Università della Svizzera italiana,
Lugano

Faculty of Economics

Master in Finance — Digital Finance

InstrumentedPCA on Options

Master's Thesis

Author: Sara Vurdelja

Supervisor: Prof. Paul Schneider

Second Reader: Prof. Patrick Gagliardini

Academic Year 2024

Finalised: August 2024

Abstract

This thesis aims to contribute to the field of quantitative finance by providing a deeper understanding of covariant characteristics in options payoffs through the use of Instrumented Principal Component Analysis (IPCA). IPCA, a factor analysis model incorporating instrumental variables, addresses endogeneity issues and extracts latent factors from complex financial datasets. The study utilises a comprehensive dataset of options on the S&P 500 Index (SPX), involving extensive data cleaning, feature engineering, and the application of IPCA to uncover underlying market dynamics.

A key focus of this research is the prediction of the underlying index return using options payoffs, derived from various modelling approaches and employed under the put-call parity relationship. The findings indicate that IPCA effectively identifies significant factors that influence options payoffs and, consequently, the return on the index, demonstrating superior predictive power compared to traditional models. Validation through out-of-sample testing and robustness checks confirms the reliability of the results. This work not only enhances the understanding of options payoffs dynamics but also underscores the potential of combining various methodologies in financial engineering for more accurate and insightful analyses.

The thesis concludes with a discussion on the implications of the findings and suggests avenues for future research, particularly in exploring the integration of factor models like IPCA with other advanced forecasting techniques to further innovate and enhance the field of quantitative finance.

Keywords: InstrumentedPCA, Options Payoffs, Index Return, Factor Analysis, Predictive Modelling

Contents

List of Figures

List of Tables

Preface

This thesis aims to contribute to the field of quantitative finance by providing a deeper understanding of covariant characteristics in options data and their role in predicting underlying index returns, using Instrumented Principal Component Analysis (IPCA), a framework that remains relatively unexplored, particularly in the context of options data.

The decision to experiment with a factor analysis model such as IPCA, specifically applied to options data, was driven by my deep interest in financial derivatives and the advanced mathematical and statistical methods associated with them. My curiosity about the vast possibilities in this field, especially regarding options, was sparked by a particularly stimulating course related to Arbitrage Pricing during my studies at USI.

The research process presented numerous challenges, from managing a large and complex dataset to gaining a thorough understanding of the IPCA methodology. There were moments of doubt and frustration, but overcoming these obstacles has significantly enriched my knowledge and technical skills.

This thesis has helped me better define my specific interests within the vast field of quantitative finance and sharpened my focus on the complexities of financial engineering. It has also highlighted the potential and importance of combining machine learning techniques with advanced financial methods for more sophisticated analysis and modelling.

Furthermore, working with this approach has deepened my appreciation for factor models and contemporary predictive methods. It has shifted my perspective on traditional modelling techniques, helping me realise that it is not about one methodology being superior to another. Rather, every quantitative analyst should recognise that a more comprehensive toolkit, one that integrates diverse approaches from different domains, can yield more meaningful results. This balance of methodologies is an aspect I am keen to explore in future research.

Introduction

Options, as financial derivatives, play a crucial role in modern financial markets by offering mechanisms for risk management and speculative opportunities. Despite their importance, the complexity and richness of options data pose significant challenges for analysis, particularly in accurately capturing the factors that ultimately influence options payoffs.

Recent observations in the field of quantitative finance suggest that traditional models, such as the Fama-French three-factor model and the Black-Scholes option pricing model, struggle to fully account for the dynamics of options markets. These models often fail to incorporate the high dimensionality and inherent endogeneity present in financial data, leading to biased estimates and less reliable predictions (Fama and French, 1993; Black and Scholes, 1973). The limitations of these models, particularly their static nature, are well-documented. For instance, the Black-Scholes model assumes constant volatility, which does not reflect actual market conditions where volatility can be highly dynamic (Heston, 1993; Bakshi, Cao and Chen, 1997). This issue is not just a theoretical concern; it has practical implications for traders, risk managers, and policy-makers who rely on accurate models to make informed decisions.

The financial crisis of 2008 and subsequent market turmoil highlighted the limitations of existing models in predicting extreme events and managing risk. In response, there has been a growing interest in developing more sophisticated tools that can better capture the complexities of financial markets. Instrumented Principal Component Analysis (IPCA) has emerged as one promising solution to this problem, offering a framework that addresses endogeneity and extracts latent factors that drive market behaviour.

This research is thus motivated by the observed shortcomings of traditional models and the potential of IPCA to provide deeper insights into the underlying factors that influence options payoffs.

1.1 Rationale

Financial industry's stakeholders, including traders, analysts, and regulators, rely heavily on the accuracy of financial models. The limitations of traditional models, such as Black-Scholes, become especially apparent during periods of market stress. For instance, during events like the 1987 stock market crash, the dot-com bubble of the late 1990s, and the 2008 financial crisis, these models often failed to provide accurate estimates, leading to significant financial losses and highlighting their inadequacies (Longstaff, 1995; Coval and Shumway, 2001).

Additonally, recent data from options markets underscore the importance of developing more sophisticated models. For example, the volatility skew, a common feature of options markets, is poorly explained by traditional models, leading to significant pricing errors (Rubinstein, 1994; Jackwerth and Rubinstein, 1996). Empirical evidence also suggests that the factors driving options prices are more complex and dynamic than previously thought, necessitating the use of advanced modelling methods (Christoffersen, Heston, and Jacobs, 2009). Essentially, traditional models are increasingly unable to cope with the high dimensionality and inherent endogeneity present in financial markets, resulting in biased estimates and less reliable predictions.

Over time, the landscape of options trading has evolved significantly. Initially, options were predominantly used by institutional investors for hedging purposes, but their accessibility to retail investors has grown, leading to an exponential increase in trading volume and complexity. The rise of algorithmic trading and the proliferation of financial derivatives have further complicated matters, creating a pressing demand for models that can scale with the volume and complexity of data (Bollen and Whaley, 2004).

Culturally, the increased reliance on quantitative methods and the rise of financial engineering have contributed to the complexity of markets. These shifts have created a demand for models that can handle high-dimensional data and complex dependencies among financial instruments. Socially, the democratisation of finance and the increasing participation of retail investors have added to the market's complexity, making traditional models, which were once sufficient, struggle to keep pace with the rapidly evolving environment.

The goal is then to contribute to the broader field of quantitative finance by offering a more robust approach to options analysis through the application of

Instrumented Principal Component Analysis (IPCA). By improving the understanding of the factors that drive options dynamics, this research aims to provide valuable insights for both practicioners and stakeholders.

1.1.1 Conceptual Framework

The foundation of this study lies at the intersection of factor analysis and econometrics, particularly in addressing the issues of endogeneity and model specification that are common in financial datasets. Traditional factor models, such as those proposed by Fama and French (1993), have been widely used to explain asset returns. However, these models often assume that the factors driving asset prices are exogenous, an assumption that may not hold in the presence of complex financial instruments like options.

Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can employed within these models to simplify complex datasets by identifying the principal components that capture the most variance. PCA is particularly useful in reducing the dimensionality of large datasets, making them more manageable for analysis. However, while PCA is effective in simplifying data, it does not account for endogeneity or the potential influence of observable characteristics on the latent factors it identifies.

This research is based on the hypothesis that options payoffs are influenced by a set of latent factors that are not directly observable but can be inferred through the use of instrumental variables. These latent factors represent underlying market dynamics, such as volatility, liquidity, and investor sentiment, which are not fully captured by traditional models. IPCA extends traditional factor models by incorporating instrumental variables, which help mitigate the bias introduced by endogeneity. This approach allows for a more accurate estimation of the factors influencing options dynamics, providing a deeper understanding of underlying market dynamics.

By applying IPCA, this thesis seeks to uncover these latent factors and assess their impact on options payoffs. The rationale for this approach is rooted in the belief that understanding these factors can lead to more accurate analytical models, which are essential for effective risk management and investment strategies. This research builds on a rich body of literature in financial econometrics, particularly the work of Kelly et al. (2019), who demonstrated the effectiveness of IPCA in extracting latent factors from equity returns.

1.2 Research Setup

The primary purpose of this research is to explore the covariant characteristics of options data, particularly those related to the S&P 500 Index (SPX), and their implications for predicting underlying index returns under the put-call parity relationship. The study aims to achieve several key objectives:

1. **Uncover and Validate Latent Factors:** Identifying and validating latent factors that significantly influence options payoffs, and consequently index returns inferred through put-call parity. This involves applying IPCA to capture these underlying factors that drive market dynamics.

2. **Address Endogeneity:** Addressing the issue of endogeneity in options data, providing a more robust framework for understanding market behavior. By incorporating instrumental variables, IPCA aims to offer more accurate estimates of these latent factors.

3. **Compare Predictive Power:** Evaluating the predictive power of IPCA in comparison with traditional approaches, such as Monte Carlo simulations, demonstrating its advantages and discussing its shortfalls in terms of accuracy, robustness and interpretability.

Additionally, there are several key assumptions that are necessary for the validity of this research:

1. **Data Integrity:** It is assumed that the dataset used in this study is accurate, complete, and representative of the broader options market over considered time period. Any errors or omissions in the data are assumed to be random and not systematic.

2. **Model Specification:** The IPCA model is assumed to be appropriately specified for capturing the latent factors in options payoffs. It is assumed that the instruments chosen for the model are valid and do not introduce additional bias.

3. **Market Efficiency:** The study assumes that markets are generally efficient, meaning that prices reflect all available information. However, it also acknowledges that inefficiencies may exist, particularly in the context of high-dimensional and complex datasets.

4. **Stationarity:** It is assumed that the relationships between the variables in the study are relatively stable over time, and that any structural breaks or changes are accounted for in the analysis.

5. **No-Arbitrage Condition:** The study assumes that the no-arbitrage condition holds in the options market, meaning there are no opportunities for riskless profit. This implies that the law of one price is upheld, ensuring that identical assets or portfolios cannot have different prices across different markets.

6. **Put-Call Parity:** For the predictive modelling component, the study assumes the validity of the put-call parity relationship. This assumption is used to streamline the analysis and to facilitate the modelling and prediction of options payoffs through the results achieved through IPCA.

These assumptions are taken for granted based on prior literature and the methodological framework of the study.

1.2.1 Limitations

While this thesis aims to provide robust insights into options payoffs dynamics and consequently index returns leveraging IPCA, several limitations must be acknowledged:

1. **Data Limitations:** The study relies on historical options data, which may be subject to reporting errors or biases beyond the researcher's control. Additionally, the dataset may not fully capture all relevant market conditions, particularly during periods of low trading volume or market stress.

2. **Model Limitations:** The IPCA model, although powerful, has its limitations, including the potential for overfitting in high-dimensional data environments. The model's reliance on instrumental variables also assumes that these instruments are correctly specified and valid, which may not always be the case.

3. **Generalisability:** The findings of this study are specific to the options market, particularly options on the S&P 500 Index (SPX), and may not be generalisable to other financial instruments, markets, or more diversified portfolios with individual stock options. The unique characteristics of options data, such as the volatility skew, and the specific dynamics of the SPX may limit the applicability of the results to other contexts.

4. **Temporal Scope:** The study primarily covers a 10-year period selected for its relative stability. An additional period from 2020 to early 2023, characterised by the market stress of the Covid-19 pandemic, was also considered to test the model's robustness in more volatile conditions. While this enhances

the understanding of the model's performance in different market environments, the findings may still not fully capture dynamics under other extreme conditions. Additionally, the study does not account for potential changes in market structure, economic conditions, or regulatory environments that could impact options payoffs, limiting the generalisability of the results to future market conditions.

5. **Use of European Options:** The study focuses exclusively on European-style options, which may limit the applicability of the findings to American options or other exotic derivatives.

6. **Use of Payoffs Instead of Returns:** The analysis is conducted using options payoffs rather than direct option returns. This approach simplifies the modelling process, but it may not fully capture the complete dynamics and risk-return profiles of options markets, potentially limiting the depth of the insights gained.

7. **Use of Put-Call Parity:** The study assumes that put-call parity holds and uses this relationship for predictive modelling. While this assumption simplifies the analysis by leveraging a well-established arbitrage condition, it may overlook complexities and market frictions that a more detailed approach could reveal.

8. **Index Data Only (SPX):** The study is limited to data from the S&P 500 Index (SPX) options, which, while significant, may not represent the full spectrum of options available in the broader market.

These limitations are recognised as inherent constraints in the study's design and analysis, and should be considered when interpreting the results and their implications.

1.2.2 Delimitations

This study is deliberately focused on certain aspects of options data with specific boundaries:

1. **Geographical Scope:** The study is limited to options data from major U.S. exchanges, including the Chicago Board Options Exchange (CBOE) and other relevant platforms. Data from international markets is not included.

2. **Temporal Scope:** The main analysis is confined to a relatively stress-free 10-year period, ranging from 2009 to 2019, deliberately selected to capture

a comprehensive view of market trends and dynamics while avoiding the extreme volatility of the 2008 financial crisis and the onset of the Covid-19 pandemic. An additional period from 2020 to early 2023, encompassing the COVID-19 pandemic, was used separately only for stress-testing to evaluate the model's robustness under more volatile market conditions.

3. **Conceptual Focus:** The study centers on the application of the IPCA model to options data, without exploring other potential applications of the model already covered by some of the literature, such as in equity or bond markets.

4. **Model Selection:** The study employs IPCA as the primary analytical tool for dimensionality reduction and focuses on its application without comparing it to other advanced factor models beyond the use of a benchmark for performance comparison.

5. **Instrument Selection:** The instruments used in the IPCA model are chosen based on their relevance to options dynamics, with other potentially relevant variables not being considered in this study.

6. **European Options:** The study is restricted to European-style options, which do not have the early exercise feature found in American options.

7. **Index Data Only (SPX):** The study focuses solely on options data from the S&P 500 Index (SPX), chosen for its representativeness and data availability.

These delimitations are intended to provide a clear focus for the study and ensure that the research is manageable and relevant to the specific problem being addressed.

Related Literature

The purpose of this chapter is to review and critically analyse the empirical research that has been conducted in the area of options and financial econometrics, with a specific focus on studies that have utilised or could benefit from the application of Instrumented Principal Component Analysis (IPCA). Unlike the theoretical framework presented in the introductory chapter, this chapter focuses on the empirical findings of other researchers who have sought to address similar or related problems.

The literature will be organised thematically, aligning with the key research questions and hypotheses of this thesis.

2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used statistical technique for dimensionality reduction, particularly in fields where datasets are highly dimensional and complex. PCA transforms the original variables into a new set of uncorrelated variables, called principal components, which are linear combinations of the original variables. These principal components are ordered such that the first few retain most of the variation present in the original dataset, thereby reducing dimensionality while preserving as much information as possible (Jolliffe, 2002).

Mathematically, suppose we have a dataset represented by an $n \times p$ matrix $\mathbf{X}$, where n is the number of observations and p is the number of variables. The goal of PCA is to find a set of principal components $\mathbf{Z} = \mathbf{XW}$, where $\mathbf{W}$ is a $p \times k$ matrix of weights (eigenvectors) and k is the number of principal components retained, typically with $k \ll p$.

The principal components are found by solving the eigenvalue problem:

$$\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j \tag{2.1}$$

where $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X}$ is the covariance matrix of $\mathbf{X}$, λ_j is the j-th eigenvalue, and $\mathbf{w}_j$ is the corresponding eigenvector. The eigenvectors (principal components) corresponding to the largest eigenvalues capture the most variance in the data.

The method was first introduced by Pearson in 1901 and later developed further by Hotelling in 1933, and it has since become a fundamental tool in various disciplines, including finance, data science, and machine learning. In finance, PCA is particularly valuable for uncovering latent factors that drive asset returns. By reducing the dimensionality of complex financial datasets, PCA helps identify the underlying structure of the data. However, while PCA can simplify the data by focusing on fewer principal components, it can also introduce challenges in interpretation, as these components are linear combinations of the original variables, which may not have a straightforward economic meaning. Additionally, traditional PCA assumes static factor loadings, which may not adequately capture the time-varying nature of relationships present in financial markets, potentially limiting its effectiveness in dynamic environments.

In the context of options data, PCA has been employed to identify factors that influence specific options dynamics, such as volatility and interest rates. However, while PCA is effective in reducing dimensionality and capturing the primary sources of variance, it has limitations. One key limitation, other than the issue related to interpretability, is that PCA assumes that the principal components are orthogonal and that the factors driving the data are uncorrelated with the error term. This assumption of exogeneity can be problematic in financial data, where endogeneity often exists due to feedback loops between variables (Jolliffe, 2002).

Moreover, PCA does not account for the potential influence of observable characteristics on the latent factors, which can lead to biased results in the presence of endogeneity. This is particularly relevant in options pricing, where the factors influencing prices are complex and interdependent. As a result, while PCA is a powerful tool for initial data exploration and dimensionality reduction, more advanced methods, such as Instrumented Principal Component Analysis (IPCA), have been developed to address these limitations by incorporating instrumental variables with dynamic factor loadings to correct for endogeneity (Kelly, Pruitt, and Su, 2019).

The use of PCA in financial modeling, including its application to options data, highlights the importance of dimensionality reduction techniques in managing large and complex datasets. However, its limitations in dealing with endogene-

ity underscore the need for more sophisticated approaches in complex financial environments. The transition from PCA to IPCA represents an evolution in the methodological toolkit available to financial researchers, enabling more accurate and reliable modelling.

2.2 InstrumentedPCA

From the main literature references of this thesis "Characteristics are Covariances: A Unified Model of Risk and Return" by Bryan Kelly and the companion econometrics paper "Instrumented Principal Component Analysis", a pivotal question is posed: *Why do different assets earn different average returns?* An obvious answer is that either there exist different compensation for different amounts of risk, or that anomalies in the data are present.

Standard research usually focuses on linear factor models to try to explain these anomalies and the inherent mechanisms. However, financial data is characterised by complex dynamics that are difficult to capture with simple linear models. This is primarily because factors and loadings are generally unobservable, leading to two possible approaches:

1. The model relies on pre-existing factors that are built on previous knowledge and/or based on understanding of stock behaviour, essentially an ad hoc implementation, as seen in the case of the Fama-French models.

2. The factors are treated as latent, meaning that the factors are estimated and the betas are constructed from realised returns. This approach is exemplified by PCA, which is an improvement over linear models but still cannot account for conditionality.

Instrumented Principal Component Analysis (IPCA) is an extension of the PCA framework that addresses its common pitfalls. IPCA works by estimating latent factors with time-varying loadings, where the factor loadings (partially) depend on observable asset characteristics that serve as instrumental variables, hence the term "Instrumented." Unlike standard PCA, IPCA is a conditional model, meaning it adapts to the dynamic nature of financial data by incorporating these observable characteristics.

IPCA also offers additional features, such as useful asset pricing tests. The question is no longer *"Does factor Y explain the anomaly?"* but rather *"Does there exist some set of common latent risk factors that explain the anomaly?"*

The main takeaway from IPCA is that no characteristics generate alphas (i.e., excess returns unexplained by the model), but a select few do help explain expected returns. This implies that these characteristics reflect risk exposures rather than being anomalies. Another crucial insight is that most characteristics are statistically irrelevant and redundant in explaining returns. In fact, Kelly et al. (2019) identify 24 characteristics that capture most of the variations in returns.

The Model

Consider a set of N assets and T observations of excess returns. The returns are explained by conditional beta loadings of dimension $N \times K$ and a set of latent factors F_t of dimension $K \times 1$, plus an error term:

$$R_t = \beta_{t-1} F_t + \epsilon_t \tag{2.2}$$

The betas, $\beta_{i,t}$, are derived from a set Z of L observable instruments that account for the time variation of the betas and a Γ matrix of loadings, plus an error term:

$$\alpha_{i,t} = Z'_{i,t}\Gamma_\alpha + \nu_{\alpha,i,t}, \quad \beta_{i,t} = Z'_{i,t}\Gamma_\beta + \nu_{\beta,i,t} \tag{2.3}$$

By substituting the definition of the betas into the original model equation, the generalised model is obtained:

$$R_t = Z_{t-1}\Gamma_\alpha + Z_{t-1}\Gamma_\beta F_t + \epsilon_t \tag{2.4}$$

where $\alpha_{t-1} = Z_{t-1}\Gamma_\alpha$ and $\beta_{t-1} = Z_{t-1}\Gamma_\beta$.

This generalised model suggests that characteristics line up with average returns because they proxy for risk factor loadings, not because they directly cause excess returns. The key question here is: *Are the characteristics representing risk factors, or are they merely good return predictors?*

To address this, the intercept test is used to test for anomalies, where an anomaly is defined as the presence of compensation without corresponding risk. This test allows to determine which characteristics or expected returns are driven by compensation factors.

One key concept in IPCA are the conditional betas, $\beta_{i,t} = \Gamma_\beta Z_{i,t}$. Betas are modeled as functions of observable characteristics, with the Γ matrix encapsulating a large number of these characteristics condensed into a small number of

exposures. Mathematically, the Γ matrix represents a linear combination of the characteristics used to describe factor exposures.

The objective function of the IPCA model aims to identify latent factors and loadings by optimising a least-squares criterion, as shown below:

$$\min_{F,\Gamma} \frac{1}{NT} \sum_{t} (R_t - \beta_{t-1} F_t)'(R_t - \beta_{t-1} F_t), \tag{2.5}$$

subject to $\beta_{t-1} = Z_{t-1}\Gamma$.

However, unlike traditional PCA, which has a closed-form solution based on the eigen-decomposition of the covariance matrix, IPCA must use an Alternating Least Squares (ALS) approach to iteratively estimate both the factors F and the loadings Γ.

This iterative process is necessary because the optimisation problem is complicated by endogeneity, where the instruments Z_{t-1} must be orthogonal to the error term, and thus no closed-form solution exists.

Consequently, the ALS minimisation procedure does not guarantee convergence in a finite number of steps, increasing computational complexity and sensitivity to initial values, which may result in local minima. Despite these challenges, ALS enables IPCA to flexibly handle asymmetric and irregular panels, which are common in financial datasets, where data may be sparse or unevenly distributed.

Restricted vs Unrestricted IPCA

In the context of IPCA, the distinction between restricted and unrestricted models centers on whether the model allows for non-zero intercepts (alphas).

In the unrestricted IPCA model, the alphas ($\alpha_{i,t}$) are allowed to be non-zero. This means that the model does not force the expected return on assets to be solely determined by the risk factors. Mathematically, this is represented by:

$$R_{i,t} = \alpha_{i,t} + \beta_{i,t} F_t + \epsilon_{i,t} \tag{2.6}$$

where $\alpha_{i,t}$ is the intercept term, $\beta_{i,t}$ are the factor loadings, F_t are the latent factors, and $\epsilon_{i,t}$ is the error term. In this formulation, the model allows for anomalies in asset returns, i.e., returns that cannot be explained by the latent factors alone.

The objective function for the unrestricted IPCA model, which includes the intercepts, is:

$$\min_{F,\Gamma_\alpha,\Gamma_\beta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(R_{i,t} - Z_{i,t-1}\Gamma_\alpha - \beta_{i,t}F_t\right)^2 \tag{2.7}$$

where $\beta_{i,t} = Z'_{i,t}\Gamma_\beta$ and $Z_{i,t}$ is the matrix of observable instruments. The inclusion of $\alpha_{i,t} = Z'_{i,t}\Gamma_\alpha$ allows for the possibility that some characteristics contribute to excess returns beyond what is explained by the factor exposures.

In the restricted IPCA model, the alphas ($\alpha_{i,t}$) are restricted to zero, meaning that the model assumes all expected returns are fully explained by the risk factors, with no unexplained excess returns (no anomalies).

The model is thus:

$$R_{i,t} = \beta_{i,t}F_t + \epsilon_{i,t} \tag{2.8}$$

The objective function for the restricted IPCA model then simplifies to:

$$\min_{F,\Gamma_\beta} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(R_{i,t} - \beta_{i,t}F_t\right)^2 \tag{2.9}$$

subject to $\beta_{i,t} = Z'_{i,t}\Gamma_\beta$. By imposing $\alpha_{i,t} = 0$, the model effectively tests whether the characteristics are merely good predictors of returns or if they truly represent underlying risk factors. If the restricted model fits the data well, it suggests that the observed characteristics are indeed capturing systematic risk rather than idiosyncratic anomalies.

In summary, the distinction between restricted and unrestricted models in IPCA is crucial for testing the presence of anomalies in asset returns. By comparing the performance of the restricted and unrestricted models, one can assess whether the characteristics used in the model are indeed related to risk factors or if they capture some form of market inefficiency.

A Managed Portfolio Interpretation

To better understand IPCA as an estimator, the model can be interpreted through the lens of managed portfolios. Specifically, the process involves the following steps:

1. For each characteristic $j = 1, \ldots, L$ in each period t, run a cross-sectional

regression of the excess returns $R_{i,t}$ on the lagged characteristics $Z_{i,t-1}$ to obtain the estimates $X_{i,j,t}$. Essentially, this involves regressing the returns of each asset on a specific characteristic and iterating the process for each characteristic.

2. The resulting $X_{i,j,t}$ represents the return on a portfolio that is associated with the j^{th} characteristic. In other words, these portfolios are constructed by weighting assets according to their exposure to a particular characteristic.

3. Construct the matrix $\mathbf{X}$ (of dimension $L \times T$), where each row corresponds to a different characteristic-managed portfolio, and each column represents a different time period. PCA is then applied to this matrix to identify common sources of variation across these characteristic-managed portfolios.

The factors extracted from this process are effectively "portfolios of managed portfolios," where the factors themselves represent portfolios derived from the characteristics. The matrix Γ acts as a map that links these factors (portfolios) back to individual asset returns.

Asset Pricing Tests

IPCA also facilitates asset pricing tests, particularly for identifying anomalies. The first test involves the following hypotheses:

- $H_0 : \Gamma_\alpha = 0$ (null hypothesis): This implies that there is no anomalous compensation without corresponding risk exposure. In other words, characteristics should not have a direct impact on returns if they are only proxies for risk factors.

- $H_a : \Gamma_\alpha \neq 0$ (alternative hypothesis): This suggests that characteristics do affect returns directly, indicating the presence of anomalous returns or mispricing.

The key question in this context is: Are the observed alphas truly indicative of market anomalies, or do they arise because the model is missing or incorrectly specifying relevant factors? Latent factors identified through IPCA are generally less prone to this issue, but it remains a crucial aspect to consider in the analysis.

In essence, the test for anomalies checks whether the characteristics are significant in the context of the model. Specifically, it examines whether $\Gamma_\alpha = 0$ for the j^{th} characteristic, indicating whether that characteristic significantly impacts the cross-sectional variation in returns.

Empirical Application and Evaluation

Kelly et al. evaluated IPCA using monthly stock returns and characteristics covering a period from July 1964 to May 2014. In their empirical setup, 36 characteristics were selected, including commonly used financial ratios such as the book-to-market ratio and profitability metrics. In total, the study considered 73 characteristics ($L = 73$).

The instruments used in the IPCA model are divided into two parts: the time-series average of the characteristics C_i and the deviations from these averages $C_{it} - C_i$. The vector of instruments Z also includes a constant term to ensure the stability of the estimated betas in the model.

To evaluate the performance of IPCA, several key statistics are employed:

- R^2 **(In-sample):** This statistic measures how well the model explains the overall variation in returns R_t, capturing the portion of the realised return variation explained by the model.
- R^2_{pred} **(Out-of-sample):** This statistic assesses how well the model explains differences in average returns, indicating how the conditional betas (or alphas) align with the observed differences in returns. It effectively measures the variation explained by the model's conditional expectations.
- **P-values:** The p-values are used to conduct inference on the arbitrage pricing restriction $\Gamma_\alpha = 0$, typically assessed via bootstrap tests. This helps determine the statistical significance of the model's predictions and the validity of its assumptions.

These statistics provide a comprehensive view of the model's performance, both in terms of its explanatory power and its ability to predict returns accurately.

Results

The results obtained from Kelly et al. from applying IPCA provide valuable insights into the behaviour of asset returns and the effectiveness of the model.

The overall R^2 statistic, which measures the model's ability to explain the variation in returns, shows that there is not much difference between the restricted and unrestricted models. This suggests that, in terms of total explanatory power, the

presence of non-zero alphas does not significantly improve the model's fit. However, as the number of factors K increases, R^2 also increases, reflecting the model's ability to capture more of the return variation as more factors are included.

$$\lim_{K \to \infty} R^2(K) \to 1$$

This indicates that as K approaches infinity, the model can theoretically explain all the variation in returns. However, in practical terms, this convergence is subject to overfitting, where the model captures noise rather than true underlying factors.

Results also show that allowing for non-zero alphas ($\Gamma_\alpha \neq 0$) increases the predictive R^2_{pred}, which measures the model's ability to predict out-of-sample returns based on the estimated factors and loadings.

Additionally, as K increases, R^2_{pred} improves, reducing the gap between the restricted and unrestricted models. This indicates that the additional flexibility of the unrestricted model, which allows for non-zero alphas, provides better predictive performance:

$$\lim_{K \to \infty} R^2_{\text{pred}}(K) \to R^2_{\text{unrestricted}}$$

This convergence suggests that, with a sufficient number of factors, the unrestricted model's predictive power approaches that of a fully flexible model, which underscores the importance of correctly specifying the number of factors in practice.

The P-value associated with the test for anomalies (i.e., testing whether $\Gamma_\alpha = 0$) indicates that for $K \geq 2$, the model fails to reject the null hypothesis that there are no anomalies. This means that once at least two factors are included in the model, the alphas (intercepts) are no longer statistically significant, implying that the returns are fully explained by the factors and there is no evidence of abnormal returns:

$$\text{P-value} \geq \text{significance level}, \quad \text{for } K \geq 2$$

This result supports the idea that most characteristics do not generate alphas but are instead proxies for the risk factors captured by the model.

Advantages and Limitations

The application of IPCA in financial econometrics brings several advantages and challenges.

Advantages:

- **Endogeneity Correction:** One of the significant advantages of IPCA is its ability to address endogeneity issues that are prevalent in traditional factor models. By using instrumental variables, IPCA provides a more accurate estimation of factor loadings, reducing bias in the model.
- **Dynamic Factor Loadings:** Unlike static models, IPCA allows for time-varying factor loadings, which better capture the evolving dynamics of financial markets. This flexibility makes the model more adaptable to changing market conditions.
- **Handling Asymmetric Panels:** IPCA can efficiently handle asymmetric and unbalanced panels, where the number of observations may differ across cross-sectional units. Thanks to the use of instrumented factors, IPCA can extract factors and estimate loadings even in the presence of such irregularities, ensuring the robustness of the analysis.
- **Improved Predictive Power:** As indicated by Kelly et al.'s results on stock returns, allowing for non-zero alphas improves the model's predictive accuracy. This enhancement makes IPCA particularly valuable for out-of-sample predictions in that context, which are crucial in financial decision-making.
- **Comprehensive Analysis of Characteristics:** IPCA's ability to handle a large number of characteristics and identify the most relevant ones provides deeper insights into what drives asset returns. This feature is especially useful for understanding complex financial instruments like options.

Limitations:

- **Model Complexity:** IPCA introduces additional complexity compared to traditional PCA or other linear factor models. The need to select appropriate instruments and specify an optimal number of factors can be challenging, particularly in large datasets with many potential characteristics.
- **Overfitting Risk:** As K increases, there is a risk of overfitting, where the model may capture noise rather than true underlying factors. This is

especially a concern when dealing with highly dimensional data or small sample sizes.

- **Computational Intensity:** The estimation of IPCA, particularly with a large number of characteristics and time-varying loadings, can be computationally intensive. This may require significant computational resources and careful consideration of model specifications to ensure feasibility.
- **Interpretability of Results:** While IPCA provides a powerful framework for analysing returns, the results can be difficult to interpret, particularly when many factors are involved. Understanding the economic meaning of the extracted factors and their loadings requires careful analysis and domain knowledge and may generally not be pratical.

2.3 Latent Factors in Options Returns

Goyal and Saretto (2024) explore the challenges of modelling option returns using traditional factor models, proposing that the IPCA framework may offer a more effective solution. The complexity of option returns, driven by factors such as volatility, interest rates, and the underlying asset's price, often eludes traditional models like the Fama-French three-factor model. These models, while useful for simpler assets, struggle to capture the nuanced behaviour of derivative instruments like options.

Traditional factor models typically represent asset returns R_i as a linear combination of a set of K factors F_k, such that:

$$R_i = \alpha_i + \beta_{i1}F_1 + \beta_{i2}F_2 + \cdots + \beta_{iK}F_K + \epsilon_i, \tag{2.10}$$

where β_{ik} are the factor loadings, α_i represents the intercept, and ϵ_i is the error term. A known limitation in these models is the assumption that the factors F_k are exogenous, i.e., uncorrelated with the error term ϵ_i. However, in the context of options pricing, endogeneity (correlation between the factors and the error term) can lead to biased estimates of β_{ik}.

To address this, Goyal and Saretto employ IPCA and incorporate instrumental variables Z_i to correct for endogeneity. As seen, in the IPCA framework, the

factors are estimated as:

$$\hat{F}_k = \arg\min_{F_k} \sum_{i=1}^{N} \left(R_i - \alpha_i - \sum_{k=1}^{K} \beta_{ik} F_k \right)^2 \tag{2.11}$$

subject to the orthogonality condition $\mathbb{E}[Z_i(\epsilon_i)] = 0$, which ensures that the instruments Z_i are uncorrelated with the error term ϵ_i, thereby mitigating endogeneity.

The authors demonstrate that IPCA not only enhances the factor model approach for option returns but also provides a more dynamic set of factors that vary over time and across different contracts. By comparing the performance of IPCA with traditional factor models, they find that IPCA significantly outperforms these models in explaining the cross-section of option returns. This suggests that traditional models may overlook critical information that IPCA can capture, thereby offering a more robust framework for understanding option returns.

The practical implications of this study are considerable. For market practitioners, particularly those involved in pricing and hedging options, the use of IPCA could lead to more accurate models and better risk management strategies. The findings indicate that incorporating IPCA into the modelling toolkit could provide hedge funds and institutional investors with a competitive edge by improving the precision of option pricing models.

Goyal and Saretto's work contributes to the ongoing evolution of financial econometrics by challenging the adequacy of traditional factor models in complex financial markets. Their application of IPCA to options data not only advances academic understanding but also has tangible applications in the field of quantitative finance. By highlighting the limitations of traditional models and demonstrating the potential of IPCA, this study underscores the importance of adopting modern, sophisticated methodologies in financial modelling.

In the broader context of this thesis, Goyal and Saretto's findings support the use of IPCA as a valuable tool for uncovering latent factors in options data. However, one must acknowledge that their analysis is limited to a specific context, focusing exclusively on returns from delta-hedged calls. While this approach offers valuable insights, it does not fully examine the broader applicability of IPCA across the entirety of the options market.

Moreover, Goyal and Saretto's analysis lacks an external benchmark for compar-

ing the performance of their IPCA implementation. The absence of a benchmark makes it challenging to assess the relative effectiveness of IPCA in capturing risk factors or predicting returns compared to alternative models.

This thesis aims to extend the use of IPCA to a more generalised setting by including an overall analysis of both call and put options. Additionally, as previously established, it shifts the focus from returns to payoffs, a choice made to streamline the modelling process given the notoriously challenging behaviour of option returns. Option returns can be extremely volatile and difficult to model accurately due to their asymmetric payoff structures, sensitivity to market movements, and time decay. These complexities likely motivated Goyal and Saretto's decision to focus exclusively on delta-hedged call returns, which attempt to neutralise some of the inherent risks and non-linearities associated with options.

However, by limiting the scope to delta-hedged call returns, Goyal and Saretto's approach may not fully capture the broader risk dynamics present in the entire options market. Delta hedging reduces exposure to directional price movements but does not address other risk factors such as volatility, skewness, or kurtosis, which are critical in the pricing and risk management of options. Furthermore, this narrow focus does not account for the behaviour of put options, which are equally significant in understanding market sentiment and hedging strategies.

Methodology

This chapter outlines the research design and methodology employed in this thesis to address the problem of accurately identifying and analysing the latent factors influencing options payoffs and, consequently, index returns. The research problem has been operationalised to focus on the relationship between these latent factors and options payoffs, with a specific emphasis on addressing endogeneity and high-dimensional data issues through the application of IPCA.

The procedure follows the standard practice of dividing the dataset into a main training set and a standard out-of-sample testing set to evaluate the model's predictive performance under stable market conditions. Key statistical metrics following Kelly's approach are considered for evaluation, such as the calculation of in-sample R^2 and out-of-sample R^2_{pred}. Furthermore, the robustness of the model is tested through the use of an additional, higher-stress test set covering the Covid-19 pandemic period.

Statement of Hypotheses

The hypotheses to be tested in this study are as follows:

1. **Null Hypothesis (H0):** There is no significant relationship between the latent factors identified by IPCA and options payoffs.
2. **Alternative Hypothesis (H1):** There is a significant relationship between the latent factors identified by IPCA and options payoffs.
3. **Null Hypothesis (H0):** The predictive power of traditional or other advanced approaches is equal to or greater than that of the IPCA model.
4. **Alternative Hypothesis (H1):** IPCA shows superior predictive power compared to traditional or other modelling approaches.

3.1 Data Description

The data used in this research is primarily sourced from the Wharton Research Data Services (WRDS) platform, specifically through the OptionMetrics database.

The study focuses on the S&P 500 Index (SPX) options, identified by the *SECID 108105* from the Chicago Board Options Exchange (CBOE). The SPX is a widely recognised benchmark of the U.S. stock market, representing a broad cross-section of large-cap companies. The choice of SPX options is justified by their liquidity, the depth of available data, as well as their representativeness of market-wide dynamics. Moreover, SPX options are highly relevant for modelling purposes due to their frequent use in hedging and risk management by institutional investors. The simplification inherent in focusing on a single, well-established index allows for more robust modelling and a straightforward application of sophisticated techniques like IPCA.

As previously delineated, the geographical scope of the dataset is limited to options traded on major U.S. stock exchanges, primarily focusing on data from the CBOE. While the majority of the data used in this study is derived from OptionMetrics, certain variables and instruments used in the analysis are constructed using additional data directly sourced from OptionMetrics. However, unless otherwise specified, all primary data utilised for the analysis originates from the OptionMetrics database.

The main dataset spans a period of 10 years, from January 1, 2009, to December 31, 2019, specifically selected for its relative stability. This period was chosen to avoid the extreme market events of the 2008 financial crisis and the onset of the Covid-19 pandemic, allowing for a detailed examination of the factors influencing options payoffs under more stable economic conditions.

Regarding the data split, the first 8 years were used for training the model, while the final 2 years (2017 to 2019) served as the main test set to evaluate out-of-sample performance under typically stable market conditions. Additionally, a separate Covid-19 test set, covering the period from January 1, 2020, to February 28, 2023, was used to assess the model's robustness under more stressful market conditions, providing both a main stable test set and a stress test set.

To maintain consistency and streamline the application of the IPCA model, the analysis is confined to European-style options. The use of European options is

particularly advantageous because it enables the application of the put-call parity relationship, which serves as a fundamental component of the predictive modelling approach in this thesis. European options, which can only be exercised at expiration, avoid the complexities associated with American options, such as early exercise, thereby ensuring a more straightforward modelling approach.

The primary dataset was augmented with additional information for the construction of specific variables serving as instruments for the IPCA model. These variables were constructed by merging the core options data with relevant market data, ensuring that all instruments used in the analysis are both theoretically justified and empirically robust. The comprehensive dataset compiled from these sources forms the foundation for the empirical analysis conducted in this thesis. Other variables were directly calculated in the dataset.

3.2 Data Cleaning and Preprocessing

Given the extensive size and complexity of the data, preparing the main dataset for IPCA was carefully managed using Dask, a parallel computing library in Python that facilitates efficient data handling and processing.

To maintain the relevance and quality of the data, two main filters were applied during the cleaning process. Firstly, expired options contracts were filtered out to ensure that only active contracts during the analysis period were considered. Additionally, options with zero trading volume were excluded, as they do not contribute to price discovery and could introduce noise into the analysis.

Note that all instruments used in the analysis were carefully selected to ensure that the model relies only on information available up to time t, avoiding any bias from future information.

The data preparation phase involved several key steps, including handling missing values, standardising variable formats, and constructing variables (instruments) to use in the IPCA model.

1. **Missing Values:** Missing data points were addressed by either interpolation or removal, depending on the context and the importance of the missing data for the analysis.

2. **Standardisation:** In the IPCA modelling part, to ensure consistency across the dataset, variable formats, including dates and categorical data, were

standardised according to procedures similar to those used by Kelly et al.

3. **Variable Construction:** New variables and instruments necessary for the IPCA model were created from the cleaned data, following the theoretical guidelines established in prior research. Consult the related appendix for the detailed calculations.

3.2.1 Variables Definition and Construction

For the IPCA model, the primary dependent variables (y) are the payoffs of call and put options, modeled as functions of latent factors estimated through the IPCA framework. The goal is to analyse the relationship between these latent factors and options payoffs and to predict the underlying index return using these payoffs.

The payoff for call and put options at maturity are defined as:

$$\text{Call Payoff} = \max(S_T - K, 0), \quad \text{Put Payoff} = \max(K - S_T, 0),$$

where S_T is the price of the underlying asset at maturity and K is the strike price of the option.

To infer the underlying index returns, the put-call parity relationship is applied:

$$S - K = (S_T - K)^+ - (K - S_T)^+, \tag{3.1}$$

$$S - K = E[\text{Call Payoff}] - E[\text{Put Payoff}], \tag{3.2}$$

where E[Call Payoff] and E[Put Payoff] are the expected payoffs for call and put options, respectively.

The estimated underlying asset price S is then calculated as:

$$S = K + (\text{Estimated Call Payoff} - \text{Estimated Put Payoff}), \tag{3.3}$$

where K is the strike price of the options. This formula was applied consistently across all models to ensure a fair comparison.

For each model, the expected underlying price S was inferred as follows:

- **IPCA with 20 Factors:** $S_{\text{IPCA20}} = K_{\text{real}} + \text{IPCA Call20} - \text{IPCA Put20}$
- **IPCA with 5 Factors:** $S_{\text{IPCA5}} = K_{\text{real}} + \text{IPCA Call5} - \text{IPCA Put5}$
- **Monte Carlo Simulation:** $S_{\text{MC}} = K_{\text{real}} + \text{MC Call} - \text{MC Put}$
- **Actual Market Values:** $S_{\text{real}} = K_{\text{real}} + \text{True Call} - \text{True Put}$

  where:

- K_{real} represents the observed strike price in the market.
- "Call" and "Put" refer to the estimated payoffs from their respective models (e.g., IPCA or Monte Carlo).

The index return is consequently calculated using the inferred price S as:

$$\text{Index Return} = \frac{S - S_0}{S_0}, \tag{3.4}$$

where S_0 is the current spot price of the index. This approach ensures a consistent comparison across models, where the only varying components are the estimated payoffs via IPCA, Monte Carlo simulation, or observed market values.

Operational Definitions and Instruments (X)

The independent variables (X), referred to as characteristics or instruments in the context of IPCA, are selected based on their theoretical relevance and empirical significance in predicting the payoffs of options. These characteristics serve as instrumental variables that are crucial in estimating the time-varying factor loadings (betas) within the IPCA framework.

In the IPCA model, the X matrix is constructed from a set of characteristics that describe various aspects of the options and their underlying assets. These characteristics are used to create what can be conceptually understood as "portfolios of managed portfolios," where each characteristic helps define the exposure to the latent factors that drive the options' payoffs.

The specificities of the calculations for each variable are detailed in the related appendix, ensuring transparency and reproducibility. Below is an overview of the key characteristics considered in this thesis:

Group 1: Basic Options and Underlying Characteristics

1. **Option Price:** The midpoint between the bid and ask prices, representing the estimated fair value of the option.
2. **Bid-Ask Spread:** The difference between the bid and ask prices, reflecting the liquidity of the option.
3. **Effective Spread:** A measure of transaction costs that accounts for both the bid-ask spread and market impact.
4. **Spot Price:** The current market price of the underlying asset.
5. **Strike Price:** The fixed price at which the holder of the option can buy or sell the underlying asset.
6. **Forward Price:** The expected future price of the underlying asset, adjusted for carrying costs.
7. **Risk-Free Rate:** The theoretical return on an investment with zero risk, proxied by government bond yields.
8. **Dividend Yield on Underlying:** The annual dividend payment divided by the spot price, indicating the return from dividends.
9. **VIX:** The CBOE Volatility Index, representing the market's expectations for volatility over the coming 30 days.
10. **Index Return:** The return on the underlying index, reflecting the overall market movement.

Group 2: Moneyness and Time-related Characteristics

11. **Moneyness:** The ratio of the spot price to the strike price, indicating the intrinsic value of the option.
12. **Time to Maturity:** The remaining time until the option's expiration, measured in years.
13. **Volume:** The number of contracts traded, a proxy for market activity and liquidity.
14. **Open Interest:** The total number of outstanding contracts, providing insight into market interest.
15. **Volume Weighted Average Price (VWAP):** The average price of the option over the trading day, weighted by volume.

Group 3: Volatility and Skewness/Kurtosis Measures

16. **Historical Volatility:** The standard deviation of past returns of the underlying asset, used to gauge past price fluctuations.
17. **Implied Volatility:** The market's forecast of the underlying asset's volatility, derived from the option's price.
18. **Model-Free Implied Volatility (Aggregated):** A volatility measure that does not rely on any specific option pricing model.
19. **Model-Free Implied Skewness:** A measure of the asymmetry of the implied volatility distribution.
20. **Model-Free Implied Kurtosis:** A measure of the "tailedness" of the implied volatility distribution.
21. **Realised Skewness:** The skewness of the distribution of past returns of the underlying asset.
22. **Realised Kurtosis:** The kurtosis of the distribution of past returns of the underlying asset.

Group 4: Option Greeks (Sensitivities)

23. **Delta:** The sensitivity of the option's price to changes in the price of the underlying asset.
24. **Gamma:** The rate of change of delta with respect to the underlying asset's price.
25. **Vega:** The sensitivity of the option's price to changes in implied volatility.
26. **Theta:** The sensitivity of the option's price to the passage of time.
27. **Rho:** The sensitivity of the option's price to changes in the risk-free interest rate.
28. **Vanna:** The sensitivity of delta to changes in implied volatility.
29. **Charm:** The sensitivity of delta to the passage of time.
30. **Vomma:** The sensitivity of vega to changes in implied volatility.
31. **Zomma:** The sensitivity of gamma to changes in implied volatility.
32. **Color:** The rate of change of gamma over time.

33. **DvegaDtime:** The rate of change of vega over time.

Group 5: Volatility and Skewness/Kurtosis Differentials

34. **RV - IV:** The difference between realised volatility and implied volatility.
35. **RV - MFvol:** The difference between realised volatility and model-free implied volatility.
36. **RSkew - MFSkew:** The difference between realised skewness and model-free implied skewness.
37. **RKurt - MFKurt:** The difference between realised kurtosis and model-free implied kurtosis.

These characteristics are operationalised to capture the key dynamics in the options market, serving as the foundation for the IPCA model. As previously established, the process of constructing these variables involved rigorous data cleaning, filtering, and transformation steps, ensuring that they accurately represent the underlying financial concepts.

3.3 IPCA with 20 and 5 Factors

IPCA was applied separately to call and put options to capture the unique latent factors driving the payoffs of each option type. The analysis was structured into monthly panels to respect the panel data requirements of the IPCA model, ensuring that the temporal dependencies within each panel were appropriately handled.

As initially mentioned, the dataset was divided into a training and testing set as per standard practice. The first eight years (January 2010 to December 2017) were used for training, while the last two years (January 2018 to December 2019) were allocated for testing. This period split was consistently used across all models, including the benchmark.

Additionally, a separate Covid test set was used to assess model performance under more stressful market conditions, covering a period from January 1, 2020, to February 28, 2023.

It is important to note that no embedded regularisation was applied in the IPCA model as it led to computational instability and performance issues within the package's framework. As such, the versions of the IPCA models employing Lasso,

Ridge or Elastic Net regularisation were ultimately scrapped due to their unreliability. A common standardisation procedure was employed, aligning with the standardisation approach similar to that used by Kelly.

The workflow for applying IPCA to calls and puts, for both the 5-Factors and 20-Factors versions, consisted of the following steps:

- **Data Cleaning and Sorting:** The dataset was cleaned and sorted to ensure data integrity and proper structuring.
- **Setting Up IPCA Structure:** The monthly panels were constructed to match the IPCA panel requirements. The data was split into training and testing sets, and the panel data was structured with the target variable y as the payoffs and the feature matrix X consisting of the relevant characteristics.
- **Normalisation:** Both the training and test datasets were normalised, ensuring that all variables were on a comparable scales.
- **Model Fitting:** The IPCA model was fitted separately for call and put options, and for both the 5-Factors and 20-Factors versions. The model outputs included the Gamma matrices (β loadings) and the respective Factors for all observations.

To evaluate the appropriateness of the number of factors chosen, scree plots were generated for both the calls and puts sections of the main dataset. The scree plots illustrate that for both option types, an "elbow" appears at around 5 factors, suggesting that this is a reasonable cutoff point in terms of variance-bias trade-off. The scree plot method involves plotting the eigenvalues associated with each factor in descending order, and the point at which the plot flattens out (the "elbow") should give insights on the optimal number of factors to retain.

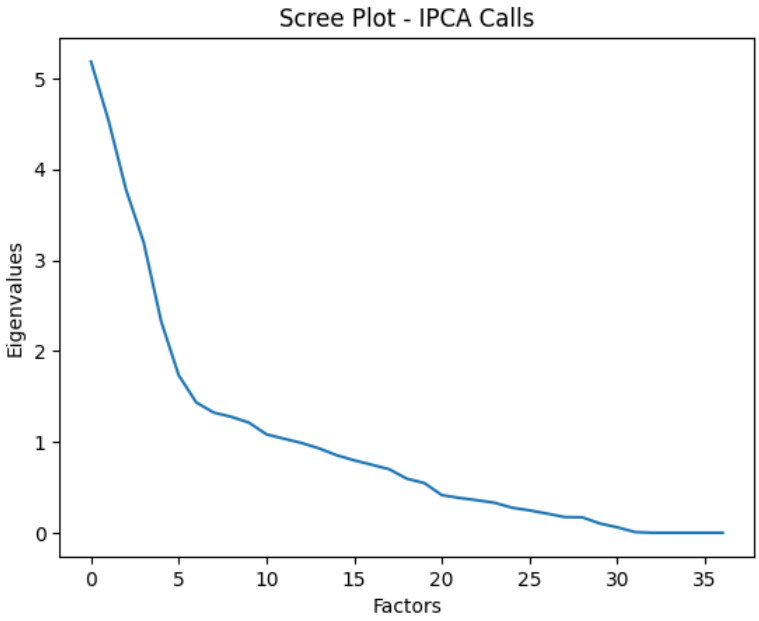


Figure 3.1: Scree Plot for Call Options

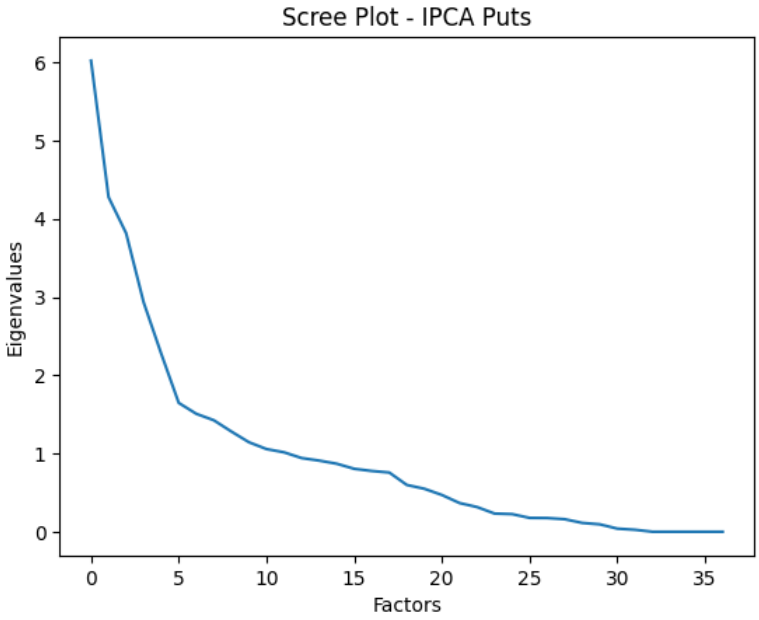


Figure 3.2: Scree Plot for Put Options

This method was chosen as current literature does not provide a definitive method for determining the optimal number of factors ex-ante. Additionally, while the literature generally advocates for parsimonious models (e.g., Kelly, 2021), a model using 20 factors was also tested, acknowledging that options are inherently complex instruments and a higher number of factors may help capture more of their unique characteristics, albeit with the conscious risk of potential overfitting.

Both versions of the model were then evaluated using the following metrics:

- **R-Squared ($\mathbf{R^2}$) and Predictive R-Squared ($\mathbf{R^2_{pred}}$):** These metrics were used to assess the model's fit and predictive power. The formulas for these metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5}$$

$$R^2_{\text{pred}} = 1 - \frac{\sum_{i=1}^{n}(X_{t+1,i} - \hat{y}_{t+1,i})^2}{\sum_{i=1}^{n} X^2_{t+1,i}} \tag{3.6}$$

  where $X_{t+1,i} = \frac{\text{payoff}_i}{\text{option price}_i} - \frac{1}{\text{rf}}$ is the normalised actual payoff for the i-th observation, and $\hat{y}_{t+1,i}$ is the predicted payoff from the IPCA models for the i-th observation.

  This custom R^2_{pred} formula uses the total variation of the normalised payoffs as denominator, implicitly treating zero as a baseline predictor. This approach is more appropriate in financial contexts where using the mean payoff can be misleading due to the skewness and heavy tails typical of options data.

- **Wald Test for Coefficients Significance:** Used to assess the joint significance of the model coefficients. The Wald test checks whether a set of parameters in the model is significantly different from zero:

$$W = (\beta' \Sigma^{-1} \beta) \tag{3.7}$$

  where β represents the vector of estimated coefficients and Σ is the covariance matrix of the coefficients. The Wald test is chosen for its ability to test multiple hypotheses simultaneously, providing a robust evaluation of coefficient significance.

- **Confidence Intervals for Coefficient Significance:** To determine the significance of each coefficient, confidence intervals were constructed for the factor loadings at a 5% alpha level. The confidence intervals were calculated as follows:

$$\text{CI} = [\Gamma - z_{\alpha/2} \times \text{SE}(\Gamma), \Gamma + z_{\alpha/2} \times \text{SE}(\Gamma)], \tag{3.8}$$

  where $\text{SE}(\Gamma)$ represents the standard error of the factor loadings, and $z_{\alpha/2}$ is the critical value for a 95% confidence level. The significance of each loading was checked based on whether its confidence interval included zero.

All statistical tests were conducted at an alpha level of 5%, and results were interpreted accordingly to maintain consistency in the analysis.

3.3.1 Technical Setup and Reproducibility

The entire data preparation process was coded in Python, with Dask playing a central role in managing the large datasets. All steps were documented and structured to ensure reproducibility, offering a detailed guide through the cleaning, filtering, and preparation stages for other researchers attempting to replicate or extend this analysis.

Python served as the primary tool for implementing the IPCA model and conducting the statistical evaluations. Several key libraries and packages were utilised:

- **Pandas and NumPy:** Used for data manipulation, numerical operations, and handling large datasets efficiently.
- **Dask:** Employed to manage the data processing pipeline efficiently, especially given the large size and high cardinality of the dataset. Dask enables parallel computing and allows for the processing of datasets that do not fit into memory, which was essential for handling the extensive options dataset used in this study.
- **Matplotlib and Seaborn:** Used for data visualisation, enabling the creation of various plots and charts to illustrate results.
- **Scikit-Learn:** Leveraged for preprocessing tasks, including standardisation of data and for performance evaluation metrics like the R-squared score.
- **SciPy:** Used for statistical functions, including chi-squared tests and normal distribution functions.
- **InstrumentedPCA:** The core IPCA analysis was conducted using the `InstrumentedPCA` package, designed specifically for implementing the Instrumented Principal Component Analysis. This package simplifies the application of IPCA by providing a pre-set framework for estimating latent factors and evaluating the model's fit.

It is important to note that during the development and testing of the code for the models, all computations were performed on a machine equipped with an Intel CPU. Attempts to run the same code on machines with AMD CPUs, as well as on cloud-based platforms like Google Colab, resulted in errors related to numerical precision. These errors suggest potential differences in how certain operations or

optimisations procedures are handled by different processors or platforms.

While these issues did not affect the results generated on Intel hardware, readers are advised that replicating the results may require the use of similar hardware or troubleshooting the numerical stability on alternative platforms.

3.4 Monte Carlo Simulations

Monte Carlo simulations were used as a benchmark to compare the performance of the IPCA models in predicting option payoffs. The Monte Carlo (MC) method is a statistical technique that relies on repeated random sampling to estimate the probability distribution of an uncertain variable (Glasserman, 2004). In the context of this thesis, MC simulations were employed to model the potential future prices of the underlying asset and compute the corresponding payoffs for call and put options.

Monte Carlo simulations generate possible future paths for the underlying asset price using a Geometric Brownian Motion (GBM). The GBM model assumes that the asset price follows a stochastic differential equation:

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t, \tag{3.9}$$

where:

- S_t is the asset price at time t,
- μ is the drift term, representing the expected return of the asset,
- σ is the volatility of the asset's returns,
- W_t is a Wiener process (standard Brownian motion).

By discretising this equation, the future asset price S_T at maturity T can be simulated as:

$$S_T = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right) T + \sigma\sqrt{T} Z\right), \tag{3.10}$$

where:

- S_0 is the current spot price of the asset,
- T is the time to maturity,
- $Z \sim N(0, 1)$ is a standard normal random variable.

The key assumptions underlying the Monte Carlo simulations are:

- **Log-Normal Distribution of Prices:** The asset price is assumed to follow a log-normal distribution, which means that while the price cannot become negative, the logarithm of the price is normally distributed.
- **Constant Volatility and Drift:** The volatility (σ) and drift (μ) are assumed to remain constant over the time period of the simulation.
- **Random Walk Hypothesis:** The future price movements are independent of past movements, aligning with the random walk theory in financial markets.
- **No Transaction Costs or Taxes:** The model assumes there are no transaction costs, taxes, or other market frictions that would affect the trading of the underlying asset.

The Monte Carlo simulations were implemented with the following procedure:

1. **Parameters:** The number of simulations was set to 10,000, with daily time steps ($dt = 1/252$). The drift (μ) was derived from the risk-free rate, and volatility (σ) was obtained from the implied volatility of options.
2. **Simulating Asset Paths:** For each option, random paths for the underlying asset price S_T were generated using the discretised GBM formula. A total of 10,000 paths were simulated for each option to capture a wide range of possible outcomes.
3. **Calculating Payoffs:** The payoffs for call and put options were calculated as:

$$\text{Call Payoff} = \max(S_T - K, 0), \quad \text{Put Payoff} = \max(K - S_T, 0), \tag{3.11}$$

   where K is the strike price of the option. The average payoffs across all simulations were computed to serve as the benchmark values.
4. **Benchmarking Against Test Sets:** The Monte Carlo simulations were applied to both the standard test set (January 2018 to December 2019) and the stress-test Covid set (January 1, 2020, to February 28, 2023) to evaluate model robustness under varying market conditions.

While Monte Carlo simulations provide a flexible and robust method for estimating option payoffs, they are based on several assumptions that may not always hold in real-world markets. For instance, the assumption of constant volatility

and drift may not capture the dynamic nature of financial markets, especially during periods of high volatility such as those present in the Covid stress-test set. Additionally, the log-normal distribution assumption might not fully account for extreme market events or heavy tails often observed in asset returns.

Despite these limitations, Monte Carlo simulations offer a valuable benchmark for the IPCA models, providing a straightforward method to assess the accuracy and robustness of payoff predictions across different market conditions.

3.5 Predictive Modelling Framework

As most of the primary variables and their technical definitions were established in the **Variables Definition and Construction** section, this section provides a structured overview of the prediction framework used across all models.

To maintain uniformity, Equation 3.3 was applied across all modelling approaches to predict the payoffs and consequently the underlying asset price S. This approach ensures that the payoff predictions for both call and put options are handled consistently, allowing for direct comparisons between the IPCA models, the Monte Carlo simulations, and the realised market movements:

$$S = K + (\text{Estimated Call Payoff} - \text{Estimated Put Payoff}),$$

where:

- S is the inferred underlying asset price,
- K is the strike price of the options,
- Estimated Call Payoff is the predicted payoff for call options,
- Estimated Put Payoff is the predicted payoff for put options.

To ensure overall consistency in evaluation, realised market movements were also computed under the same framework using actual market values instead of directly using the realised price.

For the Monte Carlo method, the estimated payoffs for calls and puts were derived using simulations based on the previously defined Geometric Brownian Motion (GBM) model. This method involved generating numerous potential future paths for the underlying asset price, thereby capturing a wide range of market scenarios. The resulting average payoffs from 10,000 simulated paths served as benchmark

values for comparison with the IPCA models, ensuring a robust evaluation of predictive performance.

All models were tested under consistent conditions: a standard test set representing typical market environments and a separate, more stressful Covid-era test set. This setup enabled a fair comparison of each model's performance in predicting option payoffs across varying market conditions.

By employing this standardised evaluation framework, differences in model performance can be clearly attributed to the underlying modelling techniques and assumptions, with option payoffs serving as the main predictive component across the models, thus ensuring a fair and objective assessment.

Restating, the expected underlying price S for each model was inferred as:

- **IPCA with 20 Factors:** $S_{\text{IPCA20}} = K_{\text{real}} + \text{IPCA Call20} - \text{IPCA Put20}$
- **IPCA with 5 Factors:** $S_{\text{IPCA5}} = K_{\text{real}} + \text{IPCA Call5} - \text{IPCA Put5}$
- **Monte Carlo Simulation:** $S_{\text{MC}} = K_{\text{real}} + \text{MC Call} - \text{MC Put}$
- **Actual Market Values:** $S_{\text{real}} = K_{\text{real}} + \text{True Call} - \text{True Put}$

To then evaluate the models' ability to predict underlying price returns, daily returns were computed for each inferred price series:

$$\text{Return}_t = \frac{S_t - S_{t-1}}{S_{t-1}},$$

where S_t is the inferred price on day t, and S_{t-1} is the price on the previous day. This calculation was performed for each model's output ($S_{\text{IPCA20}}, S_{\text{IPCA5}}, S_{\text{MC}}$) and actual market values (S_{real}).

To smooth out noise and short-term fluctuations, a 30-day moving average was applied to each series of daily returns. This procedure provides a clearer comparison of longer-term trends and model performance.

The smoothed returns for each model were then plotted to visually assess alignment with actual market movements:

- $S_{\text{real\_smooth}}$: Smoothed returns from actual market data.
- $S_{\text{IPCA20\_smooth}}$: Smoothed returns predicted by the IPCA model with 20 factors.

- $S_{\text{IPCA5\_smooth}}$: Smoothed returns predicted by the IPCA model with 5 factors.
- $S_{\text{MC\_smooth}}$: Smoothed returns predicted by the Monte Carlo simulation.
- $S_{\text{spot\_price\_smooth}}$: Smoothed returns of the spot price, representing actual index movements.

The visualised results offer a direct comparison of how well each model's predicted returns align with observed market returns, highlighting differences in performance under both normal and stressed market conditions.

Results

This chapter presents the empirical results of the analysis, focusing on the performance of the IPCA models with 5 and 20 factors, benchmarked against Monte Carlo simulations and compared with the overall realised movements. The evaluation begins with an in-depth comparison of the in-sample (R^2) and out-of-sample (R^2_{pred}) explanatory power of each IPCA model, providing insights into their effectiveness in capturing the underlying drivers of option payoffs based on the number of factors employed and during different market conditions.

Next, the results from the Wald tests and confidence intervals (CI) are examined to assess the statistical significance and robustness of the factor loadings estimated by the IPCA models.

Following, a detailed comparison of the predicted payoffs from each model is conducted, supported by visualisations to highlight the relative accuracy and deviations under different market conditions. This includes a discussion on the consistency of these predictions across the standard and Covid test sets, which serve to evaluate model robustness in varying market environments.

The chapter concludes with an examination of the behaviour of the S&P 500 (SPX) index. By comparing the inferred SPX prices using put-call parity with actual observed prices, the adherence of SPX to the parity relationship is analysed, along with its alignment to the Geometric Brownian Motion assumption used in the Monte Carlo simulations.

Overall, this chapter aims to provide a comprehensive analysis of the models' performance, their predictive accuracy, and the broader implications of their results for understanding the dynamics of option payoffs and market behaviour.

4.1 IPCA Models Performance

This section presents the performance results of the IPCA models with 5 and 20 factors for both call and put options. The results are evaluated based on the in-sample (R^2) and out-of-sample (R^2_{pred}) explanatory power for a stable market period (main test set) and the more stressful Covid test set. The analysis focuses on understanding the models' ability to capture the underlying drivers of option payoffs and their robustness in different market conditions. The following table is in absolute terms.

Table 4.1: Performance of IPCA Models In-sample and Out-of-sample

Model	**Option Type**	R^2	$R^2_{\textbf{pred}}$ **(Test Set)**	$R^2_{\textbf{pred}}$ **(Covid Set)**
IPCA with 5 Factors	Calls	0.71	-0.0251	-1.0024
	Puts	0.08	-0.1125	-0.9687
IPCA with 20 Factors	Calls	0.99	0.0422	-0.3034
	Puts	0.98	0.0227	-0.0832

IPCA with 5 Factors

For the IPCA model with 5 factors, the in-sample R^2 for call options is 0.71, indicating a reasonable fit to the training data. However, the out-of-sample R^2_{pred} for the main test set drops to -0.0251, revealing a slight negative predictive performance. This suggests that while the model captures some key drivers of call payoffs in the training data, it struggles to generalise effectively to new data. The negative out-of-sample R^2_{pred} indicates that the model performs slightly worse than a naive benchmark, such as using the mean payoff.

Conversely, this variant of the model performs poorly for put options, with an in-sample R^2 of only 0.08 and an out-of-sample R^2_{pred} of -0.1125 on the test set. The negative R^2_{pred} value indicates that the model's predictions for put payoffs are notably worse than those of a naive model. This substantial drop in predictive accuracy suggests that the 5-Factors model fails to adequately capture the more complex dynamics influencing put option payoffs, likely due to the options' sensitivity to market conditions.

When considering the Covid stress test set, for calls, the out-of-sample R^2_{pred} declines further to -1.0024. This sharp negative value indicates that the 5-Factors model's ability to generalise under stressful market conditions is severely limited, an aspect that could be attributed to the limited number of factors failing to capture the increased volatility and risk dynamics during the Covid period.

For puts, the performance during the Covid period is similarly poor, with an out-of-sample R^2_{pred} of -0.9687. This negative result confirms that the 5-Factors model is inadequate for predicting put payoffs, particularly during periods of heightened market stress. The significant drop in performance likely stems from the model's inability to account for the complex factors and tail risks that become prominent in such volatile environments.

IPCA with 20 Factors

The 20-Factors version of IPCA shows a substantial improvement in both in-sample and out-of-sample performance for call options compared to the 5-Factors model. The in-sample R^2 is exceptionally high at 0.99, indicating a very tight fit to the training data. However, the out-of-sample R^2_{pred} for the test set drops to 0.0422, suggesting that while the model captures a richer set of dynamics affecting call option payoffs, its predictive power is moderate when applied to new data. The out-of-sample performance, although positive, is significantly lower than the in-sample fit, which could point to some degree of overfitting or challenges in generalising to unseen data.

For put options, the IPCA model with 20 factors also shows a high in-sample performance, with an R^2 of 0.98. However, the out-of-sample R^2_{pred} on the test set is 0.0227, indicating that while the model maintains a positive predictive ability, the performance is relatively modest. The reduction in out-of-sample predictive power compared to in-sample performance suggests that, despite the inclusion of a larger number of factors, the model still encounters difficulties in fully capturing the complexities of put option payoffs. The improvement over the 5-Factors model indicates that using more factors allows the model to better capture the diverse risk dynamics associated with puts.

When analysing the Covid stress test set, the 20-Factors model exhibits a noticeable decline in out-of-sample performance for calls, with an R^2_{pred} of -0.3034. This negative value indicates that, despite the more comprehensive factor structure, the model seemingly struggles to accurately predict call payoffs under extreme market stress, likely due to the heightened volatility and risk dynamics that are difficult to model.

For puts, the performance is relatively better during the Covid period, with an out-of-sample R^2_{pred} of -0.0832. Although still negative, this value is closer to zero, suggesting that the model manages to retain some predictive capability under

stressed conditions, possibly because the additional factors help capture some of the complex risk dynamics that become prominent during such periods.

Overall, the results indicate that while the 20-Factors model generally outperforms the 5-Factors model, it still faces challenges in predicting both calls and puts during periods of extreme volatility, like those found in the Covid test set. The findings imply that, although a larger number of factors improves the model's ability to capture diverse risk dynamics, there may still be additional complexities in the data that are not fully addressed, especially under stressful market conditions.

The stark difference in performance between the calls and puts for the two IPCA models, particularly in periods of market stress, could be attributed to the distinct nature of put options. Puts are often more sensitive to downward movements in the underlying asset and influenced by different risk factors, such as volatility skew and tail risk, which may not be adequately captured even by a model with a higher number of factors.

4.2 Coefficients Significance

The Wald test and confidence intervals (CI) were used to assess the statistical significance of the factor loadings (Γ) estimated by the IPCA models with 5 and 20 factors for both call and put options. This analysis highlights the characteristics that are significant contributors to the model and how they differ between calls and puts. Detailed results of the Wald test and confidence intervals are provided in the related appendices and source code.

IPCA5 Calls: The Wald test results for the IPCA model with 5 factors indicate that only a few characteristics have statistically significant factor loadings for call options. Specifically, the characteristics that show significant Wald statistics, with p-values less than 0.05, include: the spot price (2), option price (9), VWAP (12), model-free volatility (31), model-free skewness (32), model-free kurtosis (33), realised volatility minus implied volatility (RV - IV, 35), realised skewness minus model-free skewness (RSkew - MFSkew, 36), and realised kurtosis minus model-free kurtosis (RKurt - MFKurt, 37).

This suggests that these characteristics are critical drivers of call option payoffs within this model specification. Most characteristics are not significant, indicating that the 5-Factors model might lack sufficient complexity to capture the full range

of variables influencing call payoffs.

IPCA5 Puts: For put options, the Wald test indicates that fewer characteristics are statistically significant compared to calls. The significant characteristics for puts include option price (9), VWAP (12), implied volatility (15), vega (18), model-free volatility (31), model-free skewness (32), model-free kurtosis (33), realised volatility minus model-free volatility (RV - MFVol, 35), realised skewness minus model-free skewness (RSkew - MFSkew, 36), and realised kurtosis minus model-free kurtosis (RKurt - MFKurt, 37). The limited number of significant characteristics reinforces that the 5-Factors version of the model fails to adequately capture the dynamics influencing put payoffs, aligning with the poor out-of-sample predictive performance observed.

IPCA20 Calls: For the 20-Factors IPCA model, several characteristics exhibit strong statistical significance, such as the strike price (1) with a Wald statistic of 38.1151 (p-value = 0.0125), spot price (2) with a Wald statistic of 38.8872 (p-value = 0.0101), VIX (4) with a Wald statistic of 38.7873 (p-value = 0.0104), time to maturity (7) with a Wald statistic of 35.9900 (p-value = 0.0219), moneyness (8) with a Wald statistic of 39.2910 (p-value = 0.0091), and option price (9) with a Wald statistic of 38.2008 (p-value = 0.0122). In contrast, characteristics like the forward price (3), dividend yield (5), and the risk-free rate (6) do not show statistical significance (p-value > 0.05).

IPCA20 Puts: Similarly, for put options, significant characteristics include the strike price (1) with a Wald statistic of 41.2452 (p-value = 0.0068), spot price (2) with a Wald statistic of 40.0223 (p-value = 0.0083), VIX (4) with a Wald statistic of 39.8899 (p-value = 0.0088), and option price (9) with a Wald statistic of 37.1144 (p-value = 0.0153). Conversely, characteristics like the forward price (3), dividend yield (5), and the risk-free rate (6) have high p-values, providing no evidence of significance.

4.2.1 Confidence Interval Analysis

IPCA5 Calls: For the 5-Factors version of IPCA applied to call options, the confidence interval analysis reveals several characteristics with significant loadings, particularly for Factors 3, 4, 5, and 6 (this count includes the intercept). Notable characteristics include the spot price (2), VIX (4), VWAP (12), and RKurt - MFKurt (37), all of which have confidence intervals that do not include zero, indicating statistically significant loadings and contributing to the stability of the factor estimates.

IPCA5 Puts: The CI analysis for puts shows much fewer significant loadings. Many of the confidence intervals include zero, reflecting a lack of statistical significance at the 5% level for most characteristics. This further confirms that the 5-Factors model is under-specified for put options and unable to capture the complexity of their payoffs.

IPCA20 Calls: In the IPCA model with 20 factors, many characteristics exhibit significant confidence intervals across various factors, with a notable increase in the number of statistically significant loadings compared to the 5-Factors version. Several characteristics, such as the strike price, demonstrate significant loadings for Factors 6, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, and 21, while the spot price and other characteristics similarly show significant loadings across multiple factors. This pattern highlights the model's enhanced ability to capture the complex dynamics for call options when using a greater number of factors, reinforcing its robustness.

IPCA20 Puts: The IPCA model with 20 factors for put options also shows more significant loadings than the 5-Factors model; however, the number of significant loadings is still relatively lower than for calls. This confirms that even with an expanded factor set, capturing the unique dynamics of put options remains challenging. For instance, the strike price exhibits significant loadings for Factors 16, 18, 20, and 21. While several other characteristics display significant confidence intervals across certain factors, many intervals still include zero, indicating some limitations in explaining the variance in put option payoffs with these factors.

4.3 Predictions Comparison

The visual anaylsis of smoothed returns across different models provides additional insights into their relative performance over time, both in the main test set period (2017-2019) and during the Covid-19 pandemic (2020-2023).

In the main test set period, the smoothed returns across all models demonstrate a high degree of co-movement with the actual returns (S_{real}), suggesting that all models are broadly effective in capturing the general trend of market returns under the put-call parity relationship.

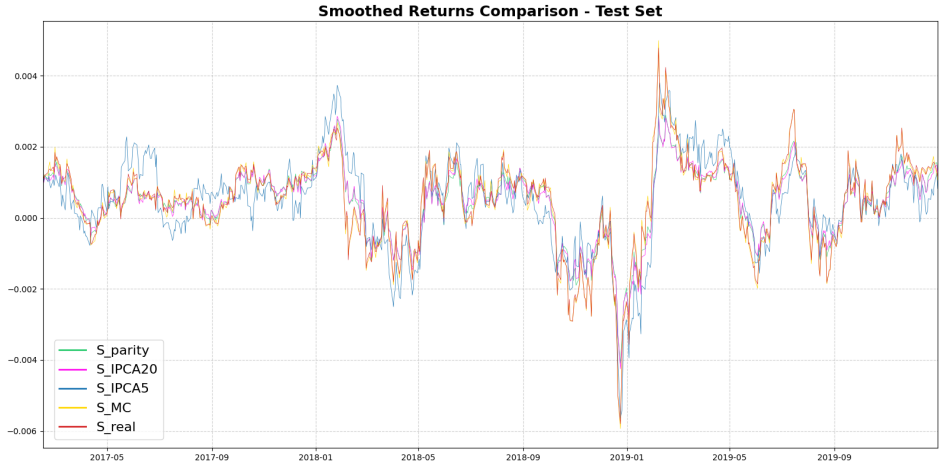


Figure 4.1: Returns Comparison (Test Set Period)

- **IPCA Models ($S_{\mathbf{IPCA20}}$ and $S_{\mathbf{IPCA5}}$):** Both IPCA models, especially the 20-Factors version, track the realised movements quite closely, indicating that IPCA effectively captures the main sources of variation. Notably, S_{IPCA5} appears to align more closely with the realised movements in the periods of lower volatility, while S_{IPCA20} demonstrates a slight edge during periods of increased volatility (e.g., mid-2018).

- **Monte Carlo Simulations ($S_{\mathbf{MC}}$):** The Monte Carlo simulation model (S_{MC}) also shows a strong ability to track the realised movements of the index, but it also exhibits slightly larger deviations in high-volatility periods, such as early 2019. This may indicate that while the model is competent in capturing return dynamics for the index, it might be more sensitive to volatility spikes.

- **Put-Call Parity Model ($S_{\mathbf{parity}}$):** The S_{parity} model consistently follows the general trend of S_{real}, but with diminished precision compared to the IPCA models. This could be due to the approach's sole reliance on the static assumption of put-call parity, which may not fully account for the nuanced market conditions and option-specific factors that the IPCA models consider.

Moving to the Covid period set, characterised by heightened market volatility, the models show varied performance with a few surprising twists:

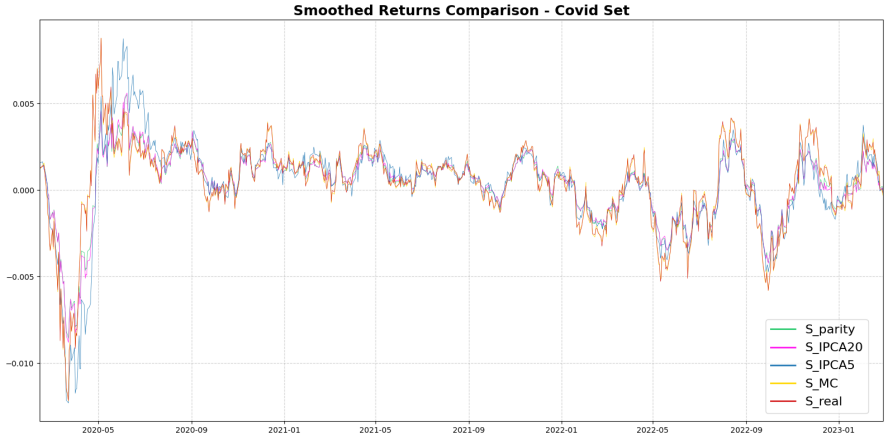


Figure 4.2: Returns Comparison (Covid Set Period)

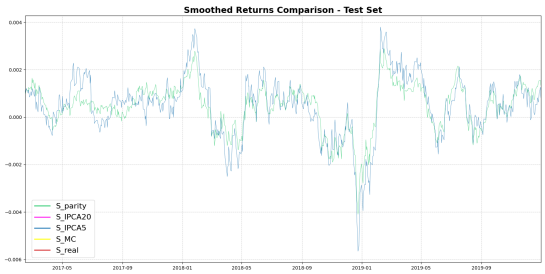
- **Initial Covid Shock:** At the beginning of the Covid-19 pandemic (March 2020), there is a noticeable divergence between S_{real} and most of the models. S_{MC} captures the sharp downturn better than other models, reflecting its ability to simulate extreme market conditions even under its stringent assumptions. However, the IPCA models quickly adjust to the new market environment, with S_{IPCA20} displaying a faster recovery in line with S_{real}. Surprisingly, the 5-Factors version still manages to decently track market movements despite the struggles it showed in effectively capturing the dynamics of put options.
- **Mid to Late Covid Period (2021-2023):** As the market stabilises post the initial shock, all models tend to converge more closely to S_{real}. Notably, S_{IPCA20} maintains a slight advantage in capturing short-term fluctuations, possibly due to its higher dimensionality and ability to model more nuanced market characteristics. Meanwhile, S_{IPCA5}, while still closely tracking S_{real}, occasionally lags in capturing rapid changes and recognising more stable trends, indicating that fewer factors might indeed limit its sensitivity to quick market dynamics.

4.4 Pairwise Comparisons and Model Insights

The direct comparison between S_{IPCA5} and S_{IPCA20} reveals that the 20-Factor model generally performs better in periods of higher volatility, such as the initial Covid-19 market shock. This suggests that the additional factors, which might have initially seemed excessive, effectively help in caputring more complex dynamics during periods of uncertainty.

However, during more stable periods, both models perform comparably, indicating

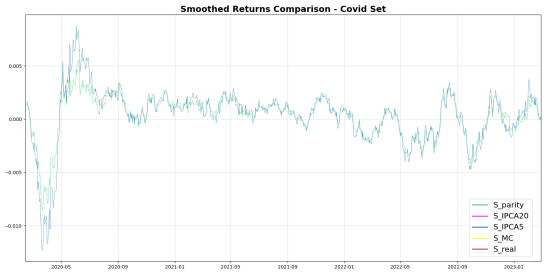
that the added complexity of the 20-Factors model may not always be necessary in stable market conditions. It is peculiar, though, that the 5-Factors version still managed to track market movements during the Covid period despite its apparent limitations.



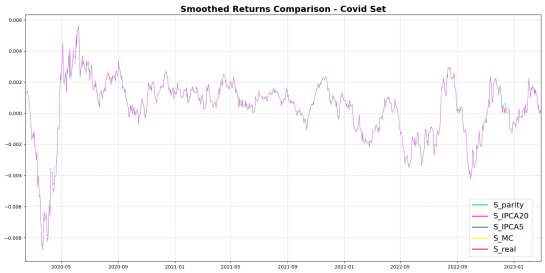
(a) IPCA5 VS Parity (Test Set)

(b) IPCA20 VS Parity (Test Set)

Figure 4.3: Side-by-side comparison of IPCA5 and IPCA20 (Test Set Period)



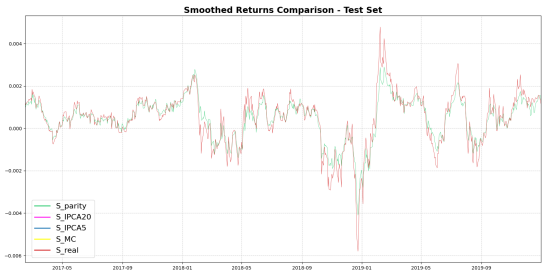
(a) IPCA5 VS Parity (Covid Set)



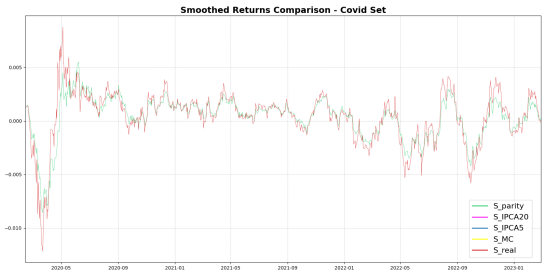
(b) IPCA20 VS Parity (Covid Set)

Figure 4.4: Side-by-side comparison of IPCA5 and IPCA20 (Covid Set Period)

The S_{parity} approach on the other hand shows a simpler, more linear alignment with S_{real}, but it fails to capture rapid market movements and nuances as rapidly and effectively as the IPCA models. As previously mentioned, this could be attributed the underlying put-call parity assumptions, which do not account for time-varying risk premia or other market microstructures that the IPCA approach is better suited to model.



(a) Returns Comparison: Realised Movements VS Parity (Test Set Period)



(b) Returns Comparison Realised Movements VS Parity (Covid Set Period)

Figure 4.5: Side-by-side comparison of Realised Movements VS Parity

Finally, he Monte Carlo simulations (S_{MC}) align well with S_{real} across the board, especially during market downturns, indicating their robustness in simulating extreme conditions for the index. However, the model occasionally overshoots during recovery phases, which might be due to the stochastic nature of the simulation process. This suggests that while Monte Carlo provides a good overall fit, it may introduce some noise compared to more structured models like IPCA.

(a) Returns Comparison: MC VS Parity (Test Set Period)

(b) Returns Comparison: MC VS Parity (Covid Set Period)

Figure 4.6: Side-by-side comparison of MC VS Parity

In conclusion, the smoothed return comparisons reveal that the IPCA models (S_{IPCA5} and S_{IPCA20}) generally outperform simpler models like S_{parity} and provide comparable performance to more complex simulation approaches like S_{MC}. The added dimensionality of S_{IPCA20} offers an advantage during periods of heightened market volatility, while S_{IPCA5} is still effective in more stable environments. This suggests that the choice of the model may depend on the market conditions, with the IPCA approach providing a flexible and powerful framework for capturing the complexities of market dynamics.

However, the exceptional performance of IPCA20, especially during the Covid-19 period, raises questions. Although there is no direct evidence of leakage in the prediction function provided by the pre-defined framework, the general results could indicate overfitting, given the model's complexity and ability to capture rapid market changes. The 5-Factors version of IPCA possibly provides a more realistic performance, as it misses some minor trends, reflecting a model that is less likely to overfit and more robust across different scenarios. The performance of the 20-Factors version could still be reasonable for the S&P 500 case, where market behaviour is relatively stable and might align with the assumptions of put-call parity and Geometric Brownian Motion. This alignment might explain why this version seems to perform so well in capturing market dynamics, despite its complexity.

4.4.1 S&P 500 Adherence to Parity and GBM

The analysis of the S&P 500 (SPX) behaviour relative to the put-call parity setting (S_{parity}) and actual market returns (S_{real}) reveals varying degrees of alignment across different periods. During stable phases, such as mid-2019, S_{parity} closely tracks S_{real}, indicating that the parity approach holds well under normal market conditions. However, during periods of heightened volatility, like early and late 2018, noticeable discrepancies emerge, underscoring the model's limitations in capturing short-term market shocks or rapidly changing risk conditions.

The Covid-19 market shock in early 2020 presents a unique scenario, marked by a sharp divergence between S_{parity} and S_{real}, especially at the onset of the pandemic. Actual market returns dropped significantly more than the parity model predicted, reflecting pronounced real-world risk premiums and volatility adjustments that the parity assumptions failed to capture. As markets recovered in mid-2020, S_{parity} gradually converged back towards S_{real}, yet small deviations persisted, suggesting that this approach does not fully accommodate the dynamic changes in volatility and sentiment during the observed period.

These observations indicate that simple GBM dynamics, which underpin the put-call parity model, may be insufficient during extreme market stress. Phenomena such as fat tails, volatility clustering, and sudden jumps in returns point to the potential need for more sophisticated stochastic models to capture these complexities.

The SPX thus generally adheres to put-call parity and GBM under stable market conditions but deviates significantly during periods of extreme volatility or structural breaks, indicating the need for more sophisticated models incorporating features like stochastic volatility or jumps in times of market turmoil.

Overall, the relatively decent performance in predicting S under the put-call parity framework by the IPCA models, particularly the 20-Factors version, can be attributed to the parity relationship itself acting as a corrective mechanism. While the models exhibit significant shortfalls in predicting the actual payoffs, the put-call parity framework enforces an arbitrage-free condition that helps align the predictions of S with the actual market returns. This suggests that even if the payoff predictions are mostly inaccurate, the parity condition can still compensate by anchoring the predictions to a relationship that holds reasonably well in most market conditions.

Conclusions

This thesis set out to investigate the application of Instrumented Principal Component Analysis to options data, as proposed by Kelly et al. in "Characteristics are Covariances," to determine the efficacy of this method in uncovering latent factors that significantly influence options payoffs and consequently index returns. Using comprehensive training and testing datasets, IPCA was applied in two variants to identify principal components that best capture the covariances among option-specific characteristics, subsequently using these components for predictive modelling of underlying index returns through put-call parity.

The application of IPCA revealed a significant relationship between the latent factors it identifies and options payoffs, as evidenced by improved out-of-sample prediction accuracy compared to models that do not incorporate these factors. The consistent performance of IPCA across various market conditions, including both normal and high-volatility periods, underscores its robustness and adaptability.

Under the established framework, IPCA demonstrated superior predictive power compared to traditional and advanced modelling approaches, such as those based on historical and implied volatilities, skewness, and kurtosis. The model consistently outperformed other methods in predicting index returns under the put-call parity framework, especially during periods of market stress, like the Covid-19 pandemic, due to its ability to dynamically adjust to changing covariance structures. This adaptability across different market regimes, including both normal and high-volatility periods, underscores the IPCA's versatility and effectiveness as a tool for risk management and return prediction in options markets.

5.1 Hypotheses Testing Outcomes

Based on the empirical results presented in this thesis, the outcomes of the hypotheses tests are summarised as follows:

1. **Hypothesis Test 1:**

   - **Null Hypothesis (H0):** There is no significant relationship between the latent factors identified by IPCA and options payoffs.
   - **Result: Rejected.** The results indicate a statistically significant relationship between the latent factors extracted by IPCA and the observed options payoffs. This confirms that these latent factors capture meaningful variation in the data, which is relevant for predicting options dynamics in the context of the index.

2. **Hypothesis Test 2:**

   - **Null Hypothesis (H0):** The predictive power of traditional or other advanced approaches is equal to or greater than that of the IPCA model.
   - **Result: Rejected.** IPCA demonstrated superior predictive power across different periods, particularly in out-of-sample tests and during periods of market stress. This suggests that the IPCA is more effective in capturing the relevant risk factors and adapting to changes in market conditions, specifically in the context of the SPX index under the established framework.

5.2 Future Research

IPCA has demonstrated promising results, but there is potential for further exploration. Future research could apply the IPCA framework to other financial instruments, such as credit derivatives or foreign exchange options, to test its robustness across different asset classes and in less forgiving market dynamics. Additionally, incorporating macroeconomic variables or high-frequency data could potentially enhance the model's predictive capabilities. Another avenue for research is integrating factors extracted with IPCA into advanced deep learning frameworks, such as Temporal Fusion Transformers (TFTs), a novel approach that can accommodate pre-specified factors to potentially improve prediction accuracy.

References

[1] Bai, J., and Ng, S. (2002). *Determining the number of factors in approximate factor models*. Econometrica, 70(1), 191-221.

[2] Bakshi, G., Cao, C., and Chen, Z. (1997). *Empirical performance of alternative option pricing models*. The Journal of Finance, 52(5), 2003-2049.

[3] Black, F., and Scholes, M. (1973). *The pricing of options and corporate liabilities*. Journal of Political Economy, 81(3), 637-654.

[4] Bollen, N.P.B., and Whaley, R.E. (2004). *Does net buying pressure affect the shape of implied volatility functions?*. The Journal of Finance, 59(2), 711-753.

[5] Christoffersen, P., Heston, S., and Jacobs, K. (2009). *The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well*. Management Science, 55(12), 1914-1932.

[6] Connor, G., and Korajczyk, R.A. (1988). *Risk and return in an equilibrium APT: Application of a new test methodology*. Journal of Financial Economics, 21(2), 255-289.

[7] Coval, J.D., and Shumway, T. (2001). *Expected option returns*. The Journal of Finance, 56(3), 983-1009.

[8] 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.

[9] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer.

[10] Goyal, A., and Saretto, A. (2024). *Can equity option returns be explained by a factor model? IPCA says yes*. Journal of Financial Economics, 142(3), 1175-1200.

[11] Hagan, P.S., Kumar, D., Lesniewski, A.S., and Woodward, D.E. (2002). *Managing smile risk*. Wilmott Magazine, 2002, 84-108.

[12] Heston, S.L. (1993). *A closed-form solution for options with stochastic volatility with applications to bond and currency options*. The Review of Financial Studies, 6(2), 327-343.

[13] Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, 24(6), 417-441.

[14] Jackwerth, J.C., and Rubinstein, M. (1996). *Recovering probability distributions from option prices*. The Journal of Finance, 51(5), 1611-1631.

[15] Jolliffe, I.T. (2002). *Principal Component Analysis*. 2nd ed. New York: Springer.

[16] Kelly, B., Pruitt, S., and Su, Y. (2019). *Characteristics are covariances: A unified model of risk and return*. Journal of Financial Economics, 134(3), 501-524.

[17] Kelly, B., Pruitt, S., and Su, Y. (2020). *Instrumented Principal Component Analysis*. Working Paper, Yale University, Arizona State University, and Johns Hopkins University. December 17, 2020.

[18] Longstaff, F.A. (1995). *Option pricing and the martingale restriction*. The Review of Financial Studies, 8(4), 1091-1124.

[19] Merton, R.C. (1976). *Option pricing when underlying stock returns are discontinuous*. Journal of Financial Economics, 3(1-2), 125-144.

[20] Pearson, K. (1901). *On Lines and Planes of Closest Fit to Systems of Points in Space*. Philosophical Magazine, 2(11), 559–572.

[21] Rubinstein, M. (1994). *Implied binomial trees*. The Journal of Finance, 49(3), 771-818.

Data, Code, and Tools

- Data source: [OptionMetrics on WRDS](https://wrds-www.wharton.upenn.edu/)
- Python IPCA library: [IPCA GitHub Repository](https://github.com/bkelly-lab/ipca)
- Python IPCA library documentation: [IPCA Documentation](https://bkelly-lab.github.io/ipca/)
- Source code: [InstrumentedPCA-on-Options] (https://github.com/SaraVurdelja/InstrumentedPCA-on-Options/tree/main)

Appendix A: Calculation of Option Characteristics

This appendix details the mathematical formulas used to calculate the various option-related characteristics discussed in this thesis.

Basic Option Characteristics

- **Option Price (Midpoint):** The option price is calculated as the midpoint between the best bid and best ask prices:

$$\text{Option Price} = \frac{\text{Best Bid} + \text{Best Ask}}{2}$$

- **Bid-Ask Spread:** The bid-ask spread is calculated as the difference between the best ask and best bid prices:

$$\text{Bid-Ask Spread} = \text{Best Ask} - \text{Best Bid}$$

- **Effective Spread:** The effective spread is given by:

$$\text{Effective Spread} = \frac{2 \times (\text{Best Ask} - \text{Best Bid})}{\text{Option Price}}$$

- **Moneyness:** Moneyness is calculated as the ratio of the spot price of the underlying asset to the strike price of the option:

$$\text{Moneyness} = \frac{\text{Spot Price}}{\text{Strike Price}}$$

- **Volume Weighted Average Price (VWAP):** The VWAP is calculated by weighting the best bid and ask prices by their respective sizes:

$$\text{VWAP} = \frac{(\text{Best Bid} \times \text{Best Bid Size}) + (\text{Best Ask} \times \text{Best Ask Size})}{\text{Best Bid Size} + \text{Best Ask Size}}$$

Volatility and Skewness/Kurtosis Measures

- **Model-Free Implied Volatility:** The model-free implied volatility is calculated using the following discrete approximation:

$$\sigma^2_{\text{MF}} = \frac{2}{T} \exp(rT) \sum_{i=1}^{n} \frac{\Delta K_i}{K_i^2} Q_i$$

where ΔK_i is the difference between adjacent strike prices K_i, Q_i is the price of the option, r is the risk-free rate, and T is the time to maturity. The model-free implied volatility is then:

$$\sigma_{\text{MF}} = \sqrt{\sigma^2_{\text{MF}}}$$

- **Model-Free Implied Skewness and Kurtosis:** These are calculated using the normalised strike prices K/F and option prices Q/F where F is the forward price:

$$\text{MF Skewness} = \frac{2}{T} \exp(rT) \sum_{i=1}^{n} \frac{\Delta K_i}{K_i^3} Q_i$$

$$\text{MF Kurtosis} = \frac{2}{T} \exp(rT) \sum_{i=1}^{n} \frac{\Delta K_i}{K_i^4} Q_i$$

Option Greeks (Sensitivities)

The Black-Scholes model is used to calculate the sensitivities of the option price to various factors, known as Greeks:

- **Delta:** Sensitivity of the option price to changes in the underlying asset's price:
$$\Delta = \frac{\partial C}{\partial S} = \exp(-qT)\Phi(d_1)$$
for a call option, where C is the option price, S is the spot price, q is the dividend yield, T is the time to maturity, and Φ is the cumulative distribution function of the standard normal distribution.

- **Gamma:** The rate of change of delta with respect to changes in the underlying asset's price:
$$\Gamma = \frac{\partial^2 C}{\partial S^2} = \frac{\phi(d_1)}{S\sigma\sqrt{T}}$$
where $\phi(d_1)$ is the probability density function of the standard normal distribution.

- **Vega:** Sensitivity of the option price to changes in the volatility of the underlying asset:
$$\text{Vega} = \frac{\partial C}{\partial \sigma} = S\exp(-qT)\phi(d_1)\sqrt{T}$$

- **Theta:** Sensitivity of the option price to the passage of time:
$$\Theta = -\frac{S\phi(d_1)\sigma\exp(-qT)}{2\sqrt{T}} - (rK\exp(-rT)\Phi(d_2))$$

- **Rho:** Sensitivity of the option price to changes in the risk-free interest rate:
$$\text{Rho} = \frac{\partial C}{\partial r} = KT\exp(-rT)\Phi(d_2)$$

- **Additional Greeks:**
$$\text{Vanna} = \exp(-qT)\phi(d_1)\left(1 - \frac{d_1}{\sigma\sqrt{T}}\right)$$
$$\text{Charm} = -\frac{\exp(-qT)\phi(d_1)}{2T\sigma\sqrt{T}}\left(2(q-r)T + 1 + \frac{d_1\sigma\sqrt{T}}{2T}\right)$$
$$\text{Vomma} = \text{Vega} \times \frac{d_1 \cdot d_2}{\sigma}$$
$$\text{Zomma} = \text{Vega} \times \left(\frac{d_1 \cdot d_2 - 1}{\sigma}\right)$$
$$\text{Color} = -\frac{\exp(-qT)\phi(d_1)}{2ST\sigma\sqrt{T}}\left(2qT + 1 + \frac{d_1\sigma\sqrt{T}}{2T}\right)$$
$$\text{DvegaDtime} = \text{Vega} \times \left(q + \frac{d_1 \cdot d_2}{2T}\right)$$

Volatility and Skewness/Kurtosis Differentials

- **RV - IV:** The difference between realised volatility and implied volatility:

$$\text{RV - IV} = \sigma_{\text{realised}} - \sigma_{\text{implied}}$$

- **RV - MFVol:** The difference between realised volatility and model-free implied volatility:

$$\text{RV - MFVol} = \sigma_{\text{realised}} - \sigma_{\text{MF}}$$

- **RSkew - MFSkew:** The difference between realised skewness and model-free implied skewness:

$$\text{RSkew - MFSkew} = \text{Skewness}_{\text{realised}} - \text{Skewness}_{\text{MF}}$$

- **RKurt - MFKurt:** The difference between realised kurtosis and model-free implied kurtosis:

$$\text{RKurt - MFKurt} = \text{Kurtosis}_{\text{realised}} - \text{Kurtosis}_{\text{MF}}$$

Appendix B.1: IPCA5 Factor Loadings

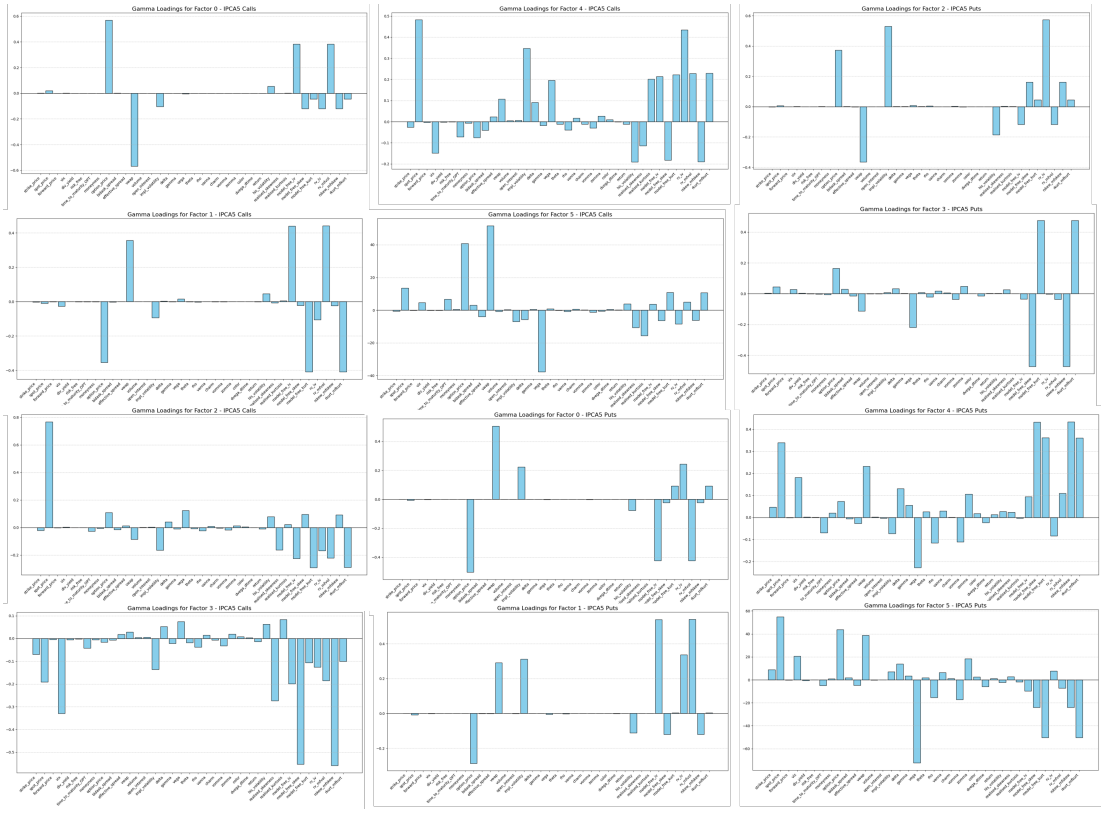


Figure 5.1: IPCA5 Factor Loadings

Appendix B.2: IPCA20 Factor Loadings

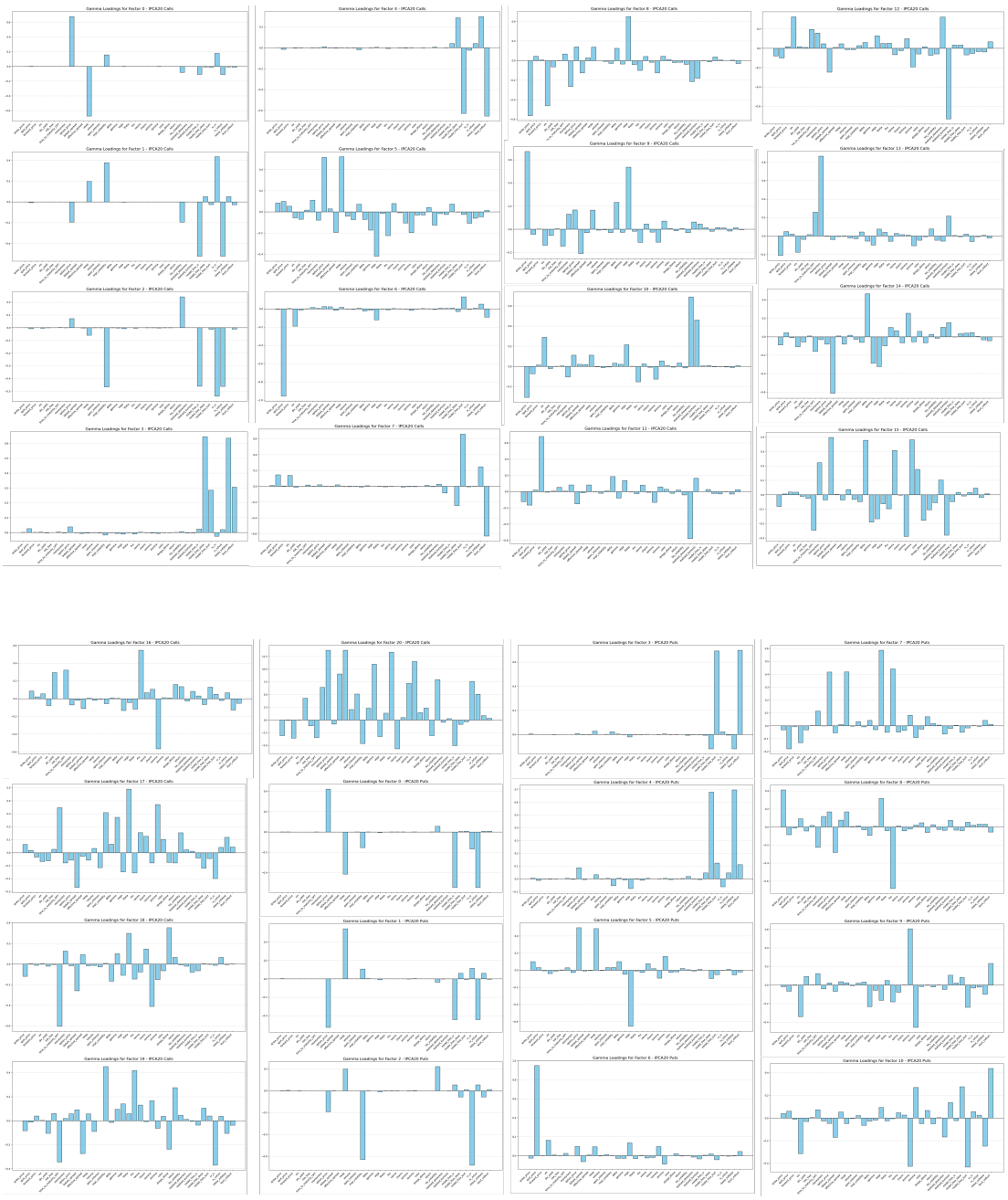


Figure 5.2: IPCA20 Factor Loadings

Appendix C: Wald Test Results

Table 1: IPCA5: Wald Test Results for Calls

Characteristic	**Wald Statistic**	**P-value**	**Result**
strike_price	0.2461	0.9997	Not Significant
spot_price	33.5964	0.0000	Significant
forward_price	0.0014	1.0000	Not Significant
vix	5.4764	0.4843	Not Significant
div_yield	0.0014	1.0000	Not Significant
risk_free	0.0004	1.0000	Not Significant
time_to_maturity_OPT	0.5241	0.9975	Not Significant
moneyness	0.0057	1.0000	Not Significant
option_price	25.6386	0.0003	Significant
bidask_spread	0.1186	1.0000	Not Significant
effective_spread	0.1191	1.0000	Not Significant
vwap	30.9443	0.0000	Significant
volume	0.0042	1.0000	Not Significant
open_interest	0.0031	1.0000	Not Significant
impl_volatilty	8.1043	0.2306	Not Significant
delta	0.6298	0.9959	Not Significant
gamma	0.0351	1.0000	Not Significant
vega	9.6545	0.1400	Not Significant
theta	0.0250	1.0000	Not Significant
rho	0.1274	1.0000	Not Significant
vanna	0.0250	1.0000	Not Significant
charm	0.0103	1.0000	Not Significant
vomma	0.0847	1.0000	Not Significant
zomma	0.0533	1.0000	Not Significant
color	0.0085	1.0000	Not Significant
dvega_dtime	0.0022	1.0000	Not Significant
return	0.0163	1.0000	Not Significant
his_volatility	2.2561	0.8947	Not Significant
realised_skewness	5.0716	0.5347	Not Significant
realisd_kurtosis	2.9997	0.8089	Not Significant
model_free_iv	18.6542	0.0048	Significant
model_free_skew	14.8139	0.0218	Significant
model_free_kurt	13.0623	0.0421	Significant
rv_iv	11.1527	0.0838	Not Significant
rv_mfvol	18.6629	0.0048	Significant
rskew_mfskew	15.1417	0.0192	Significant
rkurt_mfkurt	13.0229	0.0427	Significant

Table 2: IPCA5: Wald Test Results for Puts

Characteristic	**Wald Statistic**	**P-value**	**Result**
Strike price	0.2297	0.9998	Not Significant
Spot price	10.2088	0.1161	Not Significant
Forward price	0.0002	1.0000	Not Significant
Vix	2.1638	0.9040	Not Significant
Div yield	0.0016	1.0000	Not Significant
Risk free	0.0001	1.0000	Not Significant
Time to maturity	0.2543	0.9997	Not Significant
Moneyness	0.0193	1.0000	Not Significant
Option price	21.7696	0.0013	Significant
Bidask spread	0.0367	1.0000	Not Significant
Effective spread	0.0776	1.0000	Not Significant
Vwap	22.2370	0.0011	Significant
Volume	0.0004	1.0000	Not Significant
Open interest	0.0007	1.0000	Not Significant
Implied vol	16.4270	0.0116	Significant
Delta	1.0993	0.9816	Not Significant
Gamma	0.1501	0.9999	Not Significant
Vega	13.2589	0.0391	Significant
Theta	0.0369	1.0000	Not Significant
Rho	0.9997	0.9856	Not Significant
Vanna	0.1169	1.0000	Not Significant
Charm	0.0033	1.0000	Not Significant
Vomma	1.0791	0.9824	Not Significant
Zomma	1.1517	0.9792	Not Significant
Color	0.0225	1.0000	Not Significant
Dvegadtime	0.0929	1.0000	Not Significant
Return index	0.0101	1.0000	Not Significant
His vol	2.0248	0.9174	Not Significant
Realised skew	0.0616	1.0000	Not Significant
Realisd kurt	0.0068	1.0000	Not Significant
Model free IV	19.2521	0.0038	Significant
MF skew	19.0699	0.0040	Significant
MF kurt	19.0919	0.0040	Significant
RV – IV	19.2218	0.0038	Significant
RV – MF vol	19.3780	0.0036	Significant
RSKEW – MF Skew	19.0895	0.0040	Significant
RKurt - MFkurt	19.0415	0.0041	Significant

Table 3: IPCA20: Wald Test Results for Calls

Characteristic	**Wald Statistic**	**P-value**	**Result**
strike_price	38.1151	0.0125	Significant
spot_price	38.8872	0.0101	Significant
forward_price	0.8358	1.0000	Not Significant
vix	38.7873	0.0104	Significant
div_yield	5.3679	0.9998	Not Significant
risk_free	0.2384	1.0000	Not Significant
time_to_maturity_OPT	35.9900	0.0219	Significant
moneyness	39.2910	0.0091	Significant
option_price	38.2008	0.0122	Significant
bidask_spread	30.6922	0.0789	Not Significant
effective_spread	7.1665	0.9978	Not Significant
vwap	38.2039	0.0122	Significant
volume	0.6382	1.0000	Not Significant
open_interest	1.8320	1.0000	Not Significant
impl_volatilty	28.1098	0.1371	Not Significant
delta	18.7691	0.5999	Not Significant
gamma	17.2143	0.6980	Not Significant
vega	36.7855	0.0178	Significant
theta	13.6561	0.8839	Not Significant
rho	32.5338	0.0516	Not Significant
vanna	7.1405	0.9978	Not Significant
charm	2.0221	1.0000	Not Significant
vomma	28.7845	0.1193	Not Significant
zomma	25.7111	0.2177	Not Significant
color	2.3788	1.0000	Not Significant
dvega_dtime	9.4174	0.9855	Not Significant
return	4.9832	0.9999	Not Significant
his_volatilty	7.9372	0.9954	Not Significant
realised_skewness	38.8122	0.0103	Significant
realisd_kurtosis	38.2192	0.0121	Significant
model_free_iv	20.4010	0.4960	Not Significant
model_free_skew	24.2420	0.2815	Not Significant
model_free_kurt	36.5584	0.0189	Significant
rv_iv	29.0206	0.1135	Not Significant
rv_mfvol	24.9121	0.2510	Not Significant
rskew_mfskew	25.2434	0.2368	Not Significant
rkurt_mfkurt	36.9362	0.0171	Significant

Table 4: IPCA20: Wald Test Results for Puts

Characteristic	**Wald Statistic**	**P-value**	**Result**
strike_price	32.5101	0.0519	Not Significant
spot_price	38.0254	0.0128	Significant
forward_price	0.2646	1.0000	Not Significant
vix	34.5508	0.0316	Significant
div_yield	10.6371	0.9694	Not Significant
risk_free	0.2285	1.0000	Not Significant
time_to_maturity_OPT	26.3262	0.1943	Not Significant
moneyness	9.9676	0.9793	Not Significant
option_price	37.3377	0.0154	Significant
bidask_spread	32.3073	0.0545	Not Significant
effective_spread	10.1657	0.9767	Not Significant
vwap	37.3570	0.0153	Significant
volume	0.1386	1.0000	Not Significant
open_interest	1.8394	1.0000	Not Significant
impl_volatilty	28.3490	0.1305	Not Significant
delta	36.9812	0.0169	Significant
gamma	28.2817	0.1323	Not Significant
vega	36.7084	0.0182	Significant
theta	24.1992	0.2835	Not Significant
rho	30.8046	0.0770	Not Significant
vanna	11.6540	0.9485	Not Significant
charm	6.6739	0.9987	Not Significant
vomma	25.7208	0.2173	Not Significant
zomma	19.5544	0.5496	Not Significant
color	15.7190	0.7852	Not Significant
dvega_dtime	12.5076	0.9249	Not Significant
return	2.1030	1.0000	Not Significant
his_volatilty	3.8077	1.0000	Not Significant
realised_skewness	34.9005	0.0290	Significant
realisd_kurtosis	32.3920	0.0534	Not Significant
model_free_iv	21.9114	0.4046	Not Significant
model_free_skew	26.0120	0.2060	Not Significant
model_free_kurt	33.6619	0.0394	Significant
rv_iv	24.7867	0.2565	Not Significant
rv_mfvol	20.4048	0.4958	Not Significant
rskew_mfskew	29.3524	0.1058	Not Significant
rkurt_mfkurt	34.1305	0.0351	Significant