



中国股票风险因子模型 白皮书



清华大学 全球证券市场研究院
Institute for Global Securities Market, Tsinghua University



Institute for Interdisciplinary
Information Core Technology
交叉信息核心技术研发院



财富引擎科技
www.WealthEngine.cn

中国股票风险因子模型

清华大学全球证券研究院

摘要：

近年来，基于因子体系的建模方法在宏观经济及多种资产大类（股票、信用债、大宗商品、外汇、利率以及衍生品）受到了各国学者的重视。该领域的研究也取得了长足的进步和丰厚的成果，因子体系逐渐成为当前头部金融机构精细化管理投资流程，风险控制以及投后归因的主流工具。为助力国内金融基础设施建设，我们对我国的金融市场进行了实证研究，改良了传统因子体系构建方法。本文旨在探索中国股票风险因子模型体系的构建，深度结合中国金融市场环境实际，逐步校验并调整因子模型的假设。相比于传统的股票市场因子体系，本文构建的中国股票风险因子模型摒弃了市值加权的回归分析方法，使用更为稳健的换手加权 EM 算法估测，并以此为基础，在申万行业因子之外臻选出 12 个具有强解释力且直观稳定的风格因子。我们使用 2012-2022 年中国 A 股上市公司股票的实际数据，使用模型进行实际风险控制校验，该因子体系相比于传统因子体系在控制波动率以及尾部风险均有显著提升，有助于制定更精准的风险控制与投资决策。

关键词：因子模型、EM 算法、风险

1. 因子的基本概念

因子（Factor）的概念源于 CAPM 理论，该理论描述了在无交易成本、投资者理性决策的假设下，资产的预期超额收益和市场预期超额收益线性相关的关系。CAPM 模型是最简单的因子模型，其中市场超额收益对应的市场组合称之为“市场因子”。以此为契机，S. A. Ross（1976）提出了著名的套利定价理论（Arbitrage Pricing Theory, APT），进一步扩展定义了包含多个因子的市场一般定价模型。

为了进一步介绍因子，本文先介绍更宽泛的概念：

特征（Characteristics）：任何个股级别的数据都是个股的特征。

信号（Signal）：能够对收益有预测能力的特征，称之为信号。

因子（Factor）：经过进一步优选的，满足特定要求的信号，称之为因子。

因子暴露（Factor Exposure / Factor Loading）：它描述个股超额收益对因子组合超额收益的反应。通常在模型中记作 β 。在一个统一的因子模型中，因子暴露与因子之间，存在一一对应的关系。

因子模拟投资组合（Factor-Mimicking Portfolio）：构建一个动态调仓的投资组合，使这个投资组合的收益模仿/跟随目标因子。

在主流的因子分析框架的假设下，因子有两个重要性质：1. 因子是个股联动的重要因素；2. 个股的期望收益及当期收益，完全由因子的预期收益及当期收益贡献。¹

具体来说，APT 认为：

$$E_t[r_{t,t+1}] = \beta_t \cdot E_t[\lambda_{t,t+1}] \quad (1.1)$$

其中， E_t 表示期望算子， $\lambda_{t,t+1}$ 代表因子收益（ $K \times 1$ ）， $r_{t,t+1}$ 代表个股收益（ $N \times 1$ ）， β_t 代表因子暴露（ $N \times K$ ）。

和 CAPM 类似，多因子模型假设资产的预期收益率由公式（1.1）右侧的一系列因子的预期收益率和资产在这些因子上的暴露决定。然而，金融市场本身非常复杂，其中包含大量的仅依靠当前已经列出的因子所

¹ 基于 APT 框架（S. A. Ross, 1976），个股收益和因子之间线性相关，任何没有共同因子暴露的个股的收益都会相互独立。值得注意的是，因子未必有很高的预期收益，甚至预期收益未必是正的。

无法捕捉到的随机变动，这种问题当然可以通过寻找更多的因子来捕捉更多的变动性，但这种方法仍然不可能完全捕捉到金融市场的所有变动性，金融市场上的完全无理由的噪声是普遍存在的，因此实际市场模型，可以通过引入一个定价误差（pricing error）项来描述。因此，（1.1）式通常也可被具体写作：

$$r_{t,t+1} = \alpha_{t,t+1} + \beta_t \cdot \lambda_{t,t+1} + \varepsilon_{t,t+1} \quad (1.2)$$

其中 $\alpha_{t,t+1}$ 为资产的实际预期收益率和多因子模型隐含的预期收益率之间的定价误差， $\varepsilon_{t,t+1}$ 代表个股特异性收益（ $N \times 1$ ）。如果 $\alpha_{t,t+1}$ 显著偏离零，则代表了某个可以通过套利而获得超额收益的机会；这也同时说明由于某些原因，市场对该资产出现错误定价（mispricing），从而导致其实际预期收益率和多因子模型下的预期收益率出现了偏离。²

根据风险和收益，我们可以将因子粗略地分为两类：

- （1）风险因子：因子风险高，收益低。
- （2）Alpha 因子：因子风险低，收益高。

指数化投资重点强调投资组合对于不同风险因子的敏感程度，而 α 水平一般为主动型投资组合所重点考量。

2. 风险模型的意义

金融业务的关键难点在于风险决策。如何系统性的度量与处置风险，进而做出优秀的风险决策，是金融行业面临的核心挑战。在科学的风险建模方法中，因子法是当前国际金融市场上普遍采用的系统性解决方案，因子法建模，具有清晰明了、简洁灵活、容易扩展的特征，在国内、国际金融市场的研究和实践中都有着广泛的应用。

基于因子的投资分析与建模框架，被学术界与业界统一采用。经过几十年的努力，学术界提交了丰富的金融理论模型和计量经济统计工具来计算资产的价值，并发现了广泛接受和接受的价值、盈利能力、动量和低波动性要素。自股权投资基金出现至今，这些学术成果一直为业界

² 这个系统里 α_t 并不是必要的，因为 α_t 可以被视为是一个风险极低收益极高的特殊因子 f_0

所应用。事实上，因子投资的兴起得益于资产定价的学术研究。同时业界的因子投资活动也为已有理论的反复检验和新理论的提出提供了充足的数据，促使学术界不断提出新的研究成果。

这个过程中，因子模型本身，也在不断改进和创新。多因子模型作为替代 CAPM 的主流替代方法之一，已逐渐发展成为股权投资的有效手段。当前国际金融市场上大量的资产管理机构都在使用风险因子模型来管理他们资产组合，同时也培育了以 MSCI³, Axioma⁴, Northfield⁵, SunGard⁶等为代表的第三方通用风险因子模型体系的商业模式。

到目前为止，多因子模型一直应用于各种类型的资产投资中，特别是在股票、债券、商品期货和加密货币领域中有所体现。最新趋势从因子的角度分析了各种类型资产收入的潜在驱动因素和逻辑，并将要素用于类别之间的资产构成。对于因子投资主体来说，了解现有方法有助于更好地适应和接受因子投资的未来发展和变化。

3. 现有风险因子模型介绍

与解释并预测资产收益的收益率模型不同，在因子投资中，风险模型的主要目标是准确预测资产收益的协方差矩阵，为风险控制提供依据。在风险模型中，因子的个数往往远小于资产的个数，因此能达到降维的效果。根据风险模型，我们可以得到资产协方差矩阵与因子协方差矩阵的关系：

$$\Sigma = \beta \Sigma_{\lambda} \beta' + \Sigma_{\varepsilon} \quad (3.1)$$

其中， Σ 为资产的协方差矩阵（N 阶）， Σ_{λ} 为因子协方差矩阵（N

³ MSCI 公司，一家总部位于美国纽约的知名金融公司，提供多样化的市场投资指数服务以及专业化的多资产组合分析工具是其重要的服务模式。2004 年，MSCI 公司收购了 Barra 公司并成立了 MSCI Barra，专业提供资产因子模型体系服务，随后 2010 年，MSCI 公司进一步收购了 RiskMetrics 公司，进一步丰富了其在多资产的收益、风险建模方面的专业化工具体系。

⁴ Axioma 公司总部同样位于纽约，1998 年成立，是一家专业从事投资组合优化决策的软件服务商，投资组合经理们使用相关工具由于组合生成、组合监控、虚拟持仓模拟与风险评估，管理交易成本和持仓再平衡，作为其分析、模拟与决策的基础，Axioma 构建了相当全面的风险因子体系。

⁵ Northfield 公司成立于 1985 年，专业致力于面向资产管理的风险预测业务，其对于短期、中期、长期的风险评估，引入不同的数据源，采用不同的风险建模方式，从而形成一套覆盖相对完整的风评估体系。

⁶ SunGard 公司，总部位于美国宾夕法尼亚，最初是一家从事数据服务的软件公司，1997 年收购了交易与风险管理软件设计商 Infinity Financial Technology，进军风险管理与分析业务，并通过后续的一系列收购，构建起强大的数据服务、风险分析、资产管理的能力。

阶)， Σ_ε 为随机误差矩阵（N 阶），由于资产收益率中的随机误差相互独立， Σ_ε 为对角矩阵。 $\beta = [\beta_1, \beta_2, \dots, \beta_N]$ ，为因子暴露矩阵（N×K）。在海外，有很多应用于计算资产协方差矩阵并进行风险控制的多因子风险模型。第三节中，我们从市场常用模型中选取最新版刻画中国 A 股股市的 CNE6（CNLT）模型作为例子，详细介绍风险模型的搭建和求解方法，意在使读者更好地理解风险模型，体会风险模型计算协方差矩阵的优点及其对风险控制的重要意义。

3.1 市场常用模型估计方法

基于求解股票协方差矩阵的目标，根据式（3.1），我们需要估计三个参数： β 、 Σ_λ 和 Σ_ε 。估计过程主要分为两步：第一步，计算因子暴露矩阵 β 。第二步，通过回归得到因子收益率序列 λ 。第三步，在第二步的基础上计算 Σ_λ 和 Σ_ε 。

3.1.1 计算因子暴露矩阵 β

首先，我们回顾经典的 Fama- MacBeth 截面回归方法：

1. 时序回归（N 次）：在时刻 t ，对每个资产 i 的历史时序数据，分别进行时序回归，估计资产 i 在所有因子上的暴露 $\hat{\beta}_i$ 。
 2. 截面回归（T 次）：用 $\hat{\beta}_i$ 作为自变量，资产超额收益率 $r_{i,t,t+1}$ 作为因变量，对每个时刻 t 进行截面回归，得到各因子收益率时序 $\hat{\lambda}_{i,t,t+1}$ 。
- 进一步即可计算 Σ_λ 和 Σ_ε 。

对于因子暴露矩阵 β 的求解，传统做法是通过 N 个时间序列回归，得到每个资产 i 在全部因子上的暴露 $\hat{\beta}_i$ ，这也是 Fama-MacBeth 两步回归中第一步所采用的做法。然而， $\hat{\beta}_i$ 作为生成的回归变量本身就带有一定误差，将其作为第二步回归的自变量，会产生计量经济学中的变量误差（errors in variables, EIV）问题。

因此，为了规避上述问题，对于风格因子，市场常用多因子模型直

接使用公司特征作为风格因子暴露的原始值（如，直接取用 EP 值原始数据），并对之进行标准化处理。这样，对每一个时间点 t ，我们可以直接得到因子暴露矩阵 β ：

$$\beta = \begin{bmatrix} 1 & \beta_1^{I_1} & \cdots & \beta_1^{I_P} & \beta_1^{S_1} & \cdots & \beta_1^{S_Q} \\ 1 & \beta_2^{I_1} & \cdots & \beta_2^{I_P} & \beta_2^{S_1} & \cdots & \beta_2^{S_Q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \beta_N^{I_1} & \cdots & \beta_N^{I_P} & \beta_N^{S_1} & \cdots & \beta_N^{S_Q} \end{bmatrix} \quad (3.2)$$

矩阵中， β 的下标 $1 \sim N$ 代表 N 个资产分别对应的因子暴露，上标指代不同因子。各资产对国家因子的暴露均为 1。 $\beta_i^{I_1} \sim \beta_i^{I_P}$ 对应 P 个行业因子暴露，均为哑变量，每个资产 i 有且仅有一个 $\beta_i^{I_k} = 1$ 。 $\beta_i^{S_1} \sim \beta_i^{S_Q}$ 对应 Q 个风格因子暴露，由公司特征原始值经过标准化得到。具体而言，假设市场组合在任何风格因子上都应该是中性的，那么市场组合对各因子的暴露为零。因此，先对各因子暴露减去市值加权平均值，使之满足：

$$\sum_{i=1}^N s_i \beta_i^{S_q} = 0, q = 1, \dots, Q \quad (3.3)$$

s_i 表示资产 i 的市值权重。再对各风格因子暴露分别除以其标准差，风格因子暴露的标准化就完成了。

3.1.2 计算因子收益率序列 λ

根据 Menchero and Lee（2015）的方法，我们进行多次横截面回归，得出因子收益率序列。截面回归模型如下：

$$r_{t,t+1} = \beta_t \lambda_{t,t+1} + \varepsilon_{t,t+1} \quad (3.4)$$

其中， β_t 为 t 时刻下的因子暴露矩阵 ($N \times K$) ⁷：

$$\beta_t = \begin{bmatrix} 1 & \beta_1^{I_1} & \cdots & \beta_1^{I_P} & \beta_1^{S_1} & \cdots & \beta_1^{S_Q} \\ 1 & \beta_2^{I_1} & \cdots & \beta_2^{I_P} & \beta_2^{S_1} & \cdots & \beta_2^{S_Q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \beta_N^{I_1} & \cdots & \beta_N^{I_P} & \beta_N^{S_1} & \cdots & \beta_N^{S_Q} \end{bmatrix} \quad (3.5)$$

$r_{t,t+1}$ 为 t 到 $t+1$ 时段 N 支股票（相对无风险收益率）的超额收益 ($N \times 1$)；

⁷ 3.1.1 给出了 β 矩阵的具体计算方法，此处不再赘述。

$\lambda_{t,t+1}$ 为 t 到 $t+1$ 时段各因子收益率 ($K \times 1$)； $\varepsilon_{t,t+1}$ 为 t 到 $t+1$ 时段 N 支股票的特质性收益率 ($N \times 1$)。为了简便，下文省略时间下标 t 与 $t+1$ 。

由于国家因子暴露和 P 个行业的因子暴露之间存在共线性，造成(3.4)解不唯一，因此对行业因子给定限制：

$$s_{I_1}\lambda_{I_1} + s_{I_2}\lambda_{I_2} + \cdots + s_{I_P}\lambda_{I_P} = 0 \quad (3.6)$$

其中， λ_{I_p} 为 t 到 $t+1$ 时段行业因子收益率， s_{I_p} 为 t 时刻 I_p 行业内所有股票按市值加权算出的权重之和。根据约束条件 (3.6) 可以构造约束矩阵 C (Ruud 2000)：

$$\begin{bmatrix} \lambda_C \\ \lambda_{I_1} \\ \vdots \\ \lambda_{I_P} \\ \lambda_{S_1} \\ \vdots \\ \lambda_{S_Q} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & -\frac{s_{I_1}}{s_{I_P}} & -\frac{s_{I_2}}{s_{I_P}} & \cdots & -\frac{s_{I_{P-1}}}{s_{I_P}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \lambda_C \\ \lambda_{I_1} \\ \vdots \\ \lambda_{I_{P-1}} \\ \lambda_{S_1} \\ \vdots \\ \lambda_{S_Q} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (3.7)$$

式 (3.7) 等号右边的矩阵即为 $K \times (K-1)$ 阶约束矩阵 C 。

对于 (3.4)，使用 WLS 估计。在市场常用模型中，假设个股特质性收益率的方差与个股市值成反比，即 $Var(\varepsilon^i) \propto \frac{1}{ME^i}$ ，其中 ME 代表流通市值。基于这个假设，令 $w^i = \sqrt{\frac{1}{ME^i}}$ ，构造如下权重矩阵 W ：

$$W = \begin{bmatrix} w^1 & 0 & \cdots & 0 \\ 0 & w^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w^N \end{bmatrix} \quad (3.8)$$

有了 W 以及 C 之后，我们利用带约束条件的最小二乘法求解 (3.4)

t 到 $t+1$ 时段下的因子收益率 $\hat{\lambda}$ 。

$$\text{令 } \begin{bmatrix} \lambda_C \\ \lambda_{I_1} \\ \vdots \\ \lambda_{I_{P-1}} \\ \lambda_{S_1} \\ \vdots \\ \lambda_{S_Q} \end{bmatrix} = \gamma, \quad \begin{bmatrix} \lambda_C \\ \lambda_{I_1} \\ \vdots \\ \lambda_{I_P} \\ \lambda_{S_1} \\ \vdots \\ \lambda_{S_Q} \end{bmatrix} = \lambda, \text{ 则 (3.7) 可重写为}$$

$$\lambda = C\gamma \quad (3.9)$$

根据加权最小二乘法（WLS），我们需求解

$$\min_{\lambda_t} \|W^{-1}r - W^{-1}\beta\lambda\|^2 \quad (3.10)$$

将（3.9）代入（3.10），得到

$$\min_{\gamma} \|W^{-1}r - W^{-1}\beta C\gamma\|^2 \quad (3.11)$$

式（3.11）对 γ 求一阶导，令一阶导等于 0，解得

$$\hat{\gamma} = (C'\beta'W^{-2}\beta C)^{-1}C'\beta'W^{-2}r \quad (3.12)$$

将（3.9）代入（3.12）得

$$\hat{\lambda} = C(C'\beta'W^{-2}\beta C)^{-1}C'\beta'W^{-2}r \quad (3.13)$$

对于给定的时间窗口 $[t-k/2, t+k/2)$ ，可以通过 k 次截面回归求解出每个时刻下各因子收益率，得到各因子的收益率时间序列。

3.1.3 计算协方差矩阵 Σ_{λ} 和 Σ_{ϵ}

通过 3.1.2，我们得到各因子的收益率时间序列。进一步，根据（3.4）将每个时刻的 $\hat{\lambda}_t$ 代入回归模型，我们可以得到每个时刻下个股的特质性收益率：

$$\hat{\epsilon}_{t,t+1} = r_{t,t+1} - \beta_t \cdot \hat{\lambda}_{t,t+1} \quad (3.14)$$

将每个时刻下个股的特质性收益率序列组合成 N 支个股的特质性收益率时间序列，我们便得到因子收益率时序与特质性收益率时序。进而，我们可以估计出因子收益率的样本协方差矩阵 $\hat{\Sigma}_{\lambda}$ 和个股特质性收益率协方差矩阵 $\hat{\Sigma}_{\epsilon}$ 。实证经验表明，上述对协方差矩阵的简单估计并不准确。因此，我们还需要通过一些统计手段对之进行调整。

对 $\hat{\Sigma}_{\lambda}$ 主要使用特征因子调整法进行调整。首先，对 $\hat{\Sigma}_{\lambda}$ 进行特征分解，分解得到对角矩阵 $\hat{D}$ 与特征向量矩阵。其中，对角矩阵 $\hat{D}$ 的对角元素表示特征因子组合的方差。⁸实证表明，特征因子组合的方差估计并不准

⁸ 特征因子组合是以分解得到的每个特征向量作为权重，将因子组合看作单个资产，加权

确，估计的特征因子组合收益率方差越小，事后检验就越偏离真实值。因此，我们利用自助抽样法（bootstrap）再取样，得到再取样的对角矩阵 $\tilde{D}$ 与 $\hat{D}$ 之间的偏差，以此来近似 $\hat{D}$ 与未知的真实对角矩阵 D 之间的偏差，并以此估计偏差作为修正系数对 $\hat{\Sigma}_\lambda$ 进行修正。

而对于 $\hat{\Sigma}_\varepsilon$ ，研究表明，基于历史数据的估计也并不准确。样本特质性波动率低的股票波动性被低估，样本特质性波动率高的股票波动性被高估。因此，可以采用贝叶斯收缩法进行调整。其具体做法是将全部股票按市值大小分为十组，计算每组市值加权的平均特质性波动率，并将此平均波动率作为各组股票的“先验”。调整后的个股特质性波动率等于“先验”的平均波动率与个股特质性波动率的加权平均。经此调整，样本特质性波动率向先验数据“收缩”，样本估计稳定性增强。

经上述调整，协方差矩阵的估计值更接近实际值，模型估计风险的准确性进一步提高。将 β 与调整后的 Σ_λ 和 Σ_ε 代入（3.1），可求得个股收益率协方差矩阵 Σ 的估计值。

3.2 因子介绍

CNE6（CNLT）模型中包含了 1 个国家因子、16 个风格因子以及 32 个行业因子⁹。其中，风格因子可归为 8 类：规模因子(Size)、波动率因子(Volatility)、流动性因子(Liquidity)、价值因子(Value)、质量因子(Quality)、动量因子(Momentum)、成长因子(Growth)、分红因子(Dividend Yield)。

3.2.1 规模因子(Size)

规模因子（Size）由市值（LNCAP）和中等市值（MIDCAP）组成。

- 市值：股票市值的自然对数。
- 中等市值：代表股票的非线性市值。对市值因子求立方，并以回归权重将其对市值因子正交化，最后进行缩尾和标准化处理。

配置得到的资产组合。

⁹ 行业依据全球工业分类标准（GICS）划分。

3.2.2 波动率因子(Volatility)

波动率因子（Volatility）由 Beta 因子（HBETA）和残余波动率（Residual Volatility）构成。

- Beta 因子：代表无法被市场因子所解释的市场风险。以 252 天为半衰期，在最近 504 个交易日窗口内，将股票收益率（减去无风险收益）对全样本市值加权收益率（减去无风险收益）进行时序回归，取回归系数作为 Beta 因子。¹⁰
- 残余波动率（Residual Volatility）：由历史 Sigma（HSIGMA）、日超额收益标准差（DASTD）、累积收益范围（CMRA）三个因子构成。
 - 历史 Sigma：以计算 Beta 因子的方法进行相同回归，计算残余收益的波动率。
 - 日标准差：以 42 天为半衰期，计算过去 252 个交易日的超额收益波动率。
 - 累积收益范围：过去 12 个月累积对数超额收益的最高值与最低值之差。

3.2.3 流动性因子(Liquidity)

流动性因子（Liquidity）由月换手率（STOM）、季换手率（STOQ）、年换手率（STOA）和年化交易量比率（ATVR）构成。

- 月换手率：最近 1 个月交易股票百分比的对数。
- 季换手率：最近 3 个月月交易股票平均百分比的对数。
- 年换手率：最近 12 个月月交易股票平均百分比的对数。
- 年化交易量比率：以 63 天为半衰期，最近 252 天窗口内，日交易股票百分比的指数加权和。

¹⁰ 为了减少非同步性和自相关性的影响，收益率以 4 天为单位采取聚合处理。

3.2.4 价值因子 (Value)

价值因子由账面市值比 (BTOP)、盈利率 (Earnings Yield) 和长期反转 (Long Term Reversal) 构成。

- 账面市值比 (BTOP)：最新报告普通股账面价值/现市值
- 盈利率 (Earnings Yield) 由现金盈利价格比 (CETOP)、EP 比 (ETOP)、企业倍数 (EM)、分析师预测 EP 比 (ETOPF) 构成。
 - 现金盈利价格比：过去 12 个月现金利润/现市值
 - EP 比：过去 12 个月利润/现市值
 - 企业倍数：上一财年息税前利润 (EBIT) / 现企业价值 (EV)
 - 分析师预测 EP 比：未来 12 个月预测利润/现市值
- 长期反转 (Long Term Reversal)：由长期相对强度 (LTRSTR) 和长期历史 Alpha (LTHALPHA) 构成。
 - 长期相对强度：首先以 260 天为半衰期，在过去 1040 天的窗口内，计算股票超市场收益对数的指数加权和，作为非滞后的长期相对强度。接着滞后 273 天，计算 11 天窗口内非滞后值的等权平均值，再取负数，以此作为长期相对强度。
 - 长期历史 Alpha：首先，以 260 天为半衰期，在过去 1040 天的窗口内，按与计算 Beta 因子一样的方法 (CAPM) 取时序回归的截距项，作为非滞后的长期历史 Alpha。接着滞后 273 天，计算 11 天窗口内非滞后值的等权平均值，再取负数，以此作为长期历史 Alpha。

3.2.5 质量因子 (Quality)

质量因子由杠杆率 (Leverage)、盈利变动率 (Earnings Variability)、盈利质量 (Earnings Quality)、盈利能力 (Profitability)、投资能力 (Investment Quality) 五个二级因子构成。

- 杠杆率 (Leverage)：由市场杠杆率 (MLEV)、账面杠杆率 (BLEV) 和资产负债率 (DTOA) 三个三级因子构成。

■ 市场杠杆率：

$$\frac{\text{上一交易日普通股市值} + \text{上一财年优先股} + \text{上一财年长期负债}}{\text{上一交易日普通股市值}}$$

■ 账面杠杆率：

$$\frac{\text{上一财年普通股账面值} + \text{优先股账面值} + \text{长期负债账面值}}{\text{上一财年普通股账面值}}$$

■ 资产负债率：

$$\frac{\text{上一财年总负债}}{\text{上一财年总资产}}$$

- 盈利变动率（Earnings Variability）：由收入变动率（VSAL）、盈利变动率（VERN）、现金流变动率（VFLO）、预测每股收益变动率（ETOPF_STD）四个三级因子构成。

- 收入变动率：过去 5 年年收入标准差/平均年收入。
- 盈利变动率：过去 5 年年净利润标准差/平均年净利润。
- 现金流变动率：过去 5 年年现金流标准差/平均年现金流。
- 预测每股收益变动率：未来 12 月每股盈利预测标准差/现股价。

- 盈利质量（Earnings Quality）：由资产负债表盈余（ABS）和现金流盈余（ACF）两个三级因子构成。

- 资产负债表盈余：基于资产负债表的盈余（ACCR_BS）通过资产负债表项目的连续变化和对折旧项目的处理计算得到。资产负债表盈余（ABS）= - 基于资产负债表的盈余（ACCR_BS）/总资产。
- 现金流盈余：基于现金流量表的盈余（ACCR_CF）通过对最新现金流量表项目和折旧项目的处理计算得到。现金流盈余（ABS）= - 基于现金流量表的盈余（ACCR_CF）/总资产。

- 盈利能力（Profitability）：由资产周转率（ATO）、毛利（GP）、毛利率（GPM）、资产收益率（ROA）四个三级因子构成。
 - 资产周转率：过去 12 个月收入/最新报告的总资产
 - 毛利：（上一财年收入-销货成本（COGS））/上一财年总资产
 - 毛利率：（上一财年收入-销货成本（COGS））/上一财年收入
 - 资产收益率：过去 12 个月净利润/最新报告的总资产

- 投资能力（Investment Quality）：由总资产增长率（AGRO）、发行股份增长率（IGRO）、资本支出增长率（CXGRO）三个三级因子构成。
 - 总资产增长率：将过去 5 年总资产对时间回归，得到回归系数；将回归系数除以平均总资产再取负数，得到总资产增长率。
 - 发行股份增长率：将过去 5 年发行股份数对时间回归，得到回归系数；将回归系数除以平均发行股份数再取负数，得到发行股份增长率。
 - 资本支出增长率：将过去 5 年资本支出对时间回归，得到回归系数；将回归系数除以平均资本支出再取负数，得到资本支出增长率。

3.2.6 动量因子(Momentum)

动量因子由相对强度（RSTR）和历史 Alpha（HALPHA）构成。

- 相对强度：首先以 126 天为半衰期，在过去 252 天的窗口内，计算股票超市场收益对数的指数加权，作为非滞后的相对强度。接着滞后 11 天，计算 11 天窗口的非滞后相对强度等权平均值，以此作为相对强度。
- 历史 Alpha：首先，以 126 天为半衰期，在过去 252 天的窗口内，按与计算 Beta 因子一样的方法（CAPM）取时序回归的截距项，作为非滞后的历史 Alpha。接着滞后 11 天，计算 11 天窗口的非滞后历史 Alpha 等权平均值，以此作为历史 Alpha。

3.2.7 成长因子(Growth)

成长因子由预测长期净利润增长率（EGRLF）、每股净利润增长率（EGRO）、每股收入增长率（SGRO）组成。

- 预测长期利润增长率：分析师预测长期（3-5 年）净利润增长率。
- 每股净利润增长率：将过去 5 年每股年净利润对时间进行回归，得到回归系数；将回归系数除以平均每股年净利润得到每股净利润增长率。
- 每股收入增长率：将过去 5 年每股年收入对时间进行回归，得到回归系数；将回归系数除以平均每股年收入得到每股收入增长率。

3.2.8 分红因子(Dividend Yield)

分红因子由股息率（DTOP）和分析师预测股息率（DTOPF）构成。

- 股息率：过去 12 个月每股分红/上月末股价。
- 预测股息率：未来 12 个月每股预测分红/现股价。

4. 我们的风险模型

4.1 模型估计方法

4.1.1 极大似然法

不同于常见风险模型采用的 OLS 估计，我们的风险模型采用极大似然法估计因子收益率序列和协方差矩阵。极大似然法的思想较为直观。已知随机样本满足某概率分布，但该分布参数未知，我们试图寻找一组参数，使在该参数条件下，样本出现的概率最大。我们把这样一组参数作为对真实参数的估计。

对于多因子模型

$$r_{t,t+1} = \beta_t \lambda_{t,t+1} + \varepsilon_{t,t+1} \quad (4.1)$$

与 3.1.1 相同，首先用公司财务和交易数据计算出 β_t 矩阵。接着用极大似然法，对因子收益率 $\lambda_{t,t+1}$ 、因子协方差矩阵 Σ_λ 和异质性收益率协

方差矩阵 Σ_ε 进行估计。

做出以下假设：

$$\lambda_{t,t+1} \sim N(0, \Sigma_\lambda) \quad (4.2)$$

$$\varepsilon_{t,t+1} \sim N(0, \Sigma_\varepsilon) \quad (4.3)$$

$\lambda_{t,t+1}$ 与 $\varepsilon_{t,t+1}$ 相互独立。

根据假设条件，由（4.1）的线性关系，我们可以推出 $r_{t,t+1}$ 的分布：

$$r_{t,t+1} \sim N(0, \beta_t \Sigma_\lambda \beta_t^T + \Sigma_\varepsilon) \quad (4.4)$$

因此，对于一组样本 $\{r_{t,t+1}^i, i = 1, 2, \dots, n\}$ ，可以构造对数似然函数：

$$l(\Sigma_\lambda, \Sigma_\varepsilon) = \log \prod_{i=1}^n \frac{1}{(2\pi)^{n/2} |\beta_t \Sigma_\lambda \beta_t^T + \Sigma_\varepsilon|^{1/2}} \times \exp\left(-\frac{1}{2} r_{t,t+1}^i (\beta_t \Sigma_\lambda \beta_t^T + \Sigma_\varepsilon)^{-1} r_{t,t+1}^i\right) \quad (4.5)$$

但是直接通过最大化似然函数的方法求解参数较为困难，因此我们引入 EM 算法，用迭代的思想求解参数 Σ_λ 和 Σ_ε 。

4.1.2 EM 算法如何估计参数 Σ_λ 和 Σ_ε

EM 算法分 E 步和 M 步两个步骤。在 E 步中，给定参数值¹¹和样本收益率，计算出因子收益率的后验分布。在 M 步中，利用 E 步得出的因子收益率的后验分布，最大化对数似然函数，得到新的参数值。重复 E 步和 M 步直至参数收敛，即可获取参数 Σ_λ 和 Σ_ε 估计值。具体公式推导见附录 2。

4.1.3 计算因子收益率 $\lambda_{t,t+1}$

由 4.1.2，我们通过 EM 算法估计出了参数 Σ_λ 和 Σ_ε 。在本节，利用已知的 β_t 、 Σ_λ 和 Σ_ε ，可计算出因子收益率 $\lambda_{t,t+1}$ 。

由（4.2），（4.4）， $\lambda_{t,t+1}$ 与 $r_{t,t+1}$ 服从联合正态分布：

$$\begin{bmatrix} \lambda_{t,t+1} \\ r_{t,t+1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_\lambda & \Sigma_\lambda \beta_t^T \\ \beta_t \Sigma_\lambda^T & \beta_t \Sigma_\lambda \beta_t^T + \Sigma_\varepsilon \end{bmatrix}\right) \quad (4.6)$$

¹¹ 在第一次迭代中需设定参数的初始值。

由附录 1,

$$\begin{aligned}
 E(\lambda_{t,t+1}|r_{t,t+1}) &= \mu_f + \Sigma_{\lambda r} \Sigma_r^{-1} (r_{t,t+1} - \mu_r) \\
 &= \Sigma_{\lambda r} \Sigma_r^{-1} r_{t,t+1} \\
 &= \Sigma_{\lambda} \beta_t^T (\beta_t \Sigma_{\lambda} \beta_t^T + \Sigma_{\varepsilon})^{-1} r_{t,t+1} \quad (4.7)
 \end{aligned}$$

或者进一步将因子收益率 $\lambda_{t,t+1}$ 的条件期望简化写作:

$$E_t[\lambda_{t,t+1}|r_{t,t+1}] = \underbrace{\Sigma_{\lambda} \beta_t^T \cdot (\beta_t \Sigma_{\lambda} \beta_t^T + \Sigma_{\varepsilon})^{-1} r_{t,t+1}}_{\substack{\text{full multi-factor mimicking} \\ \text{portfolios } (K \times N)}} \quad (4.7)$$

4.1.4 计算调整

实证分析表明, EM 算法能够较好的解决线性回归模型的参数估计的准确性与稳定性问题, 但是线性回归模型中的诸多不符合线性回归理想假设的情况仍然需要克服, 因此在线性模型进行参数估计时, 需要对其中的相关数据进行对应调整。

(1) 异方差调整

线性回归模型中个股特质性收益率方差一致的假设, 往往在现实情况中并不成立, 有必要在估测时进行异方差调整。在第 3.1.2 中, 大多数调整模型的方法是使用加权最小二乘法 (WLS), 其中的权重变量通常选用 $\sqrt{ME}$ (这里 ME 代表流通市值), 因此模型相当于假设 $Var(\varepsilon_{t,t+1}^i) \propto \frac{1}{ME_t^i}$ 。基于这个假设, 数学上等价于在估测过程中, 将每一个观测都除以 $w_t^i = \sqrt{\frac{1}{ME_t^i}}$:

$$\frac{r_{t,t+1}^i}{w_t^i} = \frac{\beta_t^i}{w_t^i} \lambda_{t,t+1} + \frac{\varepsilon_{t,t+1}^i}{w_t^i} \quad (4.8)$$

这个假设的合理性可以用两种方法加以验证:

- i. 直接用个股全样本波动率对 $\sqrt{\frac{1}{ME}}$ 进行回归;
- ii. 通过对全样本用 EM 算法估测该因子模型, 并分析异质性收益率协方差矩阵 Σ_{ε} 的离散度进行验证。

通过实证检验, 我们发现, 由于我国市场上个股的市值差距较大,

常见模型的权重假设可能使得小盘股权重过低，从而忽略很多重要的因子联动。基于进一步的实证分析结果，我们发现 $s_t^i = \sqrt{\frac{\text{Tr}_t^i}{\text{ME}_t^i}}$ 是更好的权重选择，其中 Tr 代表换手金额， ME 代表流通市值。基于此，我们构造了新的权重调整矩阵 W ：

$$W = \begin{bmatrix} s_t^1 & 0 & \dots & 0 \\ 0 & s_t^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_t^N \end{bmatrix} \quad (4.9)$$

因此，经过异方差调整，（4.1）被重写为

$$W^{-1}r_{t,t+1} = W^{-1}\beta_t\lambda_{t,t+1} + \tilde{\varepsilon}_{t,t+1} \quad (4.10)$$

对(4.10)进行EM算法估算，我们会得到一个四元组 $(\beta_t, s_t, \Sigma_\lambda, \Sigma_\varepsilon)$ ，对模型目标——资产协方差矩阵的估计可由下式表示：

$$\Sigma = W(W^{-1}\beta_t\Sigma_\lambda\beta_t^TW^{-1} + \Sigma_\varepsilon)W = \beta_t\Sigma_\lambda\beta_t^T + W\Sigma_\varepsilon W \quad (4.11)$$

（2）协方差矩阵估计调整

通过 cluster 法和 shrinkage 法，我们可以得到更为稳健的协方差矩阵 Σ_ε 的估计。聚类（cluster）方法的核心思想是将同处一个 cluster 的多个股票的残差归并为一个相同的值进行联合估计，以进一步减少模型参数，使结果更稳健。shrinkage 法是将协方差矩阵 Σ_ε 的估计向均值进行收缩调整，从而提升协方差矩阵 Σ_ε 估计的稳定性。我们根据每只股票的观测数据量划分权重 w_i ，并根据权重大小确定收缩比例 $\tilde{\alpha}_i$ ，最后得到按比例收缩调整后的 Σ_{ε_i} 。

$$\overline{\Sigma_\varepsilon} = \text{median}(\widehat{\Sigma_{\varepsilon_i}}), \quad \overline{w} = \text{median}(w_i) \quad (4.12)$$

$$\alpha = 5\% \quad (4.13)$$

$$\tilde{\alpha}_i = \max\left(1 - (1 - \alpha) \cdot \frac{w_i}{\overline{w}}, \alpha\right) \quad (4.14)$$

$$\Sigma_{\varepsilon_i} = (1 - \tilde{\alpha}_i) \cdot \widehat{\Sigma_{\varepsilon_i}} + \tilde{\alpha}_i \cdot \overline{\Sigma_\varepsilon} \quad (4.15)$$

4.1.5 共线性调整

如果因子出现完全共线性，横截面回归模型会直接出现无法计算的问题，EM 算法也会因为出现不可唯一辨识的因子组而在进行 M 步时导致不稳定进而估计结果难以收敛。因为如果有两组甚至多组因子可以生成完全一样的 Σ_λ ，根据（4.5），对数似然函数在这些不同因子组条件下取值相同，这会导致 M 步在这些值附近跳跃，从而无法得到一个唯一的收敛值。

因此，在有共线性存在时，我们需要进行共线性调整，模型中唯一的共线性来源于行业因子和国家因子。首先，在原始 β 矩阵中选出不共线因子的子集，即剔除国家因子的 $N \times (K-1)$ 阶 β' 矩阵，该矩阵不存在共线性问题。接着将剔除国家因子后的 β' 矩阵输入 EM 算法进行对 Σ_λ' 的估计，再将 Σ_λ' 通过特定的矩阵变化转化为 Σ_λ ¹²。

4.1.6 模型估计方法的优劣对比

EM 算法也有其缺点，主要体现为其计算复杂耗时，收敛的速度相当依赖于 Σ_λ 和 Σ_ε 的初始值。作为改进，我们可以使用横截面回归的结果作为初始值，使初始值较为接近收敛位置，降低估算时间。在此改进基础上，EM 算法具有以下优势：

1. 更优解：每一步的 EM 回归迭代都会提高 likelihood，其结果一定会收敛。因此从 OLS 的回归结果出发，进行 EM 迭代，在 likelihood 的测量下一定会得到更优解。

2. 稳定性（噪音抵抗性）更强：EM 算法的稳定性远优于横截面回归，具体表现为它有更强的噪音抵抗性。它的因子模拟投资组合取决于因子组暴露和资产的协方差矩阵 ($Var(r_{t,t+1}) = \beta_t \Sigma_\lambda \beta_t^T + \Sigma_\varepsilon$)。因此，在对因子组微调的时候，资产间的联动性和模拟投资组合变动都不大， $Var(r_{t,t+1})$ 的变化不大，原有因子的因子模拟投资组合受到的影响也不会大。反观横截面回归方法，加入/删减因子均可能导致 $(\beta_t^T \beta_t)^{-1}$ 发生剧烈改变。

¹² 详见附录 3。

3. 更好应对数据缺失：金融市场数据集中缺失值是非常常见的，如果不能得到恰当处理，传统的 OLS 算法会对缺失值产生较大的调整，而，EM 算法的稳定性则能够在数据采集不完全的情况下依然可以给出良好的估计结果。

4. 假设更合理：在 OLS 回归中，模型假设 β_t 与 $\varepsilon_{t,t+1}$ 独立；而本模型假设 $\varepsilon_{t,t+1}$ 与 $\lambda_{t,t+1}$ 相互独立。在金融市场实践中， β_t 与 $\varepsilon_{t,t+1}$ 独立的假设是很少成立的。资产收益、因子暴露、残差等变量常常出现较为异常的特性，比如残差波动与某些因子暴露相关。在实践意义上， $\varepsilon_{t,t+1}$ 与 $\lambda_{t,t+1}$ 相互独立的假设更为合理。

对于 OLS 和 EM 算法下，残差的相关性对比，我们设计了一下测试：

1. 在测试中使用了真实的因子暴露数据，我们使用 2017 年-2020 年三年的数据，随机抽取 500 支股票在六个技术面因子上的暴露作为 β 。

2. 使用标准正态分布生成模拟的因子收益。

3. 使用 β 和随机正态分布模拟残差，使得生成的残差与因子暴露的相关性不为 0。具体测试中，我们使用某个因子的三次方来作为残差的非独立部分，然后通过调整系数来调节残差中独立与非独立的比例，如当残差中非独立的比例为 80%时，总残差= $0.8 * (\text{非独立残差}) + 0.2 * (\text{独立残差})$ 。

4. 使用模拟的因子收益和残差，获得模拟股票收益。

5. 将真实因子暴露与模拟股票收益分别送入 EM 与 OLS 方法进行回归，二者分别给出预测残差。

6. 分别计算两种方法得到的残差与模拟生成的精确残差的相关性，进行比较。

7. 改变模拟残差中独立/非独立部分的比例，可以获得两条 相关性 vs 比例的曲线。

我们将结果总结为图 1-图 4：

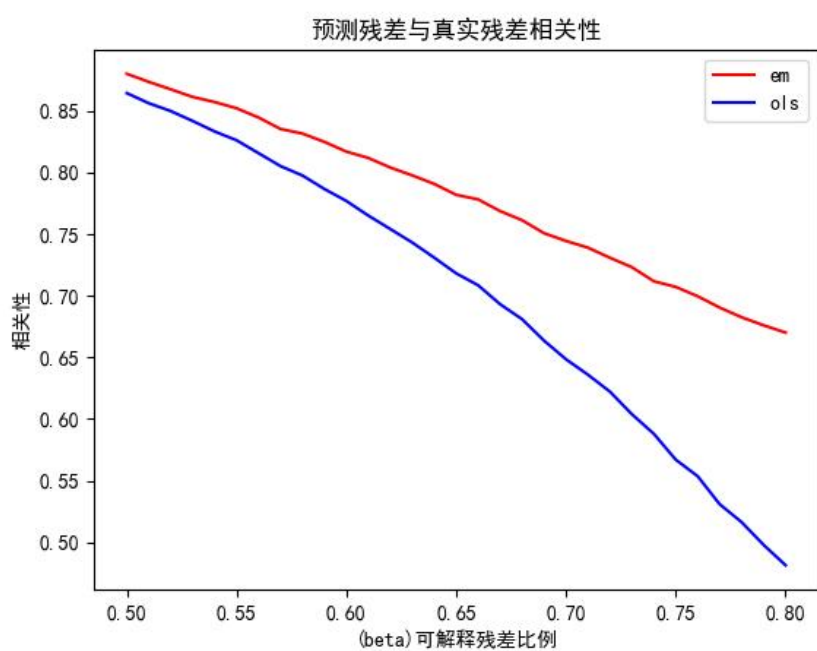


图 1 预测残差与真实残差相关性（针对 Beta 因子的估计）

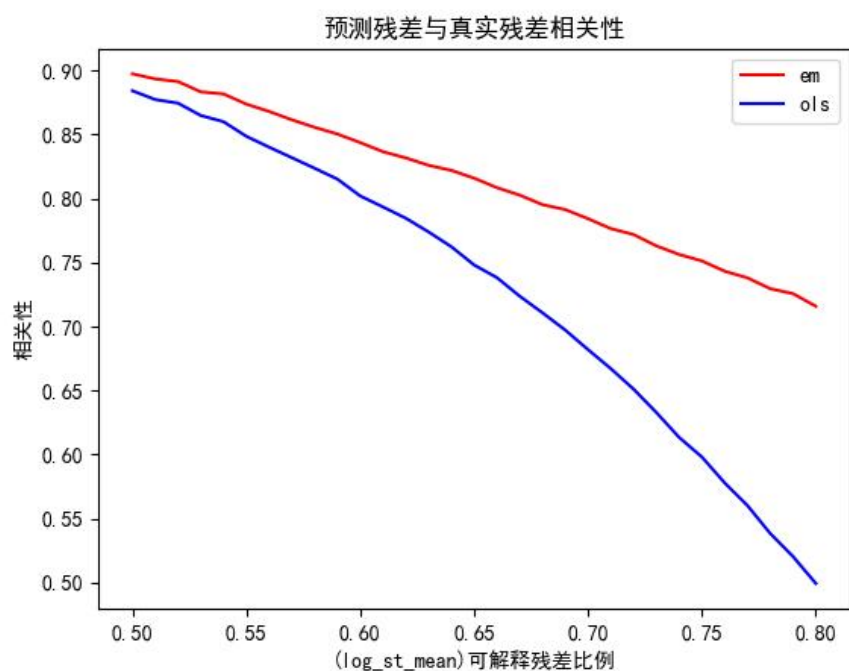


图 2 预测残差与真实残差相关性（针对年化交易量比因子的估计）

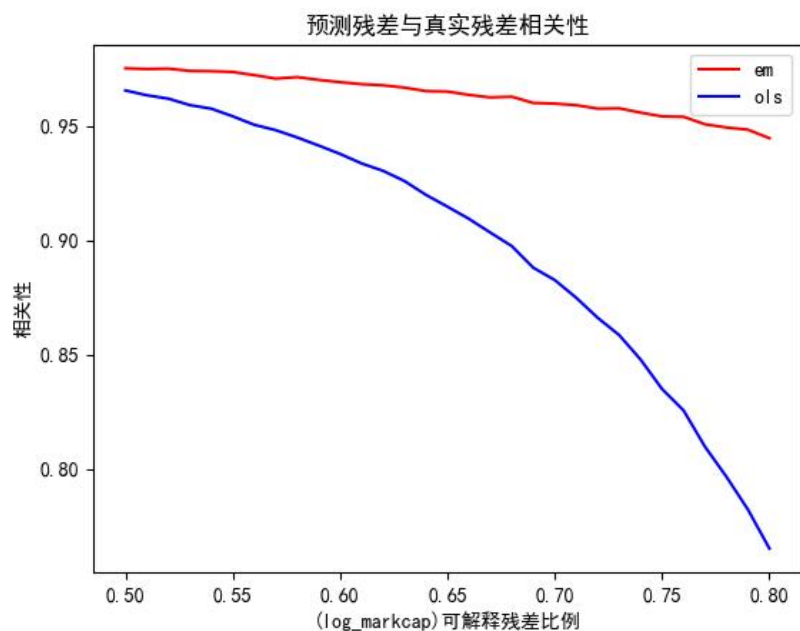


图 3 预测残差与真实残差相关性（针对市值因子的估计）

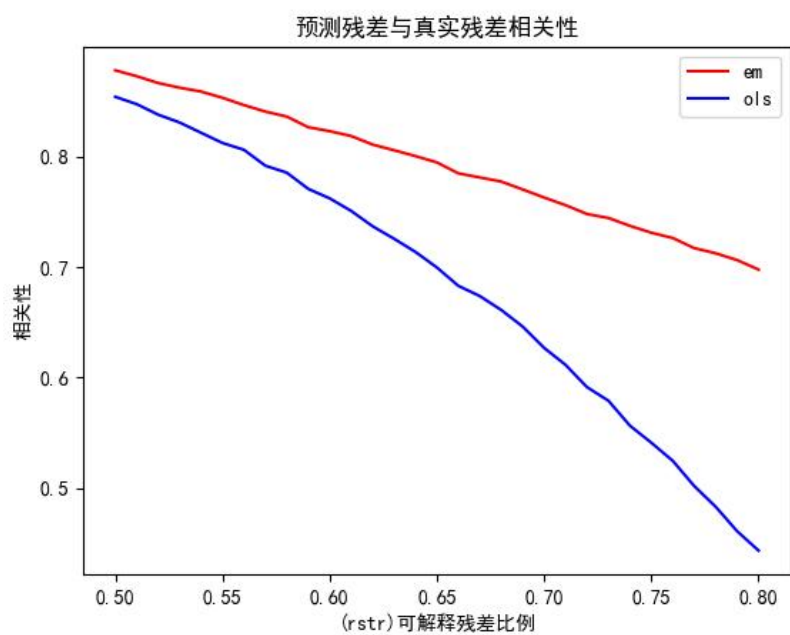


图 4 预测残差与真实残差相关性（针对短期反转因子的估计）

图 1 至图 4 中横坐标为残差中非独立的可解释部分的比例，括号中标明非独立残差由哪个因子暴露的三次方得到，如（beta）表示残差中的非独立可解释部分是由 beta 因子暴露的三次方得到的。

纵坐标为回归计算预测的残差与真实残差的相关性，具体计算方法为：在每一种比例下，两种回归方法加上模拟生成的精确值，每支股票均可以得到三个残差的时间序列，两种回归方法的时间序列分别与精确值的时间序列计算相关性，于是每支股票每种方法都可以得到两个相关系数，两种方法分别取 500 支股票的相关系数的 mean，于是在每种比例下，每种方法得到一个相关系数的 mean 值。然后我们连续改变残差中独立/非独立部分的比例，得到两条曲线。

如图 1 至图 4 所示，与 OLS 估计的残差相比，EM 方法估计的残差与真实残差的相关性更强。更高相关性意味着 EM 算法计算的残差能够更好的反应真实残差。这也说明了 EM 算法优势。

4.2 风险因子模型的评价标准

一个好的风险因子模型应该满足以下几个标准：

（1）风险因子度量准确

计算因子的数据集应尽量少包含缺失数据，同时保证因子数据的信息不损失。较少缺失值保证了计算过程中未缺失信息的利用效率，这能确保尽量多的信息纳入了模型之中。

（2）风险因子体系能准确地刻画金融资产联动关系

风险因子对资产收益率的解释应该足够高、整个风险因子模型对资产收益联动的变化关系也应该有很好的解释力。这保证了基于风险因子体系的模型能明确展示出市场的现状和金融体系正在发生的变化。

（3）风险因子模型能正确指导决策

风险因子模型需要正确地刻画风险，基于此给出的投资建议需要明确、一致且连续。这要求风险因子值要尽量少的出现极值，而且风险因子值不能有过分的跳跃，风险因子模型得出的结果也要求比较稳定。

(4) 风险因子体系开放、易扩展

风险因子体系应该是开放的，需要很容易地容纳新的因子，从而适应世界的变化。

4.3 中国风险因子模型的评价方式

将风险因子评价指标进一步细化为可度量的指标，并针对中国风险因子模型进行了量化评估：

- 1.完备性：中国风险因子模型使用每天有因子值的股票数量与上市股票数量之比计算的覆盖率来衡量因子数据的完备程度。利于模型参数估计，要求筛选因子组合的覆盖率高于 50%。
- 2.正态性：使用因子值的 5%至 95%分位数衡量因子极值的分布情况，因子回报分布尽量接近正态分布。
- 3.连续性：使用因子值月度自相关系数 AR1 大小来衡量因子值的连续性。AR1>0.95 为连续性较好。
- 4.相关性：使用因子相关系数矩阵来判断因子间的相关程度。在每一个方向的因子组中，尽可能取其中跨组相关性小的因子。
- 5.解释能力：中国风险因子模型还使用单因子加行业因子衡量单个因子的解释力，使用包含全部因子模型的 log_likelihood 衡量模型解释资产联动的能力。

4.4 中国风险因子描述

4.3.1.规模

4.3.1.1.规模

表 1 规模因子描述统计

Factor	描述				
LOGCAP	流通市值的自然对数 (log(size))				
MIDCAP	总市值 (总市值三次方与 log 市值回归的残差)				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
LOGCAP	1.00	0.99	23049.55		
达标或采用	Yes	Yes	Yes	Yes	Yes
MIDCAP	1.00	0.98	9977.05		
达标或采用	Yes	Yes		Yes	

规模：规模因子是代表股票回报相关的重要来源，并捕捉大型股和小型股之间的回报差异。我们用市值的对数来衡量规模。

上表中汇报了规模因子下的三级因子：流通市值的自然对数和总市值。其中，流通市值的自然对数覆盖率为 100%，AR1 为 0.99，LL 提升为 23049.55；总市值覆盖率为 100%，AR1 为 0.99，LL 提升为 23049.55；根据评价标准，流通市值的自然对数达到了完备性、连续性和解释能力的标准；总市值达到了完备性、连续性的标准。最终我们选择了流通市值的自然对数因子。该因子由 Fama 和 French（1993）提出。常见于近期资产定价论文比如 Chen 等（2023）等。

4.3.2.波动率

4.3.2.1.Beta

表 2 Beta 因子描述统计

Factor	描述				
BETA	BETA 值（股票超额收益率 和 市值加权指数(沪深 300)收益率的回归系数）				
	覆盖率	AR1	LL 提升	CUFM	IGSM
BETA	0.99	0.92	36631.86		
达标或采用	Yes		Yes	Yes	Yes

Beta: 捕捉市场因素无法解释的市场风险。为了更好地理解这一因素，请考虑一个完全投资的长期投资组合，该组合强烈倾向于高贝塔股票。直观地说,这个投资组合比beta等于1的投资组合具有更大的市场风险。这一额外的市场风险通过贝塔系数的正敞口来捕捉。市场因素和贝塔因素之间的时间序列相关性通常很高，因此在本例中，这两个风险源相互强化。相比之下，如果投资组合投资于低贝塔股票，那么贝塔和市场因素的风险就会部分抵消，正如预期的那样。

上表中汇报了 Beta 因子下的三级因子：**BETA 值**。其中，BETA 值的覆盖率为 99%，AR1 为 0.92，LL 提升为 36631.86。根据评价标准，BETA 值达到了完备性、连续性和解释能力的标准。Beta 因子是经常使用等因子，如 Gu 等（2021），使用 Beta 作为检验潜在因子的基础模型。

4.3.2.2.波动率

表 3 波动率因子描述统计

Factor	描述				
CumulativeRange	累计幅度（过去 12 个月超额收益的离差）				
DailySTD	日度标准差（股票日收益率在过去 252 日波动率）				
Hsigma	历史标准差（计算 beta 时残差收益率的波动率，回归用 252 交易日数据，移动平均滞后期 63 天）				
LogDailySTD	log(日度标准差)				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
CumulativeRange	0.99	0.91	13738.85		
达标或采用	Yes			Yes	
DailySTD	0.98	0.93	44615.36		
达标或采用	Yes			Yes	
Hsigma	0.99	0.96	35422.46		
达标或采用	Yes	Yes		Yes	
LogDailySTD	0.98	0.93	49705.25		
达标或采用	Yes		Yes		Yes

波动率：由四个三级因子组成：每日超额收益的波动率、每日剩余收益的波动性以及过去 12 个月股票超额收益的累积范围。由于这些描述符倾向于与贝塔因子以及尺寸因子高度共线，但程度较小，但仍很重要，因此该因子与这两个因子正交。

上表中汇报了规模因子下的三级因子：累计幅度（过去 12 个月超额收益的离差），日度标准差（股票日收益率在过去 252 日波动率），历史标准差，（计算 beta 时残差收益率的波动率，回归用 252 交易日数据，移动平均滞后期 63 天），log(日度标准差)。其中，累计幅度覆盖率为 99%，AR1 为 0.91，LL 提升为 13738.85；日度标准差覆盖率为 98%，AR1 为 0.93，LL 提升为 44615.36；历史标准差覆盖率为 99%，AR1 为 0.96，LL 提升为 35422.46；log(日度标准差)覆盖率为 98%，AR1 为 0.93，LL 提升为 49705.25。根据评价标准，累计幅度达到了完备性的标准；日度标准差达到了完备性的标准；历史标准差到了完备性、连续性的标准；log(日度标准差)达到了完备性的标准。最终我们选择了 log(日度标准差)因子。与之类似等因子也常用于论文中，比如 Jegadeesh 等(2019)和 chen（2023）等。

4.3.3.流动性

4.3.3.1 周转率

表 4 周转率因子描述统计

Factor	描述				
AnnualShareTurnover	年度换手率（过去 1 年股票换手率求和的对数值）				
AnnualizedTradedValueRatio	换手率的年度加权平均（过去 252 日换手率指数加权移动平均）				
LogAnnualizedTradedValueRatio	换手率的指数移动平均，半衰期为 21 个交易日				
MonthlyShareTurnover	月度换手率（过去 1 月股票换手率求和的对数值）				
QuarterlyShareTurnover	季度换手率（过去 1 季度股票换手率求和的对数值）				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
AnnualShareTurnover	0.95	0.98	15444.32		
达标或采用	Yes	Yes		Yes	
AnnualizedTradedValueRatio	0.95	0.97	15034.02		
达标或采用	Yes	Yes		Yes	
LogAnnualizedTradedValueRatio	0.98	0.95	48291.99		
达标或采用	Yes			Yes	
MonthlyShareTurnover	0.99	0.83	16951.75		
达标或采用	Yes			Yes	
QuarterlyShareTurnover	0.98	0.94	18361.41		
达标或采用	Yes		Yes		Yes

流动性捕获因相对交易活动而产生的回报差异，通过在选定的后续窗口内交易的总流通股份额来衡量。

上表中汇报了规模因子下的三级因子：年度换手率（过去 1 年股票换手率求和的对数值）、换手率的年度加权平均（过去 252 日换手率指数加权移动平均）、换手率的指数移动平均，半衰期为 21 个交易日，半衰期为 21 个交易日、月度换手率（过去 1 月股票换手率求和的对数值）、季度换手率（过去 1 季度股票换手率求和的对数值）。其中，年度换手率覆盖率为 95%，AR1 为 0.97，LL 提升为 15444.32；换手率的年度加权平均覆盖率为 95%，AR1 为 0.99，LL 提升为 15034.02；log(ewm(换手率))覆盖率为 98%，AR1 为 0.95，LL 提升为 48291.99；月度换手率为 99%，AR1 为 0.83，LL 提升为 16951.75；季度换手率覆盖率为 98%，AR1 为 0.94，LL 提升为 18361.41。根据评价标准，年度换手率达到了完备性、连续性的标准；换手率的年度加权平均达到了完备性、连续性

的标准； $\log(\text{ewm}(\text{换手率}))$ 达到了完备性的标准；月度换手率达到了完备性的标准；季度换手率达到了完备性的标准。最终我们选择了季度换手率因子。POJANAVATEE（2020）采用了改因子检验了泰国股市的因子模型。

4.3.4. 动量

4.3.4.1. 短期反转

表 5 短期反转因子描述统计

Factor	描述				
ShortTermReversal	短期反转（最近一个月加权累积对数日收益率）				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
ShortTermReversal	0.99	0.04	24899.01		
达标或采用	Yes		Yes	Yes	Yes

短期反转，根据最近一个月加权累积对数日收益率的相对表现区分股票。它与动量因子正交。

上表中汇报了规模因子下的三级因子：短期反转。其中，短期反转的覆盖率为 99%，AR1 为 0.04，LL 提升为 24899.01。短期反转达到了完备性的标准。最终我们选择了短期反转因子。同时，在 Kelly（2021）中也展示了反转因子在因子模型中的重要作用。

4.3.4.2. 动量

表 6 动量因子描述统计

Factor	描述				
Halpaha	历史 Alpha（回归用 252 交易日数据，移动平均滞后期 63 天）				
RelativeStrength	相对强度（长期动量减短期动量，长期动量窗口 252 天滞后 121 天、短期窗口 11 天，滞后 11 天）				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
Halpaha	0.99	0.81	27404.87		
达标或采用	Yes			Yes	
RelativeStrength	0.97	0.91	26520.20		
达标或采用	Yes		Yes	Yes	Yes

动量，根据最近 12 个月的表现区分股票。在计算这一因子时，我们排除了最近一个月的回报，以避免短期逆转的影响。

上表中汇报了规模因子下的三级因子：历史 Alpha（回归用 252 交易日数据，移动平均滞后期 63 天）、相对强度（长期动量减短期动量，

长期动量窗口 252 天滞后 121 天、短期窗口 11 天，滞后 11 天）。其中，历史 Alpha 覆盖率为 99%，AR1 为 0.81，LL 提升为 27404.87；相对强度覆盖率为 97%，AR1 为 0.91，LL 提升为 26520.20；根据评价标准，历史 Alpha 达到了完备性、连续性和解释能力的标准；相对强度达到了完备性、连续性的标准。最终我们选择了相对强度因子。

4.3.5.质量

4.3.5.1. 现金流风险

表 7 现金流风险因子描述统计

Factor	描述				
AccrualsBalancesheetVersion	资产负债表应计项目 (ACCR_BS/TA)				
AccrualsCashflowVersion	现金流量表应计项目 (ACCR_CF/TA)				
GrossProfitability	资产毛利率 (主营业务收入-营业成本)/总资产				
GrossProfitMargin	销售毛利率 (主营业务收入-营业成本)/主营业务收入				
AssetTurnover	资产周转率 (Sales/TA)				
MoneyessFreeCashflow	Cashflow Moneyess:(Operating CF + Investment CF)/(gross margin * variation in sales)				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
AccrualsBalancesheetVersion	0.99	0.74	3294.50		
达标或采用	Yes			Yes	
AccrualsCashflowVersion	0.99	0.90	4389.21		
达标或采用	Yes			Yes	
GrossProfitability	0.99	0.96	6220.18		
达标或采用	Yes	Yes		Yes	
GrossProfitMargin	0.99	0.98	7391.48		
达标或采用	Yes	Yes	Yes	Yes	Yes
AssetTurnover	0.99	1.00	3774.68		
达标或采用	Yes	Yes		Yes	
MoneyessFreeCashflow	0.93	0.74	2845.54		
达标或采用	Yes				

质量：结合资产、现金流量、利润率和资产周转率来衡量质量。

上表中汇报了规模因子下的三级因子：资产负债表应计项目 (ACCR_BS/TA)、现金流量表应计项目 (ACCR_CF/TA)、资产毛利率 (主营业务收入-营业成本)/总资产、销售毛利率 (主营业务收入-营业成本)/主营业务收入、资产周转率 (Sales/TA)、Cashflow Moneyess:(Operating CF + Investment CF)/(gross margin * variation in sales)。其中，资产负债表应计项目覆盖率为 99%，AR1 为 0.74，LL 提升为 3294.50；

现金流量表应计项目覆盖率为 99%，AR1 为 0.90，LL 提升为 4389.21；资产毛利率覆盖率为 99%，AR1 为 0.96，LL 提升为 6220.18；销售毛利率覆盖率为 99%，AR1 为 0.98，LL 提升为 7391.48；资产周转率覆盖率为 99%，AR1 为 1.00，LL 提升为 3774.68；Cashflow Moneyiness 覆盖率为 93%，AR1 为 0.74，LL 提升为 2845.54；根据评价标准，资产负债表应计项目达到了完备性的标准；现金流量表应计项目达到了完备性的标准；资产毛利率达到了完备性、连续性的标准；销售毛利率达到了完备性、连续性的标准；资产周转率达到了完备性的标准；Cashflow Moneyiness 达到了完备性的标准。最终我们选择了销售毛利率因子。Sheng（2022）表明了这个因子的有效性。

4.3.5.2. Balance Sheet Low Risk

表 8 Balance Sheet Low Risk 因子描述统计

Factor	描述				
BookLeverage	账面杠杆 $((be+pe+ld)/be)$				
MarketLeverage	市场杠杆 $((me+pe+ld)/me)$				
DebttoAssetRatio	负债资产比（总负债/总资产）				
MoneyinessAssettoDebt	Asset to debt Moneyiness $(\log(Asset / Debt) / \sigma(earning / asset))$				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
BookLeverage	0.95	0.99	4942.10		
达标或采用	Yes	Yes		Yes	
MarketLeverage	0.95	1.00	6432.06		
达标或采用	Yes	Yes		Yes	
DebttoAssetRatio	0.99	0.99	5523.33		
达标或采用	Yes	Yes		Yes	
MoneyinessAssettoDebt	0.95	0.98	5158.69		
达标或采用	Yes	Yes	Yes		Yes

流动性：捕获因相对交易活动而产生的回报差异，通过在选定的后续窗口内交易的总流通股份额来衡量。

上表中汇报了规模因子下的三级因子：账面杠杆 $((be+pe+ld)/be)$ 、市场杠杆 $((me+pe+ld)/me)$ 、负债资产比（总负债/总资产）、Asset to debt Moneyiness $(\log(Asset / Debt) / \sigma(earning / asset))$ 。其中，账面杠杆覆盖率为 95%，AR1 为 0.99，LL 提升为 4942.10；市场杠杆覆盖率为 95%，AR1 为 1.00，LL 提升为 6432.06；负债资产比覆盖率为 99%，

AR1 为 0.99, LL 提升为 5523.33; Asset to debt Moneyiness 覆盖率为 95%, AR1 为 0.98, LL 提升为 5158.69。根据评价标准, 账面杠杆达到了完备性、连续性的标准; 市场杠杆达到了完备性、连续性的标准; 负债资产比达到了完备性、连续性的标准; Asset to debt Moneyiness 达到了完备性、连续性的标准。最终我们选择了 Asset to debt Moneyiness 因子。Merton (1976) 展示了信贷在资产定价中的重要性。类似的还有 Koijen (2020) 中也论证了长期债务在资产定价中的重要性。和我们类似, Hou 等 (2021) 和 Harahap (2021) 也指出了公司预期增长因子在资产定价中的重要性。

4.3.5.3. 资产增长率

表 9 资产增长率因子描述统计

Factor	描述				
CapitalExpenditureGrowth	资本支出增长率 (近五年资本支出对时间回归的斜率除平均资本支出, 取相反数)				
IssuanceGrowth	股票发行量增长率 (最近 5 年流通股本对时间回归取斜率/平均资本支出, 取相反数)				
ReturnonAssets	总资产收益率 (过去 12 个月净利润/总资产)				
ReturnonEquity	股本收益率 (过去 12 个月净利润/equity book value)				
TotalAssetGrowth	总资产增长率 (最近 5 年总资产对斜率回归取截距/平均总资产, 取相反数)				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
CapitalExpenditureGrowth	0.94	0.91	6893.65		
达标或采用	Yes		Yes	Yes	Yes
IssuanceGrowth	0.45	0.96	4789.51		
达标或采用		Yes		Yes	
ReturnonAssets	0.99	0.98	8148.48		
达标或采用	Yes	Yes		Yes	
ReturnonEquity	0.99	0.97	8349.15		
达标或采用	Yes	Yes	Yes		Yes
TotalAssetGrowth	0.98	0.98	7736.18		
达标或采用	Yes	Yes	Yes	Yes	Yes

质量结合资产、现金流量、利润率和资产周转率来衡量质量。

上表中汇报了规模因子下的三级因子: 资本支出增长率 (近五年资本支出对时间回归的斜率除平均资本支出, 取相反数)、股票发行量增

长率（最近 5 年流通股本对时间回归取斜率/平均资本支出，取相反数）、总资产收益率（过去 12 个月净利润/总资产）、股本收益率（过去 12 个月净利润/equity book value）、总资产增长率（最近 5 年总资产对斜率回归取截距/平均总资产，取相反数）。其中，资本支出增长率覆盖率为 94%，AR1 为 0.91，LL 提升为 6893.65；股票发行量增长率覆盖率为 45%，AR1 为 0.96，LL 提升为 4789.51；总资产收益率覆盖率为 99%，AR1 为 0.98，LL 提升为 8148.48；股本收益率覆盖率为 99%，AR1 为 0.97，LL 提升为 8349.15；总资产增长率覆盖率为 98%，AR1 为 0.98，LL 提升为 7736.18。根据评价标准，资本支出增长率达到了完备性的标准；股票发行量增长率达到了连续性的标准；总资产收益率达到了完备性、连续性的标准；股本收益率达到了完备性、连续性的标准；总资产增长率达到了完备性、连续性的标准。最终我们选择了总资产增长率因子和股本收益率因子。我们和 Harahap（2021）的因子选取有一致性，模型采用了股本收益率因子。

4.3.5.4.Variability

表 10 Variability 因子描述统计

Factor	描述				
Predictedstd	分析师预期 1 年盈利标准差股价比(分析师预期未来 12 个月盈利标准差/股价)				
VariationinCashFlow	现金流波动率（过去 5 年现金及等价物增加额的标准差/平均现金及等价物净增加额）				
VariationinEarnings	净利润波动率（过去 5 年净利润标准差/平均年净利润）				
VariationinSales	营业收入波动率（过去 5 年营业收入标准差/平均年营业收入）				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
Predictedstd	0.79	0.97	4738.37		
达标或采用		Yes		Yes	
VariationinCashFlow	0.93	0.60	3306.82		
达标或采用	Yes			Yes	
VariationinEarnings	0.93	0.75	3676.00		
达标或采用	Yes			Yes	
VariationinSales	0.93	0.96	3548.35		
达标或采用	Yes	Yes		Yes	

质量结合资产、现金流量、利润率和资产周转率来衡量质量。

上表中汇报了规模因子下的三级因子：分析师预期 1 年盈利标准差股价比（分析师预期未来 12 个月盈利标准差/股价）、现金流波动率（过

去 5 年现金及等价物增加额的标准差/平均现金及等价物净增加额)、净利润波动率(过去 5 年净利润标准差/平均年净利润)、营业收入波动率(过去 5 年营业收入标准差/平均年营业收入)。其中,分析师预期 1 年盈利标准差股价比为 79%, AR1 为 0.97, LL 提升为 4738.37; 现金流波动率覆盖率为 93%, AR1 为 0.60, LL 提升为 3306.82; 净利润波动率覆盖率为 93%, AR1 为 0.75, LL 提升为 3676.00; 营业收入波动率覆盖率为 93%, AR1 为 0.96, LL 提升为 3548.35; 根据评价标准, 分析师预期 1 年盈利标准差股价比数达到了连续性的标准; 现金流波动率达到了完备性的标准; 净利润波动率达到了连续性的标准; 营业收入波动率达到了完备性、连续性的标准。我们未选择该二级指标。

4.3.6. 市值

4.3.6.1. 账面市值比

表 11 账面市值比因子描述统计

Factor	描述				
BooktoPrice	账面市值比(净资产/市值)				
LogBooktoPrice	log BP				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
BooktoPrice	0.99	0.99	13847.02		
达标或采用	Yes	Yes		Yes	
LogBooktoPrice	0.99	0.98	14728.59		
达标或采用	Yes	Yes	Yes		Yes

上表中汇报了规模因子下的三级因子: 账面市值比(净资产/市值)、log BP。其中, 账面市值比覆盖率为 99%, AR1 为 0.99, LL 提升为 13847.02; log BP 覆盖率为 99%, AR1 为 0.98, LL 提升为 14728.59; 根据评价标准, 账面市值比达到了完备性、连续性的标准; log BP 达到了完备性、连续性的标准。最终我们选择了 log BP 因子。账面价值和价格一直在因子模型中非常重要, 如 Kojien (2020) 在汇率定价模型中采用了账面价值因子。

4.3.6.2. 盈利价格比

表 12 盈利价格比因子描述统计

Factor	描述				
CashEarningtoPrice	现金盈利价格比（过去 12 个月现金盈利/市值）				
EarningtoPrice	盈利市值比（过去 12 个月盈利/市值）				
EbittoEv	息税前利润企业价值比（EBIT/EV）				
LongTermRelativeStrength	长期相对强弱（股票收益率加权平均后取相反数）				
LongTermHistoricalAlpha	长期历史 alpha				
PredictedEarningtoPrice	预期 12 个月盈利市值比（预期 12 个月盈利/市值）				
PredictedEarningtoPriceFilled	predict earnings（combine first earnings）/price				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
CashEarningtoPrice	0.99	0.95	5323.74		
达标或采用	Yes	Yes		Yes	
EarningtoPrice	0.99	0.97	10514.73		
达标或采用	Yes	Yes		Yes	
EbittoEv	0.99	0.96	8996.13		
达标或采用	Yes	Yes		Yes	
LongTermRelativeStrength	0.90	0.94	5131.42		
达标或采用	Yes			Yes	
LongTermHistoricalAlpha	0.90	0.85	7399.97		
达标或采用	Yes			Yes	
PredictedEarningtoPrice	0.75	0.97	8850.87		
达标或采用		Yes		Yes	
PredictedEarningtoPriceFilled	0.99	0.96	11761.54		
达标或采用	Yes	Yes	Yes		Yes

上表中汇报了规模因子下的三级因子：现金盈利价格比（过去 12 个月现金盈利/市值）、盈利市值比（过去 12 个月盈利/市值）、息税前利润企业价值比（EBIT/EV）、长期相对强弱（股票收益率加权平均后取相反数）、长期历史 alpha、预期 12 个月盈利市值比（预期 12 个月盈利/市值）、predict earnings（combine first earnings）/price。其中，现金盈利价格比覆盖率为 99%，AR1 为 0.95，LL 提升为 5323.74；盈利市值比覆盖率为 99%，AR1 为 0.97，LL 提升为 10514.73；息税前利润企业价值比覆盖率为 99%，AR1 为 0.96，LL 提升为 8996.13；长期相对强弱覆盖率为 90%，AR1 为 0.94，LL 提升为 5131.42；长期历史 alpha 覆盖率为 90%，AR1 为 0.85，LL 提升为 7399.97；预期 12 个月盈利市值比覆盖率为 75%，AR1 为 0.97，LL 提升为 8850.87；predict earnings 覆盖率为 99%，AR1 为 0.96，LL 提升为 11761.54。根据评价标准，现金

盈利价格比达到了完备性、连续性的标准；盈利市值比达到了完备性、连续性的标准；息税前利润企业价值比达到了完备性、连续性的标准；长期相对强弱达到了完备性的标准；长期历史 alpha 达到了完备性的标准；预期 12 个月盈利市值比达到了连续性的标准；predict earnings 达到了完备性、连续性的标准。最终我们选择了 predict earnings 因子。Wasan 和 Mulchandani（2020）指出了 predict earnings 在企业金融的重要性，同时 Lalwani 和 Madhumita（2020）也指出了盈利价格比因子在因子模型中的重要性。

4.3.7.增长

4.3.7.1.收益增长率

表 13 收益增长率因子描述统计

Factor	描述				
EarningperShareGrowth	每股收益增长率（过去五年 EPS 对时间回归斜率/平均 EPS）				
EarningGrowthbyEquityValue	Earning growth / Net Asset(每股 EPS 增长)				
PredictedEarningGrowth	预测长期盈利增长率((预期净利润 - 当前净利润) / ((abs(预期净利润)+abs(当前净利润))/2))				
PredictedEarningGrowthbyEquityValue	预测长期盈利增长率((预期净利润 - 当前净利润) / Net Asset)				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
EarningperShareGrowth	0.90	0.85	2973.85		
达标或采用	Yes			Yes	
EarningGrowthbyEquityValue	0.98	0.97	3836.52		
达标或采用	Yes	Yes		Yes	
PredictedEarningGrowth	0.72	0.95	3400.40		
达标或采用				Yes	
PredictedEarningGrowthbyEquityValue	0.72	0.96	5179.76		
达标或采用		Yes			

增长：根据销售或收益增长的前景区分股票。这一因素中最重要的描述是分析师预测的长期收益增长。它还包括过去五年的销售额和收入增长。

上表中汇报了规模因子下的三级因子：每股收益增长率（过去五年 EPS 对时间回归斜率/平均 EPS）、Earning growth / Net Asset(每股 EPS 增长)、预测长期盈利增长率 A((预期净利润 - 当前净利润) / ((abs(预期净利润)+abs(当前净利润))/2))、预测长期盈利增长率 B((预期净利润 -

当前净利润) / Net Asset。其中，每股收益增长率覆盖率为 90%，AR1 为 0.85，LL 提升为 2973.85；每股 EPS 增长覆盖率为 98%，AR1 为 0.97，LL 提升为 3836.52；预测长期盈利增长率覆盖率为 72%，AR1 为 0.95，LL 提升为 3400.40；总市值覆盖率为 72%，AR1 为 0.96，LL 提升为 5179.76。根据评价标准，每股收益增长率达到了完备性的标准；每股 EPS 增长达到了完备性、连续性的标准；预测长期盈利增长率 A 未达到标准；预测长期盈利增长率 B 达到了连续性的标准。最终我们并未选择该二级因子。

4.3.7.2.营业收入增长率

表 14 营业收入增长率因子描述统计

Factor	描述				
SalesperShareGrowth	每股营业收入增长率(过去五年每股营收对时间回归斜率/平均每股营收)				
SalesGrowthbyEquityValue	Sales growth / Net Asset (分配到每一股的销售增长)				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
SalesperShareGrowth	0.90	0.97	4456.25		
达标或采用	Yes	Yes			Yes
SalesGrowthbyEquityValue	0.99	0.95	6803.03		
达标或采用	Yes	Yes			

营业收入增长率：根据销售或收益增长的前景区分股票。这一因素中最重要的描述是分析师预测的长期收益增长。它还包括过去五年的销售额和收入增长。

上表中汇报了规模因子下的三级因子：流通市值的自然对数和总市值。其中，流通市值的自然对数覆盖率为 100%，AR1 为 0.99，LL 提升为 23049.55；总市值覆盖率为 100%，AR1 为 0.99，LL 提升为 23049.55；根据评价标准，流通市值的自然对数达到了完备性、连续性和解释能力的标准；总市值达到了完备性、连续性的标准。最终我们选择了流通市值的自然对数因子。和市值因子、增长因子类似，Hou（2021）、Chen（2023）和 Kelly 等（2020）等采用了该类因子，并取得了很好等效果。

4.3.8.市场情绪

4.3.8.1 分析师调整

表 15 分析师调整因子描述统计

Factor	描述				
RevisionRatio	分析师调整比率（近 3 个月分析师调整比率加权平均）预期数据缺失值填 0				
ChangeinPredictedEarningtoPrice	分析师预测 EP 比变化（近 4 个季度分析师预测 EP 变化的加权平均）				
ChangeinPredictedEarningperShare	分析师预测 EPS 变化（近 4 个季度分析师预测 EPS 变化的加权平均）				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
RevisionRatio	1	0.56	361.39		
达标或采用	Yes			Yes	
ChangeinPredictedEarningtoPrice	0.47	0.68	7261.57		
达标或采用				Yes	
ChangeinPredictedEarningperShare	0.52	0.73	6227.45		
达标或采用			Yes	Yes	Yes

上表中汇报了市场情绪因子下的三级因子：分析师调整比率（近 3 个月分析师调整比率加权平均）、分析师预测 EP 比变化（近 4 个季度分析师预测 EP 变化的加权平均）、分析师预测 EPS 变化（近 4 个季度分析师预测 EPS 变化的加权平均）。其中，分析师调整比率覆盖率为 100%，AR1 为 0.65，LL 提升为 361.39；分析师预测 EP 比变化为 47%，AR1 为 0.68，LL 提升为 7261.57；分析师预测 EPS 比变化为 52%，AR1 为 0.73，LL 提升为 6227.45；根据评价标准，分析师调整比率覆盖率达到到了完备性的标准。分析师预测 EPS 变化因子的解释能力提升最大，最终我们选择了分析师预测 EPS 变化因子。和预测因子类似，该类因子在因子模型中应用广发，如 Wasan 和 Kalyani（2020）、Hou 等（2021）等。

4.3.9.股息

4.3.9.1 股息率

表 16 股息率因子描述统计

Factor	描述				
DividendtoPrice	股息率（近 12 个月每股股息/股价）				
PredictedDividendtoPrice	分析师预测分红价格比（预期 12 个月每股股息/股价）				
PredictedDividendtoPriceFilled	predict divided（combine first divided 12 month rolling mean）/price				
Factor	覆盖率	AR1	LL 提升	CUFM	IGSM
DividendtoPrice	1.00	0.93	4371.52		
达标或采用	Yes			Yes	
PredictedDividendtoPrice	0.51	0.96	3974.21		
达标或采用		Yes		Yes	
PredictedDividendtoPriceFilled	1.00	0.95	5758.64		
达标或采用	Yes	Yes	Yes		Yes

上表中汇报了规模因子下的三级因子：股息率（近 12 个月每股股息/股价）、分析师预测分红价格比（预期 12 个月每股股息/股价）、predict divided（combine first divided 12 month rolling mean）/price。其中，股息率覆盖率为 100%，AR1 为 0.93，LL 提升为 4371.52；分析师预测分红价格比覆盖率为 51%，AR1 为 0.96，LL 提升为 3974.21；predict divided 覆盖率为 100%，AR1 为 0.95，LL 提升为 5758.64；根据评价标准，股息率达到了完备性的标准；分析师预测分红价格比达到了连续性的标准；predict divided 达到了完备性、连续性的标准。最终我们选择了 predict divided 因子。和我们类似，Penman 和 Zhu（2019）采用了股息因子。

5. 模型风险测试

本节主要内容为对于我们模型的风险控制能力的检验结果。本文选取三个指标测试模型的风险控制能力：组合风险管理（RMV）、风险价值（VaR）和波动率的波动率（VoV），并且基于三个指标与传统因子模型的风险控制对比。本节选取全部中国上市公司股票（3725 家）2012 年 10 月底至 2022 年 10 月底进行各个指标计算、估计。

5.1. 个股风险

首先，基于中国全部上市公司数据，我们分别估计了每只股票的 VaR、RMA 和 VoV，描述性统计结果报告于表 17 至表 19 以及图 5 至图 7。其中，依次为 VaR、RMA 和 VoV 的全部中国上市公司股票（3725 家）2012 年 10 月底至 2022 年 10 月底数据计算估计结果的描述统计结果。

表 17 个股风险 VaR

index	CUFM	IGSM
count	3725	3723
mean	2.991	3.018
std	1.031	0.761
min	0.67	0.653
25%	2.384	2.546
50%	2.741	2.876
75%	3.274	3.313
max	14.042	10.912

表 17 报告了全部中国上市公司股票（3725 家）2012 年 10 月底至 2022 年 10 月底数据，根据 CUFM 以及我们的模型估计 99%VaR 的描述统计结果。一般来说，其越接近 2.58 意味着模型更接近正太分布。从均值来说，我们模型估计出的 99%VaR 为 2.991，对比 CUFM 的 3.018，我们更为接近标准正太分布的 2.58。所以相比 CUFM 而言，我们模型估计后数据更为符合一般数据估计的假设。从方差来说，我们模型估计出的 99%VaR 方差为 1.031，其略高于 CUFM 的 0.761；我们模型估计出的 99%VaR 的极小值、25%、50%、75%的分位数、极大值分别为 0.670、2.384、2.741、3.274 和 14.042，CUFM 模型估计出的 99%VaR 的极小

值、25%、50%、75%的分位数、极大值分别为 0.653、2.546、2.876、3.313 和 10.912。

表 18 个股风险 RMV

index	CUFM	IGSM
count	3725	3723
mean	0.1	0.105
std	0.021	0.014
min	0.043	0.045
25%	0.089	0.096
50%	0.096	0.103
75%	0.106	0.111
max	0.285	0.271

表 18 报告了全部中国上市公司股票（3725 家）2012 年 10 月底至 2022 年 10 月底数据，根据 CUFM 以及我们的模型估计以 10%为目标 RMA 的描述统计结果。一般来说，其越接近 0.1 意味着模型更风险控制能力越强。从分布来看，我们模型估计出的 RMA 的极小值、25%、50%、75%的分位数、极大值分别为 0.045、0.096、0.103、0.111、0.271，CUFM 模型估计出的 RMA 的极小值、25%、50%、75%的分位数、极大值分别为 0.043、0.089、0.096、0.106、0.285，我们的分布更为集中在 0.1 附近。所以相比 CUFM 而言，我们模型对于个股风险控制更强。我们和 CUFM 的 RMA 均值接近，分别为 0.105 和 0.1；我们模型估计出的 RMA 方差为 0.021，其略高于 CUFM 的 0.014。

表 19 个股风险 VoV

index	CUFM	IGSM
count	326302	321576
mean	0.027	0.025
std	0.022	0.023
min	0	0
25%	0.012	0.01
50%	0.024	0.021
75%	0.038	0.035
max	0.516	0.454

表 19 报告了全部中国上市公司股票（3725 家）2012 年 10 月底至 2022 年 10 月底数据，根据 CUFM 以及我们的模型估计 VoV 的描述统

计结果。一般来说，其越小意味着模型风险控制越好。对比结果来看，我们模型估计出的 VoV 为 0.025，对比 CUFM 的 0.027，我们模型的平均 VoV 更小。所以相比 CUFM 而言，我们模型对于风险控制更为卓越。我们模型估计出的 VoV 方差为 0.023，其略高于 CUFM 的 0.022；我们模型估计出的 VoV 的极小值、25%、50%、75%的分位数、极大值分别为 0、0.01、0.021、0.035、0.454，CUFM 模型估计出的 VoV 的极小值、25%、50%、75%的分位数、极大值分别为 0、0.012、0.024、0.038、0.516。

5.1.2. 个股风险分布

图 5 至图 7 分别绘制了全部中国上市公司股票（3725 家）2012 年 10 月底至 2022 年 10 月底估计的 VaR、RMV、VoV 个股分布图。

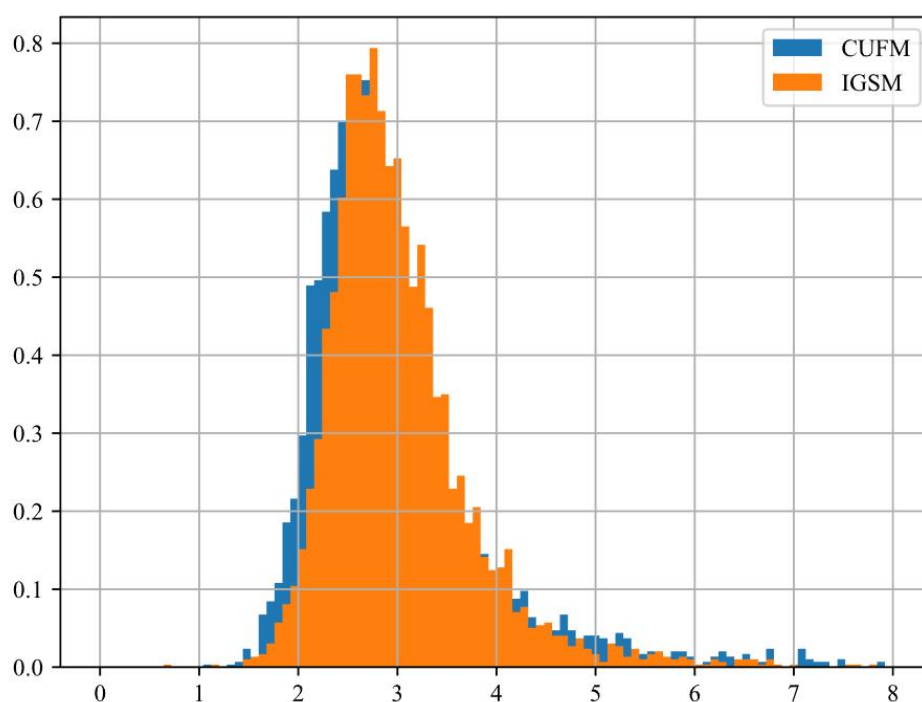


图 5 VaR 分布对比

图 5 为我们模型和 CUFM 对全部中国上市公司股票 99%VaR 的估计分布图，可以看出我们 VaR 的分布相对 CUFM 更为接近 2.58，这说明我们估计结果更为合理。这也和表 17 得出的结论一致。同时我们模型估计的 VaR 相对左偏。

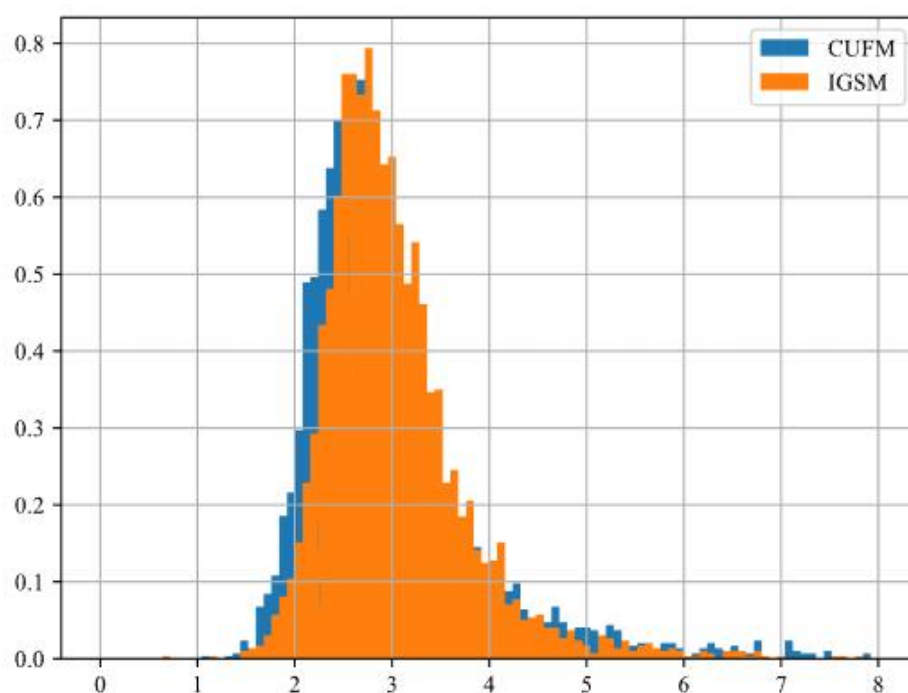


图 6 RMV 分布对比

图 6 为我们模型和 CUFM 对全部中国上市公司股票目标 10% 的 RMV 估计分布图,可以看出我们 RMV 的分布相对 CUFM 更为接近 0.1,这说明我们估计结果更为合理。这也和表 18 得出的结论一致。同时我们模型估计的 VaR 相对左偏。

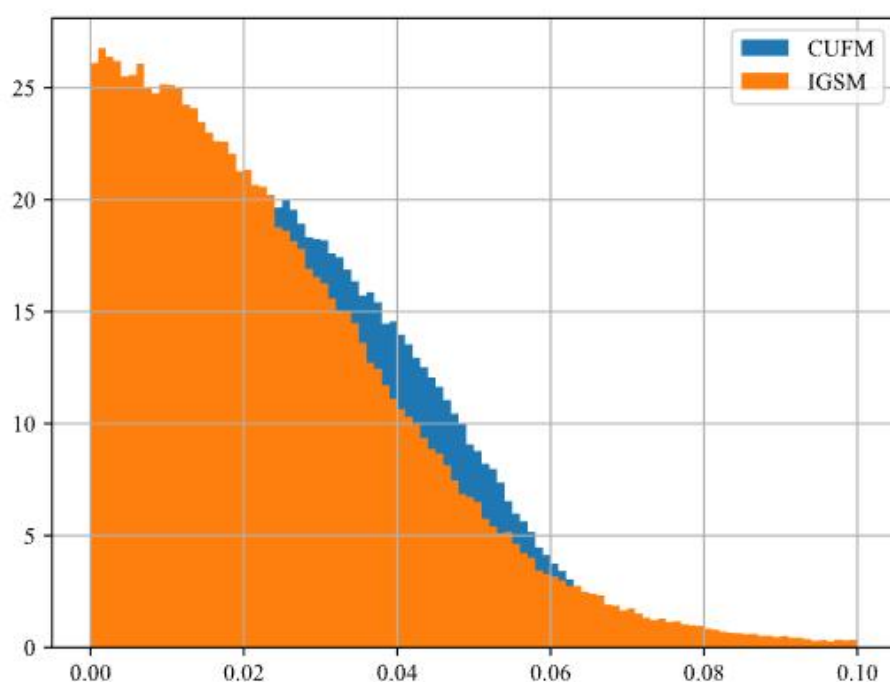


图 7 VoV 对比

图 7 为我们模型和 CUFM 对全部中国上市公司股票 VoV 的估计分布图，可以看出我们 VoV 的分布相对 CUFM 更为接近 0，这说明我们估计结果方差接近 0，这意味着我们的模型方差更为稳定。这也和表 19 得出的结论一致。同时我们模型估计的 VoV 相对左偏。

5.1.3. 个股风险管理特例分析

本节选取了几个中国股市近年来有代表性的股票、事件对我们的模型进行测试。特别的，本文选取了“獐子岛”、“贵州茅台”、“万科”和“康美药业”这几只发生重大事件的股票测试模型的风险预测能力。

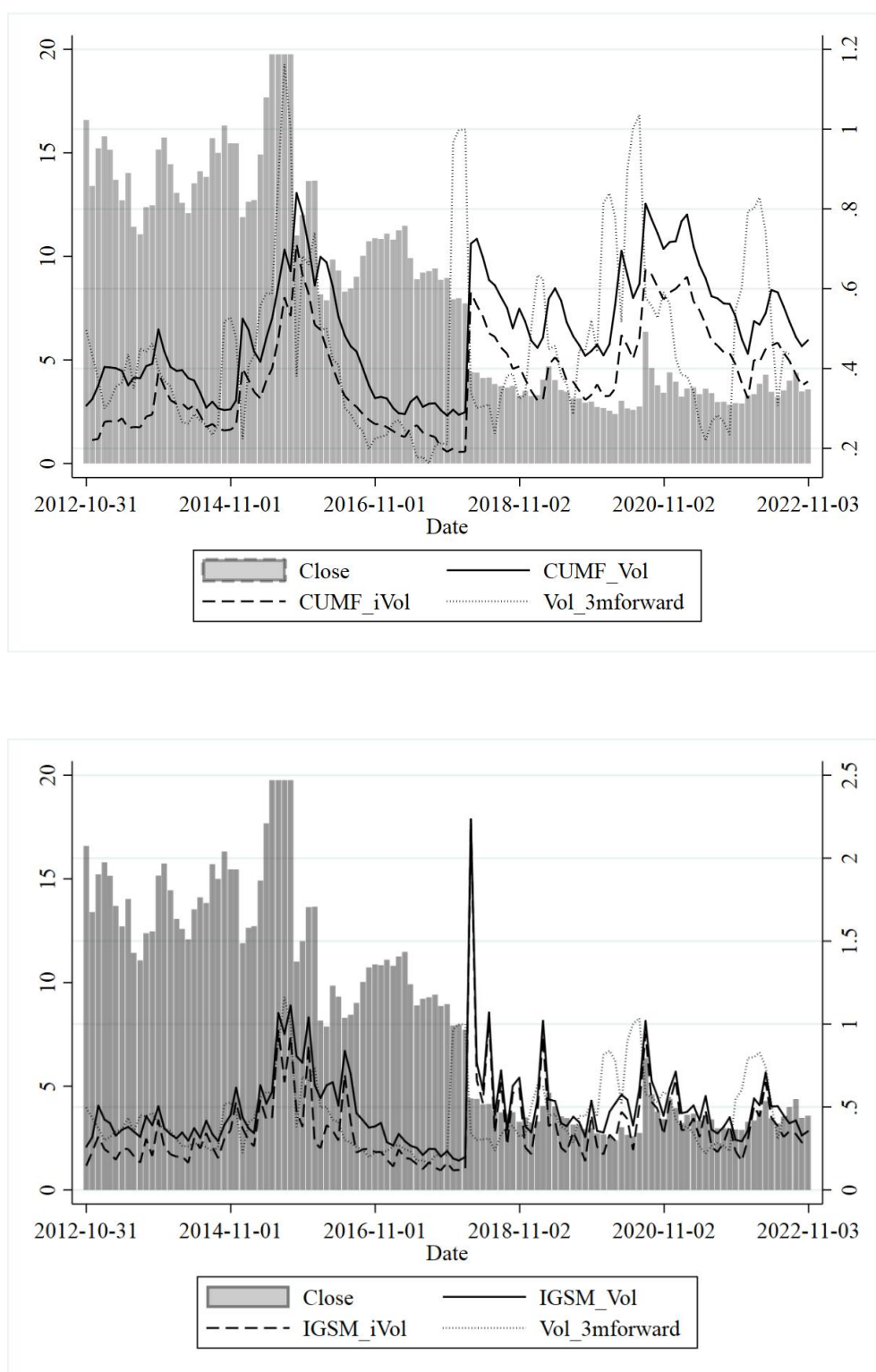


图 8 獐子岛—特异性风险

图 8 报告獐子岛在“扇贝跑了”事件中的表现以及我们模型和 CUFM 模型对于獐子岛特异性风险的折线图。2014 年 10 獐子岛发布公

告称异常的冷水造成了獐子岛的虾夷扇贝绝收;当年净利润亏损 11.95 亿元。2018 年 1 月獐子岛突发公告, 声称 2017 年降水减少, 导致饵料短缺, 再加上海水温度异常, 大量扇贝饿死。2017 年业绩大跌, 亏损 7.23 亿。2019 年 4 月、10 月, 2020 年 5 月, 獐子岛扇贝又发生多次“集体死亡/逃跑”。2020 年 6 月 24 日证监会依法对獐子岛公司信息披露违法违规案作出行政处罚及市场禁入决定。

图 8 中, 图中下面部分对应为我们模型的结果、上面部分对应 CUFM 的结果。对比两部分可以看出, 首先由于 CUFM 模型的估计的假设, 特异性波动率与总波动率之间的差距变化较小, 而我们的模型两者的差距在“扇贝跑了”事件前加大, 这暗示我们的模型具有一定的风险预测能力, 这一点从 3 个月的预期波动率也能得到类似的结果。

图 9 报告万科在“宝万之争”事件中的表现以及我们模型和 CUFM 模型对于万科特异性风险的折线图。2015 年 7 月 24 日, 宝万之争拉开序幕。2016 年 12 月 17 日, 恒大退出, 宝万之争结束。

图 9 中, 图中下面部分对应为我们模型的结果、上面部分对应 CUFM 的结果。对比两部分可以看出和图 8 类似的结果, 同样由于 CUFM 模型的估计的假设, 特异性波动率与总波动率之间的差距变化较小, 而我们的模型两者的差距在“宝万之争”事件前加大, 这暗示我们的模型具有一定的风险预测能力, 这一点从 3 个月的预期波动率也能得到类似的结果。同时, 在事后的波动率 IGSM 的估计结果也表现出更大的反应, 这都暗示了我们模型的对于风险的反应更显著。

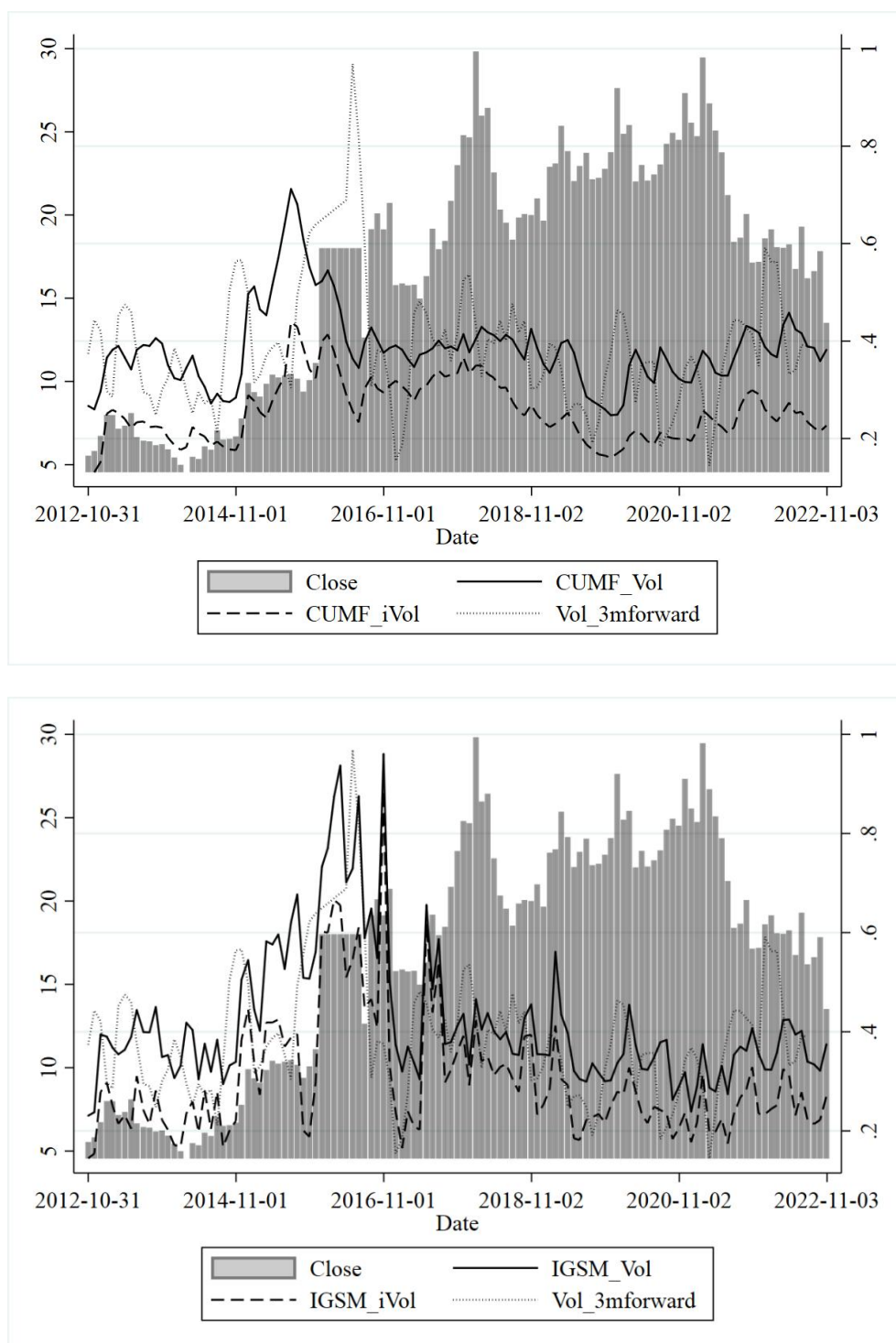


图 9 万科停牌

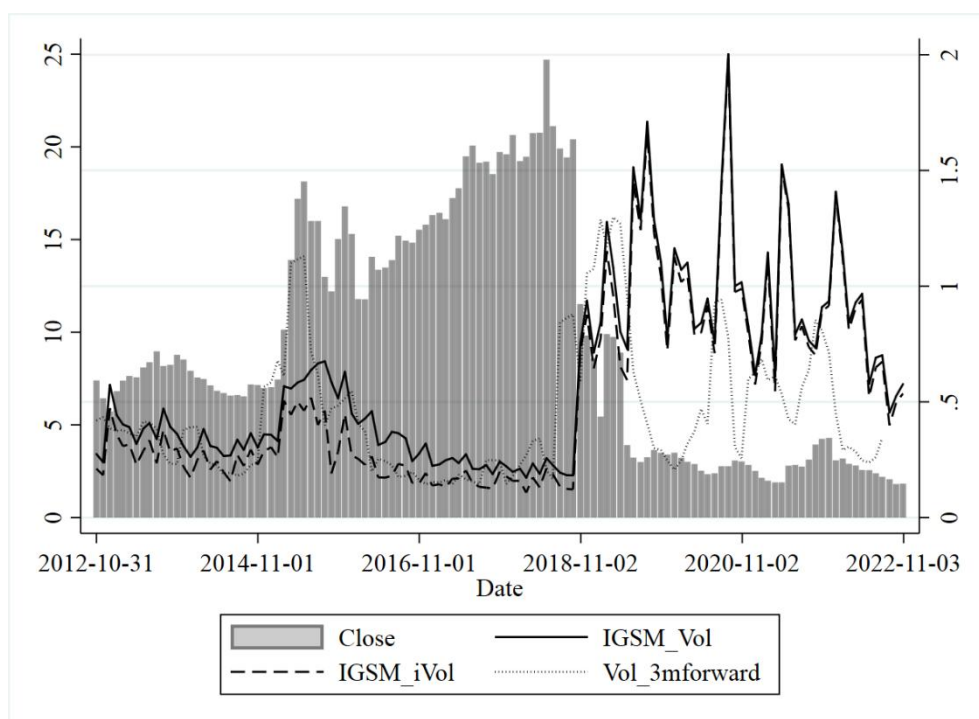
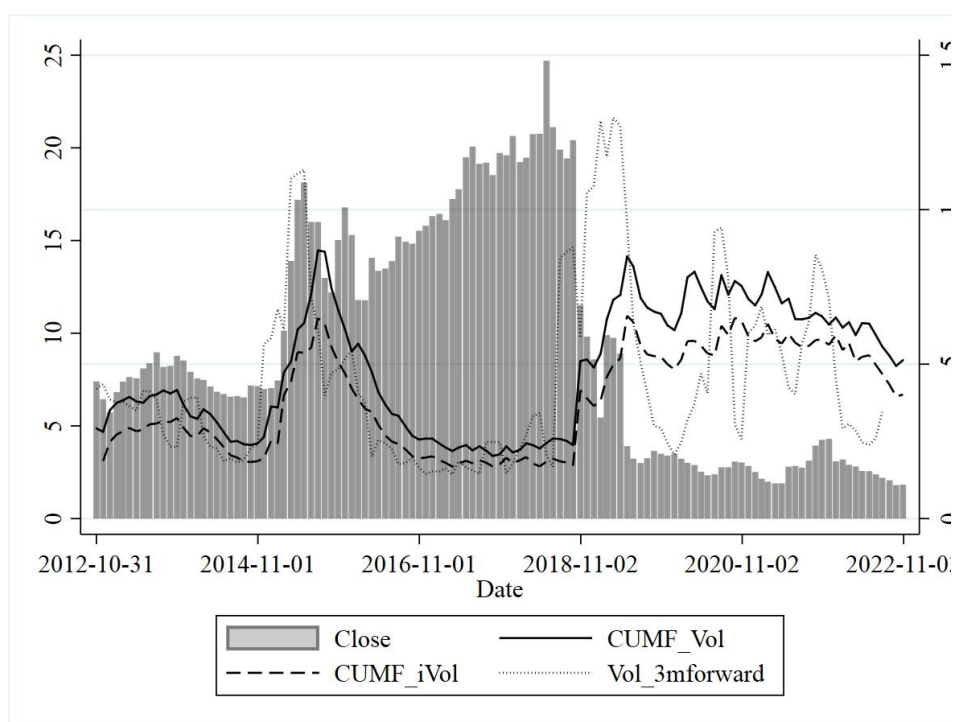


图 10 康美药业

图 10 报告了康美药业股价的表现以及们模型和 CUFM 模型对于特异性风险的折线图。2018 年 10 月 16 日，康美药业在盘中突然跌停，17 日再度闪崩跌停，17 日之后的 4 个工作日，康美市值迅速被腰斩。2020 年 5 月 13 日，中国证监会对康美药业证券违法违规案作出行政处罚和市场禁入决定，对公司和 21 名责任人处以罚款和市场禁入决定。

图 10 中，图中下面部分对应为我们模型的结果、上面部分对应 CUFM 的结果。对比两部分可以看出，同样在事件期间，相比 CUFM 模型的波动率估计结果，我们模型估计的总波动率、特异性波动率，在事前均可观察到更明显的上升。CUFM 模型的特异性波动率几乎没有变化。这表明我们的模型具有一定的预测能力，这一点从 3 个月的预期波动率也能得到类似的结果。

图 11 报告了贵州茅台股价的表现以及们模型和 CUFM 模型对于特异性风险的折线图。图 11 中，图中下面部分对应为我们模型的结果、上面部分对应 CUFM 的结果。对比两部分可以看出，在大部分时期，我们模型所估计的系统性风险的波动会更明显。平均来看，我们模型所估计的系统性风险的波动为 25%（相对均值），而 CUFM 则是 21%（相对均值）。所以我们可以相信我们模型所估计出的风险根据信息，我们对于系统性风险、异质性风险的估计反应了更多信息信息。这些信息也能转化为对于高风险事件的预测能力。

同时，我们也对比了我们模型和 CUFM 模型在不同分数总风险的表现，其中 25%、50%、75%分位数的结果分别在图 12。描述统计在表 20 中。

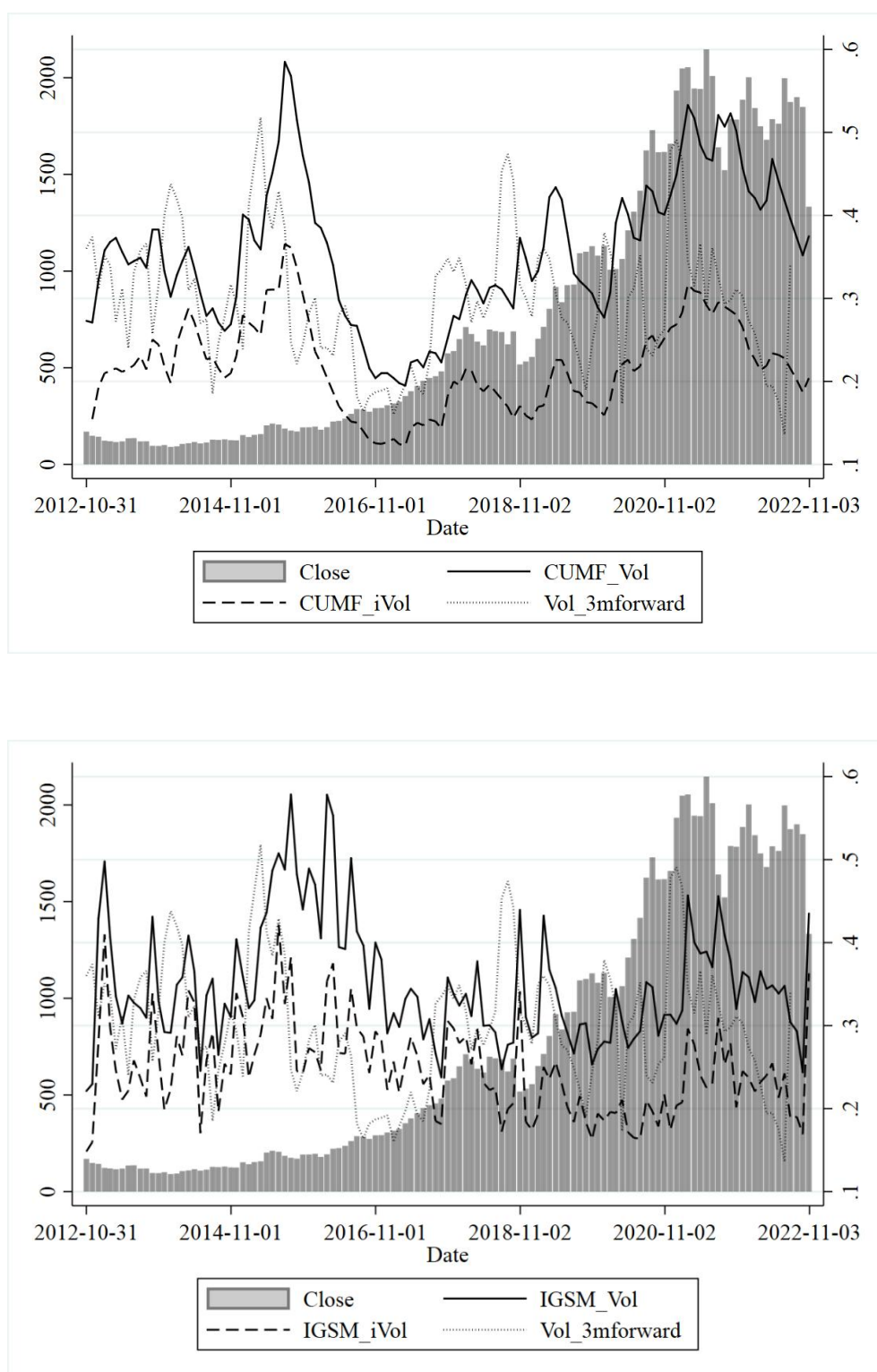


图 11 贵州茅台

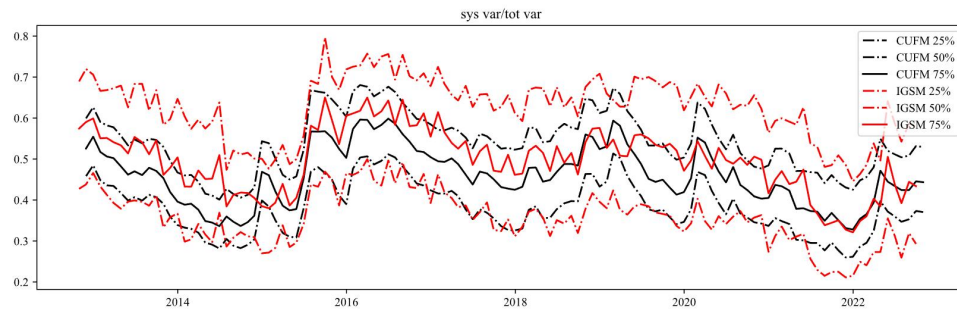


图 12 总风险

表 20 波动率描述统计

index	CUFM		IGSM	
	IDIO. VAR.	SYS. VAR.	IDIO. VAR.	SYS. VAR.
10%	4.7%	4.8%	2.3%	0.0%
25%	7.2%	6.9%	4.2%	0.0%
50%	11.8%	10.0%	8.6%	5.4%
75%	19.8%	14.4%	18.0%	9.2%
90%	30.8%	21.4%	38.2%	14.9%
mean	15.6%	12.8%	38.8%	6.9%

图 12 表明对于 75%分位数的股票，我们的模型能够估计出更低的波动率，对于 25%的股票则估计的波动率较高。表 20 展示了特异性风险和系统性风险在不同分位数的情况。系统性波动率和特异性波动率在分布表现上和总波动率一致，在较低分位数区域 CUFM 估计的结果较高，在较高分位数区域我们的结果较高。

5.2. 市场组合风险

本节我们选取了 30 个代表性的指数投资组合作为我们的测试对象，同样我们就 VaR、RMV、VoV 三个指标和 CUFM 模型进行了对比。结果汇总在表 21 至表 26，以及图 13 图 15。其中逐个指数组合的 VaR、RMV、VoV 三个指标的结果分别在表 21、表 22、表 23；VaR、RMV、VoV 三个指标的分布图在图 13 至图 15。

表 21 组合 RMV

指数简称	IGSM	CUFM	2022 年 ETF 规模	测试起始时间
沪深 300	10.0%	9.9%	47.16	2010-1
中证 500	9.8%	10.1%	12.01	2010-1
上证 50	10.4%	9.4%	17.56	2010-1
中证白酒	13.9%	11.1%	4.29	2015-01
中证 1000	10.4%	10.8%	11.34	2014-10
光伏产业	12.2%	13.2%	2.58	2019-04
中证军工	10.5%	11.4%	1.96	2014-04
CS 新能车	12.3%	12.7%	3.85	2014-12
中证红利	8.9%	9.3%	11.90	2012-12
CS 创新药	11.5%	11.2%	2.05	2019-04
300 成长	10.3%	9.3%	11.10	2012-11
全指医药	10.5%	10.7%	6.24	2012-11
中证农业	11.0%	11.3%	1.75	2015-07
全指金融	10.7%	8.9%	16.02	2012-11
CS 人工智	9.2%	9.6%	1.76	2015-08
全指信息	10.3%	10.8%	7.56	2012-11
细分化工	10.5%	10.7%	2.33	2012-11
新材料	11.3%	12.1%	4.03	2015-02
消费电子	10.4%	10.8%	2.78	2020-06
全指可选	9.7%	10.3%	6.46	2012-11
SHS 科技龙头	8.8%	9.8%	9.20	2020-07
全指能源	9.8%	10.0%	4.02	2012-11
全指材料	10.1%	10.5%	8.06	2012-11
300R 价值	10.2%	9.4%	37.35	2012-11
全指工业	10.0%	10.5%	17.23	2012-11
全指消费	11.0%	10.3%	7.78	2012-11
全指电信	9.2%	10.5%	4.18	2012-11
全指公用	10.5%	10.4%	2.71	2012-11
新华 500 指数	9.5%	10.2%	52.49	2019-11
500R 成长	9.9%	10.6%	9.46	2014-06

表 22 组合 RMV 描述统计

index	IGSM	CUMF
count	30	30
mean	10.4%	10.5%
std	1.1%	1.0%
min	8.8%	8.9%
25%	9.8%	9.9%
50%	10.4%	10.5%
75%	10.7%	10.8%
max	13.9%	13.2%

首先我们对比了组合风险管理能力，结果如表 21 所示。我们设置风险目标为 10%，波动率越接近 10%则管理越有效。我们的模型在大市值指数组合中表现出优势，如沪深 300、中证 500、上证 50 等，在部分行业指数中，我们的模型也表现更好，比如光伏产业、中证军工、CS 新能源车等。值得一提的是，我们的模型在某些行业指数中占据明显优势，比如新材料等。

表 22 报告了表 21 结果的汇总，从 30 个投资组合方差的均值来看，我们的模型对比传统因子能多控制 0.1%的波动率。我们模型的 RMV 方差和 CUFM 基本一致。从分布来看，我们的模型估计的 RMV 分布更为离散。同时，图 13 也能看看出我们的 RMV 接近 10%的组合略多，也更为左偏。

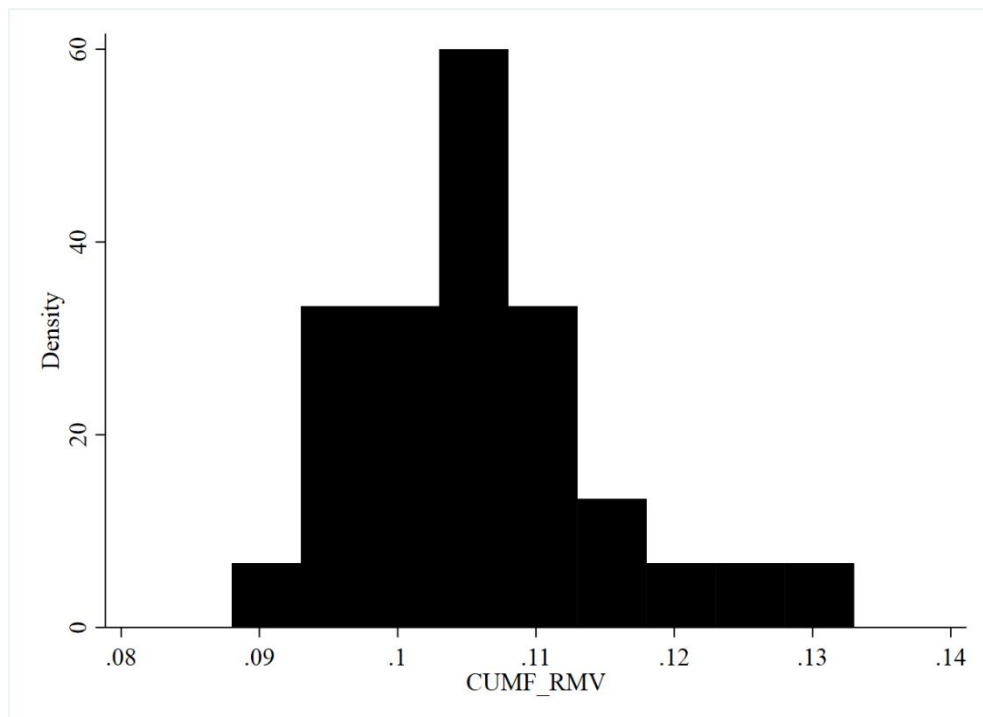
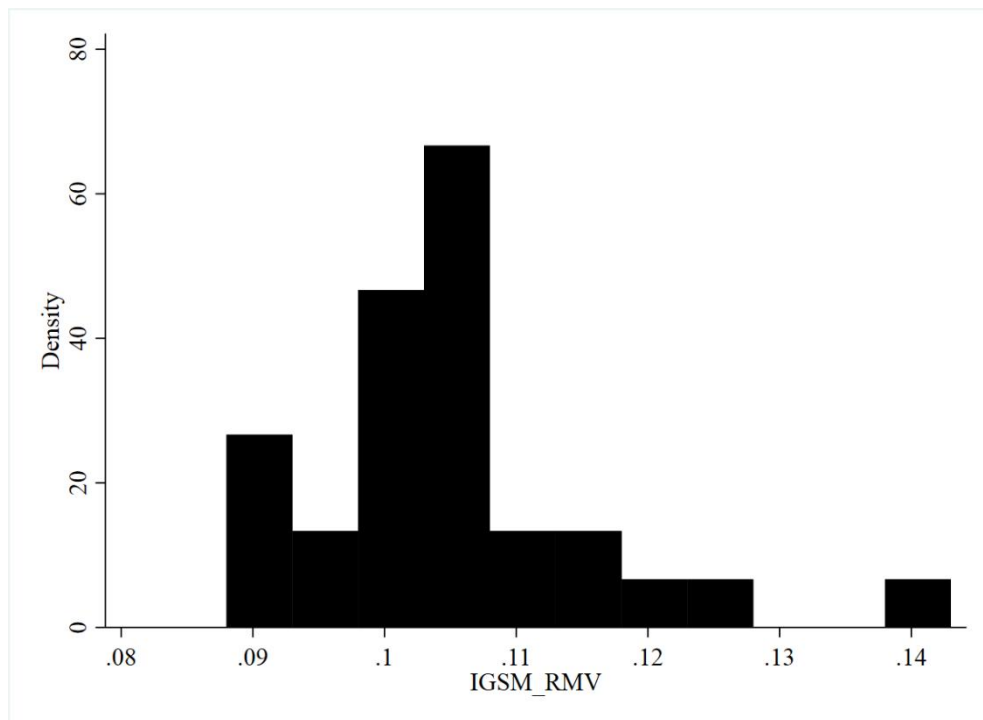


图 13 组合 RMV 分布

表 23 组合 VaR

指数简称	IGSM	CUFM
上证 50	3.44	2.59
沪深 300	2.80	2.39
细分化工	2.92	3.42
中证 1000	2.95	2.83
中证 500	2.61	2.97
300 成长	3.26	2.61
300R 价值	3.20	2.54
中证红利	2.36	2.55
中证农业	2.97	2.94
全指能源	2.65	2.78
全指材料	2.62	3.06
全指工业	2.64	2.45
全指可选	2.76	2.88
全指消费	3.71	2.68
全指医药	2.85	2.91
全指金融	3.23	2.31
全指信息	2.71	2.83
全指电信	2.59	3.06
全指公用	2.69	3.30
中证军工	2.87	3.33
CS 新能车	2.65	3.54
中证白酒	4.60	2.72
CS 人工智	2.70	2.81
光伏产业	2.67	2.97
CS 创新药	2.86	2.67
消费电子	2.63	2.74
SHS 科技龙头	2.23	2.42
新华 500 指数	1.84	2.13
500R 成长	2.76	2.99
新材料	2.98	3.08

表 24 组合 VaR 描述统计

index	IGSM	CUFM
count	30	30
mean	2.86	2.82
std	0.48	0.33
min	1.84	2.13
25%	2.65	2.60
50%	2.76	2.82
75%	2.97	2.98
max	4.60	3.54

接下来我们对比了组合风险的 VaR，结果如表 23 所示。我们选取的 99% 的 VaR，VaR 越接近 2.58 则越合理。VaR 的结果暗示我们和 CUFM 的模型估计结果比较右偏，大部分的组合结果都右偏于 2.58。在大部分的行业指数里，我们的模型的 VaR 更合理，比如全能能源等。值得一提的是，我们的模型在某些行业指数中占据明显优势，比如 CS 新能源车等。

表 24 报告了表 23 结果的汇总，从 30 个投资组合 VaR 的均值来看，我们的模型对比 CUFM 模型更为接近 2.58。我们模型的 VaR 方差更小。从分布来看，我们的模型估计的 VaR 分布更为集中。同时，图 14 也能看看出 IGSM 模型估计的 VaR 和 CUMF 的分布差距不大。这都暗示着我们模型在风险估计上的优势。

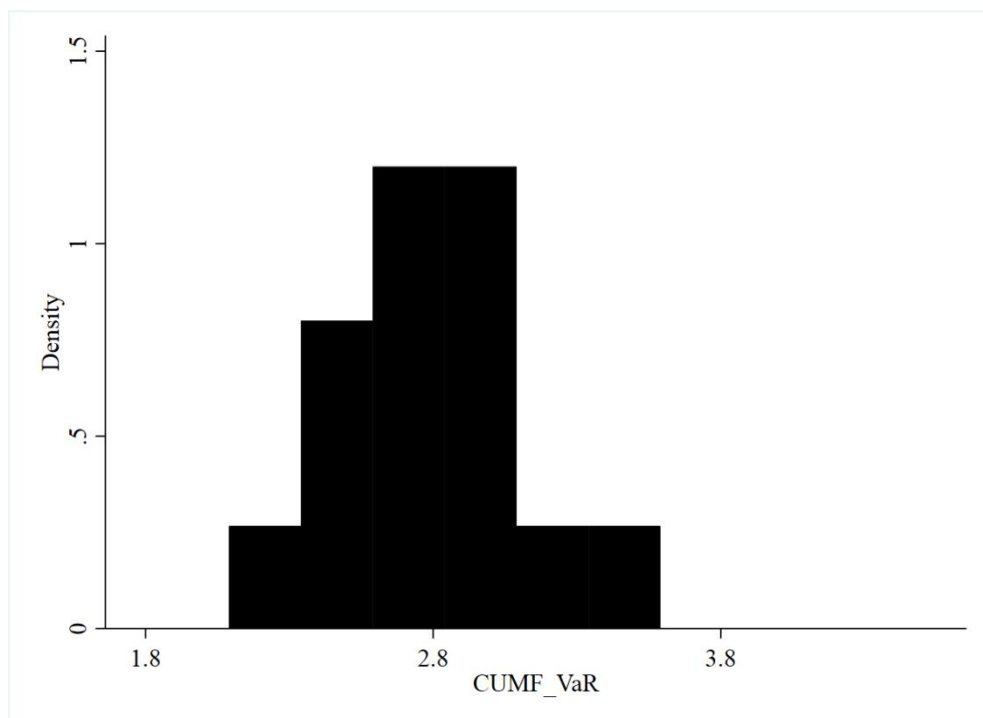
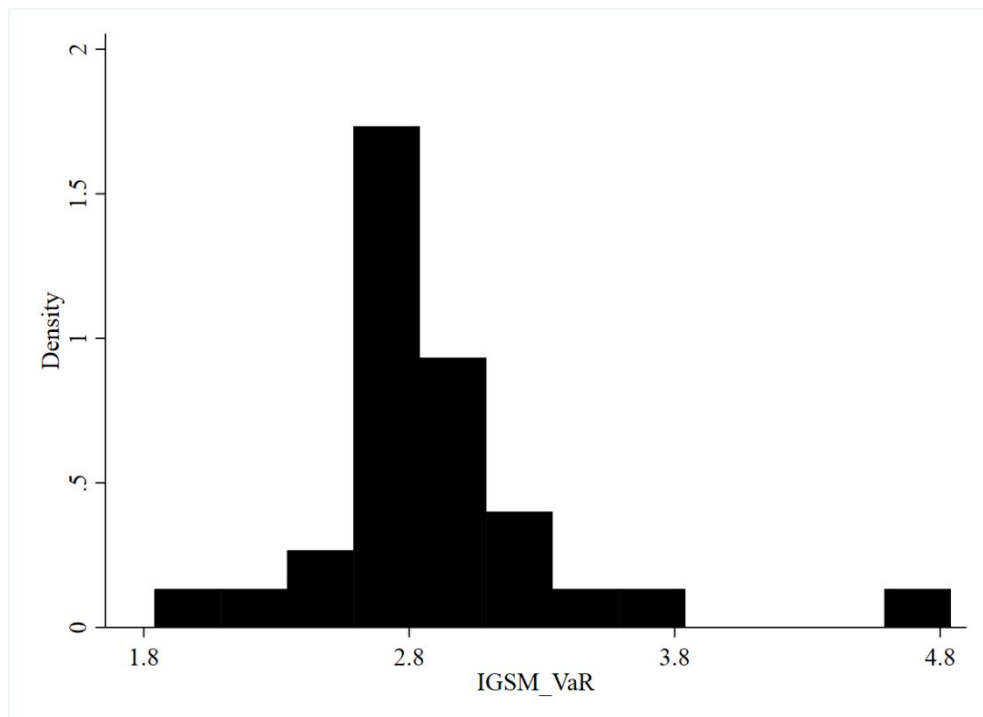


图 14 组合 VaR 分布

表 25 组合 VoV

指数简称	IGSM	CUFM
上证 50	2.44%	1.94%
沪深 300	2.43%	2.49%
细分化工	2.49%	3.14%
中证 1000	3.29%	3.85%
中证 500	2.62%	3.02%
300 成长	2.49%	1.84%
300R 价值	2.65%	2.22%
中证红利	1.83%	2.27%
中证农业	2.12%	2.48%
全指能源	2.41%	2.44%
全指材料	2.28%	2.99%
全指工业	2.82%	2.28%
全指可选	2.55%	2.98%
全指消费	2.70%	2.73%
全指医药	2.48%	2.90%
全指金融	2.72%	1.93%
全指信息	2.49%	3.06%
全指电信	1.94%	2.62%
全指公用	3.16%	3.18%
中证军工	2.28%	3.25%
CS 新能源车	1.97%	2.82%
中证白酒	3.07%	2.65%
CS 人工智	2.24%	2.84%
光伏产业	2.90%	2.89%
CS 创新药	2.15%	2.03%
消费电子	3.12%	2.77%
SHS 科技龙头	1.55%	1.39%
新华 500 指数	1.27%	1.41%
500R 成长	2.73%	3.45%
新材料	2.10%	3.05%

表 26 组合 VoV 描述统计

index	IGSM	CUFM
count	2684	2684
mean	2.46%	2.71%
std	1.87%	2.06%
min	0.00%	0.00%
25%	0.97%	0.98%
50%	2.10%	2.20%
75%	3.61%	4.15%
max	13.01%	10.94%

最后我们对比了组合风险的 VoV，结果如表 25 所示。从结果来我们估计的指数 VoV 较低，如沪深 300、中证 1000 等。表 26 报告了表 25 结果的汇总，从 30 个投资组合 VoV 的均值来看，我们的模型对比 CUFM 模型更小。我们模型的 VoV 波动率更小，这意味着我们均值相对较小这个结果，在指数之间相对稳健。从分布来看，我们的模型估计的 VoV 分布更为集中。同时，图 15 也能看看出我们的 VoV 稍微左偏。

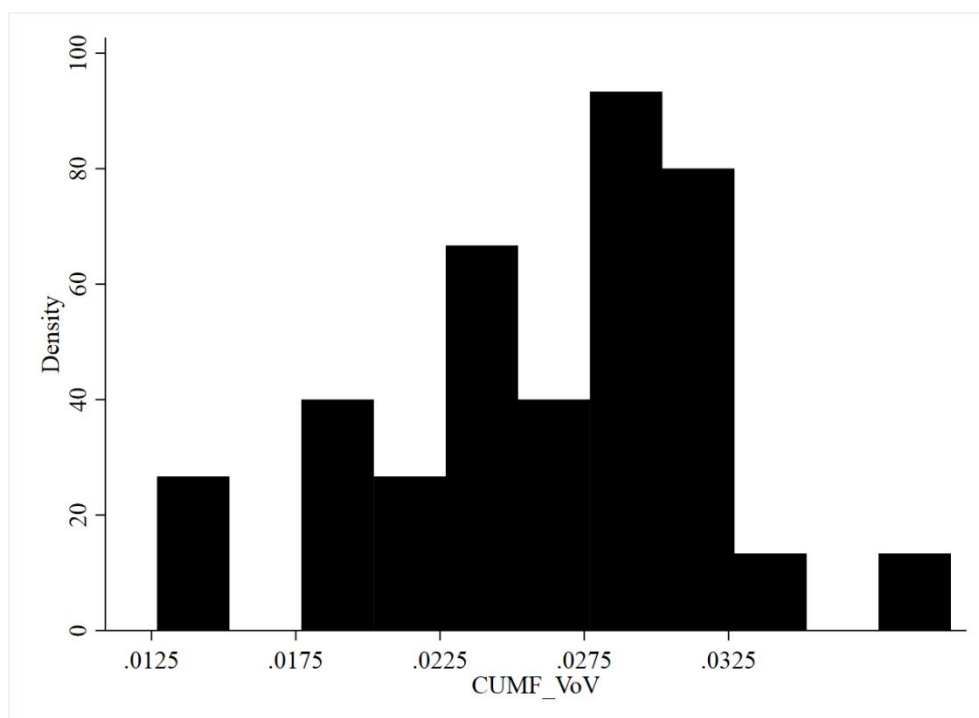
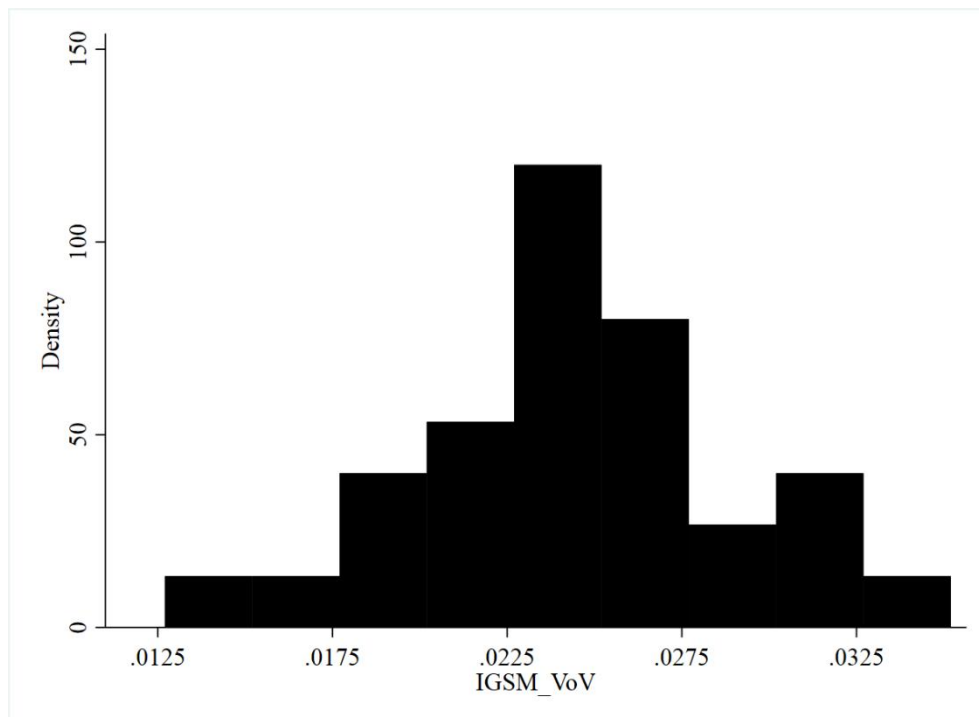


图 15 组合 VoV 分布

5.3. 因子组合风险

本节我们选取了 16 个 CUFM 模型以及 14 我们模型的的风格因子投资组合，分别检验了 VaR、RMV、VoV 三个指标下我们模型和 CUFM 模型异同。指数汇总的分布如图 16 至图 21 所示。其中图 16 至图 18 为 CUFM 模型风格因子组合的对比；图 19 至图 21 为我们模型风格因子组合的对比。

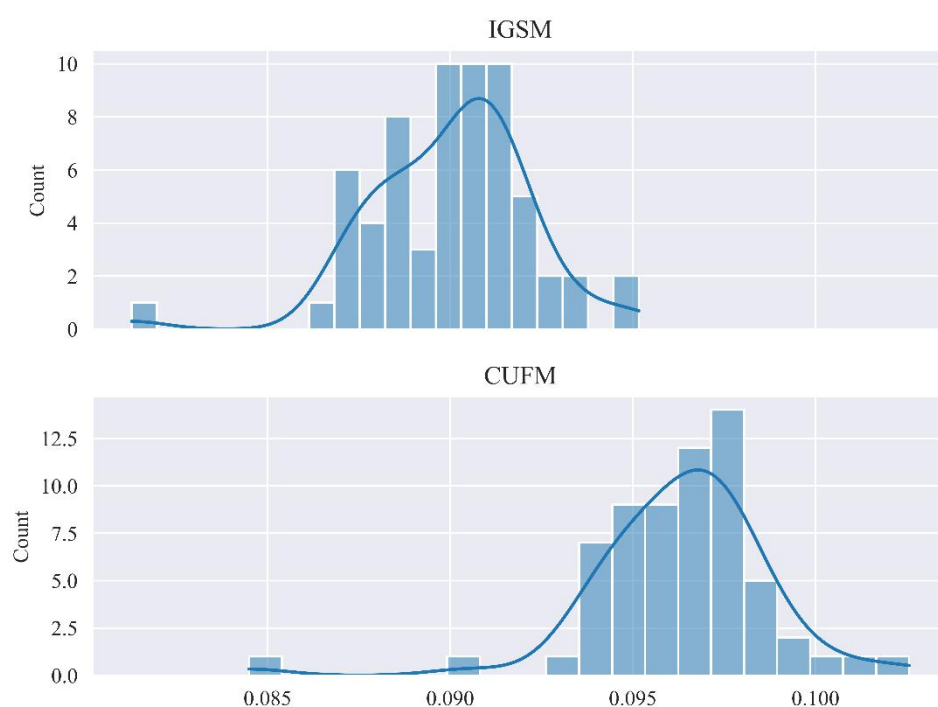


图 16 CUFM 因子组合 RMV

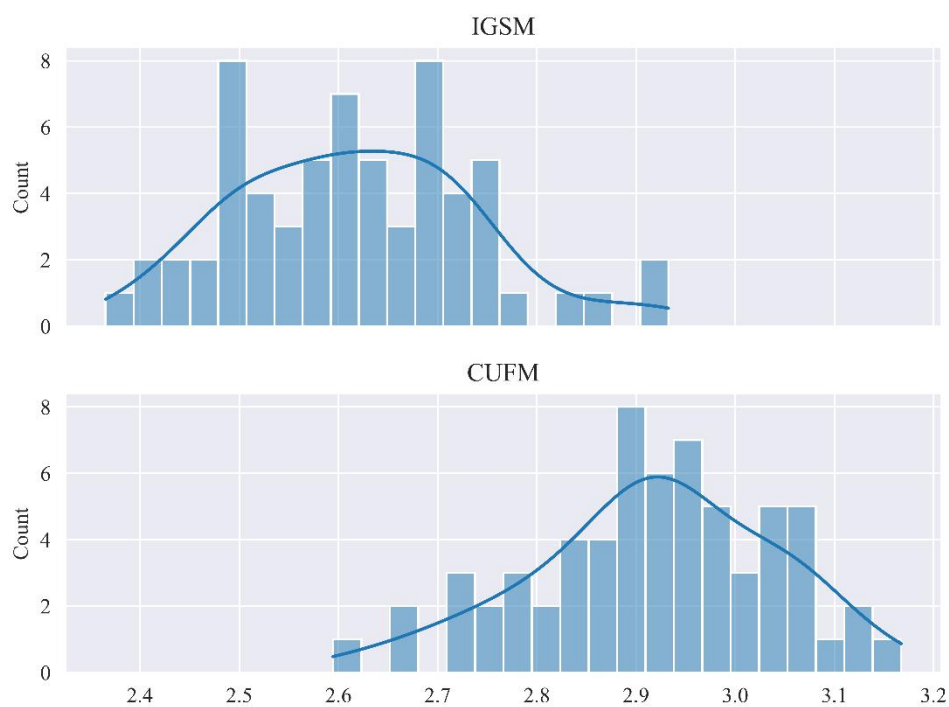


图 17 CUFM 因子组合 VaR

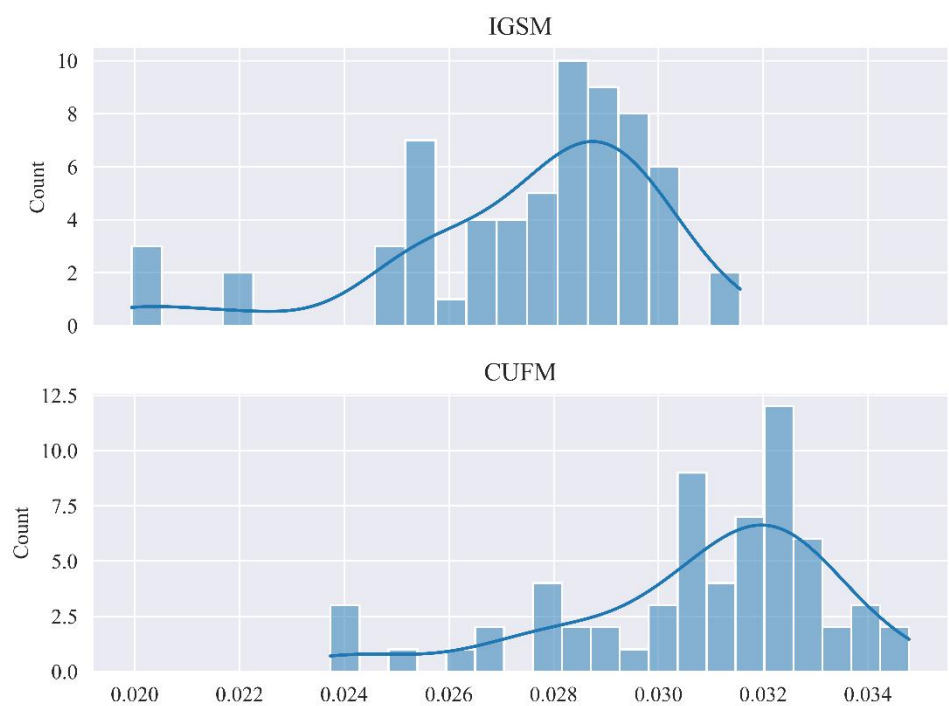


图 18 CUFM 因子组合 VoV

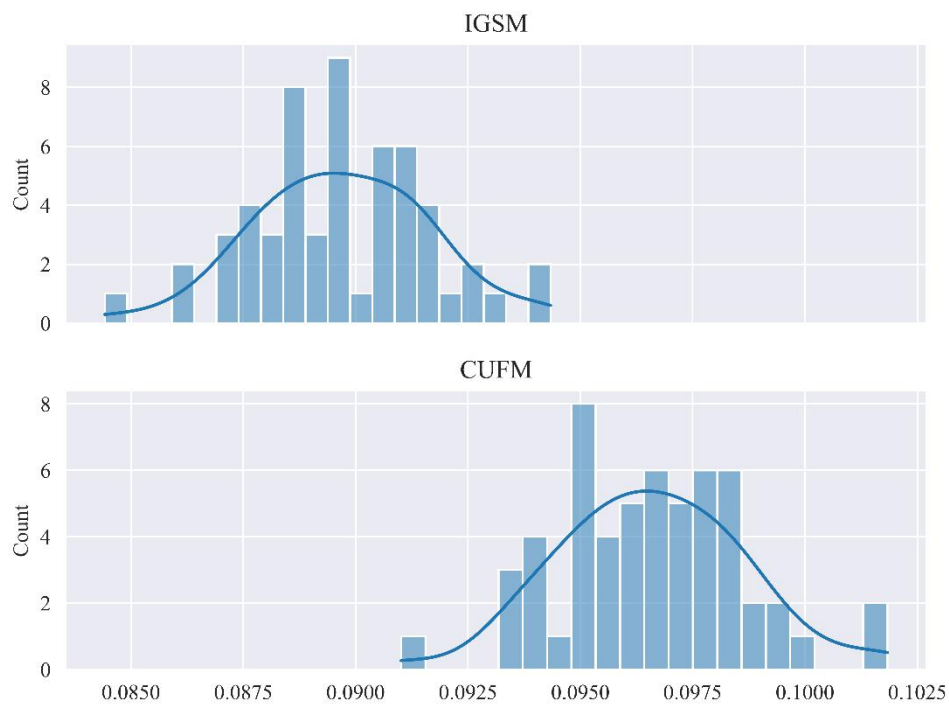


图 19 我们模型因子组合 RMV



图 20 我们模型因子组合 VaR

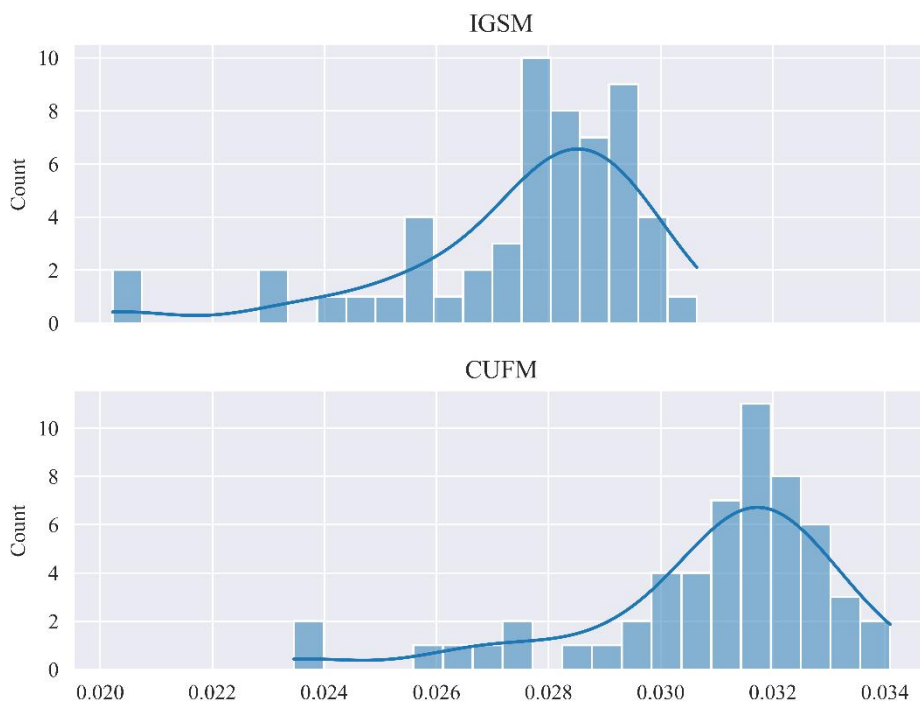


图 21 我们模型因子组合 VoV

图 16 和图 19 分别展示了我们模型和 CUFM 的风格因子组合 RMV 的分布，可以看出，我们的结果比较偏左，CUFM 模型比较偏右。从 VaR 来看，综合图 17 和图 20，我们模型 VaR 比较接近 2.58，这表明我们模型的因子组合的风险估计较为合理。图 18 和图 21 展示了我们模型和 CUFM 的风格因子组合的 VoV 分布。我们估结果 VoV 较小，这展现了我们模型更好的风险控制能力。

基于我们对于模型风格因子风险相关指标的测试，我们发现，IGSM 模型的风险预测、估计相比 CUMF 的结果更为合理。

参考文献

- Chen, Luyang, Markus Pelger, and Jason Zhu, ‘Deep Learning in Asset Pricing’, Management Science, 2023
- E. F. Fama and K. R. French, “Common risk factors in returns on stocks and bonds”, Journal of Financial Economics, vol. 33, no. 1, pp. 3–56, 1993.
- Gu, Shihao, Bryan Kelly, and Dacheng Xiu, ‘Autoencoder Asset Pricing Models’, Journal of Econometrics, 222.1 (2021), 429–50
- Harahap, Qahfi Romula Siregar Defi Desvita, ‘Influence Current Ratio, Debt to Equity Ratio and Total Asset Turnover on Return on Equity in the Transportation Sector Industry’, International Journal of Business Economics (IJBE), 2.2 (2021)
- Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, ‘An Augmented q -Factor Model with Expected Growth*’, Review of Finance, 25.1 (2021), 1–41
- Jegadeesh, Narasimhan, Joonki Noh, Kuntara Pukthuanthong, Richard Roll, and Junbo Wang, ‘Empirical Tests of Asset Pricing Models with Individual Assets: Resolving the Errors-in-Variables Bias in Risk Premium Estimation’, Journal of Financial Economics, 133.2 (2019), 273–98
- Kelly, Bryan, Seth Pruitt, and Yinan Su, ‘Instrumented Principal Component Analysis’, Working Paper, 2020
- Kelly, Bryan T., Tobias J. Moskowitz, and Seth Pruitt, ‘Understanding Momentum and Reversal’, Journal of Financial Economics, 140.3 (2021), 726–43
- Koijen, Ralph SJ and Yogo, Motohiro, ‘Exchange Rates and Asset Prices in a Global Demand System’, Working Paper, 2020
- Lalwani, Vaibhav, and Madhumita Chakraborty, ‘Multi-Factor Asset Pricing Models in Emerging and Developed Markets’, Managerial Finance, 46.3 (2019), 360–80
- Menchero, J. and J. -H. Lee, “Efficiently combining multiple sources of alpha”, Journal of Investment Management, vol. 13, no. 4, pp. 71–86, 2015.
- MSCI, BAARA China A Total Marker Equity Model for Long-Term Investors, 2018.
- Penman, Stephen H., and Julie Zhu, ‘An Accounting-Based Asset Pricing Model

and a Fundamental Factor', SSRN Electronic Journal, 2019

POJANAVATEE, Sasipa, 'Tests of a Four-Factor Asset Pricing Model: The Stock Exchange of Thailand', The Journal of Asian Finance, Economics and Business, 7.9 (2020), 117–23

Stephen A Ross, "The arbitrage theory of capital asset pricing", Journal of Economic Theory, vol. 13, no. 3, pp. 341-360, 1976.

Sheng, Jinfei, 'Asset Pricing in the Information Age: Employee Expectations and Stock Returns', Working Paper, 2022

Wasan, Pratibha, and Kalyani Mulchandani, 'Corporate Governance Factors as Predictors of Earnings Management', Journal of General Management, 45.2 (2020), 71–92

邓瓔函，BAARA 中国市场模型（CNE6）解读，西南证券，2019 年 2 月。

石川，《因子投资：方法与实践》，电子工业出版社，2020 年 9 月。

宋旻，BAARA 风险模型（CNE6）之纯因子构建与因子合成 ——多因子模型研究系列之九，渤海证券研究所，2019 年 6 月。

附录 1

若向量 x_a, x_b 服从联合正态分布，则

$$E(x_a|x_b) = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$

$$Var(x_a|x_b) = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

证明：

n 维正态分布的密度函数为

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right\}$$

其指数部分为

$$-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) = -\frac{1}{2}x^T \Sigma^{-1}x + x^T \Sigma^{-1}\mu + const \quad ①$$

令 $x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$

其协方差矩阵为 $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$ ，精度矩阵 $\Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$

因为 Σ^{-1} 为对称矩阵， Λ 也为对称矩阵，则 $\Lambda_{aa}, \Lambda_{bb}$ 为对称矩阵， $\Lambda_{ab} = \Lambda_{ba}^T$

给定 $x_b, x_a|x_b$ 服从正态分布， $f(x_a|x_b)$ 的指数部分为

$$\begin{aligned} & -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \\ &= -\frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}(x_b - \mu_b) \\ & \quad - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{ba}(x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{bb}(x_b - \mu_b) \\ &= -\frac{1}{2}x_a^T \Lambda_{aa}x_a + x_a^T [\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)] + const \quad ② \end{aligned}$$

由①②的变量系数对应可得

$$\Sigma_{a|b} = \Lambda_{aa}^{-1} \quad ③$$

$$\mu_{a|b} = \Sigma_{a|b}[\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)] = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b) \quad ④$$

由求逆公式

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}, M = (A - BD^{-1}C)^{-1} \quad ⑤$$

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \quad ⑥$$

将⑥代入⑤得

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \quad (7)$$

$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \quad (8)$$

将⑦⑧代入④得

$$E(x_a|x_b) = \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$

将⑦代入③得, $Var(x_a|x_b) = \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

附录 2

1. 高斯分布的矩阵表示法

对于因子模型

$$r = \beta\lambda + \varepsilon$$

先验分布:

$$\lambda \sim N(0, \Sigma_\lambda)$$

$$\varepsilon \sim N(0, \Sigma_\varepsilon)$$

$$r \sim N(0, \beta\Sigma_\lambda\beta^T + \Sigma_\varepsilon)$$

r 与 λ 符合多元高斯分布:

$$\begin{bmatrix} \lambda \\ r \end{bmatrix} \sim N\left(\begin{bmatrix} \vec{0} \\ \vec{0} \end{bmatrix}, \begin{bmatrix} \Sigma_\lambda & \Sigma_\lambda\beta^T \\ \beta\Sigma_\lambda^T & \beta\Sigma_\lambda\beta^T + \Sigma_\varepsilon \end{bmatrix}\right)$$

根据附录 1, 后验分布:

$$\mu_{\lambda|r} = \mu_\lambda + \Sigma_\lambda\beta^T(\beta\Sigma_\lambda\beta^T + \Sigma_\varepsilon)^{-1}(r - \mu_r) = \Sigma_\lambda\beta^T(\beta\Sigma_\lambda\beta^T + \Sigma_\varepsilon)^{-1}r$$

$$\Sigma_{\lambda|r} = \Sigma_\lambda - \Sigma_\lambda\beta^T(\beta\Sigma_\lambda\beta^T + \Sigma_\varepsilon)^{-1}\beta\Sigma_\lambda^T$$

$$\lambda|r \sim N(\mu_{\lambda|r}, \Sigma_{\lambda|r})$$

$$\mu_{r|\lambda} = \mu_r + \beta\Sigma_\lambda^T(\Sigma_\lambda)^{-1}(\lambda - \mu_\lambda) = \beta\lambda$$

$$\Sigma_{r|\lambda} = \beta\Sigma_\lambda\beta^T + \Sigma_\varepsilon - \beta\Sigma_\lambda^T(\Sigma_\lambda)^{-1}\Sigma_\lambda\beta^T = \Sigma_\varepsilon$$

$$r|\lambda \sim N(\mu_{r|\lambda}, \Sigma_{r|\lambda})$$

2. EM 算法实现过程

首先，初始化参数 Σ_λ 和 Σ_ε 。

对于一组样本 $\{r^i, i = 1, 2, \dots, n\}$,

E-step: 计算 $p(\lambda^i|r^i; \Sigma_\lambda, \Sigma_\varepsilon)$

$$p(\lambda^i|r^i; \Sigma_\lambda, \Sigma_\varepsilon) = \frac{1}{(2\pi)^{k/2} |\Sigma_{\lambda|r^i}|^{1/2}} \times \exp\left(-\frac{1}{2}(\lambda^i - \mu_{\lambda|r^i})^T \Sigma_{\lambda|r^i}^{-1} (\lambda^i - \mu_{\lambda|r^i})\right)^{13}$$

M-step: 最大化 $l(\Sigma_\lambda, \Sigma_\varepsilon)$

$$l(\Sigma_\lambda, \Sigma_\varepsilon) = \log\left(\prod_{i=1}^n p(r^i; \Sigma_\lambda, \Sigma_\varepsilon)\right) = \sum_{i=1}^n \log(p(r^i; \Sigma_\lambda, \Sigma_\varepsilon))$$

令 $Q(\lambda)$ 为因子序列 λ 的后验概率密度函数，

$$Q(\lambda) = p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon)$$

$$\begin{aligned} \int Q(\lambda) \log \frac{p(r, \lambda; \Sigma_\lambda, \Sigma_\varepsilon)}{Q(\lambda)} d\lambda \\ &= \int p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon) \log \frac{p(r, \lambda; \Sigma_\lambda, \Sigma_\varepsilon)}{p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon)} d\lambda \\ &= \int p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon) \log \frac{p(r; \Sigma_\lambda, \Sigma_\varepsilon) p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon)}{p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon)} d\lambda \\ &= \int p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon) \log(p(r; \Sigma_\lambda, \Sigma_\varepsilon)) d\lambda \\ &= \log(p(r; \Sigma_\lambda, \Sigma_\varepsilon)) \int p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon) d\lambda \\ &= \log(p(r; \Sigma_\lambda, \Sigma_\varepsilon)) \end{aligned}$$

因此，

$$Q_i(\lambda^i) = p(\lambda^i|r^i; \Sigma_\lambda, \Sigma_\varepsilon)$$

$$\sum_{i=1}^n \log(p(r^i; \Sigma_\lambda, \Sigma_\varepsilon)) = \sum_{i=1}^n \int Q_i(\lambda^i) \log \frac{p(r^i, \lambda^i; \Sigma_\lambda, \Sigma_\varepsilon)}{Q_i(\lambda^i)} d\lambda^i$$

我们的目标转化为

¹³ $p(\lambda|r; \Sigma_\lambda, \Sigma_\varepsilon)$ 为 k 维数组， k 为因子个数。

$$\begin{aligned}
 & \operatorname{argmax}_{\Sigma_{\lambda}, \Sigma_{\epsilon}} \sum_{i=1}^n \int Q_i(\lambda^i) \log \frac{p(r^i, \lambda^i; \Sigma_{\lambda}, \Sigma_{\epsilon})}{Q_i(\lambda^i)} d\lambda^i \\
 &= \operatorname{argmax}_{\Sigma_{\lambda}, \Sigma_{\epsilon}} \sum_{i=1}^n \int Q_i(\lambda^i) \log \frac{p(r^i | \lambda^i; \Sigma_{\lambda}, \Sigma_{\epsilon}) p(\lambda^i; \Sigma_{\lambda}, \Sigma_{\epsilon})}{Q_i(\lambda^i)} d\lambda^i \\
 &= \operatorname{argmax}_{\Sigma_{\lambda}, \Sigma_{\epsilon}} \sum_{i=1}^n \int Q_i(\lambda^i) \log \frac{p(r^i | \lambda^i; \Sigma_{\epsilon}) p(\lambda^i; \Sigma_{\lambda})}{Q_i(\lambda^i)} d\lambda^i
 \end{aligned}$$

由 E-step，我们可以获得 $Q_i(\lambda^i)$ ，再将 $r^i | \lambda^i \sim N(\beta \lambda^i, \Sigma_{\epsilon})$ ， $\lambda^i \sim N(0, \Sigma_{\lambda})$ 的概率密度函数代入方程，分别对参数求偏导，使一阶导数等于 0，可解得 Σ_{λ} 和 Σ_{ϵ} 。

循环 E-step 和 M-step，直至参数收敛。

附录 3

本附录探讨 EM 算法下共线性的处理与方差的分配问题。

因子暴露矩阵 $\tilde{\beta}$ 是 $N \times (K + 1)$ 阶矩阵，假设 $\tilde{\beta}$ 的列秩为 K ，不失一般性地，假设 $\tilde{\beta}$ 从左往右前 K 列互相独立，因此我们有

$$\tilde{\beta} = \beta (I_K \quad \varphi) \quad (1)$$

其中， β 是 $\tilde{\beta}$ 的前 K 列， φ 是一个 K 维列向量。

保持资产的方差协方差矩阵不变，即

$$\beta \Sigma_{\lambda} \beta^T = \tilde{\beta} \tilde{\Sigma}_{\lambda} \tilde{\beta}^T \quad (2)$$

我们知道

$$\operatorname{Var}(r) = \beta \Sigma_{\lambda} \beta^T + \operatorname{Var}(\epsilon) \quad (3)$$

其中 r 是资产收益，可写为

$$r = \beta f^T + \epsilon \quad (4)$$

上式中 f 为因子收益，是一个 K 维列向量， ϵ 是残差项。如果我们可以保持加入共线性的国家因子之后残差不变，那么(3)式中的 $\operatorname{Var}(\epsilon)$ 就不

会改变的，则显然资产的方差协方差矩阵也是不变的，(2)式也就自然成立。因此，我们可以将(2)式的条件进一步明确为

$$\tilde{\beta} \tilde{f}^T = \beta f^T \quad (5)$$

现在详细讨论 β 与 f ，在我们的模型中，不包含共线性的因子可分为行业(industry)与风格(style)两类，对应的，我们将 β 与 f 写为

$$\beta = (\beta^{ind}, \beta^{style}) \quad (6)$$

$$f = (f^{ind}, f^{style}) \quad (7)$$

这里我们假设行业因子有 k 个，则 β^{ind} 与 f^{ind} 均为 $N \times k$ 维的矩阵，

而 β 与 f 均为 $N \times (K - k)$ 维的矩阵。在加入共线性的国家因子以后

$$\tilde{\beta} = (\beta^{ind}, \beta^{style}, 1) \quad (8)$$

这里 $\tilde{\beta}$ 最右列的1代表一个分量全为1的 K 维列向量。容易验证，满足

下式的因子收益 $\tilde{f}$ 可以使得(5)式成立

$$\tilde{f} = (f^{ind} - z, f^{style}, z) \quad (8)$$

上面(8)式中的 z 理论上可以是任意 K 维列向量，这里我们取全部行业因子收益的等权平均，即

$$z = \frac{\sum f^{ind}}{k} \quad (9)$$

可以验证在数学上，采用(9)式中 z 的 $\tilde{f}$ 等于 f 右乘一个 $K \times (K + 1)$ 维的矩阵，设为 M

$$\tilde{f} = fM \quad (10)$$

矩阵 M 由下面一系列等式给出

$$M = (A_{K \times K} \quad B_{K \times 1}) \quad (11)$$

式中下标均表示矩阵的维数， $A_{K \times K}$ 可写为

$$A_{K \times K} = \begin{pmatrix} A_{k \times k}^{ind} & 0_{k \times (K-k)} \\ 0_{(K-k) \times k} & I_{(K-k) \times (K-k)} \end{pmatrix} \quad (12)$$

$A_{k \times k}^{ind}$ 的值是

$$A_{k \times k}^{ind} = \begin{pmatrix} \frac{k-1}{k} & -\frac{1}{k} & -\frac{1}{k} & \dots & \dots \\ -\frac{1}{k} & \frac{k-1}{k} & -\frac{1}{k} & \dots & \dots \\ -\frac{1}{k} & -\frac{1}{k} & \frac{k-1}{k} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ -\frac{1}{k} & \dots & \dots & -\frac{1}{k} & \frac{k-1}{k} \end{pmatrix} \quad (13)$$

(11)式中的 $B_{K \times 1}$ 可写为

$$B_{K \times 1} = \begin{pmatrix} B^{ind} \\ B^{style} \end{pmatrix} \quad (14)$$

其中 B^{ind} 是一个分量全部为 $1/k$ 的 k 维列向量, B^{style} 是一个 $(K-k)$ 维的零向量。

在获得因子收益的变换方式以及 M 的取值以后, 由于 Σ_λ 是因子收益的方差协方差矩阵, 可以直接得到

$$\tilde{\Sigma}_\lambda = M^T \Sigma_\lambda M \quad (15)$$

发布机构



指数中心



金融科技实验室



财富引擎(北京)科技有限公司