

7461560-autoencoders-for-international-equity-markets

AVTORJI: Bo Yuan, University of Cambridge

POVZETEK: Raziskava obravnava latentne faktorske modele, kot so Principal Component Analysis (PCA), Instrumented PCA (IPCA) in autoencodere (AE), za objašnjanje vmesnih vrat državne akcij. Uporabljajo se 13 firmnih faktorski tem in 52 makroekonomske indikatorje, omejeni na obdobje od 1995 do 2024. Leto. Modeli, posebej linearna AE, podajajo učinkovito objašnjanje, vendar so njihova prediktivna moč izven skupa nekonzistentna. Linearna AE dosegla največjo prediktivno moč, s upoštevanjem teško merljivega 0.45% R^2 . IPCA se pokaže metodološko neuspešen v malih skupinah podatkov, kjer je njegova linearna predpostavka zahtevana. Modeli skupaj identificirajo vrednost, trenutno gibanje in rast GDE kot ključne faktorje za risiko.

METODA: Raziskava uporablja različne latentne faktorske modele: PCA, IPCA in različne AE. Podatki se sestavljajo iz 13 firmnih faktorskih tem in 52 makroekonomskih indikatorjev, omejenih na obdobje od 1995 do 2024. Leto. Modeli ocenjujejo vmesne vrate državne akcij. Linearna AE je najučinkovitejši, s 0.45% R^2 , vendar je njihova prediktivna moč izven skupa nekonzistentna. IPCA se ukvarja z časovno spreminjajočimi se faktorskimi lastnostmi.

KLJUČNE UGOTOVITVE:

- Linearna AE je najučinkovitejši model, s 0.45% R^2 izven skupa.
- IPCA se ukvarja z časovno spreminjajočimi se faktorskimi lastnostmi.
- Vse modele identificirajo vrednost, trenutno gibanje in rast GDE kot ključne faktorje za risiko.

POMEN ZA TRGOVANJE: Raziskava podaja učinkovite informacije za globalno akcijsko alokacijo in risiko upravljanje. Linearna AE je najučinkovitejši model, s 0.45% R^2 izven skupa, vendar je njihova prediktivna moč izven skupa nekonzistentna. Modeli skupaj identificirajo vrednost, trenutno gibanje in rast GDE kot ključne faktorje za risiko.

Autoencoders for International Equity Markets

Bo Yuan

University of Cambridge

Abstract

We examine latent factor models—Principal Component Analysis (PCA), Instrumented PCA (IPCA), and Autoencoders (AE)—to explain the cross-section of international equity index returns, using a novel group of characteristics (13 firm-level factor themes built from 153 underlying characteristics, and 52 macroeconomic indicators) spanning 1995–2024. Conditional models are powerful explainers, driving pricing errors (alphas) to near zero. However, this explanatory success does not translate into robust out-of-sample predictability: only the most parsimonious linear autoencoder achieves a consistently positive, albeit modest (up to 0.45%), predictive R^2 . We further document a previously unrecognized methodological failure of IPCA in small-N settings, where weak identification of the characteristic-loading mapping causes the model to collapse out-of-sample. Despite this divergence, interpretability analysis reveals a striking consensus: all models—linear and nonlinear—identify value, momentum, and GDP growth as the dominant drivers of risk exposure, offering guidance for taming the international factor zoo.

1 Introduction

A central question in international finance is the extent to which common factors explain the cross-section of country-level equity returns. Particularly, in recent decades, the globalization of financial markets and the proliferation of exchange-traded funds (ETFs) (Ben-David et al., 2017) and index funds have profoundly transformed the landscape of international asset allocation. The shift from active stock selection to passive, index-based strategies means that institutional investors increasingly make decisions at the country level, rather than the firm level. Answering this question is of first-order importance for global asset allocation, risk management, and macro-financial analysis. Yet, modern asset pricing methods, which have been studied on large, granular U.S. firm-level datasets (Gu et al., 2021), may face a distinct challenge when applied to international country indices—an asset class characterized by a small cross-section (small-N), high asset correlation, and a different data-generating process. The literature is defined by a central conflict between the theoretical ideal of a single, integrated global market and the empirical reality of partial segmentation (Fama and French, 2012), where local factors and country-level characteristics retain significant explanatory power. This has led to the discovery of a vast international “factor zoo” (Jensen et al., 2023), a high-dimensional array of predictors whose robustness is often unclear. The core challenge, therefore, is to identify the true latent factor structure that parsimoniously explains this complex, high-dimensional cross-section.

This paper mainly asks three questions: Can the latest generation of latent factor models price international indices, thereby explaining away anomalies? What does

their performance reveal about the critical trade-off between explanatory power and predictive ability in global markets? And finally, despite their varying complexity, do these models agree on the characteristics that drive the risk-return tradeoff?

To meet this challenge, the methodological toolkit has evolved significantly. Early prescribed factor models, such as global Fama-French, are interpretable but ad hoc. Data-driven methods like Principal Component Analysis (PCA) offered an objective way to extract latent factors, but the focus on variance rather than risk premia limited pricing ability. A significant advance came with Instrumented Principal Component Analysis (IPCA) from Kelly et al. (2019b), which connects latent factors to observable characteristics by modeling time-varying factor loadings. However, IPCA’s fundamental linearity assumption stands in contrast to the growing evidence that non-linearities are of first-order importance in asset pricing. Most recently, machine learning has introduced new paradigms, most notably the Conditional Autoencoder (CAE) of Gu et al. (2021), which can be viewed as a non-linear generalization of IPCA. While autoencoders have demonstrated remarkable success in pricing U.S. individual stocks, their performance in the international country-index setting is an open question.

We conduct a “horse race” between PCA, IPCA, and a suite of conditional autoencoder models, from a simple linear specification (CA0) to deeper non-linear architectures (CA1-CA3). Our analysis is grounded in a uniquely powerful dataset covering 37 developed and emerging market country indices from January 1995 to December 2024. To construct time-varying risk exposures, we combine two frontier comprehensive data sources: the Global Factor Data (GFD) library, which provides

13 equity factor themes built from firm-level characteristics, and the Global Macro Database (GMD), which supplies 52 country-specific macroeconomic indicators. This combination of 65 predictors allows us to model risk exposures using an exceptionally rich information set that captures both micro-level equity signals and macro-level economic conditions. We evaluate these models on two distinct and equally critical dimensions: (1) their in-sample explanatory power (the ability to “kill alphas”) and (2) their out-of-sample (OOS) predictive power (the ability to forecast returns).

First, we find that conditional models are powerful in explaining cross-sectional returns. They successfully attribute contemporaneous return variation to time-varying risk exposures, effectively “killing” the alpha left by models. For instance, with a single latent factor, the number of statistically significant pricing errors falls from 25 under PCA to just 4 under the linear autoencoder. Second, this explanatory success clarifies why robust predictive power is scarce. In a highly integrated market where risk is well-explained, little predictable alpha remains. We find that only the most parsimonious linear autoencoder (CA0) achieves a positive, albeit modest, out-of-sample predictive R^2 (up to 0.45%). This result, while small in magnitude, is economically significant given the well-documented difficulty of out-of-sample return prediction. More complex non-linear AEs and the linear IPCA fail to produce reliable OOS forecasts in this setting. Third, we uncover a methodological boundary for IPCA, which fails in this small-N setting (for example, for $K = 4$ achieving an out-of-sample Total R^2 of just 7.3% versus 67.7% for PCA), a problem that has not been previously documented in this context. This study makes three primary contributions. Finally, despite this divergence in performance, an “interpretability

consensus” emerges: the models—linear and non-linear—robustly identify Value, Momentum, and GDP-related characteristics as the most important drivers of risk exposure.

For the international asset pricing literature, we provide comparisons between modern latent factor models on country indices, highlighting a difference in model performance between IS explanation and OOS prediction. For the financial machine learning literature, we test the generalizability of IPCA/autoencoders and find that in “small N (countries), large P (characteristics)” settings as ours, parsimony (in the CA0) trumps complexity (in deep AEs) for OOS robustness; complex non-linear structures are prone to overfitting. Finally, for the factor pricing literature, we provide robust, consensus-based guidance for “taming” the international factor zoo.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 details the factor models and the empirical methodology. Section 4 describes the reliable dataset. Section 5 presents the main empirical results on model performance. Section 6 explores model interpretability through feature importance analysis. Section 7 concludes.

2 Literature Review

2.1 International Asset Pricing

A key question in international asset pricing is whether a unified pricing kernel exists that can explain the cross-section of asset returns across countries (Solnik, 1974; Adler and Dumas, 1983). From a theoretical perspective, this connects to the structure of the global stochastic discount factor (SDF), while from a practical view, it is of relevance for global asset allocation and risk management.

Early theoretical advancements, extending the Capital Asset Pricing Model (CAPM) to a global setting, were built on the assumption of fully integrated markets. The canonical International CAPM (ICAPM) posits that investors hold a world market portfolio and require compensation for global market risk and, potentially, currency risk (Sercu, 1980; Stulz, 1981b; Dumas and Solnik, 1995). In this idealized world, a single global pricing kernel should price all assets, regardless of where they are issued.

Empirical reality, however, reveals a more complex picture of partial segmentation (Stulz, 1981a; Errunza and Losq, 1985). Barriers to international investment—such as capital controls, informational frictions, and heterogeneous investor preferences—can create local risk premia. This tension is starkly illustrated by the historical behavior of emerging markets (EMs), which have exhibited high average returns, high predictability, and low correlations with developed markets (DMs) (Harvey, 1995; Bekaert and Harvey, 1995), consistent with a segmentation premium. As these markets liberalize, their cost of capital tends to fall, reflecting a transition

from a local to a more global pricing kernel (Henry, 2000).

This conflict between integration and segmentation manifests in several key empirical challenges. First, the failure of simple global models implies that the correlation structure of international equities is complex, driven by significant common trends (Kasa, 1992; Eun and Shim, 1989) but also by strong cross-sectional variation at the country level (Bekaert and Harvey, 1997; Karolyi and Stulz, 2003). Studies find that local, country-specific factors retain significant explanatory power beyond global factors (Griffin, 2002). Second, a vast literature documents the predictability of international returns, identifying numerous predictive sources, including macroeconomic risk (Ferson and Harvey, 1999), political and financial risk (Erb et al., 1996; Avramov et al., 2012), and characteristic-based anomalies such as value and momentum (Asness et al., 1997; Richards, 2012; Angelidis and Tessaromatis, 2016).

This proliferation of predictors has created an international “factor zoo”, raising a critical high-dimensional challenge, particularly as many factors fail to replicate (Jensen et al., 2023). The field thus requires a methodology that can (1) handle a high-dimensional set of country-level characteristics, (2) extract hopefully a parsimonious set of latent factors, and (3) flexibly model the (potentially non-linear) relationship between these characteristics and an asset’s exposure to risk.

2.2 The Linear Toolkit and Its (Linearity) Limitation

The methodological toolkit used to address this challenge has evolved through three main stages. The first stage involves prescribed linear factor models, such as the global extensions of the Fama-French model (Fama and French, 1993, 2012). While

highly interpretable, this approach is ad hoc and struggles to provide a parsimonious representation of the high-dimensional factor zoo.

The second stage employs data-driven latent factor models, with Principal Component Analysis (PCA) as the canonical tool (Chamberlain and Rothschild, 1982; Connor and Korajczyk, 1986). PCA avoids the need to pre-specify factors, but it suffers from well-known drawbacks: (1) it is a variance-maximizing procedure, and its factors are not guaranteed to align with priced risks (Lettau and Pelger, 2020); (2) it is a “black box,” lacking economic interpretability; and (3) its factor loadings are typically static and unconditional, failing to capture time-varying risk exposures.

The third stage, and the current state-of-the-art for linear models, is Instrumented Principal Component Analysis (IPCA) (Kelly et al., 2019b). IPCA elegantly solves the limitations of PCA by modeling an asset’s factor loading, $\beta_{i,t}$, as a linear function of its observable characteristics, $z_{i,t}$. This framework is data-driven, allows for conditional, time-varying betas, and restores interpretability by linking the latent factors to observable characteristics.

However, the primary limitation shared by all these models—from Fama-French to IPCA—is their fundamental linearity assumption. This assumption is increasingly challenged by evidence from the machine learning (ML) literature. A seminal contribution by Gu et al. (2021) conducts a comprehensive “horse race” of ML methods for return prediction in the U.S. stock market. Their robust finding is that non-linearities and characteristic interactions are of first-order importance, with methods like neural networks and trees substantially outperforming linear models out-of-sample. This raises a critical and unanswered question: If non-linearity is a dominant force in U.S.

firm-level data, does the linear suffice for the complex, macro-driven domain of international country indices, or is a non-linear approach necessary?

2.3 The Non-Linear Frontier, Autoencoders, and Our Contribution

The most promising tools for answering this question emerge from the non-linear frontier of machine learning. Beyond return prediction, researchers are now using ML to estimate the SDF directly (Chen and Xiu, 2023) and to sift the factor zoo for robust predictors. A particularly innovative architecture is the conditional autoencoder (AE), introduced to finance by Gu et al. (2021).

The AE asset pricing model can be understood as a direct non-linear generalization of IPCA. Like IPCA, it maps observable characteristics ($z_{i,t}$) to latent risk exposures ($\beta_{i,t}$), but it does so using the flexible architecture of a neural network. This allows it to capture the complex, non-linear interactions that Gu et al. (2021) found to be crucial. This non-linear methodology has demonstrated superior performance in pricing U.S. stocks (Gu et al., 2021) and in other specific domains like the Korean market (Kim et al., 2023).

This leads to a clear and important gap in the literature. While some studies have applied general ML techniques to international stocks (Tobek and Hronec, 2021; Hanauer and Kalsbach, 2023), to date, the powerful conditional autoencoder framework has not been systematically applied to model the cross-section of global equity indices. This gap is significant because most advanced asset pricing tests are conducted on firm-level data, whereas country indices represent the primary asset

class for global macro investors.

This paper fills this gap. We provide the first systematic comparison of the linear (IPCA) and non-linear (Autoencoder) latent factor models in the unique domain of international country indices (an $N = 37$, $P = 65$ panel). Our contributions are fourfold:

1. We conduct the first “horse race” for conditional latent factor models (PCA vs. IPCA vs. AE) in the cross-section of international country indices.
2. We contribute to the literature on international return predictability by showing that, in this domain, only the autoencoder framework provides robust and economically meaningful out-of-sample (OOS) predictive power. The linear model fails to deliver OOS predictability in our setting.
3. We provide new evidence on the integration vs. segmentation puzzle by demonstrating that conditional models (both AE and IPCA) are vastly superior to unconditional models (PCA) at explaining the cross-section of returns. By incorporating country-level characteristics, our conditional models “kill” most pricing anomalies, substantially reducing realized pricing errors.
4. We address the “factor zoo” and the “black box” critique by providing a detailed interpretability analysis. We identify the specific economic drivers that matter most, finding that Value, Momentum, and GDP-related characteristics are the most robust drivers of risk and return across all model specifications, both linear and non-linear.

3 Methodology

This section outlines the latent factor models used to study the cross-sectional predictability of international equity returns, moving from linear models with constant loadings to nonlinear conditional models. We begin with a unified notation system, then introduce Principal Component Analysis (PCA) and Instrumented Principal Component Analysis (IPCA), before describing our conditional autoencoder architecture and its implementation. These models are employed to extract underlying systematic factors from the high-dimensional space of country-level equity returns and associated characteristics, aiming to explain variations in expected returns and identify mispricing.

3.1 Notation

Let $i = 1, \dots, N$ index assets (here, MSCI country indices) and $t = 1, \dots, T$ index months. The $N \times 1$ vector of excess returns at time $t + 1$ is denoted $\mathbf{R}_{t+1} = (R_{1,t+1}, \dots, R_{N,t+1})^\top$. For each asset i at time t , let $\mathbf{x}_{i,t} \in \mathbb{R}^P$ denote a vector of P observable characteristics, and $\mathbf{X}_t \in \mathbb{R}^{N \times P}$ the corresponding cross-sectional matrix.

We model returns via a K -factor conditional linear model:

$$\mathbf{R}_{t+1} = \mathbf{B}_t \mathbf{f}_{t+1} + \boldsymbol{\varepsilon}_{t+1}, \quad (1)$$

where $\mathbf{B}_t \in \mathbb{R}^{N \times K}$ contains factor loadings $\boldsymbol{\beta}_{i,t}^\top$ for each asset i , $\mathbf{f}_{t+1} \in \mathbb{R}^K$ is the vector of factor realizations, and $\boldsymbol{\varepsilon}_{t+1}$ is idiosyncratic noise with $\mathbb{E}[\boldsymbol{\varepsilon}_{t+1} \mid \mathcal{I}_t] = \mathbf{0}$.

3.2 Models

3.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is the canonical latent factor model and serves as a static baseline with constant loadings $\mathbf{B}_t \equiv \mathbf{B}$. Let $\mathbf{R} \in \mathbb{R}^{N \times T}$ denote the matrix of demeaned excess returns. PCA extracts K orthogonal factors that capture the maximum variance in returns by solving:

$$\max_{\mathbf{B} \in \mathbb{R}^{N \times K}} \text{tr}(\mathbf{B}^\top \widehat{\text{Cov}}(\mathbf{R}) \mathbf{B}) \quad \text{s.t.} \quad \mathbf{B}^\top \mathbf{B} = \mathbf{I}_K,$$

yielding $\mathbf{B}$ as the top K eigenvectors of the sample covariance matrix. Factor realizations are obtained by projection:

$$\hat{\mathbf{f}}_{t+1} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{R}_{t+1}.$$

While simple and robust, PCA has two important limitations in an asset pricing context: (i) the extracted factors are purely statistical objects with no necessary link to economic risk premia, and (ii) both factors and loadings are unconditional and fixed, ignoring time-varying information in observable asset or country characteristics.

3.2.2 Instrumented Principal Component Analysis (IPCA)

Instrumented Principal Component Analysis (IPCA), introduced by Kelly et al. (2019b), generalizes PCA by allowing loadings to vary as functions of observable characteristics. Specifically, for asset i at time t with characteristics $\mathbf{x}_{i,t}$, factor load-

ings are modeled as

$$\beta_{i,t} = \mathbf{\Gamma}^\top \mathbf{x}_{i,t}, \quad \mathbf{\Gamma} \in \mathbb{R}^{P \times K},$$

so that the return model becomes

$$\mathbf{R}_{t+1} = \mathbf{X}_t \mathbf{\Gamma} \mathbf{f}_{t+1} + \boldsymbol{\varepsilon}_{t+1}. \quad (2)$$

Here $\mathbf{X}_t = \begin{bmatrix} x_{1,t} & \cdots & x_{P,t} \end{bmatrix}^\top \in \mathbb{R}^{N \times P}$ stacks the characteristics across assets, and $\mathbf{f}_{t+1} \in \mathbb{R}^K$ are latent factors. Given $\mathbf{\Gamma}$, factor realizations can be recovered by OLS:

$$\hat{\mathbf{f}}_{t+1} = (\mathbf{B}_t^\top \mathbf{B}_t)^{-1} \mathbf{B}_t^\top \mathbf{R}_{t+1}, \quad \mathbf{B}_t = \mathbf{X}_t \mathbf{\Gamma}.$$

Estimation proceeds by minimizing the in-sample sum of squared residuals:

$$\min_{\mathbf{\Gamma}} \sum_{t \in \mathcal{T}_{\text{train}}} \left\| \mathbf{R}_{t+1} - \mathbf{X}_t \mathbf{\Gamma} \hat{\mathbf{f}}_{t+1} \right\|_2^2,$$

where $\hat{\mathbf{f}}_{t+1}$ depends on $\mathbf{\Gamma}$ through $\mathbf{B}_t$. This creates a non-linear recursive problem typically solved by an alternating least squares (ALS) procedure as below: iteratively update factor realizations $\{\hat{\mathbf{f}}_{t+1}\}$ given current $\mathbf{\Gamma}$, then update $\mathbf{\Gamma}$ given those factors, until convergence.

Algorithm 1: Alternating Least Squares (ALS) for IPCA

Data: Asset returns $\mathbf{R}_{t+1}$, characteristics $\mathbf{X}_t$

Result: Estimated characteristic weights Γ , factor realizations $\{\mathbf{f}_{t+1}\}$

Initialize $\Gamma^{(0)}$ (e.g., using PCA or randomly);

for $k = 1, 2, \dots$ *until convergence* **do**

for *each* $t \in \text{TrainPeriod}$ **do**

$$B_t^{(k)} \leftarrow X_t \Gamma^{(k-1)};$$

$$\hat{f}_{t+1}^{(k)} \leftarrow (B_t^{(k)\top} B_t^{(k)})^{-1} B_t^{(k)\top} R_{t+1};$$

$$\hat{R}_{t+1}^{(k)} \leftarrow B_t^{(k)} \hat{f}_{t+1}^{(k)};$$

$$\varepsilon_{t+1}^{(k)} \leftarrow R_{t+1} - \hat{R}_{t+1}^{(k)};$$

 Compute objective: $L^{(k)} = \sum_{t \in \text{TrainPeriod}} \left\| \varepsilon_{t+1}^{(k)} \right\|^2$;

 Update $\Gamma^{(k)}$ using gradient descent or another optimization method;

By construction, IPCA retains the variance-extraction spirit of PCA while incorporating conditioning information through $\mathbf{X}_t$. This produces time-varying, economically interpretable loadings and factors, offering a more flexible and theoretically grounded alternative to static PCA.

3.2.3 Conditional Autoencoder

We extend the conditional autoencoder of Gu et al. (2021), which comprises:

- A **Beta network** $\mathcal{B}_\theta$: maps characteristics $\mathbf{x}_{i,t}$ to factor loadings $\beta_{i,t} = \mathcal{B}_\theta(\mathbf{x}_{i,t})$;
- A **Gamma network** $\mathcal{G}_\phi$: maps portfolios to factor realizations $\mathbf{f}_{t+1} = \mathcal{G}_\phi(\cdot)$, typically constrained to be linear for interpretability.

Our implementation. We adapt this framework to the international equity context using MSCI country indices. Two complementary datasets serve as inputs:

- Country-specific macroeconomic indicators (originally annual; frequency alignment described in Section 4);
- Global equity factors constructed from characteristic-sorted portfolios at the firm level and aggregated to the country level.

These jointly capture macro-level economic variation and micro-level cross-sectional patterns grounded in firm fundamentals. As discussed in (Gu et al., 2021), the neural network employed in this setting adopts the structure of an autoencoder, which is particularly well-suited for learning low-dimensional latent representations of high-dimensional asset return data.

The Beta network produces country-level, time-varying factor loadings as a (potentially nonlinear) function of both macroeconomic states and global equity factor returns. The Gamma network linearly combines two sets of economically meaningful portfolios—those formed by sorting countries on macroeconomic indicators and those based on global equity factors—to generate interpretable latent factor realizations. The final fitted returns are:

$$\hat{\mathbf{R}}_{t+1} = \mathbf{B}_t(\theta) \mathbf{f}_{t+1}(\phi),$$

and parameters (θ, ϕ) are estimated by:

$$\min_{\theta, \phi} \sum_{t \in \mathcal{T}_{\text{train}}} \|\mathbf{R}_{t+1} - \mathbf{B}_t(\theta) \mathbf{f}_{t+1}(\phi)\|_2^2 + \lambda_\beta \Omega(\theta) + \lambda_f \Omega(\phi),$$

where $\Omega(\cdot)$ denotes regularization terms (e.g., L1/L2, dropout).

Figure 1 from Gu et al. (2021), illustrates the architecture of the conditional autoencoder. We follow this structure in our implementation. The architecture efficiently captures nonlinear and hierarchical dependencies among the inputs. On the left-hand side, factor loadings β_{t-1} (in green) at time $t-1$ depend on the corresponding asset-level characteristics $\mathbf{X}_t$ (in yellow) through a nonlinear activation function $g(\cdot)$ applied in the hidden layers. In our setting, the input $\mathbf{X}_t$ contains $P = 65$ characteristics for each of $N = 37$ country indices. Each row of yellow neurons therefore represents the $P \times 1$ vector $\mathbf{x}_{i,t}$ of characteristics for a single country i . On the right-hand side, the corresponding factors $\mathbf{f}_t$ at time t are linear combinations of the neurons in the second input layer. In our implementation, this input consists of the $N = 37$ individual country excess returns (in red). The key difference from Gu et al. (2021) is that, while the original model uses portfolio-level characteristics, returns, and factors, we work directly at the individual-country level for all.

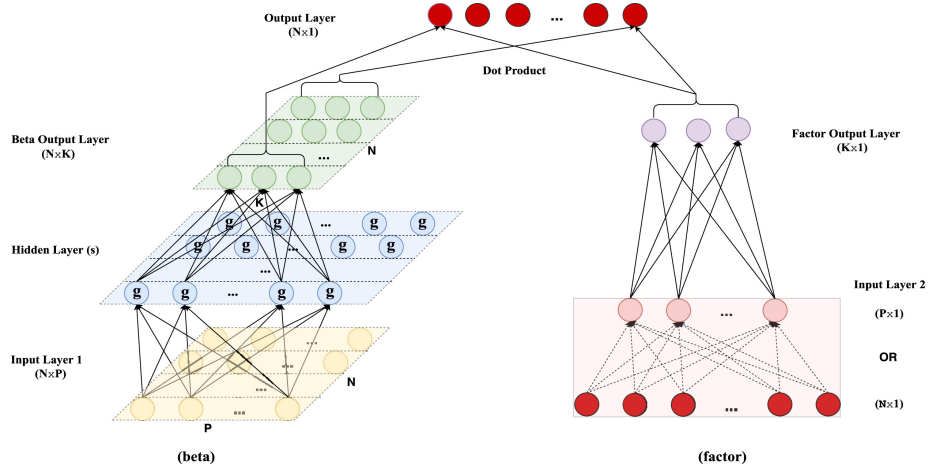


Figure 1: Autoencoder architecture Gu et al. (2021)

We explore multiple configurations, denoted $\text{CA}[L]_{-}[K]$, where L is the number of hidden layers in the Beta network and K is the number of latent factors. Unless otherwise stated, the Gamma network is linear to maintain interpretability. Training uses the Adam optimizer with early stopping based on validation loss, mini-batch updates, and standardization of all inputs at each time t .

3.3 Evaluation

Model performance is assessed using both in-sample (IS) and out-of-sample (OOS) metrics, including: Predictive R^2 ; Total R^2 ; Sharpe ratio on the spread portfolio based on the prediction models; Alpha (average pricing errors).

4 Data

4.1 Data Sources and Coverage

Our empirical analysis employs a comprehensive dataset spanning 659 months from 1970/01 to 2024/11, covering 37 developed and emerging market countries. The dataset integrates three primary data sources to construct a unified framework for cross-country equity return prediction. The overarching research design is to investigate how latent factors perform in international equity pricing. Specifically, our test assets are MSCI country-level indices, chosen for their high degree of cross-country comparability in the international asset pricing literature. From the microeconomic perspective, equity returns are influenced by firm-level fundamentals, whereas from the macroeconomic perspective, they are shaped by country-level economic state variables. Our predictors are designed to capture both angles, combining two of the most comprehensive and up-to-date databases available: the Global Factor Data (GFD) for firm-specific factors aggregated at the country level, and the Global Macro Database (GMD) for macroeconomic indicators.

The sample starting point is determined by the joint availability of the relevant databases. The Global Financial Data (GFD) themes database provides data beginning in May 1991. However, since our empirical design requires a twelve-month lag structure ($t-12$) for the construction of predictors, the effective starting date for the GFD-based features is postponed by four years, leading to the earliest feasible observation in 1995. In parallel, the MSCI return series becomes available from February 1970, and the macroeconomic variables we employ are likewise consistently recorded

starting in 1970. Although this would in principle allow an earlier starting date based on returns and macroeconomic series alone, the combined use of GFD themes data with the required lag structure imposes a stricter constraint. For consistency across datasets and to ensure that all predictors are simultaneously available, we therefore align the beginning of our sample period at 1995.

All of our country-index returns are denominated in USD and therefore embed both underlying equity performance and contemporaneous currency movements. This is a feature of the benchmark construction rather than an empirical choice: MSCI publishes indices in USD returns. This is aligned with the international asset-pricing literature which finds that currency factors explain a sizable share of variation in cross-border portfolios and that hedging can meaningfully alter risk exposures (e.g., Campbell, Serfaty-de Medeiros, and Viceira; Lustig, Roussanov, and Verdelhan).

Appendix B reports summary statistics over our sample period from 1995.

MSCI Country Index Returns: We obtain monthly equity return data from MSCI country indices via Datastream. These indices are constructed on a consistent basis across countries and net out cross-listed securities, ensuring maximum comparability across markets. We use MSCI Total Return Indices, denominated in U.S. dollars, which account for both price changes and the reinvestment of dividends, thereby providing a more comprehensive measure of investor returns than price indices alone. This choice follows standard practice in the international asset pricing literature, where total return data are generally preferred regardless of the specific index provider. The inclusion of dividends is crucial for capturing the full scope of

returns available to investors, as dividend payouts can represent a substantial component of long-term equity performance. While price indices may offer longer historical coverage, they omit this component and thus tend to underestimate true returns.

We acknowledge certain caveats: dividend taxation varies across countries and over time, and the MSCI gross total return series does not adjust for such differences. Some institutional investors (e.g., pension funds or mutual funds) may be exempt from such taxes, while others face substantial withholding. Nonetheless, consistent with most empirical studies, we abstract from these tax considerations and adopt total return indices as a proxy for investable market performance. Specifically, we compute simple monthly returns from raw index levels as

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}},$$

where P_t is the index level at the end of month t . All returns are converted to excess returns by subtracting the contemporaneous one-month U.S. Treasury bill rate from the Kenneth R. French Data Library. This adjustment isolates equity risk premiums from the risk-free component, following standard asset pricing conventions.

Global Factor Data (GFD): We incorporate 13 characteristic-based factors from the Global Factor Data library (Jensen et al., 2023). These factors are constructed from 153 underlying firm characteristics across 93 countries and four broad regions, grouped into 13 economic themes: accruals, debt issuance, investment, low leverage, low risk, momentum, profit growth, profitability, quality, seasonality, short-term reversal, size, and value. The factors are built from characteristic-managed portfolios formed at the firm level and aggregated upward using a consistent bottom-

up methodology. Specifically, the factor return data represent excess returns (in USD) of long-short portfolios sorted by firm-level characteristics, with portfolio construction available at daily and monthly frequencies under capped value-weighted schemes. For each theme, return series are computed as equal-weighted averages of the constituent characteristic-based factors. This structure provides a comprehensive and systematic summary of cross-sectional return patterns around the world, enabling the use of economically interpretable equity factors even in an international context. Importantly, although the dataset is organized at the country and theme levels, it remains firmly grounded in firm-level information through the initial sorting and portfolio formation process. The dataset spans 1991/05 to 2024/12 at monthly frequency, providing 659 months of coverage after temporal alignment. We use the 13 thematic factors for our 37 countries (test assets).

Macroeconomic Variables (GMD): We employ 52 country-specific macroeconomic indicators from the Global Macro Database (Müller et al., 2025). In contrast to the GFD, which captures micro-level firm attributes aggregated to the country level, the GMD represents macro-level national economic conditions. The variables cover a wide range of domains, including real and nominal GDP measures, inflation dynamics, fiscal positions, exchange rates, current account balances, and crisis indicators. These series are constructed by harmonizing and integrating data from major contemporary sources—including the IMF, World Bank, and OECD—with historical records from 87 additional datasets, producing comprehensive annual time series that span from the origins of modern data collection to projected estimates for 2030.

While these macroeconomic variables are annual in frequency, direct integration with monthly asset return data is constrained by this frequency mismatch. To address this, we adopt a rolling-window assignment methodology: each country’s macroeconomic indicator from the previous calendar year is uniformly assigned across all twelve months of the subsequent year. This approach is both methodologically straightforward and economically intuitive—it mirrors the practical informational constraints faced by market participants, who typically rely on macroeconomic data with at least long reporting lag. Similar rolling-window treatments are used in Fama and French (1989) and Welch and Goyal (2008) to ensure robustness and realism when combining mixed-frequency data. Alternative interpolation-based methods, such as linearly interpolating annual data to monthly frequency, may impose unrealistic continuity or linear progression within a year, leading to spurious temporal variation. The rolling-window approach avoids these pitfalls, enhancing both robustness and interpretability of the empirical results.

4.2 Data Processing and Preprocessing

Temporal Alignment and Look-ahead Bias Prevention: To prevent look-ahead bias, we implement careful temporal alignment protocols. Macroeconomic data from year t are matched with equity returns from January through December of year t , while GFD factor returns from month t are matched with equity returns in month $t+1$. This ensures that all predictor variables are known prior to the realization of the dependent variable.

Cross-sectional Standardization: Following established practices in asset

pricing with machine learning (Gu et al., 2021; Kelly et al., 2019b; Gu et al., 2020), we address missing values through cross-sectional median imputation within each time period. We then apply rank-based normalization to all 65 predictor variables, transforming their distributions to the interval $(-1, 1)$ within each month. This procedure mitigates the influence of outliers, ensures comparability across heterogeneous variables, and preserves the cross-sectional ranking information essential for portfolio formation.

Country Coverage and Matching: We maintain consistent country coverage across all data sources through careful country code mapping, achieving 100% matching accuracy between GFD location codes, GMD country names, and MSCI index identifiers. The final sample consists of 37 countries, following Fisher et al. (2017), including major developed markets and key emerging markets. This selection represents approximately 85% of global market capitalization.

4.3 Sample Splits and Empirical Design

We partition our 659-month sample into three distinct periods to ensure robust out-of-sample evaluation: training period (1995/01–2014/12), validation period (2013/01–2014/12), and out-of-sample testing period (2015/01–2024/12). In constructing the panel, we track MSCI country index excess returns for a universe of 37 countries. Within this temporal design, the in-sample training window (1995/01–2014/12) is used for model parameter estimation and factor-loading calibration, while the out-of-sample window (2015/01–2024/12) is reserved strictly for evaluation with no parameter re-estimation; model parameters estimated in-sample are held fixed and directly

applied to the testing period.

The training period captures diverse market regimes including the dot-com bubble, global financial crisis, and European sovereign debt crisis. The validation period overlaps with the final training years to enable hyperparameter tuning, while the out-of-sample period covers the most recent decade, including the COVID-19 pandemic and subsequent recovery.

To guard against look-ahead bias, we enforce a strict timing convention linking information sets to subsequent returns. At the monthly forecasting horizon we implement

$$\text{Return}_{t+1} = f(\text{Characteristics}_t), \quad (3)$$

and, more generally, we align data as

$$\text{Monthly characteristics at } t - 1 \rightarrow \text{MSCI return at } t, \quad (4)$$

$$\text{Annual characteristics at } t - 12 \rightarrow \text{MSCI return at } t, \quad (5)$$

so that all predictors are observable before the realization of returns.

This temporal structure yields each observation containing 65 standardized predictive characteristics (52 macroeconomic and 13 firm-specific factors) and corresponding excess returns. The extensive time coverage and broad cross-sectional scope provide sufficient statistical power for training complex machine learning models while preserving economically meaningful out-of-sample evaluation periods. Together, the fixed-parameter out-of-sample application and the explicit time-alignment

protocol ensure that reported performance reflects genuine predictive content rather than artifacts of re-fitting or forward-looking information.

5 Pricing model evaluation

5.1 Statistical Performance

We examine the performance of latent factor models, focusing on their capacity to account for realized return variation and to forecast conditional expected returns. Following standard practice in the asset pricing literature (Kelly et al., 2019a; Gu et al., 2021), we adopt two complementary R^2 measures: the *Total* R^2 (R_T^2) and the *Predictive* R^2 (R_P^2). Each captures a distinct aspect of model performance: the former evaluates contemporaneous explanatory power with respect to systematic risks, while the latter assesses the accuracy of forward-looking predictions of risk compensation across assets.

The R^2 statistic provides a general measure of model performance, capturing the proportion of variation in returns explained relative to a benchmark. In the conventional formulation, this benchmark corresponds to the historical mean, so that R^2 evaluates a model’s fit relative to using the historical mean return as a predictor. However, following (Gu et al., 2021; Kelly et al., 2019b), we deviate from this approach: given that the sample mean is excessively noisy in our setting, we do not subtract it when computing R^2 . This modification avoids using the historical mean as an implicit benchmark, thereby yielding a cleaner measure of the models’ explanatory and predictive power.

The total R^2 quantifies the proportion of realized return variation explained jointly by the estimated dynamic factor loadings and contemporaneous factor re-

alizations. Formally, it is defined as

$$R_T^2 = 1 - \frac{\sum_{(i,t) \in OOS} \left(r_{i,t} - \hat{\beta}'_{i,t-1} \hat{f}_t \right)^2}{\sum_{(i,t) \in OOS} r_{i,t}^2}, \quad (6)$$

where $r_{i,t}$ denotes the excess return of asset i at time t , $\hat{\beta}_{i,t-1}$ are the estimated factor loadings based on information up to $t - 1$, and $\hat{f}_t$ are the contemporaneous factor realizations. This measure summarizes how well the model captures the systematic component of returns that co-move across assets. A higher R_T^2 indicates that the factor structure effectively explains the realized common risk exposures embedded in the cross section of returns. Total R_T^2 measures how much of the total variation in returns is captured by the model's fit of returns, given access to contemporaneous information. A higher R_T^2 implies a better overall fit and ability to explain the variance of returns.

The predictive R^2 instead evaluates the extent to which the model captures differences in conditional expected returns across assets, i.e., whether it correctly prices risk premia. It is defined as

$$R_P^2 = 1 - \frac{\sum_{(i,t) \in OOS} \left(r_{i,t} - \hat{\beta}'_{i,t-1} \hat{\lambda}_{t-1} \right)^2}{\sum_{(i,t) \in OOS} r_{i,t}^2}, \quad (7)$$

where $\hat{\lambda}_t$ denotes the estimated factor risk premia at time t (Here we use the historical mean of the estimated factor realization up to time t). While R_T^2 focuses on contemporaneous fit, R_P^2 is explicitly forward-looking: it isolates the variation in returns explained by the model-implied expected returns $\hat{\beta}'_{i,t-1} \hat{\lambda}_{t-1}$. In other words, it

measures the model’s ability to account for cross-sectional differences in average returns through risk compensation. A higher R_P^2 therefore indicates stronger empirical support for the model’s implied pricing of risk premia.

Together, these two statistics provide a complementary assessment of model performance. The total R^2 reflects how accurately the factor structure captures common variation in realized returns, serving as a measure of systematic risk fit. The predictive R^2 directly tests the model’s ability to explain cross-sectional return differences through expected risk premia. Evaluating both dimensions is crucial: a model may fit realized returns well (high R_T^2) but fail to deliver meaningful predictions of conditional expected returns (low R_P^2), or vice versa. Hence, the joint consideration of R_T^2 and R_P^2 yields a rigorous and balanced view of the empirical validity of latent factor models in asset pricing.

Total R^2

The model comparison reveals a stark divergence in performance, providing critical insights into the interplay between model architecture and dataset characteristics. In-sample (Table 1), the autoencoders with hidden layers (CA1, CA2, CA3) deliver a very high Total R^2 , consistently reaching over 80%. This demonstrates the immense flexibility of non-linear models to fit the complex patterns within the training data. The simpler linear models—CA0 and PCA—also perform admirably, explaining a substantial portion of return variation. This confirms the presence of a strong, low-dimensional factor structure within international equity index returns. The transition from in-sample to out-of-sample reveals critical insights about model generalization.

The striking in-sample result, however, is the categorical failure of Instrumented Principal Component Analysis (IPCA). This poor performance persists out-of-sample (Table 2), where we care more about a model’s generalization ability. For the out-of-sample analysis, the linear conditional autoencoder, CA0, delivers the best and most stable Total R^2 , achieving levels between 50-60%. This robust performance is followed closely by the PCA benchmark. It is important to contextualize this high magnitude of explanatory power. We are predicting returns for MSCI international equity indices, which are large and diversified collections of individual stocks. As such, much of the idiosyncratic risk in the data is averaged out, making the systematic component of returns more prominent and easier to explain, leading to R^2 values comparable to portfolio-level experiments in the US equity market

Although both IPCA and our characteristic autoencoder without hidden layers (CA0) allow for time variation in exposures, their identification pathways differ in ways that bite in a small cross section of countries. IPCA imposes the structural restriction $\beta_{it} = \Gamma^\top Z_{it}$ and identifies (f_t, Γ) by using characteristics Z_{it} as instruments for time-varying loadings β_{it} (Kelly et al., 2019b). With only $N = 37$ MSCI country indices and highly collinear country characteristics, the “first stage” ($Z \rightarrow \beta$) is weak, yielding high-variance estimates and fragile out-of-sample performance—exactly the pathologies emphasized in the weak-instrument literature (Stock and Yogo, 2002). By contrast, CA0 extracts a low-rank representation directly from returns on the γ -side (a linear autoencoder is essentially PCA) and then uses characteristics on the β -side to explain exposures rather than to identify them (Baldi and Hornik, 1989). This end-to-end low-rank objective is naturally robust when common variation is

dominated by a few strong global/regional modes. Empirically, this manifests as very high in-sample total R^2 for deeper networks but the strongest predictive R^2 for the parsimonious CA0.

The juxtaposition of the successful PCA against the failing IPCA is particularly informative. PCA’s success confirms that priced, systematic factors are indeed present. IPCA’s failure indicates that the model cannot effectively link these latent factors to the observable characteristics we provide. The primary reason for this is a fundamental mismatch between the IPCA methodology and the structure of our data: the small cross-section. The IPCA estimation procedure was explicitly developed for and validated on panels with a very large cross-section (N) of assets

For country-level equity portfolios the effective dimensionality of systematic risk is low: beyond a handful of broad directions the eigenvalue spectrum declines steeply. Adding factors beyond this “elbow” typically raises in-sample fit but harms out-of-sample performance via variance inflation. Finally, the results across all models show that performance does not monotonically improve with the number of latent factors (K). This illustrates the classic bias-variance trade-off. While adding factors reduces model bias and mechanically improves the in-sample fit, each additional factor introduces more parameters to be estimated, thereby increasing model variance and the risk of overfitting. Beyond an optimal point, the model begins to fit sample-specific noise rather than the true underlying factor structure. Formal statistical criteria are often used to guide this choice by penalizing complexity, reinforcing the principle that a parsimonious factor structure is preferable for out-of-sample robustness (Bai and Ng, 2002).

The Total R-squared results largely corroborate the findings from the predictive R-squared analysis, emphasizing the superior performance of autoencoder models in explaining overall return variation. The Total R-squared values are generally higher than the predictive R-squared values for all models as expected. This is precisely because the total R_T^2 leverages contemporaneous information to explain current returns, which is inherently a less challenging task than forecasting future returns using only past data. This highlights the distinction between explaining and predicting return variation.

Consistent with the Predictive R-squared, autoencoder models demonstrate substantially higher Total R-squared values compared to PCA and IPCA. This reinforces the notion that the feature extraction capabilities of autoencoders are highly effective in capturing the complex underlying drivers of return variation in global equity markets.

Table 1: *In-sample Total R^2 (%) comparison*

Model	Number of latent factors (K)					
	1	2	3	4	5	6
PCA	51.92	60.31	65.92	69.35	72.45	75.19
IPCA	8.43	15.19	19.41	22.76	25.78	28.40
CA₀	57.62	69.20	71.91	76.05	78.54	81.61
CA₁	61.82	71.93	87.26	81.44	90.03	82.71
CA₂	92.13	80.93	89.52	69.30	89.62	81.01
CA₃	93.48	74.83	93.81	78.87	94.74	75.87

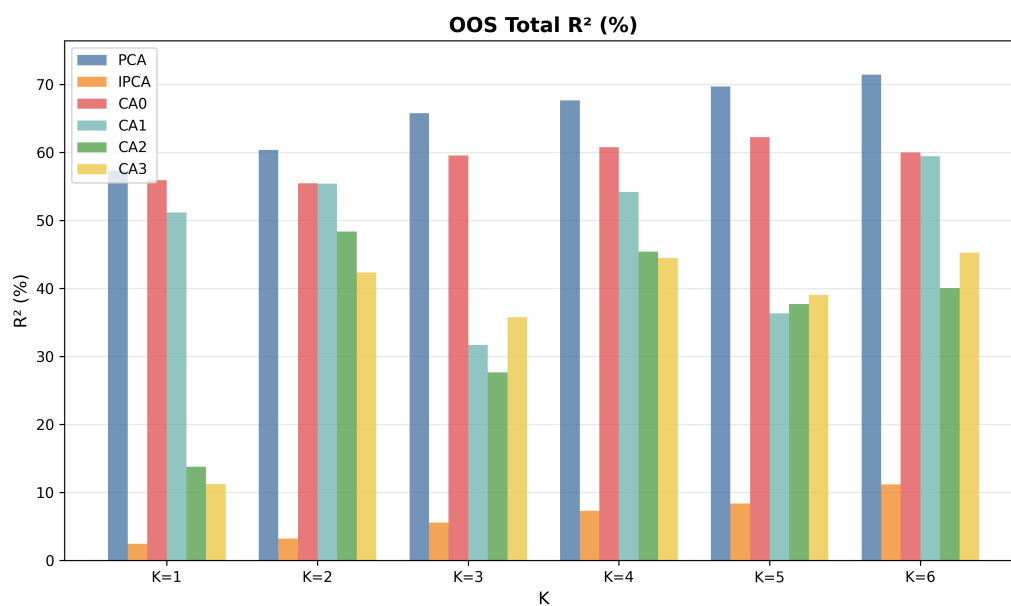
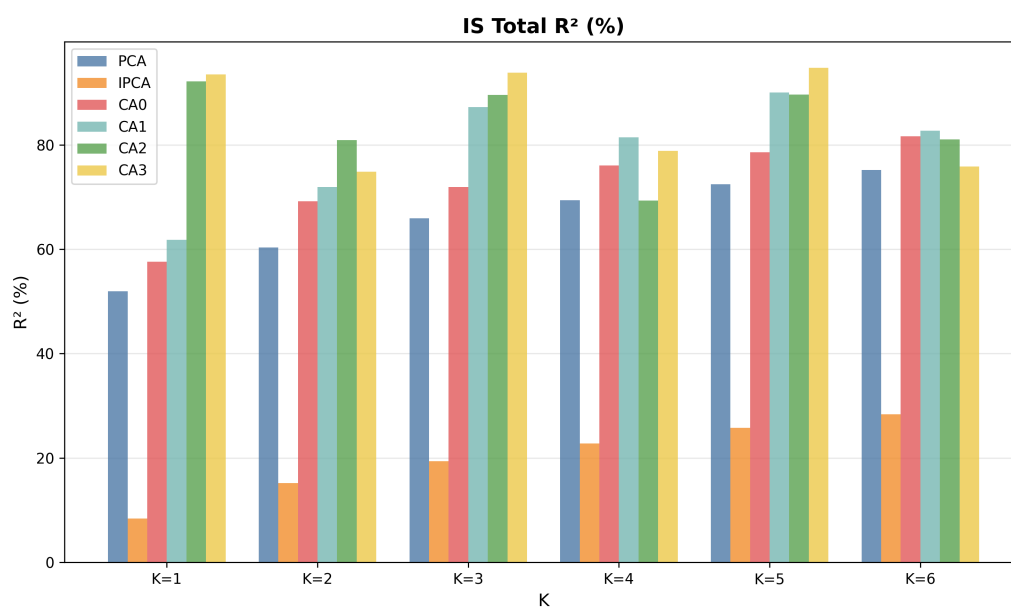


Figure 2: *Total R^2*

Table 2: *Out-of-sample Total R^2 (%) comparison*

Model	Number of latent factors (K)					
	1	2	3	4	5	6
PCA	57.29	60.38	65.80	67.67	69.70	71.45
IPCA	2.44	3.21	5.55	7.31	8.40	11.17
CA₀	55.91	55.45	59.54	60.77	62.24	60.04
CA₁	51.20	55.44	31.71	54.21	36.36	59.45
CA₂	13.80	48.39	27.66	45.44	37.69	40.07
CA₃	11.21	42.34	35.77	44.51	39.06	45.28

Predictive R^2

While the Total R^2 measures the power to explain contemporaneous systematic variation, the predictive R^2 offers a different and stricter measure of the conditional model’s forecasting utility. The followings visualise only models with a strictly positive predictive R^2 , including both in-sample and out-of-sample.

In this dimension, the autoencoder framework provides a dramatic performance improvement over the benchmarks. We find that the autoencoder with no hidden layers (CA0) is consistently the best performing model out-of-sample, achieving a predictive R^2 that exceeds 0.4%. This result is highly significant both statistically and economically, as it represents a substantial degree of predictability for monthly international index returns and dramatically outperforms the other models, whose predictive power is negligible. This finding further solidifies the conclusion that the linear conditional autoencoder architecture provides the optimal balance of explanatory power and forecasting robustness for this asset space.

Similar patterns regarding the optimal number of hidden layers and latent factors emerge. This suggests that while non-linearity is beneficial, excessive complexity

(e.g., very deep networks or too many factors) without commensurate data signal can lead to diminishing returns, or potentially overfitting.

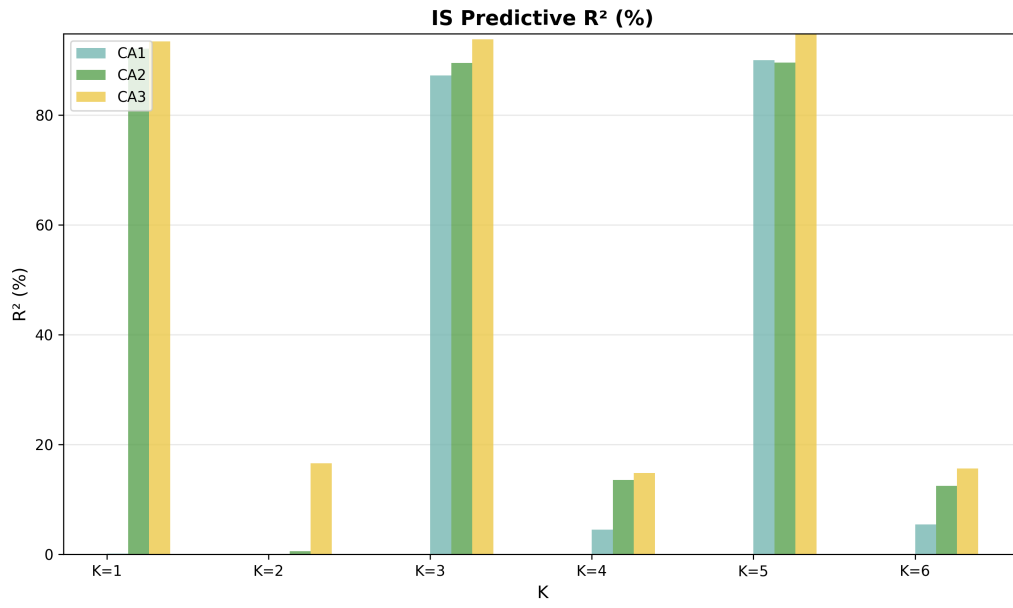
The overall conclusion from both R-squared measures is that the autoencoder framework, by flexibly incorporating characteristics and learning non-linear factor exposures, offers a significant advantage in both predicting future returns and comprehensively explaining contemporaneous cross-sectional variation in international equity returns. These findings align with the growing literature on machine learning in finance, which highlights the benefits for capturing intricate financial market dynamics.

Table 3: *In-sample Predictive R^2 (%) comparison*

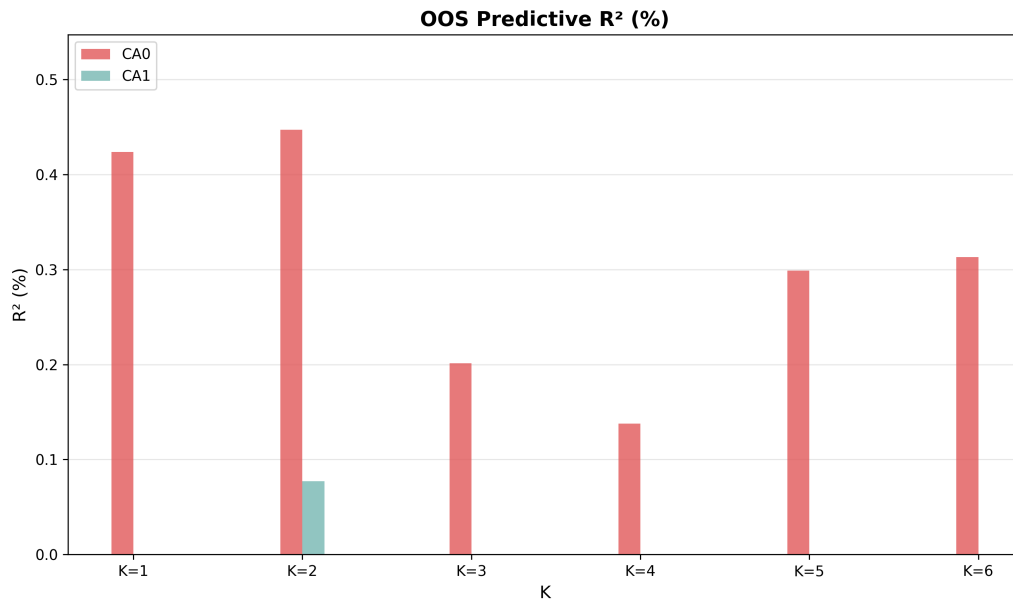
Model	Number of latent factors (K)					
	1	2	3	4	5	6
PCA	<0	<0	<0	<0	<0	<0
IPCA	<0	<0	<0	<0	<0	<0
CA₀	<0	<0	<0	<0	<0	<0
CA₁	0.12	<0	87.26	4.51	90.03	5.46
CA₂	92.13	0.58	89.52	13.54	89.62	12.49
CA₃	93.48	16.59	93.81	14.83	94.74	15.66

Table 4: *Out-of-sample Predictive R^2 (%) comparison*

Model	Number of latent factors (K)					
	1	2	3	4	5	6
PCA	<0	<0	<0	<0	<0	<0
IPCA	<0	<0	<0	<0	<0	<0
CA₀	0.42	0.45	0.20	0.14	0.30	0.31
CA₁	<0	0.08	<0	<0	<0	<0
CA₂	<0	<0	<0	<0	<0	<0
CA₃	<0	<0	<0	<0	<0	<0



(a) *In-sample*



(b) *Out-of-sample*

Figure 3: *Predictive R^2*

5.2 Economic Performance

We evaluate the out-of-sample performance of the asset-pricing models using an equal-weighted quintile long–short strategy, designed to test whether model forecasts translate into economically meaningful signals under realistic monthly rebalancing. The out-of-sample window runs from 2015 to 2024. The investable universe consists of 37 MSCI country indices. Quintiles are formed only in months with at least ten eligible countries; months that fail this condition are excluded from the portfolio time series. All inputs are lagged so that only information available through month t is used to form positions for month $t+1$, thereby avoiding look-ahead bias. Breakpoints are recomputed each month within the then-available cross-section, and portfolios are rebalanced monthly.

At the end of month t , each model produces one-step-ahead predicted excess returns $\{\hat{r}_{i,t+1}\}_i$ for the set of countries i . Countries are ranked by $\hat{r}_{i,t+1}$ and split into five quintiles $Q1_t, \dots, Q5_t$, where $Q1_t$ contains the lowest predicted returns (short leg) and $Q5_t$ the highest (long leg). Portfolios are equal-weighted within each leg,

$$w_{i,t}^L = \frac{1}{|Q5_t|} \mathbf{1}\{i \in Q5_t\}, \quad w_{i,t}^S = \frac{1}{|Q1_t|} \mathbf{1}\{i \in Q1_t\},$$

and the next-month long–short return is

$$R_{LS,t+1} = \sum_{i \in Q5_t} w_{i,t}^L r_{i,t+1} - \sum_{i \in Q1_t} w_{i,t}^S r_{i,t+1},$$

where $r_{i,t+1}$ denotes the realized excess return for country i in month $t+1$. Let

$\mu_{LS} = \frac{1}{T} \sum_{t=1}^T R_{LS,t}$ and $\sigma_{LS} = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (R_{LS,t} - \mu_{LS})^2}$ be the out-of-sample mean and standard deviation of the monthly long-short series. The annualized Sharpe ratio reported for each model is

$$\text{SR} = \frac{\mu_{LS} \times 12}{\sigma_{LS} \times \sqrt{12}} = \sqrt{12} \frac{\mu_{LS}}{\sigma_{LS}}.$$

In terms of empirical content, the quintile long-short results emphasize economic value rather than statistical fit alone. According to Table 5, PCA performs poorly out of sample in this exercise, often delivering negative or near-zero Sharpe ratios, which contrasts with the mixed picture one might obtain from in-sample or total R^2 measures. By comparison, CA_1 and IPCA translate forecasts into more robust positive long-short payoffs in the equal-weighted setting, indicating that conditional structure (and, for CA_1 , moderate nonlinearity) helps bridge from predictive fit to portfolio-level performance. The ranking is sensitive to specification and effective dimensionality, but the broad message is that models with time-varying exposures disciplined by firm- or country-level information deliver stronger economic performance than static factor extractions.

These portfolio results are diagnostic, not prescriptive. They abstract from trading costs, turnover constraints, funding and capacity limits, and other frictions. As such, they should be read as a non-implementable yet informative comparison of models' economic efficiency rather than as investable strategies.

Table 5: *Sharpe Ratios for Long-Short Portfolios (Out-of-Sample)*

Equal-weight	K					
	1	2	3	4	5	6
PCA	-0.07	-0.34	-0.53	-0.35	-0.56	-0.49
IPCA	0.57	-0.34	0.30	-0.80	-0.09	0.71
CA ₀	-0.29	0.05	-0.40	-0.03	0.21	-0.93
CA ₁	-0.25	-0.30	-0.17	-0.60	0.53	0.02
CA ₂	0.02	-0.35	0.19	-0.44	0.30	0.21
CA ₃	0.20	-0.44	0.36	-0.41	0.19	-0.35

5.3 Realized Pricing Errors (Alpha)

Following the established methodology (Gu et al., 2021), we also evaluate the models’ ability to explain the cross-section of returns by examining the magnitude and characteristics of realized pricing errors, commonly referred to as alphas.

While R^2 quantifies the overall explanatory and predictive power of a model, it does not distinguish the source of this explanatory power. Specifically, a high R^2 does not disentangle whether a model’s success stems from accurately capturing compensation for systematic risk exposures or from merely fitting patterns related to persistent mispricing. To address this ambiguity, we assess the magnitude of mispricing by testing the null hypothesis that the intercepts from our explanatory regressions—the alphas—are equal to zero. A failure to reject this null would lend support to the interpretation of the model’s factors as representing systematic risk, whereas significant alphas would indicate the presence of unexplained returns attributable to mispricing.

For this analysis, we use factors derived from information available contempora-

neously with the returns. This approach allows us to assess the degree of mispricing that remains after accounting for the identified factors within the observed period. It is crucial to distinguish this from the predictive setting employed for the out-of-sample R-squared calculations, where factors are constructed using only historical information for forecasting. The figures presented in this section thus highlight the explanatory power of the models in capturing cross-sectional variation in expected returns, given the full information set available at each month. A smaller (closer to zero) and less volatile alpha, both in magnitude and across assets, indicates a better-specified model that leaves less unexplained return variation.

Mathematically, the realized alpha for asset i at time t , $\alpha_{i,t}$, is defined as the intercept from the factor model regression:

$$R_{i,t} = \hat{\beta}_{i,t-1}^\top \hat{f}_t + \alpha_{i,t} + \epsilon_{i,t}$$

where $R_{i,t}$ is the excess return, $\hat{f}_t$ are the estimated factors (derived using contemporaneous data), and $\hat{\beta}_{i,t-1}$ are the estimated factor loadings. We examine the distribution and magnitude of these $\alpha_{i,t}$ values across different models and factor specifications.

Figure 4 through Figure 9 present the time-series average alphas for each of the 37 test assets with different number of layers, plotted against their corresponding time-series average actual returns. In these figures, each point represents a distinct country observation. The vertical axis quantifies the average monthly pricing error, or alpha, generated by a given model specification. Consequently, a model's success in explaining the cross-section of returns is visually evidenced by the clustering of

these points around the horizontal line at $\alpha = 0$. Points highlighted in red denote alphas that are statistically significant at the 5% level, providing a formal metric for the prevalence of mispricing that remains unexplained by the model.

A comparative analysis across the different modeling frameworks reveals substantial disparities in performance, particularly when contrasting traditional linear methods with more sophisticated specifications. The results for PCA, which serves as a linear benchmark, are particularly telling. For the one-factor specification ($K = 1$), the PCA model generates a strikingly large number of statistically significant pricing errors, with 25 out of 37 assets exhibiting alphas that are different from zero. This indicates a significant failure of the simple linear factor structure to capture the underlying drivers of international returns. In stark contrast, both IPCA and the various Autoencoder architectures demonstrate a markedly superior ability to price the cross-section of assets. For the same single-factor specification, the number of significant alphas is dramatically reduced to four for the IPCA and the linear Autoencoder (CA0) models, and to merely one or two for the non-linear Autoencoder variants (CA1, CA2, and CA3). This robustly demonstrates that the incorporation of instrument-driven dynamics in IPCA and the non-linear representations learned by the Autoencoders are highly effective in minimizing pricing errors.

Further investigation into the impact of model complexity yields a compelling insight regarding the optimal number of latent factors. A consistent pattern that emerges across all model families is that the most parsimonious specification, with only a single latent factor ($K = 1$), consistently produces the smallest number of significant alphas. As we increase the dimensionality of the latent factor space from

$K = 2$ to $K = 6$, we observe a general upward trend in the quantity of statistically significant pricing errors. This suggests that a single, well-constructed factor captures the most salient features of the international return cross-section, and that the inclusion of additional factors, while potentially useful for prediction, may introduce noise or lead to in-sample overfitting that manifests as larger realized pricing errors.

Finally, we examine the architectural choices within the family of Autoencoder models by varying the number of hidden layers from zero (a linear model) to three. Unlike the clear and systematic influence of the latent factor dimension, the depth of the neural network does not appear to exert a monotonic or easily discernible effect on the resulting pricing alphas. While all Autoencoder specifications outperform the PCA benchmark, the variations among them are minor and non-directional.

In conclusion, the analysis of realized pricing errors (alphas) provides strong evidence that autoencoder models, significantly reduce unexplained cross-sectional variation in international equity returns. These findings reinforce the use of machine learning in providing a more nuanced and powerful framework for understanding and mitigating mispricing in global financial markets.

5.4 Discussion

Asset pricing models serve two complementary purposes. Ex post, they are attribution engines: they explain realized returns by attributing them to priced risk factors, thereby clarifying which risks enter the stochastic discount factor (SDF) and the implied risk–return trade-off. Ex ante, they are forecasting engines: they produce conditional expected returns to be evaluated by out-of-sample (OOS) predictive per-

formance. Our empirical design mirrors this dual role via four families of goodness-of-fit statistics (in-sample and OOS, total and predictive R^2), cross-sectional pricing tests (alpha magnitudes), and the economic translation into portfolio outcomes (the Sharpe ratios of implementable long–short strategies).

Explanatory power: who explains realized comovement best? The ranking differs sharply between in-sample (IS) and out-of-sample (OOS) fit. In-sample, deeper autoencoders deliver the highest total R^2 , followed by PCA, with IPCA lagging. Out of sample, the ordering reverses in favor of robustness: PCA attains the strongest total R^2 , followed by the simple autoencoder, then deeper AEs, and finally IPCA. Intuitively, the flexible AE architectures excel at fitting contemporaneous structure within the training window, whereas PCA’s variance-maximizing construction sees better out of sample when the cross-section is small.¹ This pattern is fully consistent with the bias–variance trade-off in small- N international panels: greater in-sample flexibility improves fit but need not generalize, whereas PCA’s parsimony is comparatively stable OOS. At the same time, conditional models materially improve pricing discipline: relative to PCA, both IPCA and AEs substantially reduce realized alphas, indicating that characteristic-conditioned loadings align factor exposures with systematic risks even when OOS total fit favors PCA.

To wrap up, as per the explanatory power, the conditional latent-factor models dominate on alphas and total R^2 . This shows up in two complementary ways. First, in-sample total R^2 improves when factor loadings are allowed to vary with charac-

¹See Figure 2 for IS and OOS totals. In our tables, deeper AEs dominate IS totals, while PCA leads OOS, with CA0 close behind.

teristics. Second, across specifications, conditional latent-factor models substantially reduce pricing errors relative to an unconditional PCA benchmark. Specifically, the cross-sectional intercepts (alphas) become economically small and statistically indistinguishable from zero under IPCA and our AEs, while the unconditional PCA leaves more systematic residual structure. In other words, conditional models provide a clearer decomposition of realized index returns into a small number of latent systematic risks with time-varying exposures disciplined by country-level attributes. Consistent with the integration view of Pukthuanthong and Roll (2009), who quantify global market integration by the multi-factor R^2 and argue that a country is financially well integrated when global factors really do explain the vast bulk of its returns, our conditional latent-factor models shows that in the global equity markets, where a few pervasive factors drive a large share of the comovement across diversified country indices.

Predictive power: who forecasts expected returns best? For predictive R^2 , the evidence is even starker. In-sample, only the simple autoencoder (CA0) yields a positive predictive R^2 across our specifications. Out of sample, CA0 is the sole consistent “survivor,” delivering small but positive predictive R^2 across factor dimensions, while PCA, deeper AEs, and IPCA do not consistently clear zero.² Actually, the seminal work (Welch and Goyal, 2008) introduces an out-of-sample (OOS) benchmark, demonstrating that most canonical predictors from the literature fail to outperform the simple historical mean. These models, often appearing strong in-sample, typi-

²See Figure 3, where CA0 posts positive OOS predictive R^2 of roughly 0.1%–0.45% depending on K .

cally yield negative OOS R^2 (R_{OOS}^2) values. Within this challenging context, a small but consistently positive R_{OOS}^2 is not a sign of a weak model; rather, it is strong evidence of genuine, non-spurious predictive ability. As Campbell and Thompson (2008) argued, an R_{OOS}^2 in the range of 0.5% to 1.0% (50–100 basis points) at a monthly frequency is considered a significant achievement. Given the small cross-section (37 indices) and strong global comovement, this pattern underscores that a parsimonious conditional mapping stabilizes the bias–variance trade-off and prevents variance explosions that erode OOS forecasts. It also squares with the broader literature on the fragility of OOS return prediction and the importance of evaluating models on genuine OOS criteria rather than in-sample fit alone (Welch and Goyal, 2008).

In short, turning to forecasting, the autoencoder architecture stands out. Despite the small cross-section (37 MSCI country indices) and the high level of commonality across markets, our best-performing linear autoencoder (CA0) delivers consistently positive OOS predictive R^2 across the model configurations we report, whereas more complex nonlinear AEs (with deeper hidden layers) and IPCA do not uniformly match this OOS stability. This contrast is informative: it indicates that in a small- N panel with collinear characteristics and strong global comovement, parsimonious dimension reduction (CA0) is sufficiently flexible to capture the mapping from characteristics to conditional betas/expected returns.

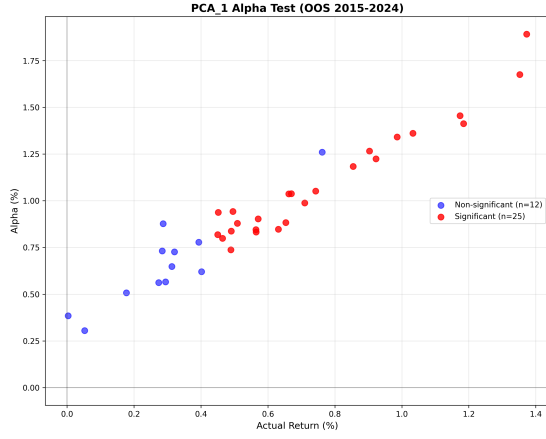
Alphas vs. portfolio Sharpe: no contradiction. Specifically, our pricing tests show that conditional latent-factor models (IPCA and AEs) drive cross-sectional alphas toward zero. That is exactly what a good explanatory model should do: once ex-

posures to priced risks are correctly specified and allowed to vary with characteristics, the intercepts vanish. In this environment, it is expected that market-neutral long-short portfolios built from small residual spreads deliver low and unstable Sharpe ratios. The long-short construction intentionally hedges away the market exposure; what remains at the country-index level are modest, noisy conditional differentials that rarely dominate residual volatility over monthly horizons. None of this undermines the model's value for risk explanation—it confirms it. These two facts are not contradictory; they simply speak to different questions. Pricing tests ask whether realized returns are consistent with a factor structure (alphas close to zero), while long-short Sharpe asks whether conditional expected return differences are large and reliable enough after implementation to dominate residual volatility. Our evidence therefore shows attribution success (low alphas) coexisting with limited economic value from relative-value long-shorts in this asset space. In fact, our evidence is entirely consistent with international asset-pricing theory: when markets are integrated, a small set of global risk factors dominate index-level returns, leaving limited room for large country-level predictable spreads.

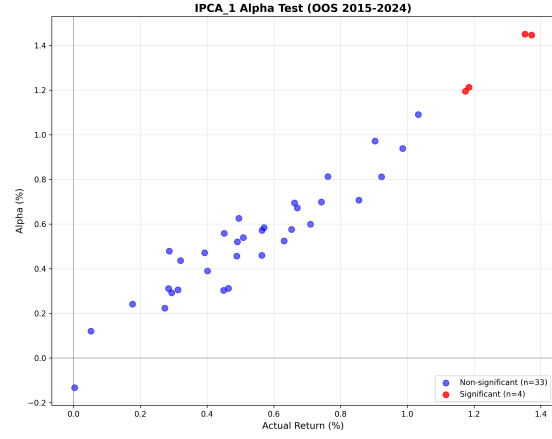
Two implications follow for economic use: (i) When markets are highly integrated, global factors explain most index-level variation (Pukthuanthong and Roll, 2009), leaving limited short-horizon cross-country spreads to monetize; hence low long-short Sharpe is the norm when alphas are statistically “killed.” (ii) If the predictive signal were sufficiently strong, the natural way to convert it into economic value at the index level is not market-neutral relative value, but probably market-timing net-long on the equity premium. In other words, when the signal primarily tracks slow-moving

global states, a buy-and-hold benchmark augmented with state-contingent tilts is the right vehicle for gains; this perspective is central to the OOS timing literature that benchmarks against the historical mean and cautions that only robust signals justify departures from buy-and-hold (Welch and Goyal, 2008; Campbell and Thompson, 2008).

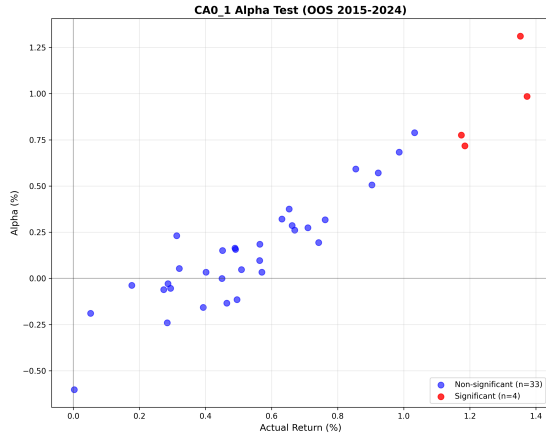
Wrap-up Our results clarify the complementary roles of conditional latent factor models in international equity at the index level. If the objective is primarily explanatory—to attribute realized returns to priced risks and to infer which characteristics proxy for risk—PCA and complicated AE are highly effective. If the objective is primarily predictive, our OOS evidence points to the value of parsimonious AE architectures (e.g., CA0) that stabilize the bias–variance tradeoff in small panels. Practically, a robust workflow is to (i) estimate a conditional AE for OOS forecasting and (ii) use PCA to interpret the learned exposures and to perform stringent alpha diagnostics. This pairing exploits the complementary strengths of representation learning and characteristic-instrumented identification. Evaluation should privilege genuine OOS procedures and inference grounded in established principles for forecast comparison rather than in-sample fit (Campbell and Thompson, 2008).



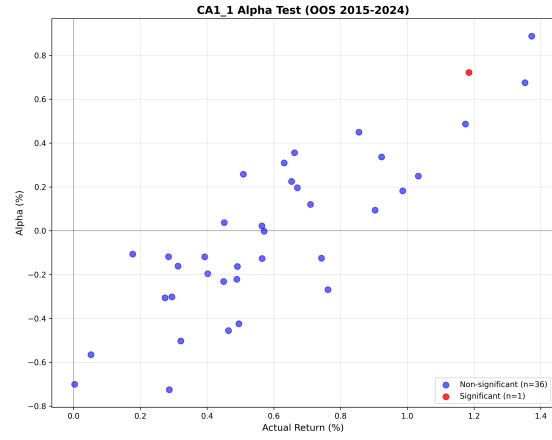
(a) PCA



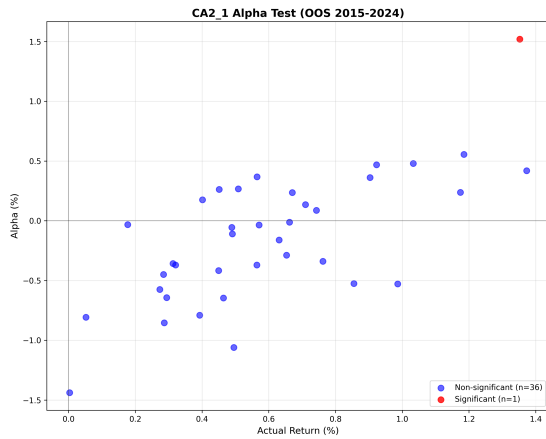
(b) IPCA



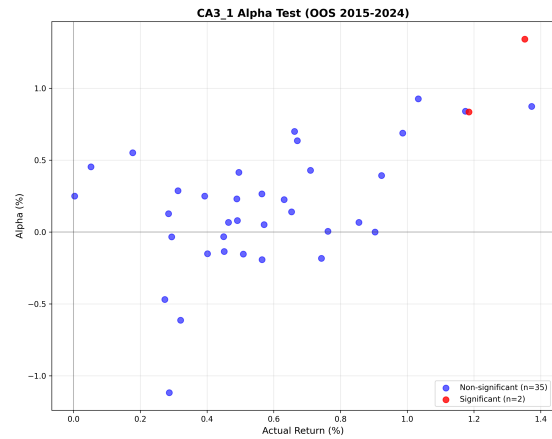
(c) CA0



(d) CA1

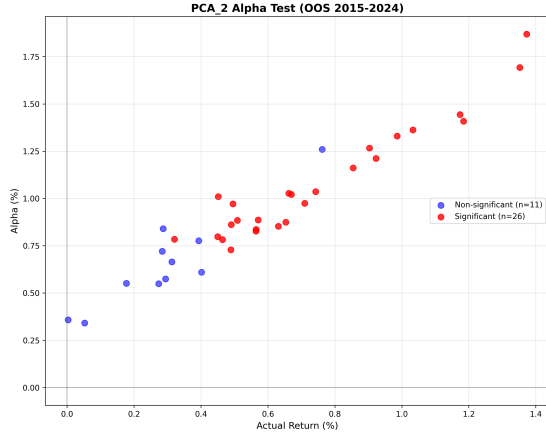


(e) CA2

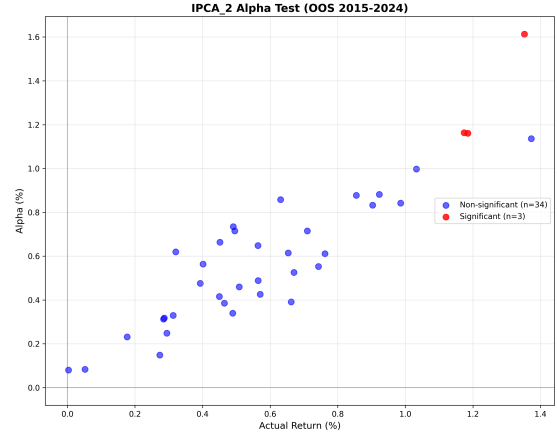


(f) CA3

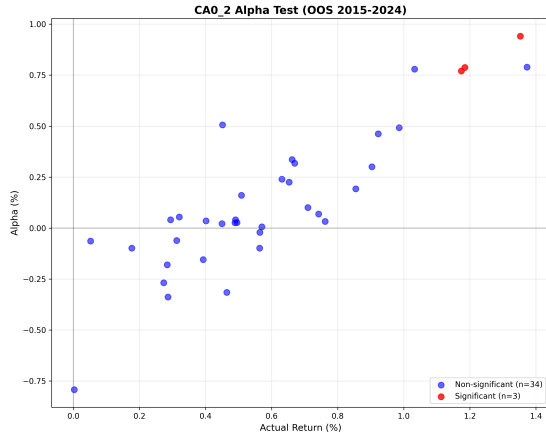
Figure 4: Six models under the same number of latent factors ($K = 1$).



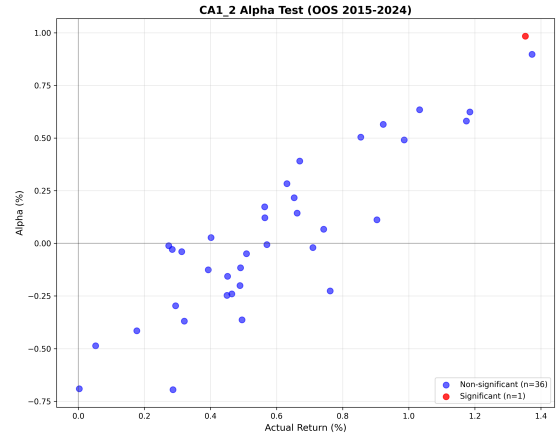
(a) *PCA*



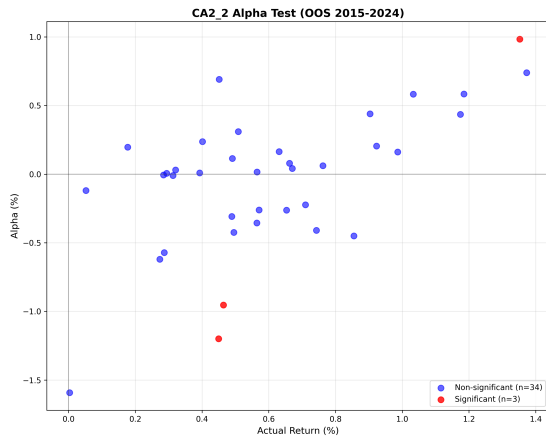
(b) *IPCA*



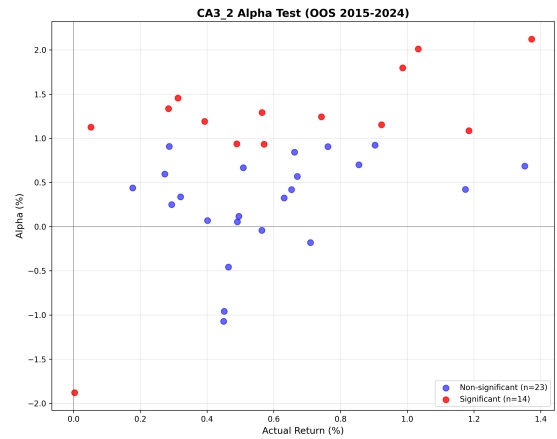
(c) *CA0*



(d) *CA1*

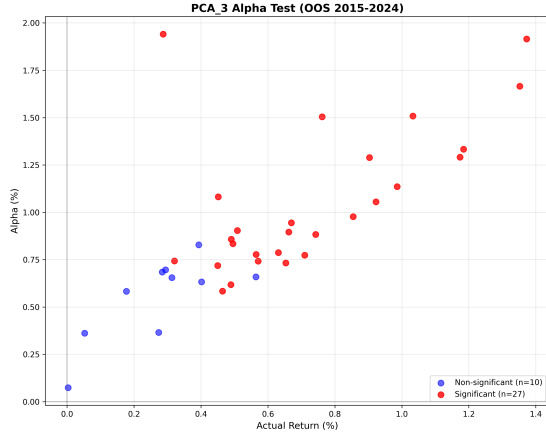


(e) *CA2*

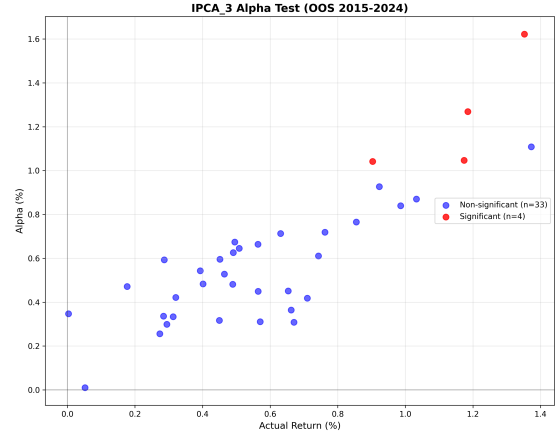


(f) *CA3*

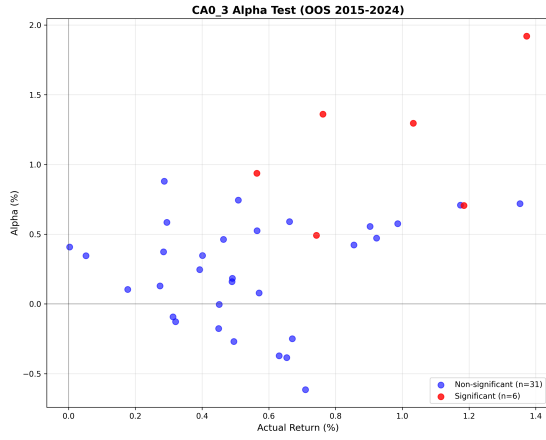
Figure 5: Six models under the same number of latent factors ($K = 2$).



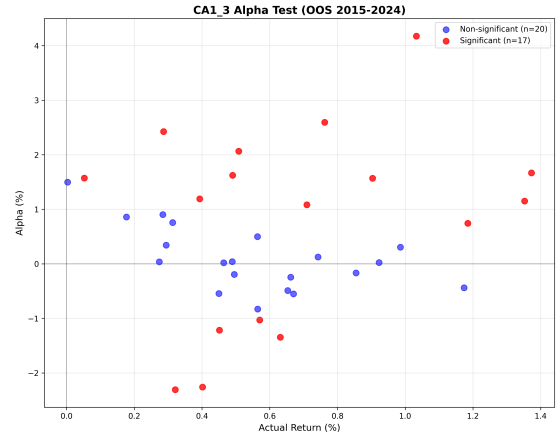
(a) *PCA*



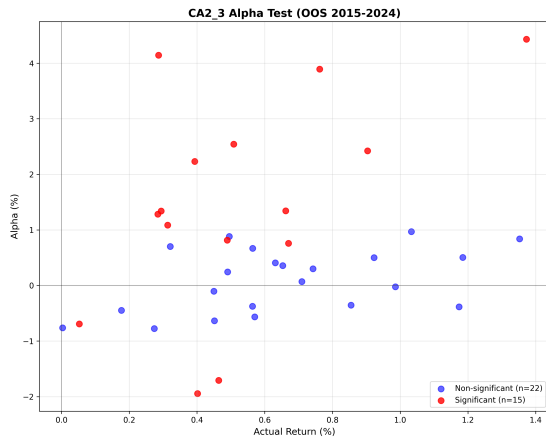
(b) *IPCA*



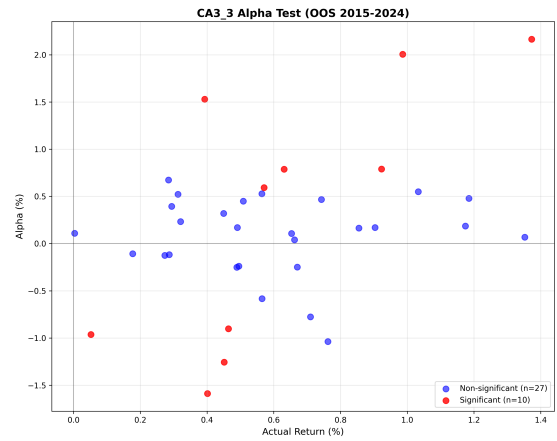
(c) *CA0*



(d) *CA1*

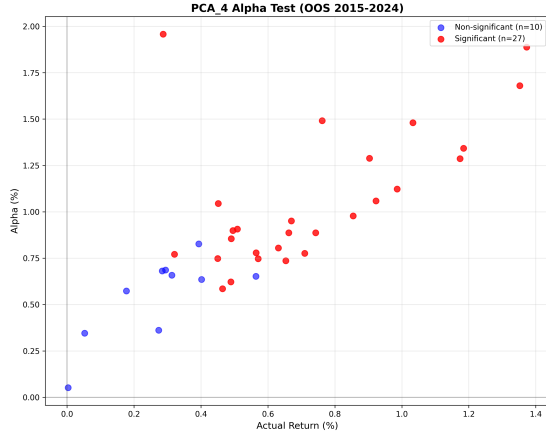


(e) *CA2*

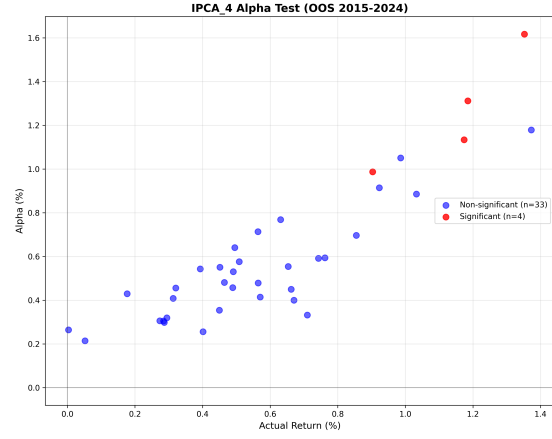


(f) *CA3*

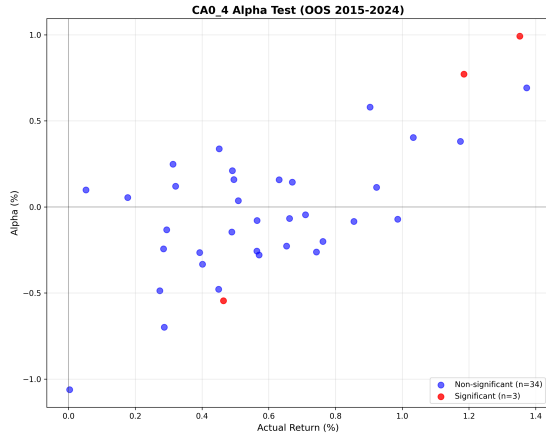
Figure 6: Six models under the same number of latent factors ($K = 3$).



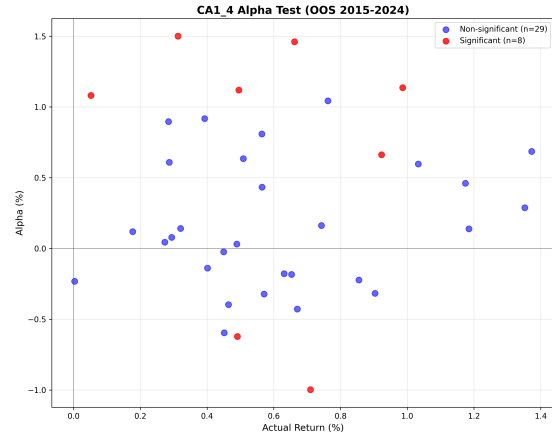
(a) *PCA*



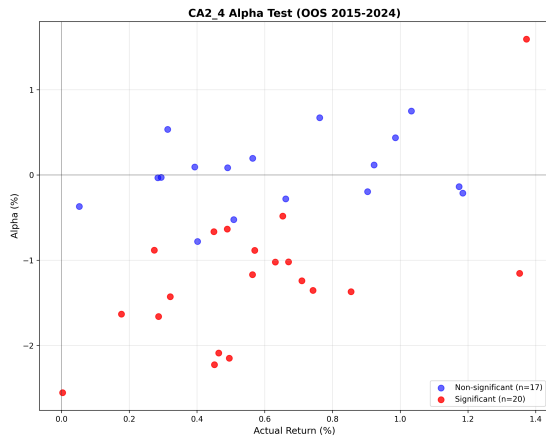
(b) *IPCA*



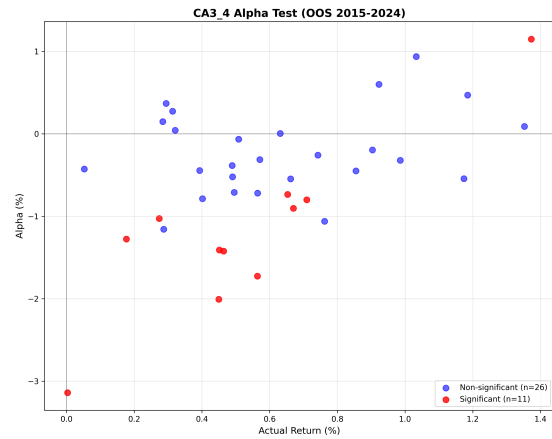
(c) *CA0*



(d) *CA1*

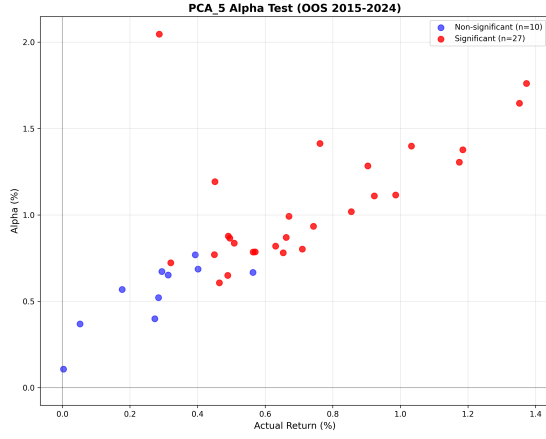


(e) *CA2*

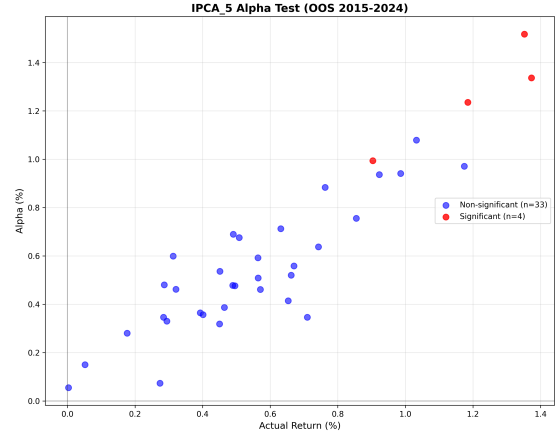


(f) *CA3*

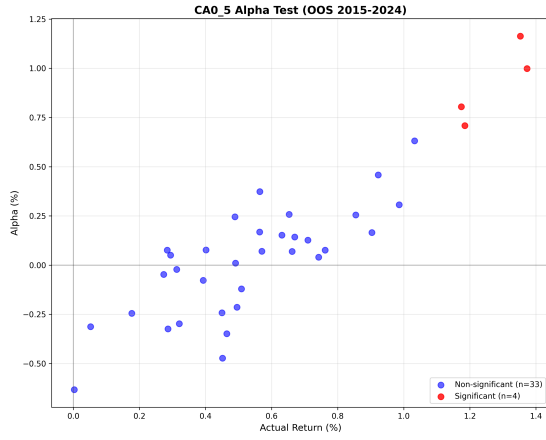
50
Figure 7: Six models under the same number of latent factors ($K = 4$).



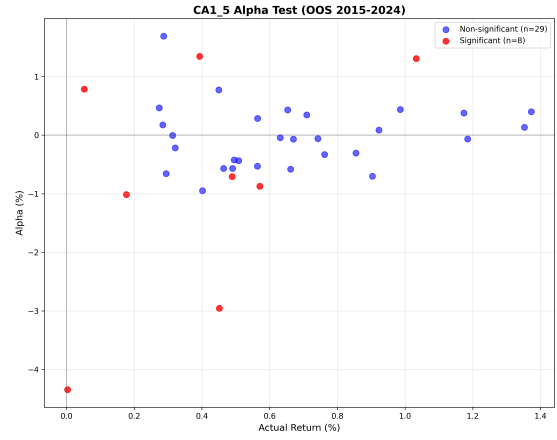
(a) PCA



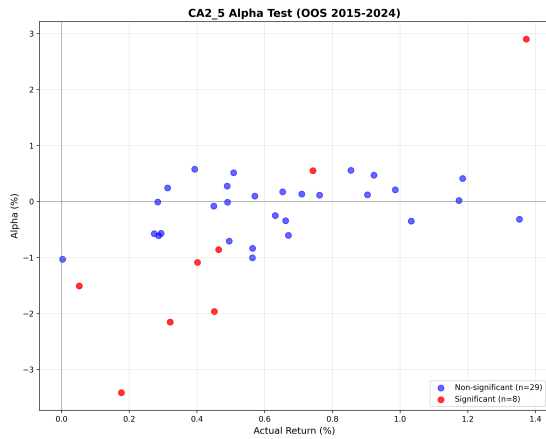
(b) IPCA



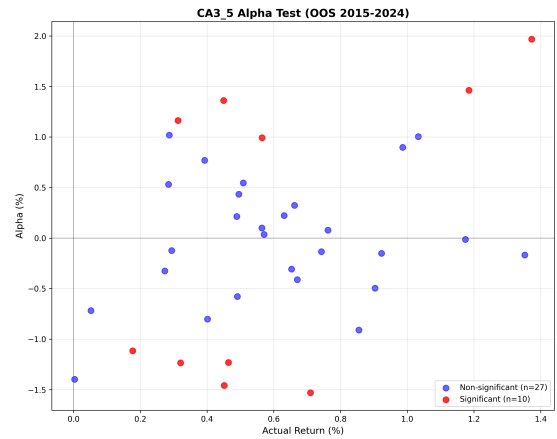
(c) CA0



(d) CA1

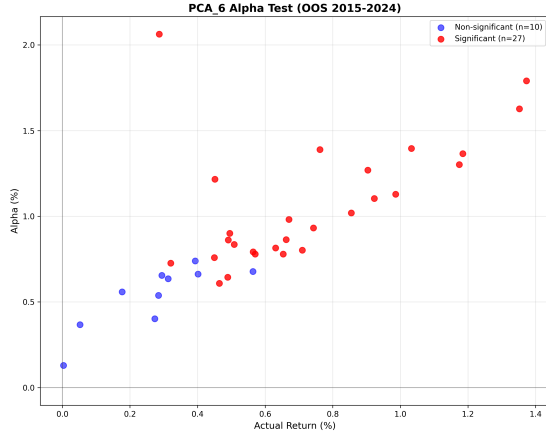


(e) CA2

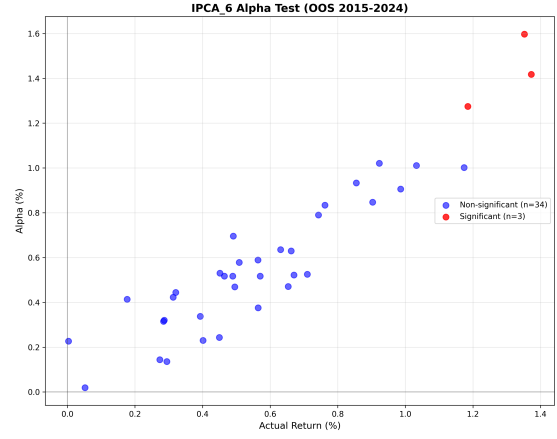


(f) CA3

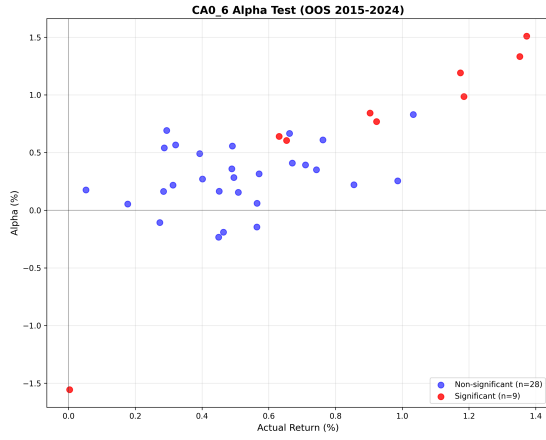
Figure 8: Six models under the same number of latent factors ($K = 5$).



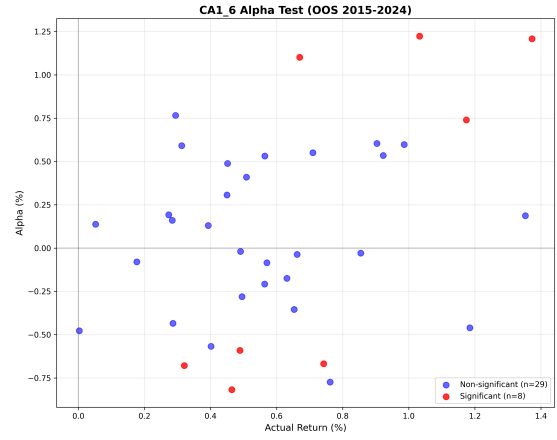
(a) *PCA*



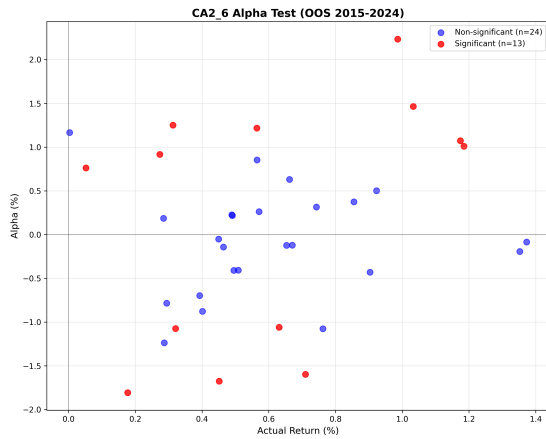
(b) *IPCA*



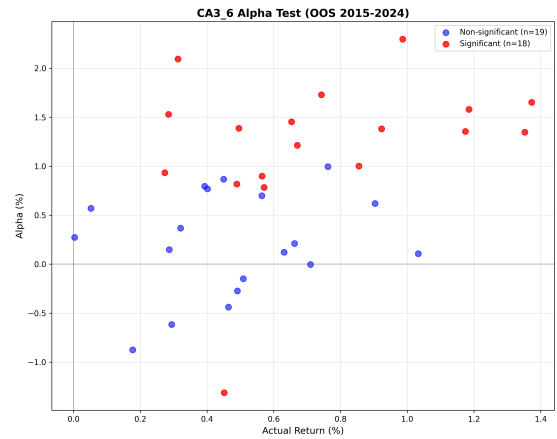
(c) *CA0*



(d) *CA1*



(e) *CA2*



(f) *CA3*

52
Figure 9: Six models under the same number of latent factors ($K = 6$).

6 Interpretability: Feature Importance

Following (Gu et al., 2021), we quantify the influence of each feature by the reduction in model fit that results from removing that characteristic from the input while keeping all learned parameters fixed. Specifically, after training the conditional autoencoder, we compute the model’s total R^2 on the training set. Then, for each characteristic i , we set all values of that characteristic to zero (with all other inputs unchanged and all network weights held fixed), recompute the predictions, and measure the decline in total R^2 . The magnitude of this decline is the variable-importance score for characteristic i . Within each specification we rescale the importance scores so they sum to one, which makes the contributions comparable across characteristics and figures. For comparability, all models are estimated with a six-factor structure ($K = 6$).

Figure 10 visualises the characteristics with the largest importance. Bars are horizontal and ordered from most to least influential. The importances are averaged over training samples, normalized to sum to one within each model, and that all models use $K = 6$. From aggregated firm-level characteristics, the value and momentum themes emerge as the most important factors; on the global side, GDP is the dominant characteristic. Notably, this ranking is stable across model classes—linear (PCA, IPCA) and nonlinear (autoencoders), and in both static and time-varying specifications—underscoring the robustness of our key variables. These variables help explain variation in both expected and realized returns, aligning closely with prior evidence on global value and momentum

Figure 11: We rank all characteristics by aggregating their ranks across models.

Top Twenty Characteristics by Variable Importance (All Models)

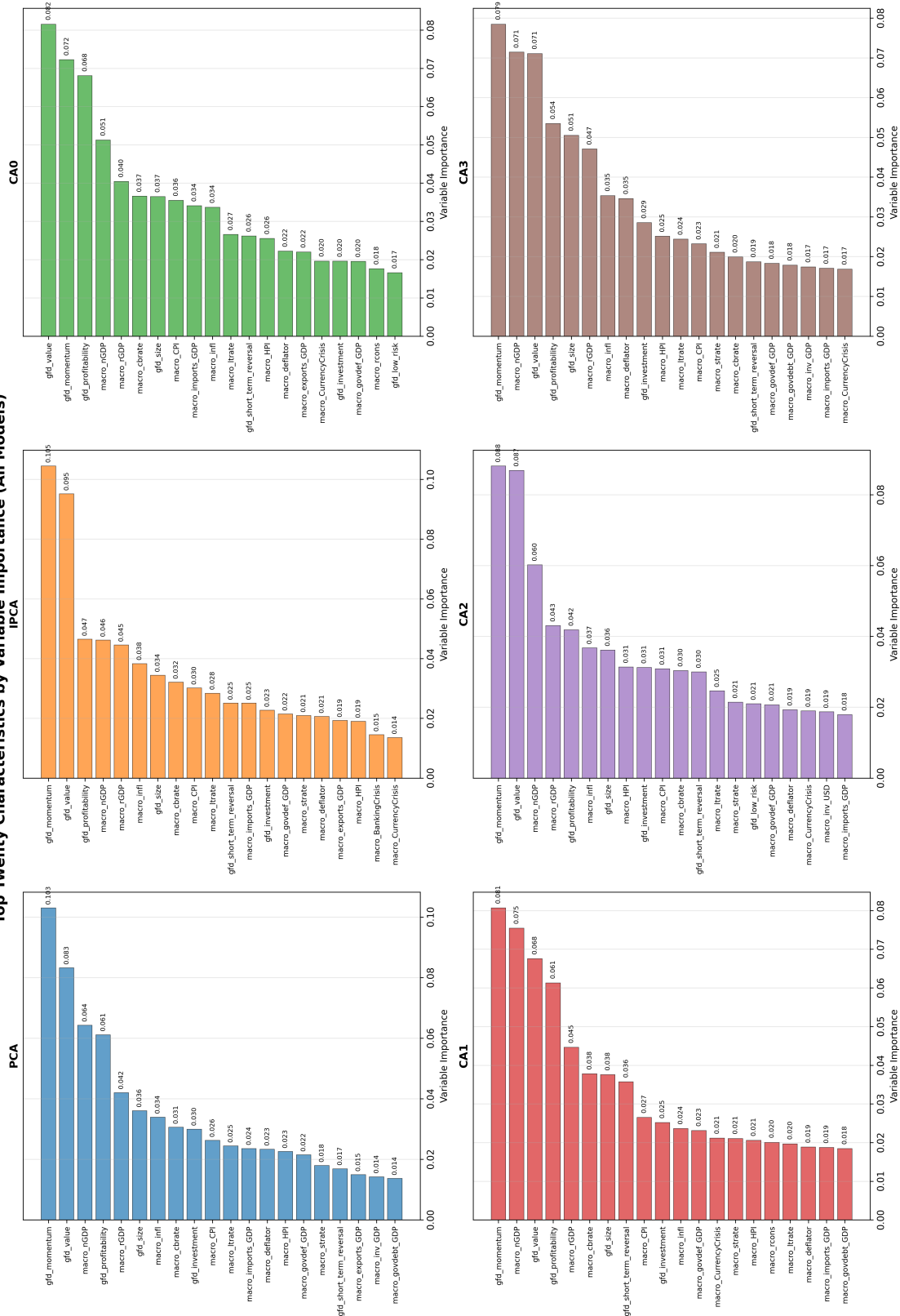


Figure 10: Topmost characteristics

Rows correspond to characteristics; columns correspond to CA_0 – CA_3 . Color intensity reflects the normalized importance within the corresponding model, with darker cells indicating greater influence. Ordering of rows is by the sum of ranks across the four models.

From the perspective of aggregated firm-level characteristics, value and momentum are the most salient thematic factors. From the perspective of macroeconomic characteristics, GDP stands out as the most influential variable. Strikingly, this pattern is consistent across all models—whether linear approaches such as PCA and IPCA or nonlinear ones such as autoencoders, and whether static or time-varying specifications. The stability of variable importance across these distinct modeling frameworks underscores the robustness of our set of key variables and highlights their relevance for explaining variations in both expected and realized returns.

Because the conditional autoencoder consists of a “beta” network producing asset-specific loadings and a “factor” network producing time-specific factors, we also report separate importances for these two components. To obtain “beta-side” importance for characteristic i , we zero out i only in the inputs to the beta network while keeping the factor-network inputs unchanged; we then measure the induced drop in total R^2 . “Factor-side” importance is computed analogously by zeroing the characteristic only in the factor-network inputs. These two rankings are shown to be broadly consistent with the overall results.

Figure 12: We replicate the heatmap using the network-specific importances, displaying side-by-side panels for the beta and factor networks. The row order matches Figure 11 to facilitate comparison. The two decompositions yield similar relative

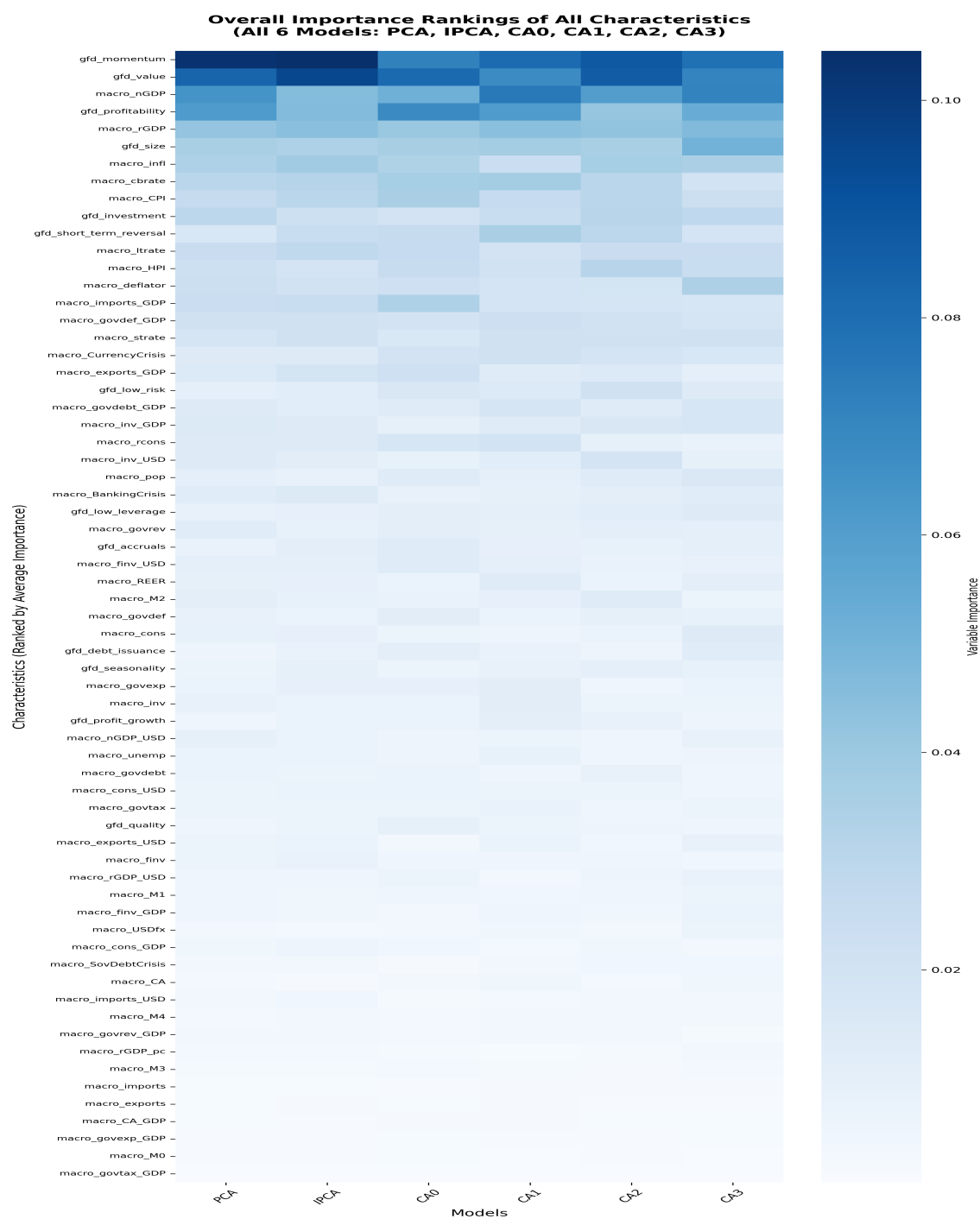


Figure 11: *Overall Importance Rankings of All Characteristics*

orderings, indicating robustness of the ranking to whether the contribution flows through loadings or through factors.

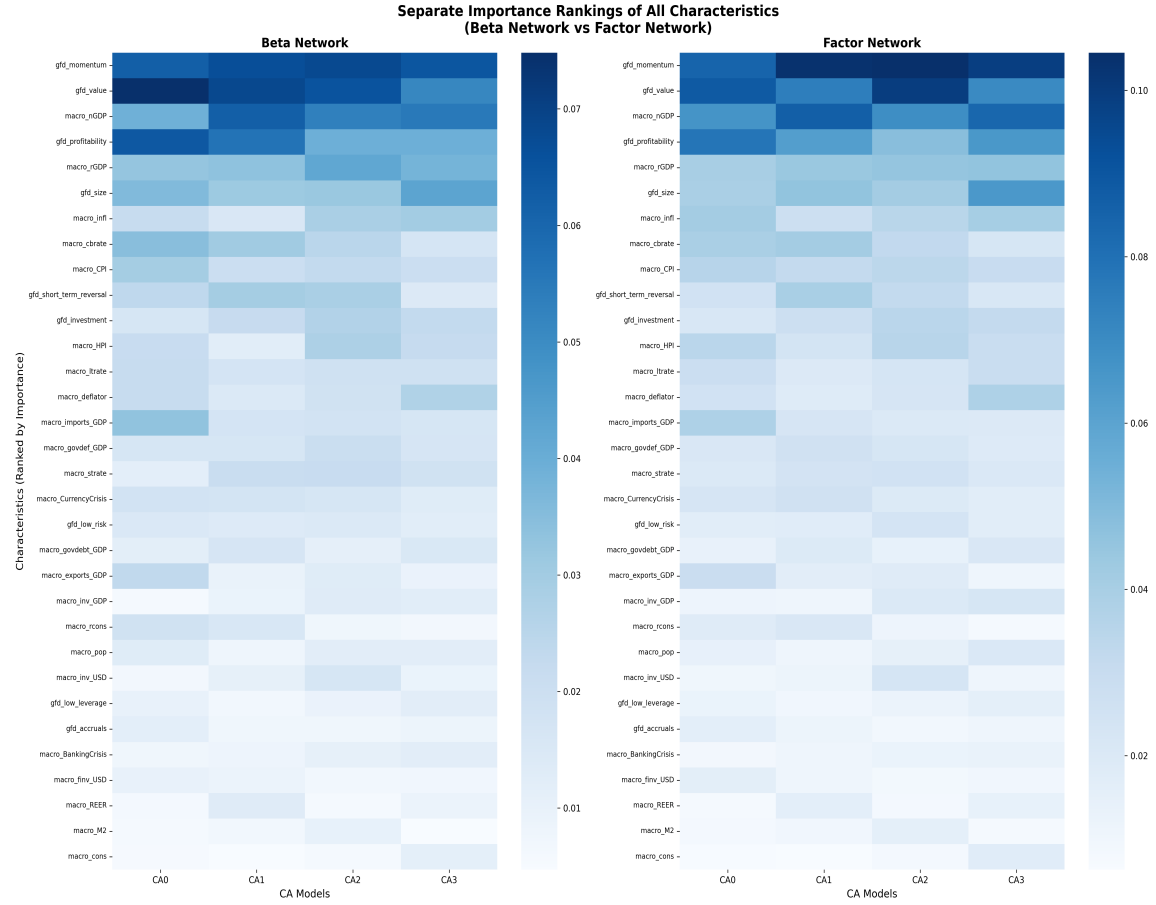


Figure 12: *Separate Importance Rankings of All Characteristics*

The experiment is conducted in-sample and uses total R^2 , the numbers speak to fit rather than out-of-sample predictive value. The alignment of rankings across model variants and across the beta and factor decompositions suggests that the broad patterns are robust.

7 Conclusions

This paper conducted an evaluation of latent factor models—unconditional PCA, conditional-linear IPCA, and a range of conditional-non-linear autoencoders—in their ability to explain and predict returns for 37 international country indices. Our results reveal a divergence between these two objectives. The core empirical finding is that modern conditional factor models are powerful engines for explanation but weak engines for prediction. We show that conditional models (both IPCA and AEs) are powerful explainers, successfully driving pricing errors (alphas) to statistically insignificant levels. However, this in-sample explanatory success fails to translate into out-of-sample predictive power. The only model to deliver a consistently positive, albeit modest, OOS predictive R^2 is the most parsimonious linear autoencoder (CA0), while more complex non-linear models overfit and IPCA fails entirely in this small cross-section environment.

Our findings carry several implications for the asset pricing literature. First, they highlight the crucial role of model complexity relative to the data environment. For country indices, the primary value of machine learning appears to be in providing a robust, data-driven framework for estimating conditional linear factor models, rather than capturing intricate non-linearities. The principle of parsimony remains paramount. Second, our work provides a clear methodological caution on the application of IPCA. Its fragility in small-N panels suggests that its use should be carefully considered outside of the large-N stock-level data where it was developed. Third, the success of conditional models in eliminating alpha, combined with the low economic returns to trading strategies, provides strong evidence on the nature of the

international stochastic discount factor. It suggests the SDF is well-approximated by a small set of global risk factors whose loadings are functions of observable country characteristics, leaving little room for exploitable mispricing. For global investment practice, our results are equally salient. The primary value of these advanced models for global asset allocators lies in risk management and performance attribution, not alpha generation. Furthermore, despite the divergence in predictive performance, our analysis provides a powerful point of consensus. Conditional models, both linear (IPCA) and non-linear (AEs), largely agree on the economic drivers of risk. Across all specifications, Value, Momentum, and GDP-related characteristics emerge as the most important variables. This consensus provides robust, data-driven guidance for “taming” the international factor zoo.

This study opens several avenues for future research. It would be valuable to apply this methodology to other international asset classes, such as currencies, where non-linearities may play a more prominent role. Future research could explore how these non-linear models perform on the full international firm-level cross-section if computational resources possible.

A Appendix

B Data: Summary statistics

Table 6: *Summary Statistics for Global Factor Data (GFD)*

Variable	Mean	Std	Min	Max	Skewness	Kurtosis
accruals	0.00	0.03	-0.36	0.32	-0.09	9.59
debt_issuance	0.00	0.02	-0.31	0.39	0.31	22.24
investment	0.00	0.03	-0.38	0.30	-0.11	14.16
low_leverage	-0.00	0.04	-0.68	0.90	1.30	56.28
low_risk	0.00	0.04	-0.34	0.28	-0.72	8.86
momentum	0.00	0.05	-0.84	0.82	-1.07	25.87
profit_growth	0.00	0.03	-0.55	0.27	-0.77	32.17
profitability	0.00	0.03	-0.32	0.28	-0.13	8.95
quality	0.00	0.04	-0.67	0.41	-1.16	26.58
seasonality	0.00	0.02	-0.39	0.35	-0.65	40.39
short_term_reversal	0.00	0.03	-0.38	0.41	0.08	14.50
size	0.00	0.04	-0.54	0.48	0.60	15.59
value	0.00	0.03	-0.38	0.38	-0.09	12.97

Table 7: *Summary Statistics for Macroeconomic Variables. This table reports the summary statistics (mean, standard deviation, minimum, maximum, skewness, and kurtosis) for the macroeconomic variables used in the analysis. The sample covers 37 countries over the period 1970/01-2024/11. All variables are obtained from the Global Macro Dataset. Detailed definitions and units of measurement for each variable are provided in Table 8.*

Variable	Mean	Std	Min	Max	Skewness	Kurtosis
BankingCrisis	0.04	0.19	0.00	1.00	4.92	22.22
CA	30218.57	22513390.81	-438607264.00	197061024.00	-10.94	196.36
CA_GDP	0.23	5.47	-32.00	30.00	0.70	3.51
CPI	72.37	67.00	0.00	1730.00	11.37	246.48
CurrencyCrisis	0.02	0.15	0.00	1.00	6.26	37.17
HPI	92.88	151.47	0.00	3622.00	16.63	340.08
M0	78792726.94	918754178.36	0.00	16009600000.00	13.92	203.00
M1	39897796.56	264981445.07	0.00	4834599936.00	11.43	156.62
M2	94378473.05	466277248.00	0.00	6136551936.00	7.51	68.24
M3	145685734.35	713906293.57	1.00	7940466688.00	6.64	49.52
M4	1806815.49	3010241.19	151.00	14937212.00	2.64	6.84
REER	103.35	27.64	34.00	424.00	3.39	28.10
SovDebtCrisis	0.01	0.10	0.00	1.00	9.64	90.90
USDFx	220.66	1369.79	0.00	15907.00	8.70	79.10
cbrate	73.48	2434.83	-1.00	103484.00	42.28	1792.68
cons	111499540.52	886257951.18	0.00	15196680192.00	11.96	156.15
cons_GDP	73.88	8.29	35.00	94.00	-1.04	2.22
cons_USD	644596.61	1854532.32	1517.00	24488710.00	7.53	69.58
deflator	73.25	77.83	0.00	2152.00	14.62	342.13
exports	42173530.62	306537862.44	0.00	5107564544.00	11.28	146.69
exports_GDP	40.71	35.73	3.00	229.00	2.65	8.00
exports_USD	213480.62	348631.32	957.00	3269328.00	3.77	19.63
finv	50705613.98	422855723.76	0.00	7365757952.00	12.14	159.91
finv_GDP	23.78	5.09	11.00	53.00	0.94	1.89
finv_USD	190429.59	497283.17	617.00	6586721.00	6.93	61.85
govdebt	75738746.99	547717719.45	0.00	9763210240.00	12.52	180.13
govdebt_GDP	54.08	41.31	2.00	447.00	2.31	10.72
govdef	-3799786.29	39778903.19	-948997696.00	43371392.00	-15.97	295.48
govdef_GDP	-2.27	5.01	-45.00	29.00	-0.33	9.48
govexp	42611272.67	316976495.06	0.00	5367538688.00	11.67	151.46
govexp_GDP	26.65	11.48	0.00	75.00	0.65	0.28
govrev	39887420.18	285583429.73	0.00	4621979648.00	11.13	137.92
govrev_GDP	31.26	13.23	7.00	67.00	0.23	-1.01
govtax	17286097.27	125158981.50	0.00	2151087360.00	11.35	144.83
govtax_GDP	20.63	8.31	5.00	50.00	0.73	0.47
imports	39343465.99	284520193.18	0.00	4707764224.00	10.98	136.71
imports_GDP	39.33	32.45	4.00	221.00	2.77	8.92
imports_USD	211292.73	386387.35	1063.00	4176541.00	4.89	32.99
infl	20.85	205.93	-8.00	7357.00	26.50	839.92
inv	54974169.51	465995970.32	0.00	8437082112.00	12.71	178.84
inv_GDP	24.88	5.54	9.00	55.00	0.92	1.53
inv_USD	194691.60	500290.69	677.00	6500765.00	6.81	59.64
ltrate	6.89	4.70	-1.00	34.00	1.07	1.71
nGDP	169383204.25	1371657864.18	0.00	23981160448.00	12.23	164.71
nGDP_USD	837639.91	2278176.34	1921.00	29839510.00	7.28	66.04
pop	72.75	173.81	2.00	1465.00	5.44	32.83
rGDP	175289243.59	1018263008.82	19177.00	13601876992.00	8.75	85.81
rGDP_USD	1035116.20	2496391.64	15092.00	24185476.00	5.83	39.34
rGDP_pc	1400917.42	5186404.89	6037.00	47592704.00	5.34	30.99
rcons	116745326.76	633850845.48	13.00	7936169472.00	8.23	76.04
strate	15.55	164.82	-1.00	6405.00	33.57	1245.04
unemp	6.55	4.67	0.00	34.00	2.01	6.03

Table 8: Variable Definitions (GMD)

Variable	Description
<i>Crisis Indicators</i>	
BankingCrisis	Binary indicator equal to 1 if the country experiences a banking crisis, 0 otherwise.
CurrencyCrisis	Binary indicator equal to 1 if the country experiences a currency crisis, 0 otherwise.
SovDebtCrisis	Binary indicator equal to 1 if the country experiences a sovereign debt crisis, 0 otherwise.
<i>Macroeconomic Fundamentals</i>	
CA	Current Account balance in nominal local currency.
CA_GDP	Current Account balance as a percentage of GDP.
CPI	Consumer Price Index (Index base year varies by country).
HPI	House Price Index.
M0, M1, M2, M3, M4	Monetary aggregates in nominal local currency.
REER	Real Effective Exchange Rate index.
USDfx	Exchange rate, measured as local currency units per USD.
cbrate	Central bank policy interest rate (percent).
cons	Nominal private consumption in local currency.
cons_GDP	Private consumption as a percentage of GDP.
cons_USD	Nominal private consumption in USD millions.
deflator	GDP Deflator index.
exports	Nominal exports in local currency.
exports_GDP	Exports as a percentage of GDP.
exports_USD	Nominal exports in USD millions.
finv	Fixed investment in nominal local currency.
finv_GDP	Fixed investment as a percentage of GDP.
finv_USD	Fixed investment in USD millions.
govdebt	General government debt in nominal local currency.
govdebt_GDP	Government debt as a percentage of GDP.
govdef	Government deficit (negative values) or surplus in nominal local currency.
govdef_GDP	Government deficit/surplus as a percentage of GDP.
govexp	Government expenditure in nominal local currency.
govexp_GDP	Government expenditure as a percentage of GDP.
govrev	Government revenue in nominal local currency.
govrev_GDP	Government revenue as a percentage of GDP.
govtax	Tax revenue in nominal local currency.
govtax_GDP	Tax revenue as a percentage of GDP.
imports	Nominal imports in nominal local currency.
imports_GDP	Imports as a percentage of GDP.
imports_USD	Nominal imports in USD millions.
infl	Inflation rate calculated as the percentage change in CPI.
inv	Total investment in nominal local currency.
inv_GDP	Total investment as a percentage of GDP.
inv_USD	Total investment in USD millions.
ltrate	Long-term interest rate (typically 10-year government bond yield, percent).
nGDP	Nominal GDP in local currency.
nGDP_USD	Nominal GDP in USD millions.
pop	Total population (in thousands or millions, depending on data source).
rGDP	Real GDP in local currency.
rGDP_USD	Real GDP in USD (constant prices).
rGDP_pc	Real GDP per capita.
rcons	Real private consumption.
strate	Short-term interest rate (typically Treasury bill rate, percent).
unemp	Unemployment rate (percent).

Table 9: *Summary Statistics for MSCI Country Returns*

Table 10: *Summary Statistics for MSCI Country Index Returns. This table presents descriptive statistics for the monthly returns of MSCI Country Indices. Returns are denominated in USD. The sample period ranges from 1970/01-2024/11. The statistics include the mean, standard deviation, minimum, maximum, skewness, and kurtosis for each country's equity index.*

Variable	Mean	Std	Min	Max	Skewness	Kurtosis
Austria	0.01	0.07	-0.42	0.28	-0.51	5.31
Belgium	0.01	0.06	-0.46	0.27	-0.95	8.51
Brazil	0.02	0.13	-0.67	0.81	0.40	5.24
Chile	0.01	0.07	-0.35	0.22	-0.27	1.67
Denmark	0.01	0.06	-0.28	0.29	-0.24	3.08
Finland	0.01	0.08	-0.31	0.30	-0.19	1.75
France	0.01	0.06	-0.34	0.27	-0.38	2.74
Germany	0.01	0.06	-0.35	0.22	-0.54	3.12
Greece	0.01	0.11	-0.43	0.55	0.37	3.40
Hong Kong	0.01	0.09	-0.43	0.88	0.93	12.09
Hungary	0.01	0.10	-0.51	0.50	-0.13	4.66
India	0.01	0.08	-0.33	0.36	-0.14	1.98
Indonesia	0.01	0.13	-0.43	0.94	1.66	12.50
Ireland	0.01	0.07	-0.40	0.19	-1.11	5.08
Israel	0.01	0.07	-0.32	0.27	-0.49	2.60
Italy	0.01	0.07	-0.37	0.31	-0.11	1.71
Japan	0.01	0.06	-0.24	0.24	0.12	0.90
Korea	0.01	0.10	-0.31	0.71	0.89	6.85
Malaysia	0.01	0.07	-0.30	0.50	0.56	7.59
Mexico	0.02	0.09	-0.40	0.29	-0.54	2.29
Netherlands	0.01	0.06	-0.32	0.26	-0.54	3.65
New Zealand	0.01	0.06	-0.26	0.28	-0.08	1.80
Norway	0.01	0.08	-0.40	0.32	-0.55	3.10
Peru	0.01	0.09	-0.42	0.36	-0.20	3.06
Philippines	0.01	0.09	-0.36	0.43	0.13	2.94
Poland	0.01	0.12	-0.39	1.17	2.25	22.30
Portugal	0.00	0.07	-0.34	0.28	-0.16	1.91
Singapore	0.01	0.08	-0.41	0.53	0.39	6.06
South Africa	0.01	0.08	-0.41	0.28	-0.67	3.21
Spain	0.01	0.07	-0.35	0.30	-0.26	2.69
Sweden	0.01	0.07	-0.33	0.26	-0.39	1.91
Switzerland	0.01	0.05	-0.22	0.25	-0.30	1.97
Taiwan	0.01	0.09	-0.34	0.46	0.31	2.19
Thailand	0.01	0.09	-0.34	0.43	0.12	2.90
Turkey	0.02	0.15	-0.39	0.73	0.83	2.77
UK	0.01	0.06	-0.36	0.56	0.58	11.05
USA	0.01	0.05	-0.32	0.23	-0.96	5.99

References

- Adler, M. and Dumas, B. (1983). International portfolio choice and corporation finance: A synthesis. *The Journal of Finance*, 38(3):925–984.
- Angelidis, D. and Tassaromatis, N. (2016). Global equity investing: The benefits of value, momentum, low-volatility, and size. *The Journal of Portfolio Management*, 42(3):59–69.
- Asness, C. S., Liew, J. M., and Stevens, R. L. (1997). The interaction of value and momentum strategies. *Financial Analysts Journal*, 53(2):29–36.
- Avramov, D., Chordia, T., Jostova, G., and Philipov, A. (2012). Credit risk and the cross-section of stock returns. *Journal of Financial Economics*, 105(2):331–348.
- Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.
- Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58.
- Bekaert, G. and Harvey, C. R. (1995). Time-varying world market integration. *The Journal of Finance*, 50(2):403–444.
- Bekaert, G. and Harvey, C. R. (1997). Emerging equity market volatility. *Journal of Financial economics*, 43(1):29–77.
- Ben-David, I., Franzoni, F., and Moussawi, R. (2017). Exchange-traded funds. *Annual Review of Financial Economics*, 9(1):169–189.

- Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531.
- Chamberlain, G. and Rothschild, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets.
- Chen, L. and Xiu, D. (2023). Deep learning in asset pricing. *Review of Financial Studies*, 36(1):1–45.
- Connor, G. and Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of financial economics*, 15(3):373–394.
- Dumas, B. and Solnik, B. (1995). The world price of foreign exchange risk. *The journal of finance*, 50(2):445–479.
- Erb, C. B., Harvey, C. R., and Viskanta, T. E. (1996). Political risk, economic risk, and financial risk. *Financial Analysts Journal*, 52(6):29–46.
- Errunza, V. and Losq, E. (1985). International asset pricing under mild segmentation: Theory and test. *The Journal of Finance*, 40(1):105–124.
- Eun, C. S. and Shim, S. (1989). International transmission of stock market movements. *Journal of financial and quantitative Analysis*, 24(2):241–256.
- Fama, E. F. and French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of financial economics*, 25(1):23–49.

- Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.
- Fama, E. F. and French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics*, 105(3):457–472.
- Ferson, W. E. and Harvey, C. R. (1999). Macroeconomic predictors of international equity returns. *The Review of Financial Studies*, 12(3):689–729.
- Fisher, G. S., Shah, R., and Titman, S. (2017). Should you tilt your equity portfolio to smaller countries? *Journal of Portfolio Management*, 44(1):127.
- Griffin, J. M. (2002). Are the fama and french factors global or country specific? *The Review of Financial Studies*, 15(3):783–803.
- Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.
- Gu, S., Kelly, B., and Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450.
- Hanauer, M. X. and Kalsbach, S. (2023). Predicting emerging market returns using machine learning. *Journal of International Money and Finance*.
- Harvey, C. R. (1995). Predictable risk and returns in emerging markets. *The review of financial studies*, 8(3):773–816.
- Henry, P. B. (2000). Stock market liberalization, economic reform, and emerging market equity prices. *The Journal of Finance*, 55(2):529–564.

- Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.
- Karolyi, G. A. and Stulz, R. M. (2003). Are financial assets priced locally or globally? *Handbook of the Economics of Finance*, 1:975–1020.
- Kasa, K. (1992). Common stochastic trends in international stock markets. *Journal of monetary Economics*, 29(1):95–124.
- Kelly, B., Pruitt, S., and Su, Y. (2019a). Instrumented principal component analysis. *The Review of Financial Studies*, 32(12):4821–4874.
- Kelly, B. T., Pruitt, S., and Su, Y. (2019b). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.
- Kim, J., Lee, H., and Park, S. (2023). Conditional autoencoder asset pricing in emerging markets: Evidence from korea. *Emerging Markets Review*.
- Lettau, M. and Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies*, 33(5):2274–2325.
- Müller, K., Xu, C., Lehib, M., and Chen, Z. (2025). The global macro database: A new international macroeconomic dataset. Technical report, National Bureau of Economic Research.
- Pukthuanthong, K. and Roll, R. (2009). Global market integration: An alternative measure and its application. *Journal of Financial Economics*, 94(2):214–232.

- Richards, A. J. (2012). Winner-loser reversals in national stock market indices: Can they be explained? *Journal of Finance*, 67(5):2113–2148.
- Sercu, P. (1980). A generalization of the international asset pricing model. *Revue de l'Association Française de Finance*, 1:91–135.
- Solnik, B. (1974). An equilibrium model of the international capital market. *Journal of Economic Theory*, 8(4):500–524.
- Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.
- Stulz, R. (1981a). A model of international asset pricing. *Journal of financial economics*, 9(4):383–406.
- Stulz, R. M. (1981b). A model of international asset pricing. *Journal of Financial Economics*, 9(4):383–406.
- Tobek, M. and Hronec, M. (2021). Machine learning equity premium prediction in international markets. *International Review of Financial Analysis*, 76:101737.
- Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.