UTILIZATION OF MACHINE LEARNING FOR EMPIRICAL ASSET PRICING IN EMERGING MARKETS

BY

DASTAN ADILOV

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Finance
in the Graduate School of Business
Nazarbayev University, 2024

Astana, Kazakhstan

Advisor: Francesco Rocciolo

Abstract

We perform Principal Component Regression (PCR) analysis to predict cross-sectional stock returns in emerging market economies. As a benchmark comparison, we employ OLS models and demonstrate predictive power of the machine learning based PCR model. We utilize 64 firm characteristics to determine the most significant predictors for the emerging market countries as well as for individual countries. The results, demonstrate predictive power of the PCR model over the linear regression model, showing consistent results in both the country-specific analysis and in the overall analysis of the emerging market. The most important set of predictors throughout the analysis proved to be book-to-market, sales-to-price, leverage (lev), cash flow-to-price (cfp), dividends (dy), and gross profitability (gma).

Keywords: Principal component regression, emerging market economies, machine learning, stock returns, predictors.

Utilization of Machine Learning for Empirical Asset Pricing in Emerging Markets

Emerging markets, with their unique and complicated characteristics – such as lower liquidity, higher volatility, data shortage – delivers significant challenges to the traditional empirical asset pricing models. This paper examines how machine learning-based model, utilizing a comprehensive set of predictors, can effectively predict cross sectional stock returns in the emerging market, and identify the most important predictive factors.

Emerging markets have appeared to be a crucial area of study with its unique characteristics and global importance. Unlike developed markets, which have established valuation methodologies, emerging markets lack a clear understanding of the determinants of returns. In addition, emerging markets attract a significant amount of investment, growing at a rate two or three times higher than that of developed market countries (Bruner et al., 2002). Therefore, the analysis of the emerging market provides an opportunity to identify predictors that can capture determinants of returns. Traditional asset pricing models, such as the CAPM or Fama – French factor models that are used in developed markets tend to underperform under this setting. The incorporation of profitability measures, valuation ratios, leverage indicators into the empirical studies will improve the precision of return predictability, contributing to a more in-depth analysis of the structure of emerging market. The research will allow to bridge a gap between the literature and investors, providing insights of

economic and market conditions and navigating through complex market structure.

Machine learning has become a transformative tool for addressing the challenges of high – dimensionality, which frequently arise in traditional statistical models. The research by Gu et al. (2020) in the paper "Empirical Asset Pricing via Machine Learning" demonstrated the capabilities of various machine learning models. By employing PCR, PLS, regression trees and neural networks, they highlighted that machine learning models could manage high – dimensionality and discovered nonlinear relationship between predictors. They identified momentum, volatility, and liquidity as the most important factors that contribute to the prediction of stock returns in the US. In addition, the research showed that machine learning-based models deliver a considerable amount of economic gains, as evidenced by a higher value of the Sharpe ratio in portfolio strategies. This research provided an understanding on how machine learning can surpass traditional regression – based approaches and handle the challenges of asset pricing.

Based on this framework, this paper utilizes Principal Component Regression (PCR) and Ordinary Least Squares (OLS) regression to an emerging market dataset consisting of 64 firm characteristics over a 16-year time period (2008 – 2023). The results highlight the limitation of OLS when facing high – dimensionality. The OLS model yielded an out-of-sample R^2 of -0.467% which indicates an overfitting issue and poor performance. Restricting OLS to size,

value, and momentum improved the model's performance, delivering an out-of-sample R^2 of 0.413%. In contrast, the PCR outperformed both models, achieving an out-of-sample R^2 of 0.534%.

Further analysis revealed a similar pattern, where PCR outperformed the other models in each emerging market country. For example, in China, PCR achieved an out-of-sample R^2 of 0.264% with 26 components, while full OLS model showed an out-of-sample R^2 of -3.030%. A similar result was obtained for Taiwan, South Korea, Thailand, Malaysia, Turkey, Poland, Indonesia, and Pakistan. However, for the Indian market, the OLS model restricted to three predictors, outperforms PCR model, achieving an out-of-sample R^2 of 0.919% and 0.899% respectively.

Across the analysis of the emerging market, the most influential stock-level predictors were related to valuation ratios and fundamental signals. Specifically, these included book-to-market (bm), sales-to-price (sp), leverage (lev), and cash flow-to-price (cfp). These characteristics are highly relevant for the emerging market, where financial systems are underdeveloped and markets are volatile, implying that fundamental valuation metrics are particularly important to analyze. Further analysis showed that country-specific predictors differ from the analysis of the emerging market countries as a whole. For example, China India, Taiwan, South Korea, and Poland converge on the importance of price trends such as change in momentum, 6-month momentum, and trading volume. However, Thailand highlighted the importance of

profitability and liquidity measures. This thereby, indicating the persistence of country-specific, unique market structures and the significance of the analysis.

In addition, these findings highlight the benefits of applying PCR for the reduction of overfitting, and utilization of high – dimensional data. By incorporating these results, this paper expands the literature on stock return prediction for emerging market economies, and demonstrates the capability of machine learning models in contrast to traditional approaches.

Data and Methodology

The dataset used in the paper consists of monthly stock returns for all listed financial companies from 25 emerging market economies: Kazakhstan, Brazil, Chile, China, Colombia, Czech Republic, Egypt, Greece, Hungary, India, Indonesia, Malaysia, Mexico, Pakistan, Peru, Philippines, Poland, Qatar, Saudi Arabia, South Africa, South Korea, Taiwan, Thailand, Turkey, and the United Arab Emirates. The dataset includes the largest pool of assets to avoid the sample selection bias or data snooping bias, as well as the overfitting. To be specific by increasing both the number of parameters and observations, the results are quantitatively unchanged and qualitatively identical if those firms are excluded. Monthly stock returns were collected from the Eikon Refinitiv database with a time span of 16 years, from January 2008 to December 2023. The total number of stocks included in the dataset is 11,038. For the calculation of excess returns, the risk-free rate for the emerging markets was obtained from the Kenneth R French Data Library.

In addition, the dataset was enriched with an additional set of stock-level predictors. These include, 64 firm characteristics, of which 47 are updated annually, 7 quarterly, and 10 monthly. The variables were transformed according to Green et al. (2017) and adjusted for the emerging market economies based on the availability of the data. A detailed description of these characteristics is provided in the appendix.

Throughout the analysis, the independent determinants of average returns were identified by regressing returns on all predictors simultaneously. This approach guarantees that none of the characteristics were excluded due to missing values for a certain firm in a particular month. Notably, almost all of the characteristics in the dataset have missing observation throughout the time period. To retain as much data as possible, the dataset was winsorized at the 1st and 99th percentiles, to limit the impact of outliers. In addition, the rest of the missing values were replaced with zeros. This approach was applied to all 64 predictors, ensuring that the dataset is robust and unbiased.

For the machine learning analysis, we selected Principal Component Regression (PCR) to approximate the empirical model $E_t(r_{i,t+1}) = g^*(z_{i,t})$ where $E_t(r_{i,t+1})$ represents the excess return, and $g^*(z_{i,t})$ is a vector of firm characteristics structured as a P×1.

Principal component regression

Principal component regression (PCR) is a dimension reduction technique that addresses the issue of variable selection. During the analysis of a large

number of explanatory variables, traditional regression models often produce suboptimal results. Multicollinearity among explanatory variables complicates the process of assessing the effect of each variable on the predicted variable. PCR alleviates this issue and provides stable coefficient estimates.

PCR works in two steps. First, Principal Component Analysis (PCA) transforms the original variables into their linear combinations, called components. These components preserve the original covariance structure of the data. Then, these principal components are used as new explanatory variables in the principal component regression.

Despite the advantage of dimension reduction, PCR has its own limitations. For instance, principal components are selected according to the explained variance among the predictors, but not the relevance to the predicted variable. As a result, characteristics with low variance may be dropped, even if they are significant. In addition, after the construction of the components, they are challenging to interpret.

Sample Splitting and Tuning

A crucial preliminary step for the machine learning analysis is the selection of hyperparameters. Hyperparameter tuning plays a crucial step in controlling model complexity and accuracy, as it addresses the issue of overfitting and ensures robust predictive performance. In this study, the selection of hyperparameters for the PCR is achieved through the use of a validation sample. This approach is commonly used in the literature, including

by Gu et al. (2020). Specifically, the dataset was chronologically divided into three non-overlapping subsets: training subsample, validation subsample, and testing subsample. The sample splitting scheme used in this paper follows the approach outlined by West (2006). Following this approach, the PCR model is estimated using the training and validation subsamples and evaluated using the testing subsample.

The training subsample covers 5 years (2008–2012) of the original dataset. It was used for the calculation of principal components that will serve as a foundation of the PCR model. The validation sample, covering 3 years (2013–2015), is used for the calculation of an out of sample R^2 for 40 models, and determination of the optimal number of components or hyperparameters. Finally, these results are applied for the last testing subsample (2016–2023) to evaluate the predictive performance of the PCR model.

Models Comparison and Performance Metric

In order to compare the predictive power of the PCR model, two Ordinary Least Squares (OLS) models were included in the analysis. The first model used the full set of predictors. The other model, OLS-3 was restricted to three predictors: size, value and momentum. Given the high dimensionality of the dataset, these models performed significantly worse than the PCR model.

Additionally, to evaluate the predictive performance of the PCR model, we used the out-of-sample R^2 values, obtained from the equation (1). It is calculated as a proportion of variance in the excess return explained by the

model's prediction on the testing subsample. The testing subsample is denoted as τ_3 and it was not involved in the model selection and hyperparameter selection. This measure consolidates prediction errors across firms and time into a single metric for each model.

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\epsilon\tau_3}\left(r_{i,t+1} - \hat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\epsilon\tau_3} r^2_{i,t+1}} \quad (1)$$

A critical difference of this method of calculating an out-of-sample R^2 value is that in the denominator the total sum squared excess returns is not centered around the mean. This approach deviates from the traditional method, where the denominator is benchmarked against the mean. Particularly, traditional approach suffers from bias when applied to individual stock returns. Due to the high noise in historical mean returns, benchmarked against the mean underperforms relative to naive zero forecast, thereby, lowering model performance. Therefore, to address this issue, we used equation (1).

Empirical analysis

Following the preparation of the dataset, the Principal Component Analysis (PCA) was applied to identify key components that capture a significant portion of the variability in the data. Throughout the analysis, 40 models were computed iteratively. These models were assessed using a performance metric, which guided the selection of the optimal number of components for the Principal Component Regression. Subsequently, this

number of components was applied for the regression analysis and prediction of stock returns.

The next step of the research involved a variable importance analysis. This was done to identify the drivers of the stock returns. The procedure begins with a construction of principal components on a set of predictors via PCA. After the identification of components, the loadings or weights of the original variables contributed to each component are stored in a separate matrix. Subsequently, the PCR model is estimated using principal components. Then, the loadings of the predictors are multiplied with the regression coefficients which captures the contribution of each variable to the predictive power of the model through the principal components. In order to ensure the comparability of the results, the importance scores were normalized to sum to one. The results are visualized in Figure 1, where all variables are ranked by their importance.

Results

Table 2 summarizes the results of the comparison between OLS, OLS – 3 and PCR model's predictive performance. The unrestricted OLS model for the emerging markets economies produced an out-of-sample R^2 of -0.467%. This outcome is not surprising. In the absence of regularization, OLS model becomes highly sensitive to overfitting the sample data. However, limiting OLS to three predictors – size, value, and momentum – achieves a better result, producing an out-of-sample R^2 of -0.403%. In contrast, PCR outperformed both models with an out-of-sample R^2 of 0.534%.

Table 2
Monthly out-of-sample stock-level prediction performance (percentage R^2_{OOS})

Country	Companies	Components	OLS	OLS - 3	PCR
All	11,038	2	-0.467	0.403	0.534
China	3,003	26	-3.030	-0.880	0.264
India	1,926	10	0.509	0.919	0.899
Taiwan	1,675	32	-0.506	0.730	0.794
S. Korea	1,652	35	-4.610	-1.610	0.590
Thailand	431	1	-0.709	-0.009	0.708
Malaysia	518	9	-0.384	1.040	1.080
Turkey	262	12	-0.547	0.707	1.052
Poland	288	33	-1.190	1.062	1.477
Indonesia	329	10	-0.51	0.010	1.019
Pakistan	236	14	-2.115	0.099	0.899

In this table, we report monthly out of sample R squared values for the total set of predictors using OLS, OLS - 3 using size, value, and momentum, PCR with optimal number of components. The table provides values for all emerging market countries, as well as for each country individually.

In addition to analyzing the results for the emerging market countries as a whole, individual country-specific analyses provide additional information about the predictive performance of models. For example, in China, the PCR

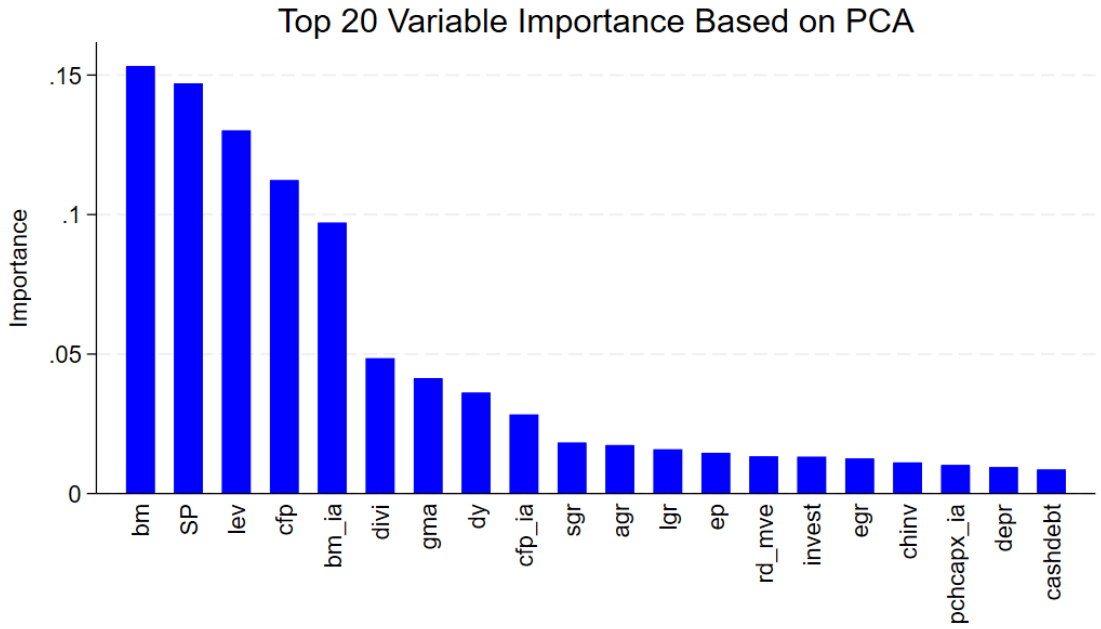
model demonstrates an out-of-sample R^2 of 0.264%, utilizing 26 components. In contrast, the full OLS and restricted OLS models achieves an out-of-sample R^2 of -3.030% and -0.880% respectively. This implies that these models are not suitable for the prediction of stock returns for the Chinese market. These findings are consistent with the results obtained by Zhao et al. (2023), where machine learning-based models outperformed linear models. In Taiwan, the results were similar to Chinese market, where the PCR model outperformed all other models with an out-of-sample R^2 of 0.794%. In South Korea, PCR with 35 components yielded an out-of-sample R^2 of 0.590%. However, OLS and OLS – 3 models yielded -4.610% and -1.610% respectively. This implies that the OLS models struggle to capture the dynamics of stock returns in a South Korean market.

In contrast, in India, the PCR was outperformed by the OLS – 3 model, achieving an out-of-sample R^2 of 0.919%. To compare, PCR model yielded an out-of-sample R^2 of 0.899%, utilizing 10 components. This may have occurred due to the linear relationship of input variables for an Indian market.

Overall, the results demonstrated consistent result throughout all emerging market countries, highlighting the significance of dimension reduction. Apart from that, the findings are consistent with the results obtained by Gu et al. (2019), where PCR model outperformed both OLS and OLS – 3 models for the US market.

We now analyze the variable importance of characteristics for the PCR model. As shown in Figure 1, the most influential stock-level predictors are book-to-market (bm), sales-to-price (sp), leverage (lev), cash flow-to-price (cfp), dividends (dy), and gross profitability (gma). Book-to-market reflects the firms' value relative to its market value of equity, providing insights about the financial health and market sentiment. Sales-to-price indicates the amount of revenue generated relative to the stock price, allowing to assess whether the stock is overvalued or undervalued. In the emerging markets context, due to the lack of market efficiency, stock prices may not reflect all of the information. However, being characterized as a fast-growing market, emerging market countries has a high sales growth. This may reflect the reason beyond the significance of the variable. Leverage reflects the debt level which is a crucial metric for the assessment of the risk. Leverage is particularly relevant for the emerging markets, as it shows the firms with a high level of debt which causes a financial distress. Thereby, affecting the expected returns. Cash flow-to-price provides an indication regarding the amount of cash generated relative to the stock price. To be specific, it tells investors how much cash they will receive for a dollar spent on a stock. The importance of this characteristic implies that the financial health of the firms in the emerging market may provide valuable insights regarding the expected stock returns. Dividends and gross profitability also point to the importance of fundamental valuation metrics.

These characteristics jointly capture key drivers of the firm performance in the emerging markets context. Specifically, valuation, risk, growth, and operational efficiency ratios are particularly relevant for the investment decision.



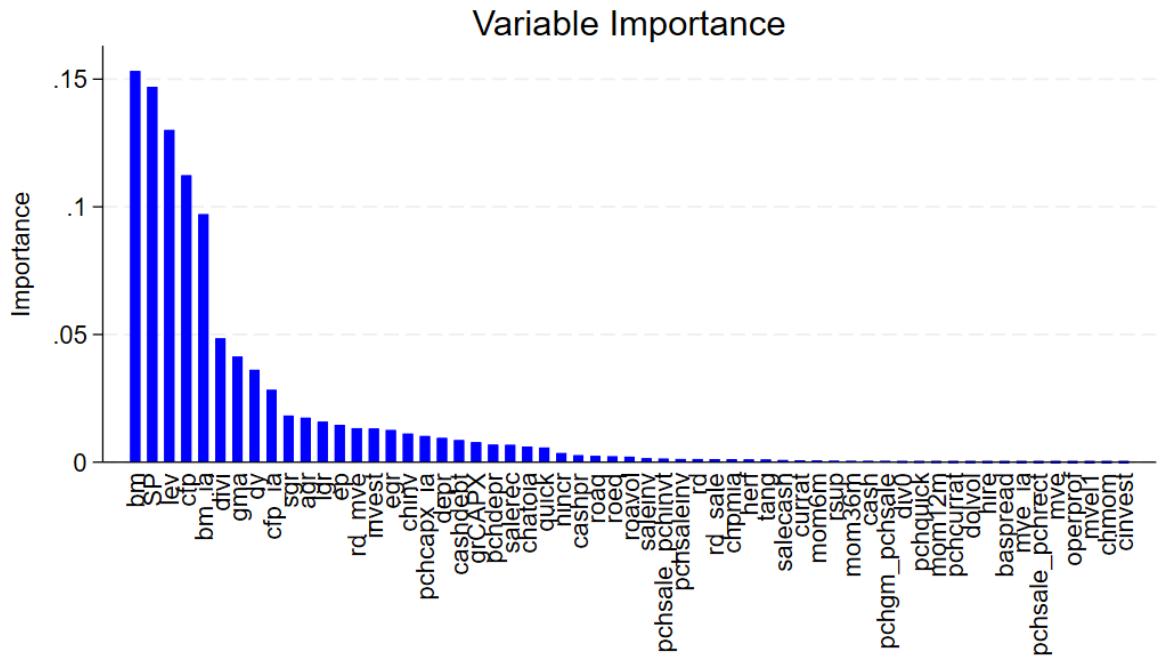


Figure 1
Variable importance for PCA
Variable importance graphs for the top – 20 and all variables for the PCR model, normalized to sum to one.

In comparison with the US market, as analyzed by Gu et al. (2019), price trends such as momentum, short-term reversal did not appear to be among the most important stock-level predictors. One of the reasons for such difference may occur due to the availability of highly liquid instruments in the US market. Thereby, investors are capable of capturing short-term inefficiencies. In addition, US investors may be more speculative, focusing on price trends. In contrast, emerging market investors might prioritize fundamental signals, which may protect during higher economic instability and external shocks.

Extending the analysis, country-specific variable importance diverges from the results obtained for all of the emerging market countries. For the countries such as India, China, Taiwan, Poland, South Korea, important characteristics are consistent. Price trends such as change in momentum, 6-month momentum, 12-month momentum, dollar trading volume, and corporate investment are universally important and dominate the ranking. The results are similar to the US market possibly indicating that in the presence of a large amount of stocks and high liquidity, price trends perform.

However, Thailand has different results. Variable importance is dominated by return on assets (roaq), return on equity (roeq), return volatility (roavol), number of earnings increase (nincr), and cash holdings (cash). These

results highlight the significance of profitability measures and liquidity. In the context of emerging markets, profitable and cash rich firms withstand economic shocks and volatile market conditions better, allowing them to access the financing in a constrained environment.

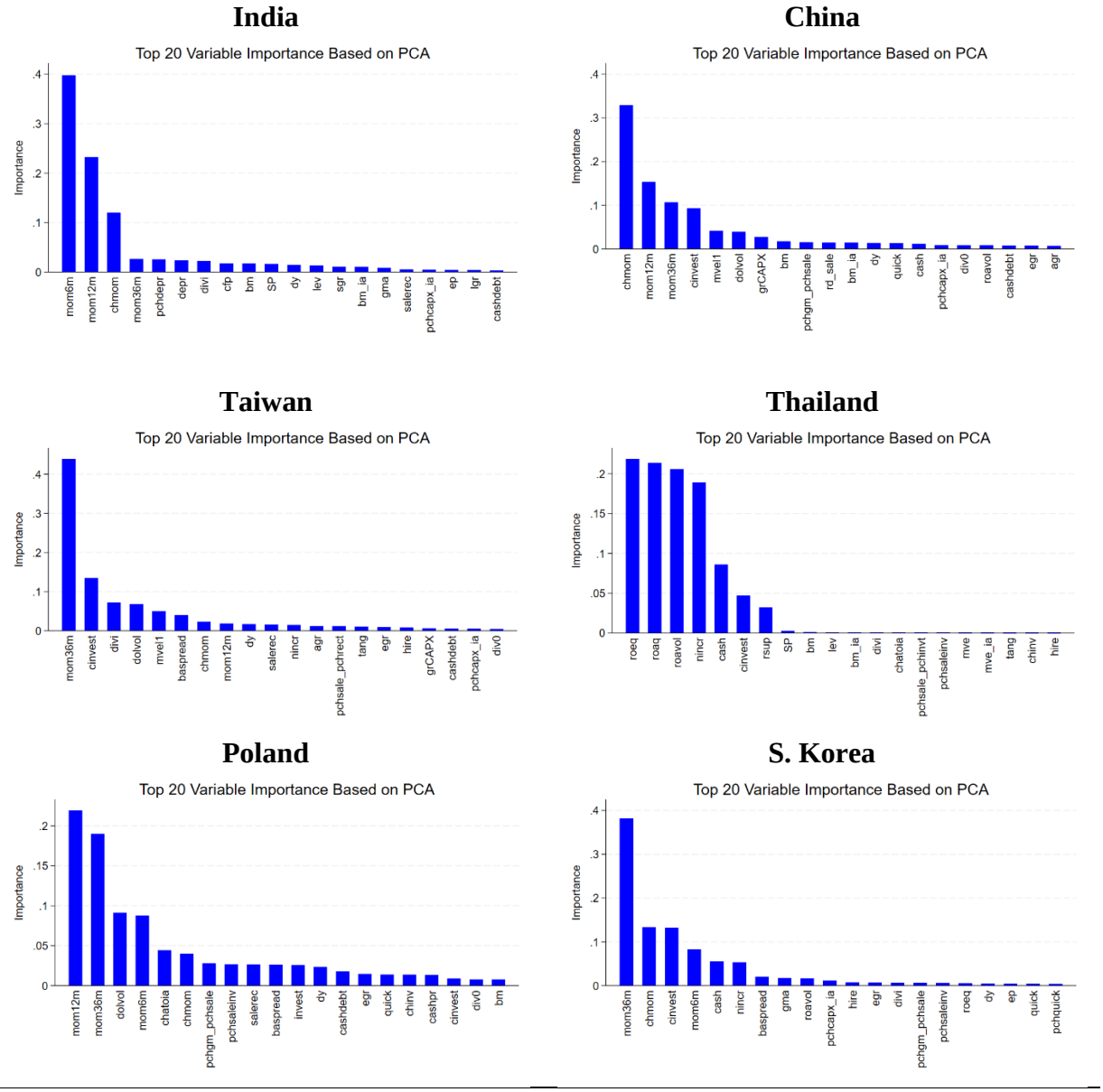


Figure 2
Variable importance for PCA, by country
Graphs illustrate country-specific variable importance for six countries: India, China, Taiwan, Thailand, Poland, and South Korea.

Conclusion

Employing the framework of return prediction, we analyzed the predictive power of the machine learning-based PCR model. Our findings demonstrated that PCR can enhance the empirical understanding of asset pricing, outperforming traditional linear models. It is particularly relevant in the context of emerging market economies, where the absence of established valuation methodology prevents investors from obtaining key insights about the drivers of stock returns. By utilizing a large set of predictors, the PCR model demonstrated consistent results and enabled the investigation of the variable importance of those predictors. Fundamental valuation metrics emerged as the most significant set of predictors, followed by operational profitability and liquidity. However, the results diverged in variability during the country-specific analysis, indicating the importance of price trend-related metrics. These findings are crucial for understanding of the drivers of stock returns in the emerging markets and contribute to the existing literature on asset pricing.

References

Bruner, R. F., Conroy, R. M., Estrada, J., Kritzman, M., & Li, W. (2002). Introduction to valuation in emerging markets. *Emerging Markets Review*, *3*(4), 310–324. https://doi.org/10.1016/S1566-0141(02)00039-0

Green, J., J. R. M. Hand, and X. F. Zhang. 2013. The supraview of return predictive signals. *Review of Accounting Studies* 18:692–730.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2274. https://doi.org/10.1093/rfs/hhaa009

West, K. D. (2006). Forecast evaluation. In G. Elliott, C. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 99–134). Elsevier.

Zhao, C., Yuan, X., Long, J., Jin, L., & Guan, B. (2023). Financial indicators analysis using machine learning: Evidence from Chinese stock market. *Finance Research Letters*, *58*, 104590. https://doi.org/10.1016/j.frl.2023.104590

Appendix

Acronym	Frequency	Definition
agr	Annual	Annual percent change in total assets.
baspread	Monthly	Monthly average of daily bid-ask spread.
beta	Monthly	Estimated market beta from weekly returns and equal-weighted market returns for 3 years ending month t-1, with at least 52 weeks of returns.
betasq	Monthly	Market beta squared.
bm	Annual	Book value of equity divided by end of fiscal year-end market capitalization.
bm_ia	Annual	Industry-adjusted book-to-market ratio.
cash	Quarterly	Cash and cash equivalents divided by average total assets.
cashdebt	Annual	Earnings before depreciation and extraordinary items divided by total liabilities.
cashpr	Annual	Fiscal year-end market capitalization plus long-term debt divided by cash and equivalents.
cfp	Annual	Operating cash flows divided by fiscal-year-end market capitalization.
cfp_ia	Annual	Industry-adjusted cfp
chatoia	Annual	2-digit SIC fiscal-year mean-adjusted change in sales, divided by average total assets.
chcsho	Annual	Annual percent change in shares outstanding.
chempia	Annual	Industry-adjusted change in number of employees.
chinv	Annual	Change in inventory scaled by average total assets.
chmom	Monthly	Cumulative returns from months 1-6 to t-1 minus months t-12 to t-7.
chpmia	Annual	2-digit SIC -fiscal-year mean adjusted change in income before extraordinary items divided by sales
cinvest	Quarterly	Change over one quarter in net PP&E divided by sales - average of this variable for prior 3 quarters; if saleq = 0, then scale by 0.01
currat	Annual	Current assets divided by current liabilities.
depr	Annual	Depreciation divided by PP&E.

divi	Annual	An indicator variable equal to 1 if the company pays dividends but did not in the prior year.
divo	Annual	An indicator variable equal to 1 if the company does not pay dividends but did in the prior year.
dolvol	Monthly	Natural log of trading volume times price per share from month t-1.
dy	Annual	Total dividends divided by market capitalization at fiscal year-end.
egr	Annual	Annual percent change in book value of equity.
ep	Annual	Annual income before extraordinary items divided by end-of-fiscal-year market cap.
gma	Annual	Revenues (revt) minus cost of goods sold (cogs) divided by lagged total assets (at).
grCAPX	Annual	Percent change in capital expenditures from year t-2 to year t.
grltnoa	Annual	Growth in long-term net operating assets.
herf	Annual	2-digit SIC fiscal-year sales concentration (sum of squared percent of sales in industry for each company)
hire	Annual	Percent change in number of employees.
ill	Monthly	Average of daily (absolute return / dollar volume)
invest	Annual	Annual change in gross property, plant, and equipment (ppegt) + annual change in inventories (invt), all scaled by lagged total assets (at)
lev	Annual	Total liabilities (lt) divided by fiscal year-end market capitalization
lgr	Annual	Annual percent change in total liabilities.
mom12m	Monthly	11-month cumulative return ending one month before month-end.
mom36m	Monthly	Cumulative returns from months t-36 to t-13.
mom6m	Monthly	5-month cumulative return ending one month before month-end.
momentum	Monthly	1-month cumulative return.
mve	Annual	Natural log of market capitalization at the end of month t-1.
mve_ia	Annual	Industry-adjusted mve.

nincr	Quarterly	Number of consecutive quarters (up to eight quarters) with an increase in earnings (ibq).
operprof	Annual	Revenue minus cost of goods sold - SG&A expense - interest expense divided by lagged common shareholders' equity.
pchcapx_ia	Annual	2-digit SIC fiscal-year mean-adjusted percent change in capital expenditures (capx).
pchcurrat	Annual	Percent change in currat.
pchdepr	Annual	Percent change in depr.
pchgm_pchsale	Annual	Percent change in gross margin (sale - cogs) minus percent change in sales (sale).
pchquick	Annual	Percent change in quick.
pchsale_pchinvt	Annual	Annual percent change in sales (sale) minus annual percent change in inventory (invt).
pchsale_pchrect	Annual	Annual percent change in sales (sale) minus annual percent change in receivables (rect).
pchsale_pchxsga	Annual	Annual percent change in sales (sale) minus annual percent change in SG&A (xsga).
pchsaleinv	Annual	Percent change in saleinv.
quick	Annual	(Current assets - inventory) / current liabilities.
rd	Annual	An indicator variable equal to 1 if R&D expense as a percentage of total assets has an increase greater than 5%.
rd_mve	Annual	R&D expense divided by end-of-fiscal-year market capitalization.
rd_sale	Annual	R&D expense divided by sales (xrd / sale).
roaq	Quarterly	Income before extraordinary items (ibq) divided by one-quarter lagged total assets (atq).
roavol	Quarterly	Standard deviation for 16 quarters of income before extraordinary items (ibq) divided by average total assets (atq).
roeq	Quarterly	Earnings before extraordinary items divided by lagged common shareholders' equity.
roic	Annual	Annual earnings before interest and taxes (ebit) minus nonoperating income (nopi) divided by non-cash enterprise value (ceq + lt - che).
rsup	Quarterly	Sales from quarter t minus sales from quarter t-4 (saleq)

		divided by fiscal-quarter-end market capitalization (cshoq * prccq).
salecash	Annual	Annual sales divided by cash and cash equivalents.
saleinv	Annual	Annual sales divided by total inventory.
salerec	Annual	Annual sales divided by accounts receivable.
sgr	Annual	Annual percent change in sales (sale).
sp	Annual	Annual revenue (sale) divided by fiscal year-end market capitalization.
tang	Annual	Cash holdings + 0.715 × receivables + 0.547 × inventory + 0.535 × PPE / total assets.