 DNNF 在成本和计算效率方面最优。

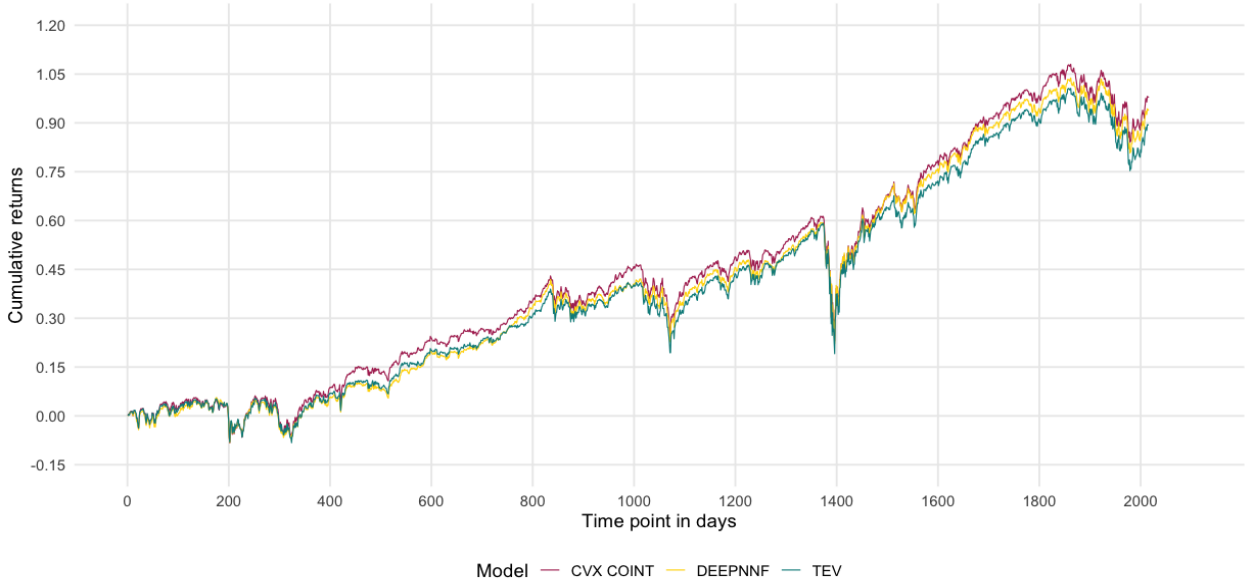


图15：指数跟踪三种建模范式下最佳跟踪投资组合的累计收益率，以及标准普尔500指数

图15 展示了标准普尔500指数与三种基准模型在2019个交易日（2014年12月至2022年8月）内的累计收益率。三种模型均能紧密跟踪指数，走势几乎同步。TEV始终最接近基准，凸显了其出色的复制精度。在市场持续上涨期间，Cvx CoInt的复合收益率始终略高，体现了其更强的风险—收益平衡能力。DNNF的走势与Cvx CoInt几乎相同，但峰值略低，同时换手率最低、计算速度最快。总体而言，这场正面对比证实了每种范式各具优势：TEV在纯粹的跟踪精度方面树立了标杆，Cvx CoInt实现了最有利的收益—风险权衡，而DNNF则展现了现代机器学习如何以无与伦比的成本和计算效率实现有竞争力的准确性。

5 结论

本综述系统地将现有方法归纳为三种范式，对指数跟踪领域的文献进行了考察：基于优化的模型、统计框架和数据驱动的机器学习方法。每种范式各具优势：基于优化的模型为精确复制基准提供了严谨的公式；统计方法利用资产收益中的结构关系；数据驱动方法则具有灵活性和可扩展性，能够捕捉非线性依赖关系及复杂的市场动态。

除概念性综述外，研究还针对应用于标准普尔500指数的29种代表性模型开展了大规模实证分析。结果显示，不同范式之间存在明显差异。跟踪误差方差（TEV）模型的复制准确性最高，实现了最低的跟踪误差和与指数的最高相关性。凸协整（Cvx CoInt）模型的风险调整后表现最强，体现为更优的 Sharpe、索提诺比率和信息比率。相比之下，固定噪声深度神经网络（DNNF）在保持较强复制准确性的同时，在换手率和计算效率方面具有显著优势。这些结果表明，方法论的最佳选择取决于投资者的目标：是最小化复制误差、最大化收益—风险权衡，还是提升可扩展性。

该综述还指出了若干值得探索的未来研究方向。重要的拓展包括针对不同市场状态进行动态模型选择，以及明确纳入交易成本和流动性

约束，并制定多期再平衡策略。此外，在这一背景下，可解释和稳健的机器学习方法仍鲜有探索，是值得进一步研究的自然方向。

为促进研究结果的可复现性和可比性，我们提供了一个开源代码库，其中包含本文所评述模型的实现以及实验框架。研究人员和从业者可利用这一资源，将新方法直接与成熟方法进行基准比较，从而推动指数跟踪策略的设计与评估不断进步。

数据声明

支持本研究结论的数据来源于彭博数据流，可在合理请求下向通讯作者获取。分析所用代码（R/Python）已通过项目的 GitHub 代码库公开https://github.com/vrindadhingra/index-tracking-review。

关于在写作过程中使用生成式人工智能和人工智能辅助技术的声明

在准备本文期间，作者使用 GPT-5 润色语言并提升稿件的清晰度。使用该工具后，作者对内容进行了审阅，并根据需要作出编辑，且对出版物内容承担全部责任。

利益披露

作者声明不存在利益冲突。

参考文献

Acosta-Gonz´alez E,Armas-Herrera R,Fern´andez-Rodr´ıguezF (2015)关于指数跟踪与统计套利中通过协整选择股票的研究：选股的作用。Quantitative Finance 15(6):1075 – 1091. https://doi.org/10.1080/14697688.2014.940604 Aguilar M, Custovic A, Huang Z, et al. (2022)利用分位数构建更优的跟踪投资组合。Investment Management and Financial Innovations 19(1):14 – 31. https://doi.org/10.21511/imfi.19(1).2022.02 Anadu K, Mathias K, Patrick M, et al. (2020)从主动投资转向被动投资：对金融稳定的风险？Financial Analysts Journal 76(4):23–39. https://doi.org/10.1080/0015198X.2020.1779498 Andriosopoulos K, Doumpos M, Papapostolou NC, et al. (2013)利用进化算法优化航运股票及货运市场的投资组合并进行指数跟踪。Transportation Research Part E: Logistics and Transportation Review 52:16 – 34. https://doi.org/10.1016/j.tre.2012.11.006, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-84872076455&doi=10.1016%2fj.tre.2012.11. 006&partnerID=40&md5=0e472a368373f02a5bf8fe47e8f306ff Anis HT, Costa G, Kwon RH (2023)基于风险分配的指数跟踪。Computers and Operations Research 154. https://doi.org/10.1016/j.cor.2023.106219 Ayll´on-Gavil´an R, Mart´ınez-Estudillo FJ, Guijo-Rubio D, et al. (2024)序数决策树的划分准则：一项实验研究。arXiv:2412.13697 Ay´on JAD, de Lourdes Sandoval Solis M, Vel´azquez RG, et al. (2024)用于指数跟踪问题的和声搜索元启发式算法。载于：Workshop on Engineering Applications，Springer，第15–26页，https: //doi.org/10.1007/978-3-031-74598-0 2 Beasley J, Meade N, Chang TJ (2003)一种用于指数跟踪问题的进化启发式算法。European Journal of Operational Research 148(3):621 – 643. https://doi.org/10.1016/S0377-2217(02)00425-3 Benidis K, Feng Y, Palomar DP (2018)用于高维金融指数跟踪的稀疏投资组合。IEEE Transactions on Signal Processing 66(1):155 – 170. https://doi.org/10.1109/TSP.2017.2762286 Bogle JC (2011)指数基金如何诞生。华尔街日报—东部版 258:A15 Boyd S, Vandenberghe L (2004)《凸优化》。剑桥大学出版社，英国剑桥 Bradrania R, Pirayesh Neghab D, Shafizadeh M (2022)指数跟踪中的状态依赖型股票选择：一种机器学习方法。Financial Markets and Portfolio Management 36(1):1–28 Canakgoz N, Beasley J (2009)用于指数跟踪和增强型指数化的混合整数规划方法。European Journal of Operational Research 196(1):384 – 399. https://doi.org/10.1016/j.ejor.2008.03.015

Cao Y, Li H, Yang Y (2022)结合随机森林与多重共线性建模进行指数跟踪。Communications in Statistics: Simulation and Computation https://doi.org/10.1080/03610918.2022. 2116050 Cesarone F, Paolo AD, Bufalo M, et al. (2025)一种用于大规模投资范围指数跟踪的基准资产主成分因子分解方法。Finance Research Letters 79:107,244. https://doi.org/10.1016/j.frl. 2025.107244 Chen C, Kwon RH (2012)用于指数跟踪的稳健投资组合选择。Computers and Operations Research 39(4):829 – 837. https://doi.org/10.1016/j.cor.2010.08.019 Corielli F, Marcellino M (2006)基于因子的指数跟踪。《银行与金融杂志》30(8):2215 – 2233. https://doi.org/10.1016/j.jbankfin.2005.07.012 Dai Z, Li L (2024)用于增强型指数跟踪的深度学习。Quantitative Finance 24(5):569–591 De Leone R (2011)用于时间序列分析的支持向量回归。载于：《运筹学会议录》2010，Springer，柏林，海德堡，https://doi.org/10.1007/978-3-642-20009-0 6, URL https://doi.org/10. 1007/978-3-642-20009-0 6 Derigs U, Nickel NH (2003)基于元启发式算法的投资组合优化决策支持：被动投资组合管理中跟踪误差最小化的案例研究。《运筹学谱》25(3):345 – 378. https://doi.org/10.1007/s00291-003-0127-5 Dose C, Cincotti S (2005)金融时间序列聚类及其在指数跟踪和增强型指数跟踪投资组合中的应用。《物理学A：统计力学及其应用》355(1):145 – 151. https://doi.org/ 10.1016/j.physa.2005.02.078 Engle RF, Granger CWJ (1987)协整与误差修正：表示、估计与检验。《计量经济学》55(2):251–276. https://doi.org/10.2307/1913236 Fama EF (1970)有效资本市场：理论与实证研究综述。《金融学杂志》25(2):383–417 Fastrich B, Paterlini S, Winker P (2014)指数跟踪中的基数约束与q-范数约束。Quantitative Finance 14(11):2019 – 2032. https://doi.org/10.1080/14697688.2012.691986 Focardi SM, Fabozzi FJ (2004)一种基于时间序列聚类的指数跟踪方法。Quantitative Finance 4(4):417 – 425. https://doi.org/10.1080/14697680400008668 Frino A, Gallagher DR (2001)标准普尔500指数基金的跟踪。《投资组合管理杂志》28(1):44 – 55. https://doi.org/10.3905/jpm.2001.319822 Gaivoronski AA, Krylov S, Van Der Wijst N (2005)最优投资组合选择与动态基准跟踪。European Journal of Operational Research 163(1):115 – 131. https://doi.org/10.1016/j.ejor.2003. 12.001 Goel A, Sharma A, Mehra A (2018)使用混合条件风险价值进行指数跟踪与增强型指数化。Journal of Computational and Applied Mathematics 335:361 – 380. https://doi.org/10.1016/j.cam. 2017.12.015 Goel A, Filipovi´c D, Pasricha P (2024)基于拓扑数据分析聚类的稀疏投资组合选择。URL https://arxiv.org/abs/2401.16920, 2401.16920 Goel A, Filipovi´c D, Pasricha P (2025)基于拓扑数据分析聚类的稀疏投资组合选择。Quantitative Finance 25(8):1261–1291. https://doi.org/10.1080/14697688.2025.2544762 Goel A, Pasricha P, Kanniainen J (2026)风险降低型稀疏指数跟踪投资组合：一种拓扑数据分析方法。《Omega》138:103,432. https://doi.org/10.1016/j.omega.2025.103432 Goodell JW, Kumar S, Lim WM, et al. (2021)金融领域的人工智能与机器学习：通过文献计量分析识别基础、主题和研究簇。《行为与实验金融杂志》32:100,577. https://doi.org/https://doi.org/10.1016/j.jbef.2021.100577 Granger CWJ (1981)时间序列数据的若干性质及其在计量经济模型设定中的应用。《计量经济学杂志》16(1):121–130. https://doi.org/10.1016/0304-4076(81)90079-8 Grassetti F (2025)优化指数跟踪：一种基于随机矩阵理论的投资组合选择方法。《物理学A：统计力学及其应用》674:130,747. https://doi.org/10.1016/j.physa.2025. 130747 Guastaroba G, Speranza M (2012)核搜索：在指数跟踪问题中的应用。European Journal of Operational Research 217(1):54 – 68. https://doi.org/10.1016/j.ejor.2011.09.004 Jensen MC (1968)共同基金在1945–1964年期间的表现。《金融学杂志》23(2) Kim S, Kim S (2020)通过深度潜在表征学习进行指数跟踪。Quantitative Finance 20(4):639 – 652. https://doi.org/10.1080/14697688.2019.1683599 Konno H, Watanabe H (1996)债券投资组合优化问题及其在指数跟踪中的应用：一种部分优化方法。《日本运筹学会杂志》39(3):295 – 306. https://doi.org/10.15807/jorsj.39.295

Krink T、Mittnik S、Paterlini S（2009）差分进化与组合搜索在受约束指数跟踪中的应用。《运筹学年鉴》172(1)：153–176。https://doi.org/10.1007/s10479-009-0552-1 Kwak Y、Song J、Lee H（2021）用于指数跟踪投资组合优化的固定噪声神经网络。《专家系统及其应用》183。https://doi.org/10.1016/j.eswa.2021.115298 Kwon RH、Wu D（2017）基于因子的稳健指数跟踪。《优化与工程》18(2)：443–466。https://doi.org/10.1007/s11081-016-9314-5 Larsen GA、Resnick BG（1998）指数化的实证见解：市值规模、分层和加权如何影响跟踪误差。《投资组合管理杂志》25(1)：51 Li N（2020）基于复合分位数回归的高维指数跟踪高效稀疏投资组合。《统计计算与模拟杂志》90(8)：1466–1478。https://doi.org/10.1080/00949655.2020.1731750 Li N、Zhu G、Niu Y 等（2025）通过Huber准则构建用于受约束指数跟踪的快速稳定投资组合。《工程经济学家》70(1-2)：57–71。https://doi.org/10.1080/0013791X.2025.2489362 Malkiel B（2003）被动投资策略与有效市场。《欧洲金融管理》9：1–10。https://doi.org/10.1111/1468-036X.00205 Markowitz H（1952）投资组合选择。《金融学杂志》7(1)：77–91 Mezali H、Beasley J（2013）用于指数跟踪和增强型指数化的分位数回归。《运筹学会志》64(11)：1676–1692。https://doi.org/10.1057/jors.2012.186 Mutunge P、Haugland D（2018）最小化基数约束投资组合的跟踪误差。Computers and Operations Research 90：33–41。https://doi.org/10.1016/j.cor.2017.09.002 Ouyang H、Zhang X、Yan H（2019）基于深度神经网络的指数跟踪。《认知系统研究》57：107–114。https://doi.org/10.1016/j.cogsys.2018.10.022 Peng X、Gong C、He XD（2023a）用于金融指数跟踪的强化学习。arXiv预印本 arXiv:230802820 Peng X、Gong C、He XD（2023b）用于金融指数跟踪的强化学习。网址：https://arxiv.org/abs/2308.02820，2308.02820 Prondzinski D、Miller M（2018）主动投资与被动投资：来自2009—2017年市场的证据。《会计与金融杂志》18(8)：119–143 Rockafellar R、Uryasev S（2002）一般损失分布的条件风险价值。《银行与金融杂志》26(7)：1443–1471。https://doi.org/10.1016/S0378-4266(02)00271-6，网址：https://www.scopus.com/inward/record.uri?eid=2-s2.0-0036076694&doi=10.1016%2fS0378-4266%2802%2900271-6&partnerID=40&md5=2bc00d8a616831f8a9c454c8ef9bbe7c Roll R（1992 18 4 13 22）跟踪误差的均值/方差分析。《投资组合管理杂志》（）：– Alvaro Rubio-Garc´ıa、Fern´andez-Lorenzo S、Garc´ıa-Ripoll JJ 等（2024）混合模拟退火算法求解指数跟踪问题的准确解。《物理学A：统计力学及其应用》639。https://doi.org/10.1016/j.physa.2024.129637 Rudolf M、Wolter HJ、Zimmermann H（1999）用于跟踪误差最小化的线性模型。《银行与金融杂志》23(1)：85–103。https://doi.org/10.1016/S0378-4266(98)00076-4 Ruiz-Torrubiano R、Su´arez A（2009）指数跟踪的混合优化方法。《运筹学年鉴》166(1)：57–71。https://doi.org/10.1007/s10479-008-0404-4 Sant'Anna LR、Filomena TP、Caldeira JF（2017）利用协整、相关性和内生投资组合选择实现指数跟踪与增强型指数化。《经济学与金融季刊评论》65：146–157。https://doi.org/10.1016/j.qref.2016.08.008 Sant'Anna LR、Caldeira JF、Filomena TP（2020a）基于Lasso的指数跟踪与统计套利多空策略。《北美经济与金融杂志》51。https://doi.org/10.1016/j.najef.2019.101055 Sant'Anna LR、de Oliveira AD、Filomena TP 等（2020b）基于协整凸重构求解指数跟踪问题。《金融研究快报》37。https://doi.org/10.1016/j.frl.2019.101356 Sant'Anna LR、Righi MB、M¨uller FM 等（2022）风险度量指数跟踪模型。《国际经济学与金融评论》80：361–383。https://doi.org/10.1016/j.iref.2022.02.032 Sant'Anna LR、Filomena TP、Guedes PC 等（2017）通过结合遗传算法和非线性规划的混合启发式方法，实现受控资产数量的指数跟踪。《运筹学年鉴》258(2)：849–867。https://doi.org/10.1007/s10479-016-2111-x Scozzari A、Tardella F、Paterlini S 等（2013）具有UCITS约束的指数跟踪问题的精确方法和启发式方法。《运筹学年鉴》205(1)：235–250。https://doi.org/10.1007/s10479-012-1207-1

Sharpe WF（1964）资本资产价格：风险条件下市场均衡理论。《金融学杂志》19(3)：425–442 Silva JCS、de Almeida Filho AT（2024）关于指数跟踪问题求解方法的系统性文献综述。IMA管理数学期刊 35(2)：163–196。https://doi.org/10.1093/imaman/dpad007 Strub O、Baumann P（2018）指数跟踪投资组合的最优构建与再平衡。《欧洲运筹学杂志》264(1)：370–387。 https://doi.org/10.1016/j.ejor.2017.06.055，网址：https://www.scopus.com/inward/record.uri?eid=2-s2.0-85025479826&doi=10.1016%2fj.ejor.2017.06.055&partnerID=40&md5=857 faa5a95faada118f9f79319c58492 Takeda A、Gotoh JY、Sugiama M（2010）作为条件风险价值最小化方法的支持向量回归及其在金融时间序列分析中的应用。第118–123页， https://doi.org/10.1109/MLSP.2010.5589245 Takeda A、Niranjan M、Gotoh Jy 等（2013）在指数跟踪投资组合中同时追求样本外表现与稀疏性。计算管理科学10(1)：21–49。 https://doi.org/10.1007/s10287-012-0158-y Tang R、Li P（2014）使用软子空间聚类方法的指数优化复制算法。载于：2014年IEEE第七届联合国际信息技术与人工智能会议（ITAIC 2014），第414–418页， https://doi.org/10.1109/ITAIC.2014.7065082 Teng Y、Yang L、Yuan K 等（2017）使用稀疏支持向量回归进行指数跟踪。《计算机科学讲义》（含《人工智能讲义》和《生物信息学讲义》子系列）10559 LNCS：293–315。https://doi.org/10.1007/978-3-319-67777-4 26 Wang M、Xu C、Xu F 等（2012）具有条件风险价值风险约束的指数跟踪问题混合0-1线性规划。《运筹学年鉴》196(1)：591–609。 https://doi.org/10.1007/s10479-011-1042-9 Wang M、Xu F、Dai YH（2018）采用分层抽样和最优分配的指数跟踪模型。《商业与工业应用随机模型》34(2)：144–157。https://doi.org/10.1002/asmb.2287 Wang YJ、Wu LH、Wu LC（2024）深度学习框架下用于指数跟踪投资组合构建与预测的综合提取方法。《超级计算杂志》80(2)：2047–2066。https://doi.org/10.1007/s11227-023-05538-z Wu D、Kwon RH、Costa G（2017）用于跟踪标准普尔500指数的基于聚类的约束方法。《国际生产经济学杂志》193：222–243。https://doi.org/10.1016/j.ijpe.2017.07.018 Wu L、Yang Y（2014）非负弹性网络及其在指数跟踪中的应用。《应用数学与计算》227：541–552。https://doi.org/10.1016/j.amc.2013.11.049 Wu L、Yang Y、Liu H（2014a）非负Lasso及其在指数跟踪中的应用。《计算统计与数据分析》70：116–126。 https://doi.org/10.1016/j.csda.2013.08.012 Wu L、Yang Y、Liu H（2014b）非负Lasso及其在指数跟踪中的应用。《计算统计与数据分析》70：116–126。https://doi.org/10.1016/j.csda.2013.08.012 Xu F、Lu Z、Xu Z（2016）基数约束指数跟踪问题的高效优化方法。《优化方法与软件》31(2)：258–271。 https://doi.org/10.1080/10556788.2015.1062891 Xu F、Li B、Ma J 等（2026）利用资产依赖结构进行基于网络的指数跟踪。《运筹学国际汇刊》33：1498–1524。https://doi.org/doi.org/10.1111/itor.70100 Yuanyuan Cao HL、Yang Y（2022）结合随机森林与多重共线性建模进行指数跟踪。《统计通讯：模拟与计算》0(0)：1–12。https://doi.org/10.1080/03610918.2022.2116050 Zapata Quimbayo C、Moreno Trujillo J（2025）基于范数约束和正则化的指数跟踪。载于：《工程应用计算机科学》。WEA 2024。《计算机与信息科学通讯》第2222卷。Springer，Cham，https://doi.org/10.1007/978-3-031-74598-0 7 Zhang C、Liang S、Lyu F 等（2020）使用自编码器进行股票指数跟踪优化。《物理学前沿》8。 https://doi.org/10.3389/fphy.2020.00388 Zhang F、Lyu Q、Wang J（2025）基于回归的指数跟踪与基于聚类的指数跟踪：一项实证研究。载于：《神经网络进展》—ISNN 2025：第19届国际神经网络研讨会，中国张掖，2025年8月22日至24日，会议论文集，第51–62页，https://doi.org/10.1007/978-981-95-1233-1 5，网址：https://doi.org/10.1007/978-981-95-1233-1 5 Zhang R、Li H、Wang J（2021）基于动态时间规整和约束k中心点聚类的指数跟踪。载于：第11届智能控制与信息处理国际会议（ICICIP 2021），第352–359页，https://doi.org/10.1109/ICICIP53388.2021.9642192 Zhang Y、De Smedt J（2024）利用沙普利加性解释和一维逐点卷积自编码器进行指数跟踪。《国际金融分析评论》95：103,487

Zheng Y, Chen B, Hospedales TM 等（2020a）基数约束下的指数跟踪：一种随机神经网络方法。载于：AAAI 2020——第34届人工智能会议，页码 1242–1249 Zheng Y, Chen B, Hospedales TM 等（2020b）基数约束下的指数跟踪：一种随机神经网络方法。载于：AAAI 人工智能会议论文集，页码 1242–1249 Zhu H, Chen Y, Wang K（2010年）一种用于指数跟踪问题的粒子群优化启发式算法。《计算机科学讲义》（含《人工智能讲义》和《生物信息学讲义》子系列）6063 LNCS（第1部分）：238–245。https://doi.org/10.1007/978-3-642-13278-031，网址 https://www.scopus.com/inward/record.uri?eid=2-s2.0-77954428629&doi=10.1007%2f978-3-642-13278-0 31&partnerID=40&md5=3529b1d0030d487aab7e36b258d4fdff