Modeling Corporate Bond Returns

DSO 585 Data Driven Consulting Final Report

Neil Nieh
Marcus Walton

Abstract

This project investigates the modeling of corporate bond returns using Instrumented Principal Component Analysis (IPCA), with the primary objective of replicating and validating the findings from Bryan Kelly (2023). Specifically, we focus on forecasting credit excess return (CER) — the portion of bond returns attributable to credit risk. Our first goal is to ensure that our in-sample and out-of-sample (OOS) results are consistent with those reported in Kelly's work, including replication of key metrics such as predictive R^2 and portfolio Sharpe ratios.

In addition to academic replication, we assess the model's practical business value. This demonstrates how both sell side and buy side firms can apply the IPCA into their business line.

Finally, we explore several approaches to improve model performance. These include engineering new features such as the change in Distance to Default (ΔD2D), and benchmarking the IPCA model against modern machine learning techniques including XGBoost and Long Short-Term Memory (LSTM) networks. These experiments help define the future potential for hybrid modeling approaches.

I. IPCA Model Explained

Traditional bond pricing approaches have attempted to map factor structures for corporate bonds by borrowing equity factor methodologies. For example, Bai, Bali, and Wen (2019, "BBW") construct bond factors via portfolio sorts on characteristics, following the Fama–French blueprint. In those models, factors like "value" or "profitability" are predefined as long–short portfolios of bonds, and bond returns are regressed on these observable factor returns. While intuitive, this factor approach can be restrictive. It assumes the chosen characteristic portfolios perfectly represent true risk factors, and it produces static factor loadings (betas) via time-series regression on those factors. These assumptions limit the models' ability to capture the complexity and time-variation in corporate bond risks. The result has been poor performance in explaining bond returns, especially out of sample.

The Instrumented Principal Component Analysis (IPCA) model is introduced to overcome these shortcomings by flexibly estimating latent factors and time-varying exposures from the data. The purpose of the IPCA model in this context is to provide a conditional factor model that more accurately captures corporate bond risk premia and dynamics, thereby bridging the gap in our understanding of bond pricing.

The IPCA methodology introduces several key innovations over traditional PCA and factor models:

- Latent factor discovery: It does not prespecify factors but lets the data reveal them, avoiding bias from using noisy proxies. Traditional models assume their chosen factors (e.g. "value" portfolio) are correct; IPCA instead finds the latent factors that truly drive returns.
- Characteristic-based betas: Bond factor loadings are modeled as a function of bond characteristics, enabling conditional, time-varying betas. In contrast, standard PCA or Fama–French regressions produce one fixed beta per bond (or require rolling regressions). IPCA's betas update with market conditions and bond-specific changes.
- High-dimensional instrument use: IPCA can handle many characteristics simultaneously, addressing the "factor zoo" problem. Instead of examining one characteristic at a time in isolation, it considers a large set of attributes together, allowing the model to identify which characteristics are most informative for explaining bond returns.

II. Business Value

The innovations of the Instrumented Principal Component Analysis (IPCA) model offer distinct business value across the financial ecosystem. Given the complex microstructure of corporate bond markets, characterized by over-the-counter (OTC) trading, limited transparency, and high barriers to entry, IPCA is especially valuable to institutional investors, where the scale and data needed to implement such models are readily available. The business value for buy-side and sell-side firms differs as follows:

<u>Buy Side:</u>

Turning Noise into Actionable Alpha

- IPCA's superior explanatory power allows asset managers to identify the true, dynamic sources of bond returns~~.~~

Bridging the Gap Between Quantitative Models and Traditional Investment Approaches

- One of the persistent challenges for quant teams is gaining the trust of traditional portfolio managers who are often skeptical of complex models. IPCA addresses this by combining machine learning capabilities with economic interpretability. This makes the output of the model more digestible and trustworthy for discretionary managers and investment committees, who can incorporate these insights into their decision-making processes without abandoning their fundamental investment philosophies.

Enhancing Portfolio Construction, Diversification, and Risk Management

- With its ability to isolate time-varying exposures, IPCA allows asset managers to construct portfolios that are not only better diversified but also more responsive to shifts in market regimes. By focusing on the most relevant and persistent return drivers, managers can build strategies that are more robust across different economic environments, enhancing both performance and downside protection.

Example: BlackRock Bond Fund Manager

- A portfolio manager at BlackRock managing a large corporate bond fund could use IPCA to distill the overwhelming universe of market signals down to the few critical factors that matter most in the current environment.

<u>Sell-Side:</u>

Sharper Pricing Models, Trading Edge, and Risk Management

- For sell-side trading desks, IPCA significantly enhances the accuracy of pricing models for corporate bonds, structured credit, and fixed-income products. By incorporating IPCA-derived factors, traders can quote prices with higher confidence, offer tighter bid-ask spreads, and improve the precision of valuations for both liquid and illiquid securities. This improved pricing accuracy directly supports higher trading volumes, stronger client flow, and better management of risk exposure.

Driving Innovation in Product Structuring and Thematic Solutions

- Beyond trading, IPCA empowers product structuring teams to create innovative, factor-based investment solutions. By identifying statistically significant, economically meaningful factors, banks can design ETFs, custom indices, or structured notes that capture specific risk premia identified by IPCA. This not only enables differentiation in a competitive market but also allows the bank to proactively offer clients new, data-driven investment strategies tailored to current market themes.

Strengthening Research, Thought Leadership, and Client Engagement

- IPCA also enhances the analytical toolkit of sell-side research teams. By providing transparent, data-driven insights into evolving market drivers, research teams can deliver more differentiated trade ideas and market commentary. This enhances client advisory conversations, deepens relationships with institutional investors, and reinforces the bank's positioning as a thought leader and trusted advisor in the fixed-income space.

Example: Goldman Sachs Structured Products

- Goldman Sachs could leverage IPCA to construct structured bond products where payouts are linked to an IPCA-optimized basket of bonds reflecting factors like issuer leverage or liquidity premiums. Not only does this lead to more relevant and efficiently priced products, but it also allows the bank's sales and research teams to bring differentiated, factor-driven ideas to clients. This enhances the bank's competitive positioning and opens new revenue streams through innovative product offerings.

III. Data Sources and Preprocessing

Data Sources:

For this project, we integrated two key datasets to model corporate bond returns using the IPCA framework:

1. ICE Fixed Income Data (Corporate Bond Data):

Our primary bond dataset was sourced from the Intercontinental Exchange (ICE), a leading provider of fixed income data services. The ICE dataset provides detailed daily end-of-day (EOD) pricing data for U.S. corporate bonds, including fields such as:

- CUSIP and ISIN identifiers
- Descriptions, maturity dates, and ratings
- Market metrics such as option-adjusted spreads (OAS), effective duration, yield to worst, and accrued interest
- Total and excess returns, along with benchmark and index data

ICE data is particularly valuable due to its granularity and accuracy in the opaque OTC bond market. Unlike TRACE, which is transaction-based, ICE leverages evaluated pricing methodologies that aim to produce more stable and liquid estimates, particularly useful for portfolio-level modeling and relative value analysis.

2. Intrinio Equities Data:

For the equities data, we used Intrinio's monthly stock data service, which offers firm-level financials, returns, and key valuation metrics. The inclusion of equities data is critical for the IPCA approach, as prior research suggests that equity markets often lead bond markets in pricing risk and information. This dataset was used to augment our feature space with issuer-specific equity variables, ~~enhancing the model's ability to capture cross-asset risk spillovers~~.

Preprocessing Steps:

1. Bond Data Cleaning and Validation:

The ICE bond data consisted of daily files that required meticulous cleaning due to variations in formatting and quality. Key preprocessing steps included:

- Handling missing data and erroneous files, such as files with incomplete headers or parsing errors, through automated filtering and manual validation.
- Converting 'As of Date' columns to proper datetime formats and ensuring chronological consistency.
- Removing or scaling return fields to obtain clean monthly excess returns, which are necessary for the IPCA regression framework.

2. Equity Data Cleaning and Feature Engineering:
For the Intrinio stock data:

- Monthly firm-level files were collected and aggregated into a master dataset, ensuring consistency in ticker identifiers.
- Key firm characteristics, including valuation ratios, profitability metrics, and leverage ratios, were extracted and merged with the bond dataset based on issuer identifiers (e.g., matching corporate bond CUSIPs to stock tickers).
- Additional adjustments were made to handle multiple classes of securities or inactive tickers.

3. Final Data Integration:
Both datasets were merged at a monthly bond-level granularity, with issuer-level equities variables broadcast across bond records where applicable. The final panel dataset allowed us to:

- Calculate bond-level excess returns relative to Treasury returns.
- Construct a wide set of features combining bond characteristics (from ICE) and firm fundamentals (from Intrinio).
- Ensure data completeness and reliability for downstream IPCA modeling.

IV. Model Performance

In-Sample Analysis:

To mirror Kelly's performance assessment, we calculate three R^2 metrics: *total* R^2*, time-series* R^2, and *cross-sectional* R^2. *Total* R^2 captures the model's overall predictive power. *Time-series* R^2 evaluates how well the model forecasts each bond's returns over time, while *cross-sectional* R^2 measures its ability to predict returns across all bonds at a given point in time.

IPCA In-Sample Performance

K	Our Model			Kelly's Model		
	Total R^2	Time-series R^2	Cross-sectional R^2	Total R^2	Time-series R^2	Cross-sectional R^2
1	48.2	33.2	33.2	41.6	33.4	23.1
2	52.5	30.7	35.6	45.0	36.5	25.8
3	55.4	31.5	40.4	46.9	39.4	28.3
4	56.8	30.8	42.0	48.6	42.2	32.0
5	57.8	31.9	43.7	49.7	43.1	33.3

Overall, our *in-sample* R^2 exceeds Kelly's—probably because we focus only on investment-grade bonds, which are less volatile than the high-yield issues in his sample. That said, our *time-series* R^2 remains lower than our *cross-sectional* R^2. Importantly, we replicate Kelly's pattern: model fit improves as we add more latent factors.

Out-of-Sample Analysis:

We experimented with different combinations of training window lengths and forecast horizons for out-of-sample testing. The **training window** refers to the number of past months used to estimate (to train) the model, while the **forecast horizon** denotes how many months ahead the model predicts returns. Across these configurations, we found that a fixed 120-month training window combined with a 25-month forecast horizon yielded results most similar to Kelly's. It's worth noting that Kelly uses an expanding training window, whereas we hold ours constant. Also, Kelly's paper does not explicitly state his chosen forecast horizon.

IPCA Out-of-Sample Performance

	Training Window	Forecast Horizon	K	Intercept	Total R^2
Kelly’s Model	Expanding	Unspecified	5	False	50.7
Our Model	120	25	5	True	36.5
				False	38.7
			4	True	39.2
			3	True	39.3
			1	True	39.1

Unlike the in-sample analysis, our out-of-sample *total R^2* is approximately 10-12 percentage points lower than Kelly’s. Additionally, we observe that the best out-of-sample R^2 performance does not occur when the **intercept = False** and the **K=5**. Instead, the optimal performance arises when K equals 3 and the intercept is set to true. However, the difference between these two configurations is not substantial. Therefore, for consistency with Kelly’s original setup, we proceed with the specification where the intercept is set to false and K equals 5 in the subsequent analysis.

<u>Adding New Features:</u>

We also experimented with adding new features to the model to assess whether they could improve its overall explanatory power. Two additional variables were introduced.

The first is Delta Distance to Default, which captures the monthly change in a firm's distance to default. We chose this feature because Distance to Default (D2D) is one of the most important predictors in Kelly’s IPCA specification. By incorporating its change over time, we aim to capture not just the static risk of default, but also its dynamics—which may further enhance the model’s ability to detect shifts in credit risk.

The second feature is Spread per Duration, which normalizes a bond’s credit spread by its duration. The motivation behind this variable is that bonds with longer maturities typically exhibit higher spreads simply due to their longer exposure to risk. By scaling spread by duration, we aim to transform it into a value-like signal, isolating the compensation for credit risk from that of interest rate risk. This allows the model to better differentiate between bonds offering true excess return potential and those with high spread due merely to long duration.

However, after incorporating these two new features, the model's explanatory power did not improve significantly. The increase in total R^2 was only 0.1, and the gains in time-series and cross-sectional R^2 appear to be largely mechanical—likely driven by the increase in the number of variables, rather than a genuine enhancement in the model's predictive ability.

V. Benchmarking ML Models

After experimenting with adding a new feature to the IPCA model, we also explored whether traditional machine learning models could effectively predict excess returns. Specifically, we focused on two approaches: XGBoost and LSTM. These models were selected for the following reasons:

- **XGBoost** is a tree-based boosting algorithm known for its strong performance on structured data and its ability to capture nonlinear relationships and feature interactions. It also handles overfitting well through regularization techniques.

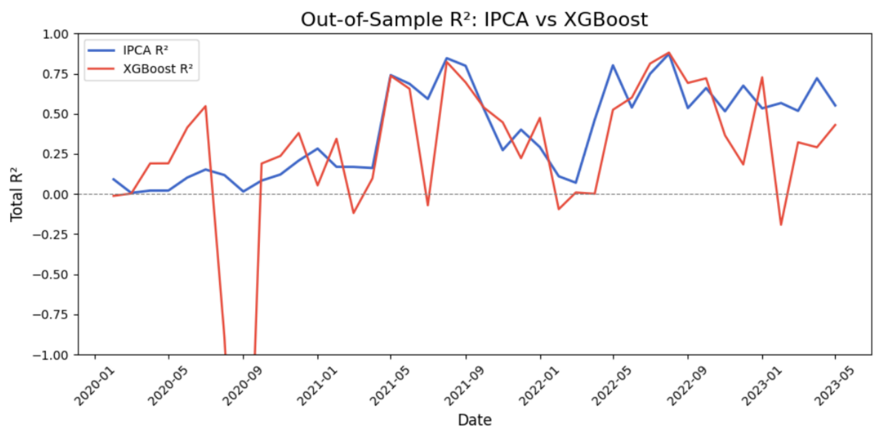
- **LSTM (Long Short-Term Memory)** is a type of recurrent neural network (RNN) designed to model sequential data. Given the temporal nature of bond returns, we assume LSTM is well-suited for capturing time-series patterns and long-term predictions.

By comparing these models to the IPCA framework, we aim to evaluate whether modern machine learning methods offer similar or additional advantages for excess return forecasting.

Machine Learning Model Comparison (In-Sample)

Models	Total R^2	Time-series R^2	Cross-sectional R^2
XGBoost	26.4	27.4	20.7
LSTM	2.08	1.47	7.69
XGBoost (after hypertuning)	36.0	34.8	20.8

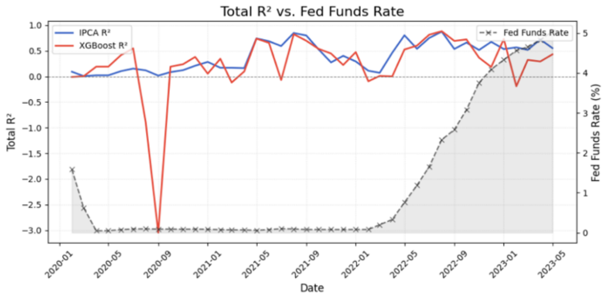
In our initial run without any tuning, XGBoost achieved a total R^2 of approximately 0.26, while LSTM performed quite poorly. As a result, we focused on hyperparameter tuning for XGBoost. Our tuning approach involved an exhaustive grid search, where we systematically evaluated all combinations of specified parameter values. Once the best-performing parameter set was identified, we used that configuration for subsequent out-of-sample analysis and compared its performance against our IPCA model. The *out-of-sample R^2* for this tuned configuration reached 0.28. While this still represents a decline compared to the in-sample performance, the drop is less pronounced than what we observed with the IPCA model.

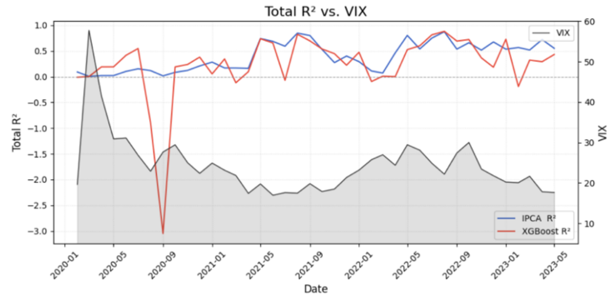


As shown in the figure above, the *total R^2* of the XGBoost model exhibits significantly greater volatility compared to the IPCA model. Notably, there are instances where XGBoost's *total R^2* falls below zero, indicating that its predictive performance is worse than using the historical average excess return as a benchmark. This is particularly evident in September 2020, when the model's predictive ability appears to have collapsed. In the following section, we investigate potential factors that may have contributed to this sharp decline in *total R^2*.

We examine whether the federal funds rate (Fed Rate) and the VIX (volatility index) influence the explanatory power of our model. The Fed Rate is a critical macroeconomic variable that directly affects discount rates, bond yields, and investor expectations—making it a fundamental driver of fixed income returns. Changes in monetary policy, especially during tightening or easing cycles, often lead to shifts in risk premia, which may impact the model's ability to capture return dynamics accurately.

We also investigate the VIX, a widely used proxy for market uncertainty and investor sentiment. Periods of elevated VIX typically coincide with spikes in risk aversion and volatility spillovers across asset classes. Given the growing evidence of bond–equity market interdependence, it is plausible that fluctuations in equity market volatility (as measured by the VIX) could disrupt the factor structure captured by our model. Furthermore, Kelly includes the VIX as one of the predictors in his model, reinforcing its potential relevance in explaining variations in corporate bond excess returns.



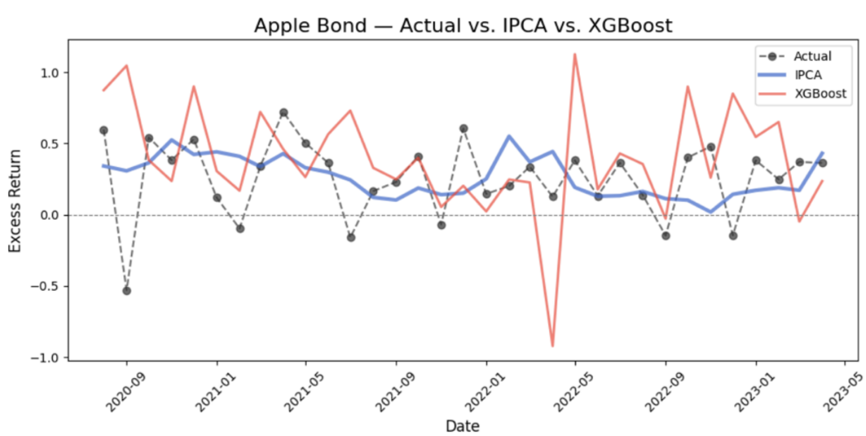


We observe that the federal funds rate appears to have little impact on the explanatory power of our model, regardless of whether the market is in a tightening or easing cycle. In particular, prior to the sharp decline in XGBoost's R^2 in September 2020, interest rates had already stabilized, suggesting that monetary policy was unlikely the primary driver of the model's failure at that point.

As for the VIX, we note that in March 2020—during the height of pandemic-driven market panic—it surged to nearly 60. This spike in market volatility corresponded with a collapse in R^2 for both the XGBoost and IPCA models, effectively dropping to zero. This indicates that extreme market conditions can overwhelm the models' predictive capabilities. However, despite incorporating these two macroeconomic indicators, we are still unable to pinpoint the exact cause behind the severe breakdown of the XGBoost model's performance in September 2020.

To better illustrate the predictive power of our models, we selected a bond to demonstrate—CUSIP 037833DF4, issued by Apple Inc., which matured on May 11, 2025, and carries a coupon rate of 1.125%.

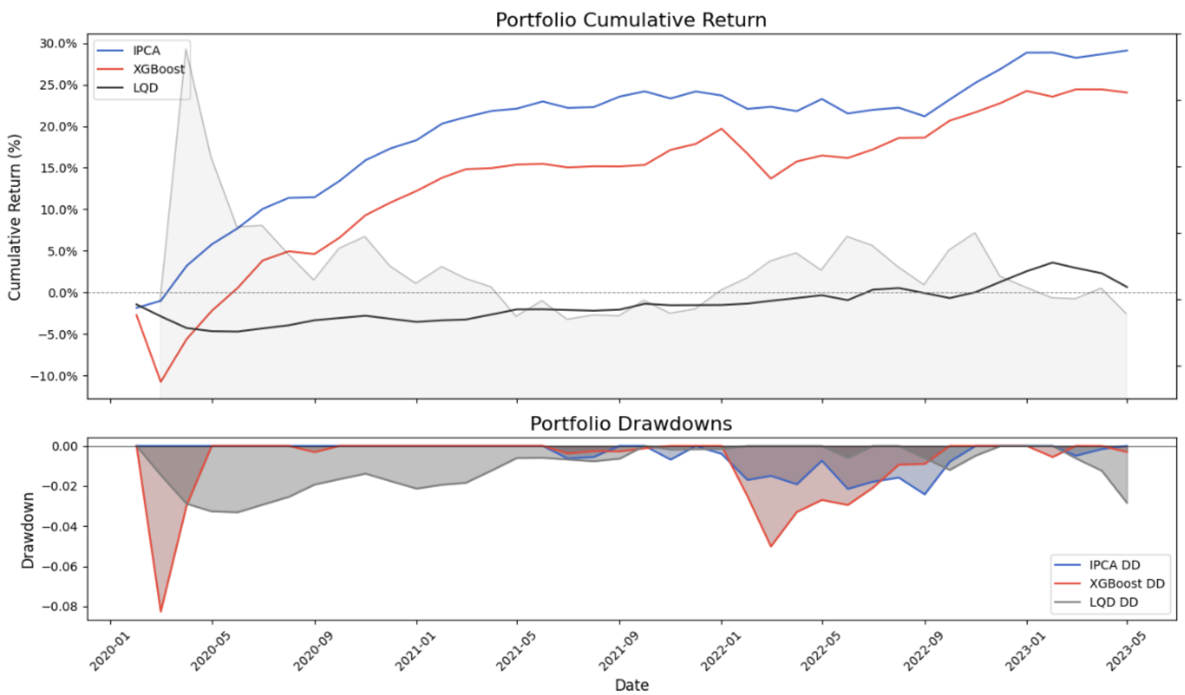
We plotted the actual returns against the model-predicted returns (i.e., Y *versus* $\hat{Y}$) for this bond. IPCA reported mean absolute error(MAE) of 0.225, lower than XGBoost' 0.334. The results clearly show that the IPCA model provides more accurate and stable forecasts compared to XGBoost. Consistent with its lower R^2, the XGBoost model exhibits noticeably greater volatility in its predictions, underscoring its weaker performance in tracking this bond's return dynamics.



VI. Portfolio Performance

Finally, following Kelly's approach, we construct a long-short portfolio and compare Sharpe ratios. We adopt an equal-weight strategy, with 10% allocated to both the long and short sides. The portfolio construction is straightforward: each month, we rank bonds based on predicted returns, go long the top 10%, and short the bottom 10%.

We also incorporate estimate trading costs into our analysis. While Bryan Kelly assumes a flat 19 basis points (bps) trading cost in his paper, he notes that this is a conservative estimate based on high-yield bonds. Since our portfolio consists solely of investment-grade bonds, we apply a lower rate of 17 bps.



We also included LQD, the iShares investment-grade corporate bond ETF, as a benchmark for comparison. Based on our backtest results, the IPCA model achieved an annualized return of 8.05% with a net-of-cost Sharpe ratio of 1.62**.** In contrast, XGBoost delivered a 6.97% annualized return and a Sharpe ratio of 0.94, both with statistically significant t-statistics.

While XGBoost clearly underperforms relative to IPCA, it still outperforms LQD by a meaningful margin. That said, we should be cautious when using LQD as a benchmark, given that it experienced extended periods of negative returns in the past—raising questions about how well it represents a reliable baseline for comparison.

	Annualized Return	**Sharpe Ratio (Net Cost)**	**t-statistics**
IPCA	8.05%	1.62	12.4
XGBoost	6.97%	0.94	5.9

VII. Conclusion and Future Improvements

In conclusion, we reviewed Bryan Kelly's paper and highlighted the advantages of the IPCA model compared to traditional factor models. We also outlined the business value of this approach for various financial institutions, including both the sell-side and buy-side.

We successfully replicated results consistent with Kelly's empirical findings, validating the model's robustness. In addition, we experimented with incorporating new features into the IPCA framework and benchmarked its performance against traditional machine learning models. Finally, we constructed a long-short portfolio, conducted backtesting, and evaluated its Sharpe ratio relative to Kelly's results.

There are, of course, several areas for further exploration. These include evaluating the correlation structure between new features and existing ones before introducing them into machine learning models, exploring hybrid machine learning techniques to fine-tune model performance, and considering alternative portfolio construction methods—rather than fixing the top and bottom 10% as the long and short legs. These represent potential directions for future research.