Master Thesis Quantitative Finance

Explaining the Cross-Section of Equity Option Returns through Unsupervised Machine Learning Methods

Author: Bas van Andel (508077)

Supervisor: Dr. M. Grith

Second Assessor: Prof. Dr. M. van der Wel

May 1, 2021

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Abstract

Limited prior research addresses the cross-section of delta-hedged equity option returns. Meanwhile, the literature on explaining stock returns is ever expanding with the introduction of novel machine learning methods. In this paper, I aim to bridge this gap in the literature by studying conditional latent factor models that use covariates in order to explain the cross-section of equity option returns and evaluating the most important characteristics that drive these models. I find that these unsupervised machine learning models significantly outperform traditional factor models and that non-linearity of the factor loadings further improves the models, especially from an economic perspective. Furthermore, the models identify the volatility risk premium and option liquidity as the most important characteristics in terms of explanatory power. However, I discover that these advanced modeling approaches are not very robust and, thus, very dependent on the data that they are applied to.

Keywords: Equity option returns, Instrumented Principal Component Analysis, conditional autoencoder, unsupervised machine learning, factor model

Contents

1 Introduction

Ever since the introduction of the Capital Asset Pricing Model and the three-factor model proposed by Fama and French (1993), research on explaining the cross-sectional differences between stock returns has become increasingly popular. However, up until today, there is only a limited number of studies aiming to explain the cross-section of equity option returns. Horenstein, Vasquez, and Xiao (2018) highlight the importance of understanding the cross-section of option returns. In practice it could prove expensive for investors to hedge their portfolio variance risk. They argue that one can hedge this variance risk by trading the factors that drive the cross-section of delta-hedged option returns. Besides, Bakshi and Kapadia (2003) demonstrate that delta-hedged option portfolios do not only exhibit the equity risk premium, but also a volatility risk premium. Thus, as equity options cannot be simply regarded as levered positions in the underlying stocks, they provide an interesting area of research.

This paper investigates whether novel methods in the field of unsupervised machine learning yield superior results compared to conventional factor models, when explaining the cross-section of equity option returns. These modeling approaches utilize covariates to explain the risk-return characteristics and could, therefore, offer enhanced explanatory power, while maintaining the economic interpretability. I evaluate the performance in terms of total R^2 and predictive R^2, in addition to the Sharpe ratio and variance. Furthermore, this paper aims to identify a subset of the most important characteristics in terms of explaining this cross-section. Finally, by splitting the test sample based on these important characteristics, into sets with materially different types of options (e.g. liquid options versus illiquid options), I challenge the robustness of these results.

For the covariates, I consider a large set of 95 option and stock characteristics, inspired by Green, Hand, and Zhang (2017), Brooks, Chance, and Shafaati (2018) and Horenstein et al. (2018). Each of the characteristics in the set has proven to be relevant in prior research and can be broadly categorized as either trading, return or balance sheet related.

In this paper, I study two unsupervised machine learning methods. The first model is a latent conditional factor model called Instrumented Principal Component Analysis (IPCA). Kelly, Pruitt, and Su (2019) introduce this new method in an effort to overcome the shortcomings of conventional observable and latent factor models, in order to explain the cross-section of stock returns. In contrast to Principal Component Analysis (PCA), in IPCA factor loadings linearly depend on observable asset characteristics (the instrumental variables). Therefore, IPCA allows for the factor model to incorporate stock characteristics into the analysis, capitalizing on previous research on factor structures. Furthermore, IPCA performs dimension reduction directly in the model.

Gu, Kelly, and Xiu (2019a) present another asset pricing model for explaining stock returns, inspired by autoencoder neural networks. The proposed Conditional Autoencoder (CA) model extends on the standard autoencoder model by including covariates and helps guide dimension reduction. However, in contrast to IPCA of Kelly et al. (2019) which assumes linearity for the factor exposures, the autoencoder allows for nonlinear functions. Consequently, the autoencoder model shares the benefits offered by IPCA, but allows for more flexibility, which could improve the results.

The aforementioned modeling approaches seem to offer a lot of potential when explaining stock returns. To be best of my knowledge, these methods have not yet been studied in the field of delta-hedged equity option returns. Hence, I aim to bridge this gap in the literature, as it could offer new insights into the performance of conditional latent factor models, while also expanding on existing literature about important characteristics in explaining equity option returns.

The empirical analysis finds that the models that use covariates, IPCA and CA, significantly outperform conventional observable and latent factor models, both from a statistical and economic perspective. For the 5-factor model, IPCA and CA demonstrate a total R^2 of 8.0% and 7.7%, respectively, compared to 4.4% for standard Principal Component Analysis. From an economic perspective, CA is yielding the highest Sharpe ratio of 2.61, for long-short portfolios constructed using its predicted returns.

Moreover, the paper finds that two characteristics are the most important in explaining the cross-section of equity option returns: the volatility risk premium and option liquidity. IPCA and CA both observe this result, but deviate in terms of the importance of the remaining, less significant characteristics. This finding reinforces the results from prior studies conducted on the subject, most notably by Goyal and Saretto (2009) and Christoffersen, Goyenko, Jacobs, and Karoui (2018). Finally, the robustness analysis shows that the models' performance is highly dependent on the sample of options considered. Especially when splitting the sample based on the aforementioned important characteristics, the model results can differ substantially.

This study contributes to the existing literature by providing evidence of the versatility of the models under different types of data. It does, however, reveal a potential shortcoming of the models in terms of their robustness to the sample. Furthermore, the paper reaches similar conclusions to prior research, but by taking a structurally different approach. This further reinforces the soundness of these findings.

The remaining structure of the paper is as follows. Section 2 elaborates further on how this paper is related to prior literature. Section 3 describes the modeling framework and methods in more details. Moreover, it discusses the testing and performance criteria. Section 4 introduces the data for the empirical study and discusses how delta-hedged option returns are constructed. Section 5 presents the findings of the comparison of the different methods. Finally, Section 6 concludes and proposes possible areas of further research.

2 Literature Review

Bakshi and Kapadia (2003) are one of the pioneers in studying delta-hedged option returns, focusing on index options. They demonstrate that on average a delta-hedged option strategy underperforms zero, where at-the-money options show the largest underperformance. Furthermore, this underperformance increases during periods of higher volatility.

Leveraging the common finding that volatility is mean-reverting, Goyal and Saretto (2009) show that large deviations of implied volatility from historical volatility often imply mispricing of an option. That is, the zero-cost strategy with a long position in an option portfolio of stocks with a large positive difference between historical and implied volatility and short in a portfolio with a large negative difference generates statistically and economically significant returns. In line with the findings of Bakshi and Kapadia (2003) for index options, Cao and Han (2013) show that individual delta-hedged equity option returns decrease with an increase in the idiosyncratic volatility of the underlying stock. This result cannot be explained by the volatility risk premium. Moreover, they note that returns decrease further for options on less liquid stocks and when option open interest is higher. Finally, Christoffersen et al. (2018) find that equity option returns are larger in the case of less liquid options, which is in contrast to option returns on less liquid stocks.

In an endeavor to summarize the aforementioned findings, Horenstein et al. (2018) aim to find the common factors that drive the cross-section of delta-hedged equity option returns. Out of the 13 considered factors, they find that a four-factor model (firm size, idiosyncratic volatility, volatility deviation and market volatility risk) can explain the cross-section and time series of option returns. Moreover, they show that traditional stock factors cannot price the cross-section of option returns. Brooks et al. (2018) perform a similar analysis, but consider a substantially larger set of factors, namely 99 underlying stock and option characteristics. Green et al. (2017) study the relevant characteristics to explain monthly stock returns and provide a large set of stock characteristics in doing so. Together, these papers form the basis of the characteristics data that I consider in my analysis.

Besides observable factor models, many research is devoted to employing latent factor models in order to explain the cross-section of returns. Chamberlain and Rothschild (1982) pioneered the use of Principal Component Analysis (PCA) as a latent factor model to simultaneously estimate factors and factor loadings from asset returns. PCA is directly related to CA, as autoencoders can be considered a nonlinear neural network counterpart to PCA (Baldi & Hornik, 1989).

Kelly et al. (2019) and Gu et al. (2019a) introduce novel methods that aim to take the favourable aspects from both observable and latent factor models, by incorporating covariates into the model. Gu, Kelly, and Xiu (2019b) expand on their research by considering a wider range of asset pricing models, but only concentrate on evaluating their pure predictability, as opposed to the risk-return characteristics. The aforementioned papers focus solely on explaining the cross-section of stock returns.

3 Methodology

This section elaborates on the modeling approaches used in the empirical analysis and outlines the evaluation framework. First, I describe the IPCA model in detail. Second, I discuss a generalization of IPCA, which concerns a conditional autoencoder (CA). Third, I introduce two benchmark models that share features of the aforementioned methods. Section 3.4 elaborates on the evaluation framework. Next, Section 3.5 is devoted to explaining the methodology behind determining the most relevant characteristics in the models. Finally, I describe the robustness checks on the models in Section 3.6.

3.1 Instrumented Principal Component Analysis

Instrumented Principal Component Analysis (IPCA), as pioneered by Kelly, Pruitt, and Su (2017), is a novel modeling approach aimed at explaining the cross-section of returns. Similar to Principal Component Analysis (PCA), IPCA treats risk factors as latent. However, unlike PCA's static loadings and inability to incorporate other data beyond returns, IPCA allows for time-varying factor loadings that depend on observable characteristics related to the asset. Hence, IPCA offers a revamped approach to standard PCA that exploits its favorable aspects, while overcoming its shortcomings. For instance, the instrumental variables used in IPCA ensure an increased economic interpretability over PCA. In doing so, IPCA is able to capitalize on the extensive literature on observable factor models and factor anomalies.

Another beneficial feature of IPCA is its embedded dimension reduction in the model. Consequently, IPCA imposes parameter parsimony and should be well-behaved when characteristics are highly correlated, noisy or spurious.

Following the definition from Kelly et al. (2019), the IPCA model is specified as

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t} f_{t+1} + \epsilon_{i,t+1},$$
$$\alpha_{i,t} = z_{i,t}^{'} \Gamma_\alpha + \nu_{\alpha,i,t},\ \beta_{i,t} = z_{i,t}^{'} \Gamma_\beta + \nu_{\beta,i,t}, \tag{1}$$

where $r_{i,t+1}$ is the excess return, $\beta_{i,t}$ the dynamic factor loading and f_{t+1} the vector of k latent factors. The time-varying factor loadings are a function of the observable asset characteristics, or instruments, denoted by L-vector $z_{i,t}$. The anomaly intercept is represented by $\alpha_{i,t}$, which is also a function of the instrumental variables. Throughout this research, I impose the no-arbitrage restriction in order to focus on the risk-return relationship of the factors. Hence, it follows that $\alpha_{i,t} = 0$ for all i and t or, alternatively, $\Gamma_\alpha = \mathbf{0}$ and $\nu_{\alpha,t,t} = 0$ for all i and t.

The L x K parameter matrix Γ_β represents a mapping from potentially high dimensional characteristics to a small number of risk factors loadings, ensuring dimension reduction takes place within the model. Simultaneously, $\nu_{\beta,i,t}$ captures any residual behavior of the loadings that is orthogonal to the instruments, as the observable characteristics might not perfectly explain the risk exposures.

3.1.1 Model Estimation

Under the zero-intercept no-arbitrage restriction, the model in Equation 1 can be further simplified as

$$r_{i,t+1} = z_{i,t}^{'} \Gamma_\beta f_{t+1} + \epsilon_{i,t+1}^{*}, \tag{2}$$

where $\epsilon_{i,t+1}^{*} = \epsilon_{i,t+1} + \nu_{\beta,i,t} f_{f+1}$ is the new error term.

Besides, I rewrite the model in vector form, such that Equation 2 becomes

$$r_{t+1} = Z_t \Gamma_\beta f_{t+1} + \epsilon_{t+1}^{*}, \tag{3}$$

where r_{t+1} is the N x 1 vector of stacked asset returns at time $t+1$, Z_t is the N x L matrix of stacked characteristics of each asset at time t, and the individual asset error terms $\epsilon_{i,t+1}^{*}$ are represented by the N x 1 vector ϵ_{t+1}^{*}.

Following Kelly et al. (2019), I define the objective function as the minimization of the sum of squared errors in the model:

$$\min_{\Gamma_\beta, F} \sum_{t=1}^{T-1} \left(r_{t+1} - Z_t \Gamma_\beta f_{t+1}\right)^{'} \left(r_{t+1} - Z_t \Gamma_\beta f_{t+1}\right). \tag{4}$$

The solution of this objective function is provided by a system of first-order conditions, without a closed-form solution. Specifically, Kelly et al. (2019) show that the f_{t+1} and Γ_β that minimize Equation 4 satisfy the following equations:

$$\hat{f}_{t+1} = \left(\hat{\Gamma}_\beta^{'} Z_t^{'} Z_t \hat{\Gamma}_\beta\right)^{-1} \hat{\Gamma}_\beta^{'} Z_t^{'} r_{t+1}, \forall t \tag{5}$$

and

$$\text{vec}(\hat{\Gamma}_\beta^{'}) = \left(\sum_{t=1}^{T-1} Z_t^{'} Z_t \otimes \hat{f}_{t+1} \hat{f}_{t+1}^{'}\right)^{-1} \left(\sum_{t=1}^{T-1} \left[Z_t \otimes \hat{f}_{t+1}^{'}\right]^{'} r_{t+1}\right). \tag{6}$$

As this system of equations does not have a closed-form solution, I resort to numerical methods to solve the system for f_{t+1} and Γ_β simultaneously.

3.1.2 Numerical Solution: Alternating Least Squares

The solution of f_{t+1} and Γ_β in Equation 5 and 6 is currently unidentified. In order to identify a unique solution, I impose three additional restrictions: (1) $\Gamma_\beta^{'} \Gamma_\beta = \text{I}_k$; (2) the unconditional second moment of f_t is a diagonal matrix with descending elements along the diagonal; (3) the mean of f_t is non-negative. These assumptions solely ensure a unique solution, without imposing any economic restrictions on the model (Kelly et al., 2019).

Alternating Least Squares Following Kelly et al. (2019), I apply the Alternating Least Squares algorithm to find numerical solutions to f_{t+1} and Γ_β. In particular, I choose the eigenvectors corresponding to the k largest eigenvalues of the characteristic-managed portfolio's second moment matrix as the initial guess for $\hat{\Gamma}_\beta$. That is, $\sum_t r_{t+1}^p r_{t+1}^{p'} = \left(\frac{Z_t^{'} r_{t+1}}{N_{t+1}}\right) \left(\frac{Z_t^{'} r_{t+1}}{N_{t+1}}\right)^{'}$ is the initial guess, where N_{t+1} is the number of non-missing observations. Section 4.4 elaborates on the usage and advantages of the managed portfolios. Kelly et al. (2019) highlight that this starting point is a close approximation to the exact solution and can, thus, significantly reduce the algorithm's run-time.

After the initialization of $\hat{\Gamma}_\beta$, I proceed by solving for $\hat{f}_{t+1}$ in Equation 5. This boils down to a least squares regression for all t. Subsequently, I plug in the obtained $\hat{f}_{t+1}$ into Equation 6 and solve for $\hat{\Gamma}_\beta$ by evaluating the least squares regression. I continue these iterations until convergence, which is reached when at any point the maximum absolute change in any element of either $\hat{\Gamma}_\beta$ or $\hat{f}_{t+1}$, for all t, is less than 10^{-6}.

3.2 Conditional Autoencoder

Even though IPCA offers various advantages over conventional methods such as observable factor models and latent factor models, it also has a potential limitation. Namely, IPCA only allows for a linear relationship between the characteristics and the factor loadings. Goyal and Saretto (2009) argue that it is unlikely for a linear factor model to explain the cross-section of option returns over any discrete time interval.

Gu et al. (2019a) introduce an extended conditional autoencoder (CA) model that bypasses this limitation by allowing for nonlinear functions for the factor exposures. Similar to IPCA (and PCA), autoencoders are classified as dimension reduction models in the field of unsupervised machine learning. They are a type of neural network, in which the outputs try to estimate the corresponding input variables. In general, neural networks proceed by passing input variables through a set of neurons in the hidden layers (i.e. encoding), which is subsequently decoded on the output layer. While the standard autoencoder solely uses asset returns as input, Gu et al. (2019a) propose a conditional extension that incorporates asset characteristics and, consequently, generalizes the IPCA model.

3.2.1 Model Estimation

In defining the conditional autoencoder model, I follow Gu et al. (2019a) for any new notation, while holding on to earlier notation and assumptions. In particular, I use the rectified linear unit ($g(y) = \max(y, 0)$) as the model's nonlinear function in the methodology and empirical results, and maintain the zero-intercept no-arbitrage restriction throughout. Moreover, I define $K^{(l)}$ as the number of neurons in each layer l, for $l = 1, \ldots, L$, and let $r_k^{(l)}$ or $z_k^{(l)}$ denote the output of neuron k in layer l for the factors and loadings, respectively.

The conditional autoencoder consists of two branches[1]: one branch to model the factor loadings, or betas, as a nonlinear function of the asset characteristics, and a second branch to model the factors using the individual asset returns. The first branch, related to factor loadings, is specified by the following set of recursive equations:

$$z_{i,t-1}^{(0)} = z_{i,t-1}, \tag{7}$$

$$z_{i,t-1}^{(l)} = \max\left(b^{(l-1)} + W^{(l-1)} z_{i,t-1}^{(l-1)}, 0\right), \; l = 1, \ldots, L_\beta, \tag{8}$$

$$\beta_{i,t-1} = b^{(L_\beta)} + W^{(L_\beta)} z_{i,t-1}^{(L_\beta)}. \tag{9}$$

Equation 7 initializes the network using the characteristic data, $z_{i,t-1}$. Equation 8 outlines

[1] Gu et al. (2019a) provide an excellent visual representation of this model in Figure 2.

the recursive transformation that occurs in each hidden layer of the inputs from the previous to the next layer. $W^{(l-1)}$ represents a $K^{(l)}$ x $K^{(l-1)}$ weighting matrix and $b^{(l-1)}$ is the $K^{(l)}$-vector of bias parameters. The objective of the neural network is to simultaneously optimize these weighting and bias parameters. In the empirical analysis, I fix the number of hidden layers, L_β, to 1, as this provides the best results for Gu et al. (2019a). Finally, Equation 9 describes how the K-dimensional factor loadings are obtained as the output of the network.

For the second input layer, related to factors, I also fix the number of hidden layers, L_f, to 1. In this way, the resulting factors are simply linear combinations or portfolios of the individual assets. The network is defined as follows:

$$r_t^{(0)} = r_t^p, \tag{10}$$

$$r_t^{(1)} = \max\left(\tilde{b}^{(0)} + \tilde{W}^{(0)} r_t^{(0)}, 0\right), \tag{11}$$

$$f_t = \tilde{b}^{(1)} + \tilde{W}^{(1)} r_t^{(1)}. \tag{12}$$

This time, the network is initialized with a set of portfolios of individual asset returns, defined as:

$$r_t^p = \frac{Z_{t-1} r_t}{N_t}, \tag{13}$$

where N_t represents the number of non-missing observations. Gu et al. (2019a) highlight the various benefits that emerge from initializing the network with a set of portfolios, instead of individual asset returns r_t. Most notably, it performs an initial dimension reduction of the data and resolves issues arising from unbalanced panels. Similar to Equation 8, Equation 11 describes the transformation through the single hidden layer, where $\tilde{W}^{(l-1)}$ and $\tilde{b}^{(l-1)}$ once again represent the weighting matrix and vector of bias parameters, respectively. Finally, the K-dimensional output is obtained from Equation 12.

Combining the output of the factor loadings (Equation 9) and factors (Equation 12) results in the final model fit for each individual asset return. More specifically, I multiply loading $\beta_{i,t-1}$ by factor f_t for every i and t. The goal of the autoencoder is to optimize the weighting and bias parameters in such way that sum of squared errors is minimized. More formally, the objective function of the autoencoder is defined as follows:

$$\min_{b,\tilde{b},W,\tilde{W}} \frac{1}{NT} \sum_{t=0}^{T-1} \sum_{i=0}^{N-1} ||r_{i,t+1} - \beta_{i,t}^{'}(b, W) f_{t+1}(\tilde{b}, \tilde{W})||^2, \tag{14}$$

where $\beta_{i,t}^{'}(b, W)$ and $f_{t+1}(\tilde{b}, \tilde{W})$ are the model outputs of Equation 9 and Equation 12, respectively, given their respective weighting and bias parameters.

3.2.2 Regularization

I apply regularization to further mitigate the risk of overfitting of the data. Regularization promotes more parsimonious models by adding a penalty term to the objective function that shrinks coefficient estimates towards zero. Consequently, it sacrifices in-sample performance in an endeavor to improve its out-of-sample results. Following Gu et al. (2019a), I apply three regularization techniques: LASSO, early stopping and random initialization. Besides, I split the data into a disjoint training, validation and test sample, as further elaborated on in Section 3.4.3.

LASSO LASSO regularization adapts the objective function in Equation 14 slightly by adding a first-order penalty term. Moreover, LASSO promotes sparsity by setting insignificant weight parameters equal to zero. The revised objective function is now defined as

$$\mathcal{L}(\theta;\cdot) = \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=0}^{N-1}||r_{i,t+1} - \beta_{i,t}^{'}f_{t+1}||^2 + \lambda\sum_{j}|\theta_j|, \tag{15}$$

where $\lambda\sum_j|\theta_j|$ represents the LASSO penalty term, with non-negative hyperparameter λ, which is tuned in the validation sample. θ summarizes all the weight parameters (i.e. weighting matrices $W^{(j)}$ and $\tilde{W}^{(j)}$, and bias parameters $b^{(j)}$ and $\tilde{b}^{(j)}$) from the networks in Equation 7 through 12.

Early Stopping As a second regularization technique, I apply early stopping. The algorithm starts with an initial guess that all weight parameters of θ are zero. At each iteration, it reduces the pricing errors in the training sample by updating the parameter estimates. Simultaneously, errors in the validation sample are calculated using these parameter estimates. This process proceeds until the validation sample errors no longer decline, even though the model continues to reduce errors in the training sample. Intuitively, early stopping ensures that parameters are shrunken towards the initial guess of zeros, since this algorithm generally terminates before full optimization of the training sample fitting errors has been achieved. Algorithm 1 provides a formal definition of early stopping.

Algorithm 1 Early Stopping

Initialization: $j = 0$, $\epsilon = \infty$ and select p (patience parameter)
while $j < p$ **do**
  Update θ (using SGD in Algorithm 2)
  Use θ to calculate the errors from the validation sample ($\epsilon^{'}$)
  if $\epsilon^{'} < \epsilon$ **then**
    $j \leftarrow 0$
    $\epsilon \leftarrow \epsilon^{'}$
    $\theta^{'} \leftarrow \theta$
  else
    $j \leftarrow j + 1$
return $\theta^{'}$

Random Seeds The final regularization method that I employ is a randomization technique that further improves the stability of the results. Specifically, I initialize the neural network using ten different random seeds and average the estimates that result from these networks. Hence, this diminishes the risk of reporting results from a local optimum in the optimization.

3.2.3 Optimization: Stochastic Gradient Descent

Gu et al. (2019a) propose the use of the stochastic gradient descent (SGD) algorithm to optimize the conditional autoencoder model from Section 3.2.1. The power of this algorithm lies in the fact that it only requires a small random subset of the data at each iteration to evaluate the gradient. While the accuracy diminishes slightly as a consequence of this approximation, considerable gain is achieved in terms of run time. In particular, I adopt the SGD algorithm as proposed by Kingma and Ba (2014) and specified in Algorithm 2.

Algorithm 2 Stochastic Gradient Descent (SGD)

Initialization: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, $m_0 = 0$, $v_0 = 0$ and $t = 0$
Tuning parameters: α (learning rate) and b (batch size)
Set up θ_0 as initial parameter vector
while θ_t is not converged **do**
  $t \leftarrow t + 1$
  $g_t \leftarrow \nabla_\theta \left[\frac{1}{b} \sum_{s \in B_t} \mathcal{L}(\theta; s)\right] |_{\theta=\theta_{t-1}}$, where B_t is the set of batch samples
  $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
  $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
  $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$
  $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$
  $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
return θ_t

Finally, I also adopt the Batch Normalization algorithm, proposed by Ioffe and Szegedy (2015). This technique bypasses the problem that arises in neural networks when the distribution of each layer's inputs changes during training, also referred to as internal covariate shift. This phenomenon can substantially increase the run-time or can result in non-optimal outcomes. Ioffe and Szegedy (2015) resolve this problem by normalizing the layer inputs for each batch. Algorithm 3 formalizes this process.

Algorithm 3 Batch Normalization

1: Input parameters: values of x over mini-batch $B = \{x_1, \ldots, x_N\}$
2: Initialization: ϵ is constant (for numerical stability)
3: Tuning parameters: γ and β
4: $\mu_B \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i$
5: $\sigma_B^2 \leftarrow \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_B)^2$
6: $\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$
7: $y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i)$
8: **return** y_i for $i = 1, \ldots, N$

3.3 Benchmark Models

Since the models that I describe in Section 3.1 and Section 3.2 share features of both observable and latent factor models, it logically follows to introduce multiple benchmark models to compare against. More specifically, in line with Gu et al. (2019a), I apply Principal Component Analysis (PCA) and an observable factor model as the benchmark models. While both benchmarks are unconditional by default, by considering characteristic-managed portfolios as test assets, I embed conditional information into the models.

3.3.1 Latent Factor Model

PCA forms a natural candidate, due to its close relation to the advanced models considered in this research. First of all, PCA is a dimension reduction technique in the field of unsupervised learning. IPCA and autoencoder described in Section 3.1 and Section 3.2, respectively, fall into the same category of machine learning methods. Second of all, IPCA is considered an extension to PCA that includes instruments or characteristics into the analysis.

In contrast to feature selection, where a subset of factors are selected from a potentially large group of characteristics, PCA is a dimension reduction technique that extracts features. In doing so, PCA solely uses the asset returns and constructs k latent factors, for a pre-specified k.

In order to find the principal components, I first define the N x N sample covariance matrix of the asset returns as $\tilde{\Sigma} = \frac{1}{T-1}\sum_{t=1}^{T}(r_t - \bar{r})(r_t - \bar{r})'$. In this equation, the returns are stacked such that $r_t = (r_{1,t}, \ldots, r_{N,t})$, and $\bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t$ represents the mean returns across the time dimension. Jolliffe (2002) shows that the covariance matrix can be decomposed into

$$\tilde{\Sigma} = \mathrm{A}\Lambda\mathrm{A}', \tag{16}$$

where A is the matrix of eigenvectors and Λ the diagonal matrix with eigenvalues λ_i. Using this eigenvector decomposition, PCA proceeds by selecting k eigenvectors in decreasing order of importance in terms of explaining the variation. That is, the eigenvalues are sorted in a descending order and the first k eigenvalues, and corresponding eigenvectors, are selected by PCA. Subsequently, I construct the k latent factors, f_t, using the eigenvector transformation for each i-th factor: $f_{i,t} = \mathrm{a}_i'(r_t - \bar{r})$, where a_i is the i-th column of matrix A and $f_t = (f_{1,t}, \dots, f_{k,t})$. Finally, the first k columns of matrix A provide the corresponding factor loadings matrix B $= (\beta_1, \dots \beta_k)$, which is static across time: $\beta_i = \beta_{i,t}$ for all $i = 1, \dots, k$ and $t = 1, \dots, T$.

3.3.2 Observable Factor Model

Besides PCA as a latent factor model, I introduce an observable factor model of k factors, for $k = 1, 2, 3, 4, 5, 6$. When considering stock returns, an obvious observable benchmark is the Fama and French (1993) three-factor model, and its more recent extensions (e.g., Fama and French (2015)). Kelly et al. (2019) and Gu et al. (2019a) indeed resort to this literature for their observable factor model benchmarks, as they only analyze stock return data. Unfortunately, the literature on equity options does not provide an equivalently common benchmark model. Therefore, I capitalize on the findings of Horenstein et al. (2018) about common factors in equity option returns to create observable factor benchmark models.

In the stock return literature, the Capital Asset Pricing Model (CAPM) is generally used as an observable factor model with only one factor: the excess market return (often proxied by the excess return of the S&P 500 Index over the risk-free return). In a similar fashion, Goyal and Saretto (2009) and Cao and Han (2013) argue that one can consider the delta-hedged return of S&P 500 Index options as a market factor for option returns. Consequently, the 1-factor observable factor benchmark model is defined as

$$r_{i,t+1} = \alpha_i + \beta_{m,i} DH_{m,t} + \epsilon_{i,t+1}, \tag{17}$$

where $DH_{m,t}$ is the delta-hedged return of the S&P 500 Index option and $\beta_{m,i}$ the associated factor loading of each asset i. The intercept, α_i, and error term, $\epsilon_{i,t+1}$, are similar in definition to the basic framework in Equation 18.

Horenstein et al. (2018) find six relevant factors to explain delta-hedged equity option returns, out of their 13 candidate factors. Moreover, they rank these factors in terms of importance. All candidate factors are created based on long-short decile portfolios. To construct k-factor models for $k = 2, 3, 4, 5, 6$, I gradually add five factors to the model in Equation 17 in order of importance, as presented by Horenstein et al. (2018). That is, the 2-factor model consists of DH_m and the size factor, which represents the long-short decile portfolio based on market capitalization of the underlying stock. The 3-factor model adds a factor related to idiosyncratic volatility, while the 4-factor model also includes a factor for the volatility differential between realized and implied volatility. Finally, the 5-factor models adds a factor related to the cash-to-assets ratio and the 6-factor model also includes a factor based on analyst earnings forecast dispersion. Appendix Section B provides a complete overview of all the k-factor observable benchmark models for $k = 1, 2, 3, 4, 5, 6$, including elaborate definitions of the factors used.

3.4 Evaluation Framework

This section introduces statistical and economic evaluation criteria used in the empirical analysis. Moreover, it describes how the dataset is divided into a training, validation and test sample, which is a common approach when evaluating machine learning methods.

In order to be able to elaborate on the evaluating criteria, I define a generic factor model that reflects each of the models in this paper as

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t}' f_{t+1} + \epsilon_{i,t+1}, \tag{18}$$

where $r_{i,t+1}$ is the excess asset return, which in this paper is represented by the delta-hedged equity option return, as defined in Section 4.2. Furthermore, f_{t+1} is the K-vector of factor returns, $\beta_{i,t}$ the associated factor loading and the intercept is given by $\alpha_{i,t}$. Finally, $\epsilon_{i,t+1}$ is the error term, with $\mathrm{E}_t(\epsilon_{i,t+1}) = \mathrm{E}_t(\epsilon_{i,t+1} f_{t+1}) = 0$, for all i and t. The cross-section is represented by $i = 1, \dots, N$ and the time dimension by $t = 1, \dots, T$.

Besides, I follow Kelly et al. (2019) in defining the risk price associated with factors, λ_t, as the expected one-step ahead return of the factors. That is, $\lambda_t = \mathrm{E}_t(f_{t+1})$. λ_t can also be expressed as a function of the stochastic discount factor[2] (Kelly et al., 2019).

3.4.1 No-Arbitrage Restriction

In evaluating the performance of methods, one can choose between comparing their pure prediction performance or solely considering the predictability and compensation that arises from

[2]This statement is correct under the following assumptions. No arbitrage, a stochastic discount factor (m_{t+1}) exists and the following equation is satisfied (for any excess return $r_{i,t}$): $\mathrm{E}_t(m_{t+1} r_{i,t+1}) = 0$. This equation in turns implies: $\mathrm{E}_t(r_{i,t+1}) = \frac{Cov_t(m_{t+1}, r_{i,t+1})}{Var_t(m_{t+1})} \left(-\frac{Var_t(m_{t+1})}{\mathrm{E}_t(m_{t+1})} \right)$, where the last fraction is equal to the risk price λ_t.

exposure to risk factors. In line with Gu et al. (2019a), I opt for the latter approach and, thus, impose the no-arbitrage restriction on all the models prior to the comparison. That is, all the models in this research are all specified without an intercept, or $\alpha_{i,t} = 0$ for all i and t in Equation 18.

In order to challenge the validity of this restriction, I test to which extend this no-arbitrage restriction is satisfied in the data. The model residuals, i.e. the difference between the actual return and the out-of-sample fitted value, are represented by

$$\hat{\epsilon}_{i,t+1} = r_{i,t+1} - \hat{\beta}'_{i,t}\hat{f}_{t+1}, \tag{19}$$

for all i and t. Intuitively, these residual terms should not be statistically different from zero for the no-arbitrage restriction to be valid. Hence, following Gu et al. (2019a), I perform t-tests for every asset and for each of the modeling approaches to evaluate the restriction.

3.4.2 Performance Evaluation

This paper evaluates the out-of-sample performance of the modeling approaches based on two statistical criteria: total R^2 and predictive R^2, and two economic criteria: Sharpe ratio and variance. I follow Gu et al. (2019a) for the definition of these model performance measures. The variables in Equation 20 and Equation 21 follow the definitions from the framework in Equation 18.

Total R^2 Total R^2 assesses the model's performance in terms of describing the riskiness of delta-hedged options. That is, it represents the fraction of variance in $r_{i,t+1}$ that is explained by $\hat{\beta}'_{i,t}\hat{f}_{t+1}$, and is defined as

$$\text{Total } R^2 = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - \hat{\beta}'_{i,t}\hat{f}_{t+1}\right)^2}{\sum_{i,t} r^2_{i,t+1}}. \tag{20}$$

Predictive R^2 Predictive R^2 evaluates the model's ability to explain variation in risk compensation and predict future returns based on past data. More formally, it is the fraction of variation in $r_{i,t+1}$ that is explained by the predictions of future returns ($\hat{\beta}'_{i,t}\hat{\lambda}_t$), and is defined as

$$\text{Predictive } R^2 = 1 - \frac{\sum_{i,t}\left(r_{i,t+1} - \hat{\beta}'_{i,t}\hat{\lambda}_t\right)^2}{\sum_{i,t} r^2_{i,t+1}}, \tag{21}$$

where $\hat{\lambda}_t$ is the sample average of $\hat{f}$ up until month t.

Sharpe ratio and Variance In addition to the two main statistical evaluation criteria, total R^2 and predictive R^2, I also evaluate the models' economic performance. That is, I follow Gu et al. (2019a) and use each model's return predictions to form decile portfolios of the options. Subsequently, I construct a monthly-rebalanced zero-cost portfolio by taking a long position in the top decile and a short position in the bottom decile.

The resulting portfolios are assessed based on their Sharpe ratios and variances. Sharpe ratio (Sharpe, 1966), defined as the average excess return over its standard deviation, is a commonly used metric by both academia and practitioners to evaluate the risk-adjusted performance of a financial asset. The variance highlights the variability in the performance and is thus also reported for completeness.

3.4.3 Training, Validation and Test Samples

Machine learning methods are often prone to overfitting, resulting in exceptional in-sample results at the cost of poor out-of-sample performance. In order to bypass this problem, it is common in the machine learning literature to divide the data into disjoint, collectively exhaustive, training, validation and test samples. The training sample is used to estimate the model parameters. If necessary, these estimates are subsequently used in the validation sample, in order to tune the hyperparameters. While the training and validation sample are sometimes combined, validation of the hyperparameters plays a crucial role when regularization is applied. Section 3.2.2 elaborates on the regularization techniques that I apply to the conditional autoencoder. Finally, the test sample is used to evaluate the model's out-of-sample performance based on the model estimates from the training and validation samples.

3.5 Determining the Relevant Characteristics

Besides the statistical and economic performance evaluation described in Section 3.4.2, an important part of the empirical analysis is concerned with identifying the relevant characteristics that drive the various models. In line with Kelly et al. (2019) and Gu et al. (2019a), I define characteristic importance in terms of contribution to the total R^2. That is, I calculate the reduction in total R^2 for each characteristic resulting from setting all values of this characteristic to zero, while keeping the others intact. While this analysis can be performed for any model, with any number of factors, I focus on identifying important characteristics in the IPCA and CA models with 5 factors. Finally, I investigate whether similar characteristics are identified for both models, what their economic interpretation is and whether this is in line with findings from prior research.

3.6 Robustness

Based on the findings from Section 3.5, I select the two most important characteristics among both the IPCA and conditional autoencoder models. Due to the potential magnitude of the explanatory power arising from these characteristics, it is advisable to challenge the robustness of the models when isolating either of these two characteristics. Specifically, I adopt the approach from Kelly et al. (2019) for large versus small stocks and break the sample in two groups (i.e. high and low values), for each characteristic. The group cut-off points are determined by the median values of the respective characteristic.

For each of the four groups, without performing any additional model estimations, I recalculate the total R^2 and predictive R^2, holding the model parameters fixed at their estimates from the original sample. A similar statistical performance in each group would imply that the models and its findings are robust. Besides the characteristics-based robustness test, I also test the models' robustness by splitting the sample at random and performing a similar analysis.

4 Data

This section introduces the option data used throughout the rest of the paper. Moreover, it describes how I construct delta-hedged option returns and provides summary statistics of the processed data. Finally, I discuss the option and stock characteristics used in the various models.

4.1 Data and Filtering Procedure

For the empirical analysis, I focus on the U.S. market and obtain daily equity option and stock data from January 1996 until December 2018. More specifically, I obtain option data on the constituents of the S&P 500 Index (S&P) from OptionMetrics and stock data from the Center for Research in Security Prices (CRSP). The constituents of the S&P represent a large part of the total market capitalization of U.S. public companies and are most likely to be associated with highly tradable and liquid options.

Following Cao and Han (2013) and Horenstein et al. (2018), I apply several filters to the option data. First, options with zero trading volumes, missing bid prices, missing underlying stock data, abnormal bid-ask spreads or mid prices below \$1/8 are excluded. I also exclude any options that violate the no-arbitrage condition. Second, only options with moneyness between 0.8 and 1.2 are kept. Moneyness is defined as the ratio of the stock price over the strike price. Third, in order to avoid the early exercise premium of American options, I exclude options where the underlying stock paid a dividend during the remaining life of the option. Lastly, as suggested by Boyer and Vorkink (2014), options with extreme prices are excluded.

At the end of each month and for each firm, I select the option closest to being at-the-money with the shortest maturity above one month, and that has an expiration date on the third Friday of the month. In practice, this results in a set of options with a maturity of approximately one and a half month. This is consistent with previous literature on delta-hedged option returns (e.g., Cao and Han (2013) and Horenstein et al. (2018)) and results in a subset with the most tradable options. Finally, in line with Brooks et al. (2018), any options with zero open interest at the end of the month, when the options are selected, are removed. This data filtering procedure results in a final sample of 67,157 option-month observations for calls.

Following the methodology, I split the data into a training, validation and test sample, maintaining a division of approximately 30%, 20% and 50%, respectively. In particular, I classify the first seven years of data from January 1996 until December 2002 as the training data set and the four years from January 2003 until December 2006 as the validation sample. The remaining data, from January 2007 until December 2018, is used as test data for the out-of-sample evaluation.

4.2 Delta-Hedged Option Returns

The return on an individual equity option or portfolio of options depends on various factors, often in practice and in the literature referred to as the option greeks. An important option greek is the delta, which represents the sensitivity of the option price to a change in the price of the underlying stock. Formally, the delta is defined as the first derivative of the option price with respect to the underlying stock price. OptionMetrics calculates the delta and other option greeks using the binomial tree, as introduced by Cox, Ross, and Rubinstein (1979).

In order to isolate the option return, irrespective of the movements of the underlying stock, one can consider the delta-hedged option return. In the case of a call option, delta hedging involves short selling the underlying stock equal to the delta value the option, such that the delta of the combined position is zero (i.e. delta neutral). As the delta of an option is varying, the size of the hedge has to be adjusted accordingly. In practice, however, it is impossible to continuously adjust the size of the hedge and, thus, I assume a daily rebalanced delta-hedge in this study.

Following the definition from Bakshi and Kapadia (2003) and Cao and Han (2013), I consider an option that is hedged (rebalanced) N times over a period $[t, t+\tau]$. In the case of a daily rebalanced delta-hedge, it follows that $N = \tau$. The discrete delta-hedged option gain, $\pi_{t,t+\tau}$, is defined as

$$\pi_{t,t+\tau} = C_{t+\tau} - C_t - \sum_{n=0}^{N-1} \Delta_{C,t_n}(S_{t_{n+1}} - S_{t_n}) - \sum_{n=0}^{N-1} \frac{a_n r_{t_n}}{365}(C_t - \Delta_{C,t_n} S_{t_n}), \tag{22}$$

where C_t is the mid price of the call option at time t, Δ_{C,t_n} represents the delta of the call option at time t_n, S_{t_n} is the underlying stock price at time t_n, r_{t_n} is the annualized risk free rate and a_n is the number of calendar days between hedges t_n and t_{n+1}.

After applying data filtering procedure of Section 4.1, the delta-hedged option gains are calculated. For each of the delta-hedged options, I consider a holding period of one month. That is, the position is initiated at the end of the month and closed at the end of the following month. Conveniently, this results in monthly return observations and avoids any potential issues that could arise with the option settlement when holding on to the position until maturity (i.e. the third Friday of the subsequent month) (Goyal & Saretto, 2009). As an additional filter, I remove any options that have less than 10 observations within this one month period. By summing up the delta-hedged option gains throughout the holding period of an option, the total option gain is calculated. Following Cao and Han (2013), I divide this total gain by $\Delta_t S_t - C_t$ in order to obtain comparable monthly delta-hedged option returns.

Table 1 shows the summary statistics of the option data. The average delta-hedged option return is negative, which is in line with previous findings on individual equity option returns (Carr and Wu (2009) and Cao and Han (2013)). By construction, the average maturity is around 50 days and the moneyness close to 100%. Figure 1 visualizes the delta-hedged option returns, highlighting the negative mean and median return, as well as the bell-shaped distribution of the returns.

Table 1: Summary statistics of the call option data

	Delta-hedged option return (%)	Days to maturity	Moneyness (%)
Mean	-0.49	50	100.33
Median	-0.56	50	100.32
Standard deviation	2.62	2	2.17
10th percentile	-3.13	46	97.44
25th percentile	-1.76	49	98.89
75th percentile	0.65	51	101.82
90th percentile	2.21	52	103.30

Note: The summary statistics in this table are based on options on the constituents of the S&P500 from January 1996 until June 2019. The delta-hedged option returns are monthly returns, where the position is initialized on the last day of the month, delta-hedged on a daily basis and unwound on the last day of the next month. The returns represents the sum of the delta-hedged option gains divided by $\Delta_t S_t - C_t$. Moneyness is defined as the ratio of the stock price over the option strike price.

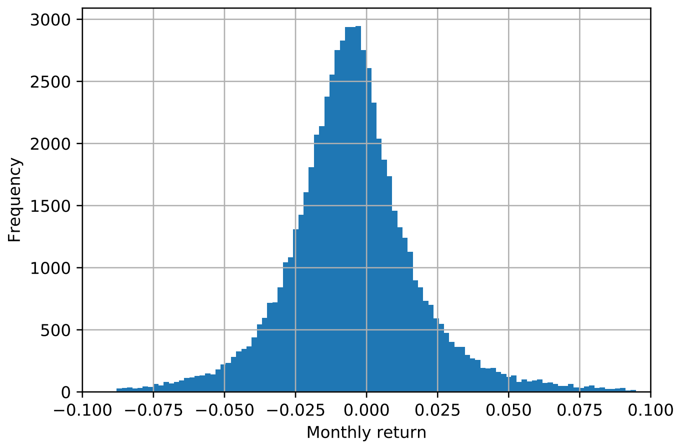


Figure 1: Distribution of delta-hedged call option returns on all S&P500 constituents from January 1996 until June 2019 (excluding outliers[3])

4.3 Characteristics

Capitalizing on existing literature about the cross-section of stock returns and option returns, I consider 95 option and stock characteristics that are used to explain delta-hedged equity option returns. More specifically, I obtain data of the option characteristics from Brooks et al. (2018) and Horenstein et al. (2018), while following the extensive list provided by Green et al. (2017) for stock characteristics.

[3]The histogram represents the center 99% of the returns. That is, values below the 0.5th percentile and above the 99.5th percentile have been excluded.

Table 2 lists all the characteristics used in the empirical analysis. Appendix Section A provides a more detailed overview, including references to the literature that introduced, and proved the relevance of, the corresponding characteristic. The characteristics data are obtained from Option Metrics, CRSP, Compustat and I/B/E/S and are monthly in nature. To the best of my knowledge, this is one of the most comprehensive sets of characteristics considered in a study on the cross-section of delta-hedged equity option returns.

In order to reduce the models' sensitivity to outliers, each of the characteristics is standardized on a scale from $[-0.5, +0.5]$. That is, for every period, I rank each of the characteristics and divide its rank by the number of non-missing observations. Next, I subtract 0.5 to move the standardized characteristics into the $[-0.5, +0.5]$ interval and balance them around zero. Kelly et al. (2019) note that such a transformation of characteristics does not qualitatively influence the results.

Table 2: Overview of option and stock characteristics

Panel A: Option Characteristics

Acronym	Characteristic	Acronym	Characteristic
VRP	Volatility risk premium	*opt_liquidity*	Option bid-ask spread
opt_demand	Option demand pressure	*opt_volume*	Option trading volume
open_interest	Option open interest		

Panel B: Stock Characteristics

Acronym	Characteristic	Acronym	Characteristic
absacc	Absolute value of *acc*	*mom12m*	12-month momentum
acc	Working capital accruals	*mom1m*	1-month momentum
aeavol	Abnormal volume around earnings announcements	*mom36m*	36-month momentum
age	Number of years of coverage on Compustat	*mom6m*	6-month momentum
agr	Asset growth	*mve*	Market capitalization
baspread	Stock bid-ask spread	*mve_ia*	Industry-adjusted *mve*
beta	Market beta	*nanalyst*	Number of analysts covering the stock
bm	Book-to-market	*nincr*	Number of earnings increases
bm_ia	Industry-adjusted *bm*	*orgcap*	Organization capital

cash	Cash to total assets	*pchcapx_ia*	Industry-adjusted percent change in CapEx
cashdebt	Cash flow to debt	*pchcurrat*	Percent change in *currat*
cashpr	Cash productivity	*pchdepr*	Percent change in *depr*
cfp	Cash flow to price	*pchgm_pchsale*	Percent change in gross margin minus percent change in sales
cfp_ia	Industry-adjusted *cfp*	*pchquick*	Percent change in *quick*
chatoia	Industry-adjusted change in asset turnover	*pchsale_pchinvt*	Percent change in sales minus percent change in inventory
chcsho	Change in shares outstanding	*pchsale_pchrect*	Percent change in sales minus percent change in receivables
chempia	Industry-adjusted change in number of employees	*pchsale_pchxsga*	Percent change in sales minus percent change in SG&A
chfeps	Change in forecasted EPS (earnings per share)	*pchsaleinv*	Percent change in sales to inventory
chinv	Change in inventory	*pctacc*	Percent accruals
chmom	Change in 6-month momentum	*pricedelay*	Explained variation of stock return by lagged market returns
chnanalyst	Change in *nanalyst*	*quick*	Quick ratio
chpmia	Industry-adjusted change in net profit margin	*rating*	S&P debt credit rating
chtx	Change in tax expense	*roeq*	Return on equity
cinvest	Corporate investment	*rd_mve*	R&D to market cap.
currat	Current ratio	*rd_sale*	R&D to sales
depr	Depreciation to PP&E	*roaq*	Return on assets
disp	Forecasted EPS dispersion	*roavol*	Earnings volatility
		roic	Return on invested capital
dolvol	Stock dollar trading volume	*rsup*	Revenue surprise
dy	Dividend yield	*sp*	Sales to price
ear	Earnings announcement return	*salecash*	Sales to cash
egr	Growth in common shareholder equity	*saleinv*	Sales to inventory
ep	Earnings to price	*salerec*	Sales to receivables

fgr5yr	Forecasted 5-year EPS growth rate	*sfe*	Scaled earnings forecasts
gma	Gross profitability	*sgr*	Sales growth
grCAPX	Growth in CapEx		
herf	Industry sales concentration	*sratio*	Systematic volatility to total volatility (*tvol*)
hire	Employee growth rate	*std_dolvol*	Volatility of liquidity (dollar trading volume)
operprof	Operating profitability	*std_turn*	Volatility of liquidity (share turnover)
idiovol	Idiosyncratic stock return volatility	*stdacc*	Volatility of accruals
ill	Illiquidity	*stdcf*	Cash flow volatility
indmom	Industry momentum	*sue*	Unexpected earnings
invest	Capital expenditures and inventory	*tang*	Asset tangibility
lev	Leverage	*tb*	Tax to book income
lgr	Growth in long-term debt	*turn*	Share turnover
maxret	Maximum daily stock return	*tvol*	Total volatility

Note: An extensive overview of the characteristics, including descriptions and data sources, is provided in Appendix Section A.

4.4 Test Assets

Most of the asset pricing literature focuses on explaining the returns of portfolios of individual assets, such as long-short portfolios of stocks. By forming portfolios of individual assets, an initial dimension reduction is performed, averaging out a large part of the idiosyncratic risk. Consequently, portfolios often exhibit lower estimation errors when used as test assets and have proved to be useful in finding common factors due to the reduced idiosyncratic variation. However, Lewellen, Nagel, and Shanken (2010) argue that asset pricing tests on portfolios are often misleading and one should instead test models on individual asset returns.

Due to the embedded dimension reduction of IPCA and CA, these modeling approaches can be applied to both portfolios and individual equity option returns (Kelly et al., 2019). As the two groups of test assets exhibit substantially different behavior, I examine both types in the empirical analysis. Specifically, I follow Kelly et al. (2019) and Gu et al. (2019a) and compare all modeling approaches on both individual delta-hedged option returns and characteristic-managed portfolios.

Each characteristic, as described in Table 2, is used to create a single managed portfolio, thus resulting in 95 managed portfolios. Formally, for a set of L instruments or characteristics, the corresponding L-vector of managed portfolio returns at time $t+1$ is defined as

$$r^p_{t+1} = \frac{Z'_t r_{t+1}}{N_{t+1}}, \tag{23}$$

where Z_t represents the N x L matrix of standardized characteristics at time t, r_{t+1} the vector of individual delta-hedged option returns at time $t+1$, and N_{t+1} is the number of non-missing return observations at time $t+1$.

It is important to highlight that no model re-estimation is required after obtaining fitted values for the individual asset returns. Instead, since the portfolio weights (derived from the standardized characteristics) are known in advance, one can simply apply the portfolio transformation of Equation 23 to the fitted returns of the individual delta-hedged option returns in order to obtain estimates for the managed portfolios.

Another benefit of using managed portfolios is in the comparison with the benchmark models. The latent and observable factor benchmark models from Section 3.3 are unconditional in nature, that is, their parameters are time invariant. While one can opt to model conditional distributions instead, this approach is often cumbersome. Consequently, Cochrane (2009) suggest the usage of characteristic-managed portfolios as an alternative to incorporating conditional information into the models. In this way, one can continue to consider unconditional moments, while exploiting the benefits of the incorporated condition information, without having any further complexity arising from time-varying parameters in the model.

Economically, managed portfolios can be considered as long-short portfolios, each sorted and weighted based on the corresponding characteristic's ranking of assets. This is in line with past research, such as Fama and French (1993) and Horenstein et al. (2018), however, instead of only using the top and bottom quantiles, all assets are included and weighted according to their rank. Besides, by using managed portfolios as test assets, I can incorporate conditional information into the unconditional benchmark models. Consequently, due the conditional nature of the advanced methods described in the methodology, it is interesting to compare the performance of these models against benchmarks when using managed portfolios. Simultaneously, the comparison between IPCA and CA is additionally evaluated using individual equity option returns as test assets.

5 Results

This section investigates the empirical findings resulting from applying the models from Section 3 to the data, as discussed in Section 4. First, I compare the models' performance from a statistical perspective, focusing on Total R^2 and Predictive R^2. Second, I use each model's return predictions to construct portfolios and evaluate their economic performance. Third, Section 5.2 tests the no-arbitrage assumption that is enforced throughout the paper. Fourth, Section 5.3 investigates the most relevant characteristics and relates this back to prior research. Finally, I test the robustness of the models by splitting the test sample on the basis of Section 5.3 results and by splitting randomly.

5.1 Performance Evaluation

5.1.1 Statistical Analysis

Table 3 reports the out-of-sample Total R^2 for each of the models. The Observable Factor Model significantly underperforms the other factor models, showing only negative R^2 numbers. Interestingly, the model's performance in sample is substantially better than its out-of-sample performance (e.g. the in sample Total R^2 for individual test assets, r_t, is 36% in the case of the full 6-factor model). This is in line with the research conducted by Horenstein et al. (2018), who find similar in sample R^2 results. Gu et al. (2019a) also report poor findings for their observable factor model, which is based on Fama-French (Fama & French, 2015). Besides the fact that it concerns out-of-sample results, they also attribute the model's poor performance to its static beta, which is proving to be a significant shortcoming compared to the characteristics-driven models.

As outlined in Section 4.4, I test each of the model using two types of test assets: individual returns (r_t) and managed portfolios (r_t^p). The results derived from managed portfolios are generally expected to be better, due to its reduced dimensionality. This phenomenon is clearly highlighted in the results for the PCA, IPCA and CA models.

PCA offers a decent out-of-sample performance for the single factor variation, but only improves marginally for the incremental multi-factor models. This shows that a single factor is able to explain a relatively large part of the variation in returns and could, therefore, indicate that there is a single characteristic that is driving the returns to a large extent.

However, when comparing the PCA to the conditional models, IPCA and CA, its shortcoming are revealed. The time-varying factor loadings and embedded dimension reduction of IPCA and CA result in a strong outperformance of these advanced models, compared to PCA. This is particularly reflected in the results for multi-factor models, e.g. $K = 4, 5, 6$.

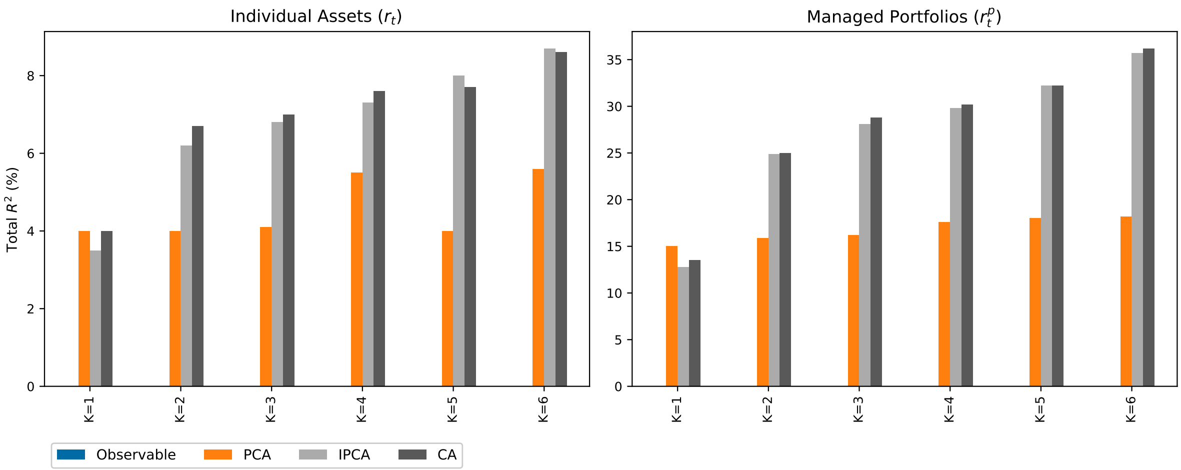
Finally, IPCA and CA perform broadly in line with a maximum Total R^2 of 8.7% and 8.6%, respectively, for the 6-factor model. Although CA does show slightly better results overall, the differences are marginal. Hence, I conclude that CA's non-linearity does not seem to offer superior explanatory power over the linear IPCA model, when considering delta-hedged equity option returns.

Table 4 reports the out-of-sample Predictive R^2. The Predictive R^2 shows how well the model can predict future returns, by solely using past data from returns and characteristics. Therefore, these results should be a good predictor of each model's economic performance, as discussed in Section 5.1.2.

Similar to the Total R^2 results, the Observable Factor Model has a very poor out-of-sample performance and is, thus, not further discussed. PCA's results also show a similar pattern: strong performance for the single factor model, but only marginal improvements for the incremental factor models. A notable difference to the findings in Table 3 is that the PCA, IPCA and CA models perform mostly in line for individual returns as test assets. In the case of managed portfolios, IPCA and CA are still the clear outperformers. Finally, CA does tend to perform better than IPCA in terms of out-of-sample predictability, which is in line with findings from Gu et al. (2019a), albeit with a smaller magnitude.

Table 3: Out-of-sample Total R^2 (%)

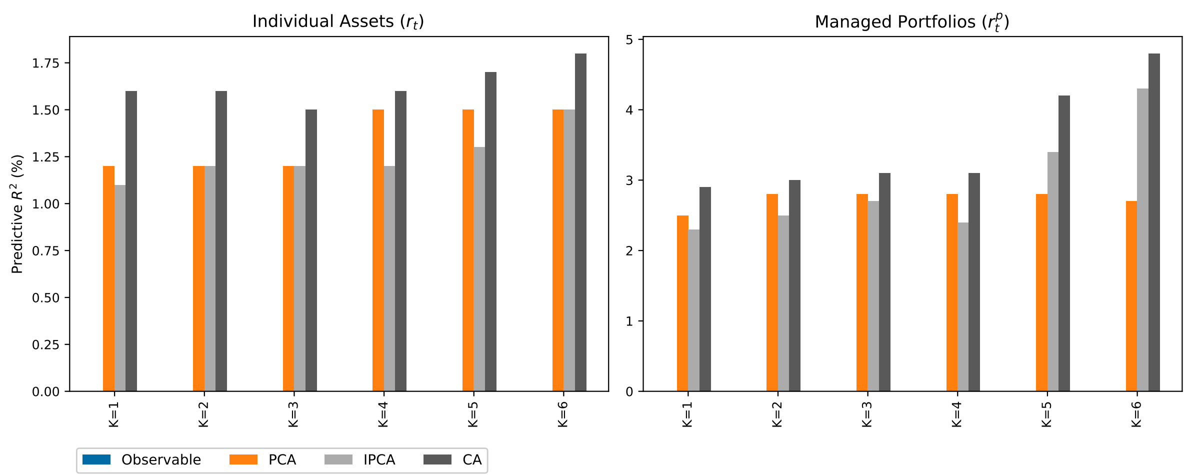
Model	Test Assets	K 1	2	3	4	5	6
Observable Factor Model	r_t	< 0	< 0	< 0	< 0	< 0	< 0
	r_t^p	< 0	< 0	< 0	< 0	< 0	< 0
PCA	r_t	2.9	3.2	3.5	4.3	4.4	5.6
	r_t^p	15.0	15.9	16.2	17.6	18.0	18.3
IPCA	r_t	3.5	6.2	6.8	7.3	8.0	8.7
	r_t^p	12.8	24.9	28.1	29.8	32.2	35.7
CA	r_t	4.0	6.7	7.0	7.6	7.7	8.6
	r_t^p	13.5	25.0	28.8	30.2	32.2	36.2



Note: Observable Factor Model relates to the models described in Section 3.3 and is further elaborated on in Appendix Section B. PCA (Principal Component Analysis) is also described in Section 3.3. An extensive description of IPCA (Instrumented Principal Components Analysis) and CA (Conditional Autoencoder) is provided in Section 3.1 and Section 3.2, respectively. The following hyperparameters are used for CA: learning rate = 0.001, batch size = 1000 and LASSO penalty = 0.001. Two groups of test assets are specified: r_t (individual assets) and r_t^p (managed portfolios). K represents the number of factors in the respective models.

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

Model	Test Assets	K: 1	2	3	4	5	6
Observable Factor Model	r_t	< 0	< 0	< 0	< 0	< 0	< 0
	r_t^p	< 0	< 0	< 0	< 0	< 0	< 0
PCA	r_t	1.21	1.21	1.19	1.54	1.53	1.52
	r_t^p	2.53	2.83	2.76	2.77	2.78	2.68
IPCA	r_t	1.07	1.17	1.16	1.17	1.33	1.50
	r_t^p	2.26	2.53	2.66	2.44	3.39	4.26
CA	r_t	1.62	1.58	1.51	1.58	1.66	1.76
	r_t^p	2.86	2.96	3.14	3.10	4.21	4.84



Note: Observable Factor Model relates to the models described in Section 3.3 and is further elaborated on in Appendix Section B. PCA (Principal Component Analysis) is also described in Section 3.3. An extensive description of IPCA (Instrumented Principal Components Analysis) and CA (Conditional Autoencoder) is provided in Section 3.1 and Section 3.2, respectively. The following hyperparameters are used for CA: learning rate = 0.001, batch size = 1000 and LASSO penalty = 0.001. Two groups of test assets are specified: r_t (individual assets) and r_t^p (managed portfolios). K represents the number of factors in the respective models.

5.1.2 Economic Analysis

Besides comparing the models from a statistical perspective, I evaluate each of the models' economic performance using Sharpe ratios and variances. As discussed in Section 3.4.2, the Sharpe ratios correspond to monthly-rebalanced zero-cost portfolios, constructed by taking a long position in the top decile and a short position in the bottom decile. These deciles are based on the model's predicted returns, linking the economic analysis directly to the Predictive R^2 results in Table 4. One would expect models with a high Predictive R^2 to deliver strong economic performance (i.e. a high Sharpe ratio).

Table 5 shows out-of-sample Sharpe ratios (Panel A) and variances (Panel B) of the aforementioned long-short decile portfolios. It is important to note that the Sharpe ratios of all models are positive, regardless of the number of factors. Even the Observable Factor Model, that failed to produce any positive Predictive R^2, yields positive Share ratios. This is a remarkable result, since the mean option return in the sample is negative, as highlighted in Table 1. Hence, although not always statistically significant, each of the models is able to predict future returns at least to some extent, from an economic perspective.

The difference in economic performance between the conditional models and the unconditional benchmark models is very clear. The Observable Factor Model and PCA perform broadly in line, while IPCA and CA significantly outperform these benchmarks. That is, the Predictive R^2 results in Table 4 are indeed a good indicator of the economic performance, except for PCA, which shows a relatively high Predictive R^2 compared to its Sharpe ratio. Finally, in line with predictability results, CA has the highest Sharpe ratios of up to 2.61 for the 5-factor model.

Panel B of Table 5 presents the portfolio variances, used to calculate the Sharpe ratio. The variance of the unconditional benchmark models is generally smaller, indicating their respective returns are substantially smaller than IPCA and CA, which have higher Sharpe ratios.

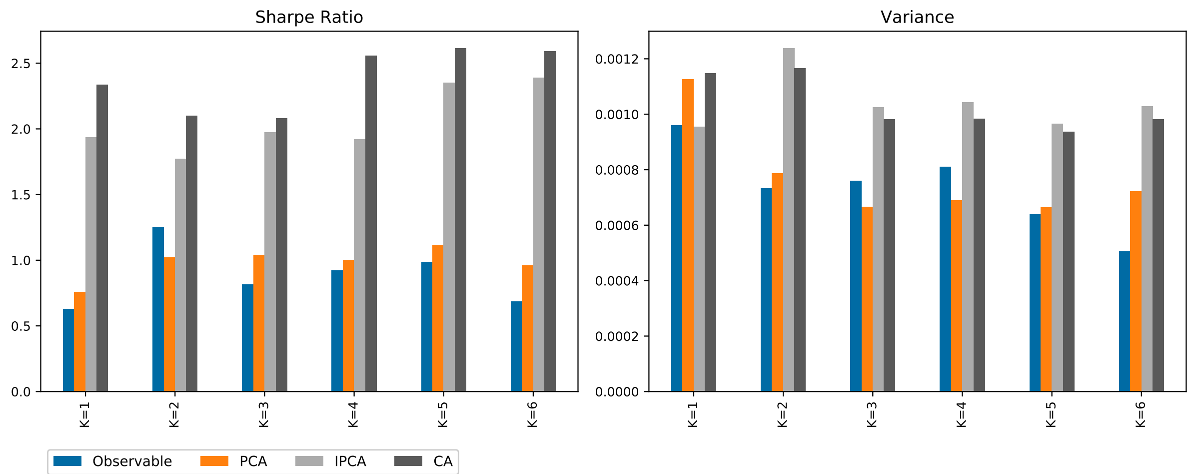
Table 5: Out-of-sample Performance Metrics of Long-Short Portfolios

Panel A: Out-of-sample Sharpe Ratios (annualized)

	K					
Model	1	2	3	4	5	6
Observable Factor Model	0.63	1.25	0.81	0.92	0.99	0.69
PCA	0.76	1.02	1.04	1.00	1.11	0.96
IPCA	1.94	1.77	1.97	1.92	2.35	2.39
CA	2.34	2.10	2.08	2.56	2.61	2.59

Panel B: Out-of-sample Portfolio Variances (annualized)

	K					
Model	1	2	3	4	5	6
Observable Factor Model	0.0010	0.0007	0.0008	0.0008	0.0006	0.0005
PCA	0.0011	0.0008	0.0007	0.0007	0.0007	0.0007
IPCA	0.0010	0.0012	0.0010	0.0010	0.0010	0.0010
CA	0.0011	0.0012	0.0010	0.0010	0.0009	0.0010



Note: Observable Factor Model relates to the models described in Section 3.3 and is further elaborated on in Appendix Section B. PCA (Principal Component Analysis) is also described in Section 3.3. An extensive description of IPCA (Instrumented Principal Components Analysis) and CA (Conditional Autoencoder) is provided in Section 3.1 and Section 3.2, respectively. Two groups of test assets are specified: r_t (individual assets) and r_t^p (managed portfolios). K represents the number of factors in the respective models. The Sharpe ratio, and its respective variance, are derived from long-short decile portfolios, which are constructed using the models' return predictions.

5.2 No-Arbitrage Results

One key assumption made throughout this paper is the no-arbitrage restriction for all the models. That is, the evaluation framework, as described by Equation 18, is specified without an intercept (i.e. $\alpha_{i,t} = 0$ for all i and t). Consequently, the models only consider exposure to risk factors in their performance analysis, and assets without any exposure should not earn any excess return.

Figure 2 depicts the out-of-sample mean pricing errors against its respective mean return. For each of these pricing errors I perform a t-test, in order to test whether the error is significantly different from zero. Following Gu et al. (2019a), I use managed portfolios, r_t^p, as test assets for this analysis, due to its lower dimensionality.

Overall the tests show sufficient evidence in favour of the no-arbitrage restriction. That is, most pricing errors are not significantly different from zero. Pricing errors for PCA tend to deviate the most from zero, with 23 out of 95 t-statistics exceeds 3 in absolute value. IPCA and CA show more robust results with only 9 and 5 t-statistics above 3, respectively. The Observable Factor Model does not have any significant pricing error, but is also known not to offer any meaningful out-of-sample result, as highlighted in Section 5.1.

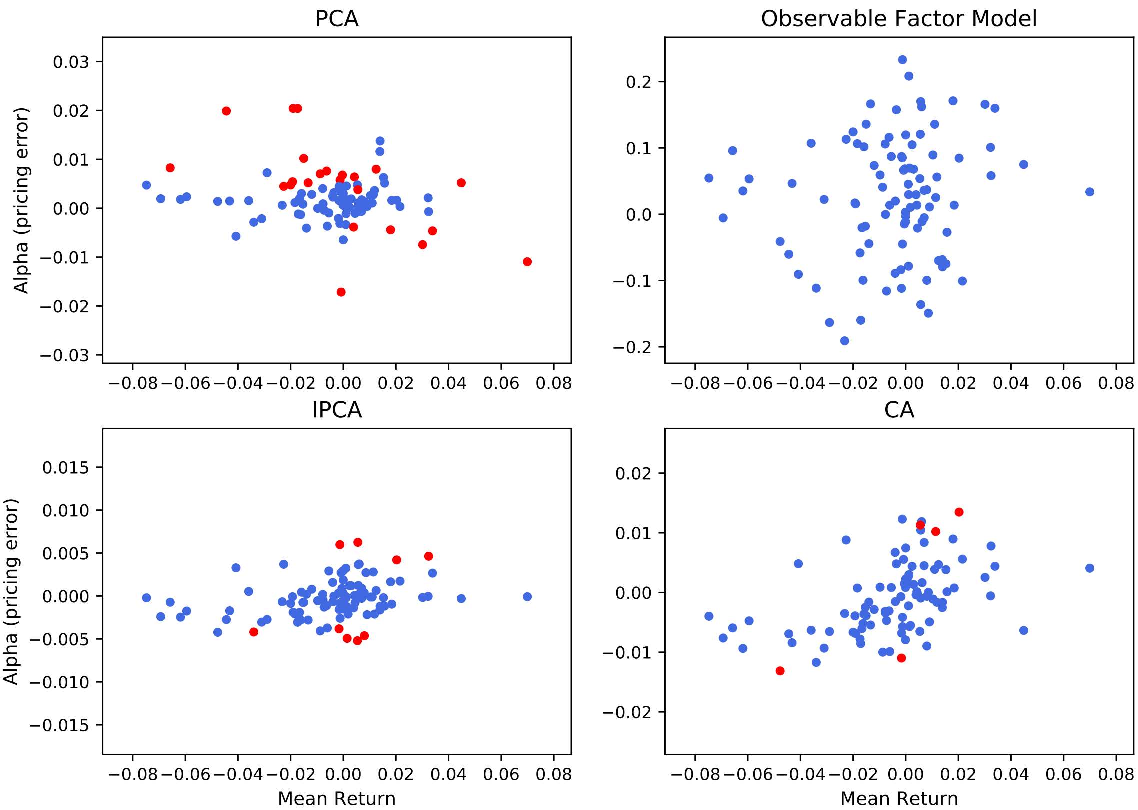
5.3 Relevant Characteristics

Besides the superior performance of the conditional models in this paper, they offer another important advantage: IPCA and CA have the ability to incorporate option and stock characteristics into the model. Consequently, one can analyze and compare the marginal contribution of each characteristic to the model's performance. I follow the approach taken by Gu et al. (2019a) and rank the characteristics based on the reduction in Total R^2 when the respective characteristic is removed from the evaluation.

Figure 3 shows the top 20 characteristics in terms of their marginal contribution, in the case of $K = 5$ factors. The performance of IPCA is very reliant on a single characteristic: the volatility risk premium (*VRP*). Other important characteristic are the option bid-ask spread (*opt_liquidity*) and the number of consecutive earnings increases (*nincr*).

Determining the relevant characteristics of CA is less straightforward. The deviation in marginal contribution among the characteristics is slightly smaller, making it harder to pinpoint the cut-off for a characteristic to be deemed important. The two most relevant characteristics, *VRP* and *opt_liquidity*, are in line with the results from IPCA, which reinforces the finding that they are important in explaining the cross-section of equity option returns.

Figure 2: Out-of-sample Pricing Errors versus Average Returns



Number of Observations	PCA	Observable Factor Model	IPCA	CA
t-statistic > 3 or t-statistic < -3	23	0	9	5
$-3 \leq$ t-statistic ≤ 3	72	95	86	90

Note: Red data points exhibit alphas with absolute t-statistics larger than 3. Blue data points have t-statistics below 3. Pricing errors are calculated as the model residuals: $\hat{\epsilon}_{i,t+1} = r^p_{i,t+1} - \hat{\beta}'_{i,t}\hat{f}_{t+1}$, for all i and t, and averaged over the time dimension.

Moreover, these two option characteristics are well-known in prior literature on equity option returns. Goyal and Saretto (2009) show this for *VRP* (i.e. the difference between historical and implied volatility), and Christoffersen et al. (2018) report significantly positive return spreads for illiquid over liquid equity options. The economic intuition is therefore clear as well. The volatility risk premium is present due to the mean reversion in the volatility of a stock. That is, in the long term the implied volatility of an option should equal its realized volatility. The positive illiquidity premium arises from the compensation required by market makers for the risk of large positions they are holding (Christoffersen et al., 2018). Finally, in line with Horenstein et al. (2018), stock characteristics appear to have limited explanatory power in the cross-section of equity option returns, as the most important characteristics are all option-related.

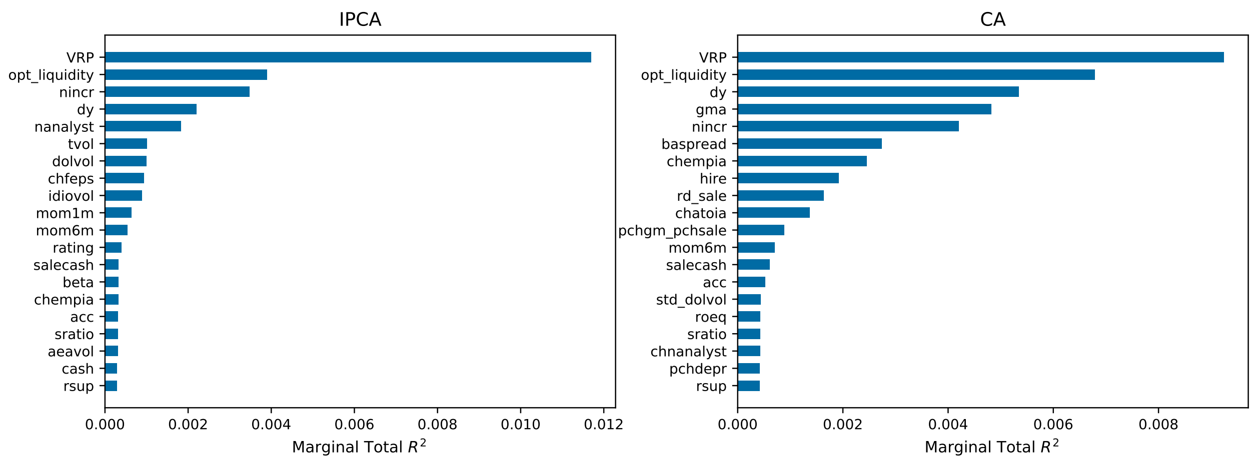


Figure 3: Top 20 Marginal Contribution to Total R^2 ($k = 5$) for IPCA (left) and CA (right).

5.4 Robustness

In this section, I perform two type of robustness tests on the IPCA and CA models. First, I capitalize on the findings from Section 5.3 by testing the models' robustness by splitting the sample based on the important characteristics. Second, I perform a similar test by splitting the sample randomly.

It is important to note that no re-estimation of the models is performed under the split samples. Instead, I keep the parameters from the models in Section 5.1 and only change the test sample in order to recalculate the performance metrics.

5.4.1 Split Samples: Important Characteristics

Table 6 reports robustness results by splitting the test sample based on high and low values for *VRP* and *opt_liquidity*. As mentioned in Section 5.3, *VRP* and *opt_liquidity* are important in explaining the cross section of equity option returns, both from a statistical and economic perspective. Therefore, it is important to evaluate the robustness of the findings when we isolate these characteristics.

The results in Table 6 once again reinforce the importance of *VRP* and *opt_liquidity*. However, it also reveals a potential weakness in the IPCA and CA models. There is a material difference between the Total R^2 and Predictive R^2 when considering test samples with above median versus below median characteristic values. More specifically, the performance of the sample with low *VRP* and low *opt_liquidity* is substantially better than the performance of the samples with high values for these characteristics. While this is observed for both models, it is especially present for IPCA, which appears to be less robust than CA.

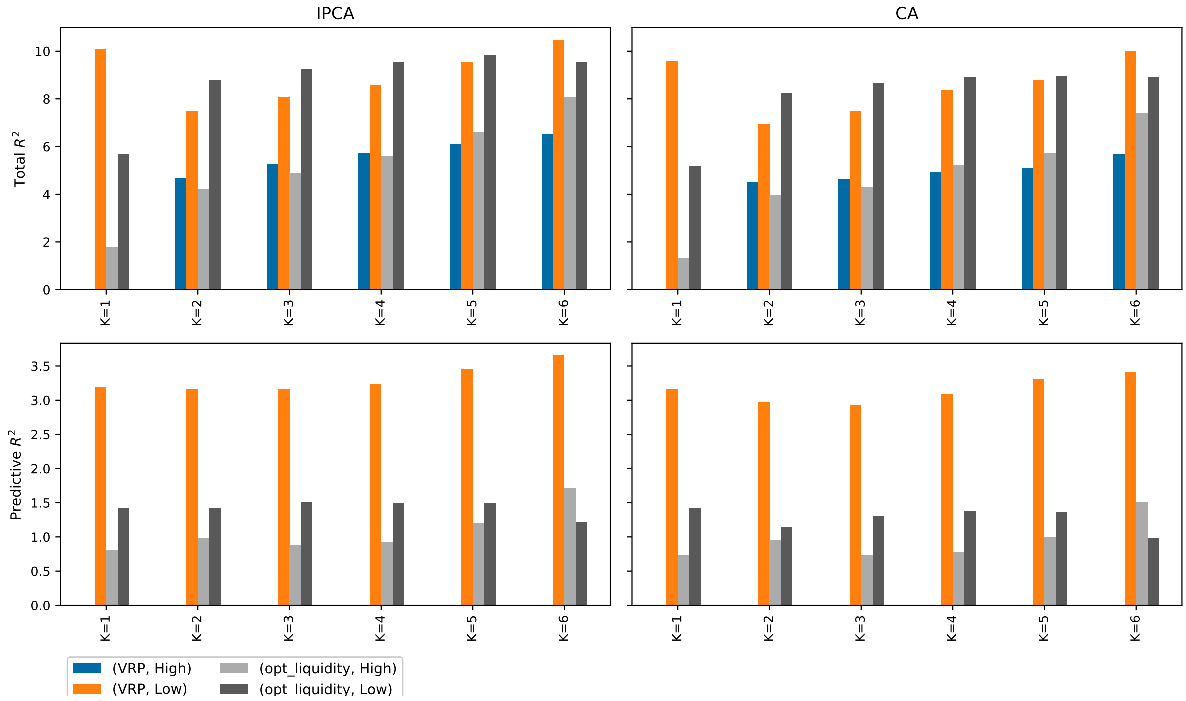
Interestingly, high explanation power seems to be related to expected negative returns of the options. Both a low *VRP* (i.e. historical volatility minus implied volatility is small or negative) and a low *opt_liquidity* (i.e. tight bid-offer option spreads, or liquid options) are on average concerned with low or negative returns, as shown by Goyal and Saretto (2009) and Christoffersen et al. (2018), respectively. Hence, I conclude from this that negative option returns are more pronounced for these characteristics and, thus, that the positive returns of long-short decile portfolios based on these characteristics are mainly driven by the short position as opposed to the long position.

5.4.2 Split Samples: At Random

As a second robustness test, I split the sample randomly into two test sets. In order for the models to be robust, one would expect minimal deviation between the performance of the two samples. Table 7 reports the robustness results under a random split. The performance of IPCA and CA is again not very robust, although to a lesser degree than in Table 6. While the Total R^2 of the two samples are relatively close, the Predictive R^2 tend to deviate, especially for CA. All in all, I conclude that the models are not very robust and, thus, extremely reliant on the training and test data used to evaluate them.

Table 6: Characteristics-Grouped Performance

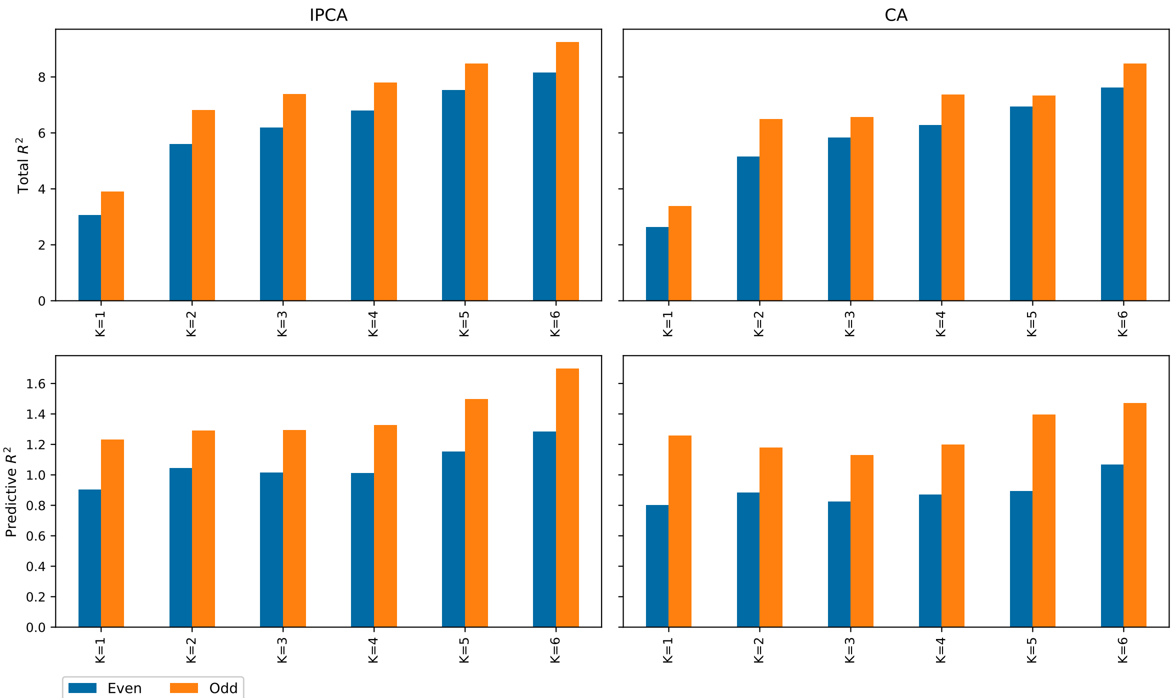
	High						Low					
	K						K					
	1	2	3	4	5	6	1	2	3	4	5	6
Panel A.1: IPCA (VRP)												
Total R^2 (%)	<0	4.66	5.27	5.74	6.11	6.53	10.09	7.49	8.05	8.57	9.55	10.47
Pred. R^2 (%)	<0	<0	<0	<0	<0	<0	3.19	3.16	3.16	3.23	3.45	3.65
Panel A.2: IPCA (opt_liquidity)												
Total R^2 (%)	1.79	4.22	4.90	5.58	6.61	8.06	5.69	8.80	9.26	9.53	9.83	9.55
Pred. R^2 (%)	0.80	0.98	0.88	0.93	1.20	1.72	1.42	1.42	1.51	1.49	1.49	1.22
Panel B.1: CA (VRP)												
Total R^2 (%)	3.56	4.50	4.63	4.92	5.08	5.67	6.43	6.93	7.48	8.39	8.78	9.99
Pred. R^2 (%)	<0	<0	<0	<0	<0	<0	3.16	2.97	2.93	3.08	3.30	3.42
Panel B.2: CA (opt_liquidity)												
Total R^2 (%)	1.34	3.97	4.29	5.21	5.73	7.40	5.17	8.26	8.67	8.93	8.94	8.90
Pred. R^2 (%)	0.17	0.14	0.13	0.17	0.21	0.26	1.42	1.14	1.30	1.38	1.36	0.98



Note: Panel A.1 and Panel B.1 display results for a test sample with above median (high) and below median (low) Volatility Risk Premia (VRP) for IPCA and CA, respectively. Panel A.2 and Panel B.2 display results for a test sample with above median (high) and below median (low) option liquidity (opt_liquidity) for IPCA and CA, respectively.

Table 7: Random Split Performance

	Even						Odd					
	K						K					
	1	2	3	4	5	6	1	2	3	4	5	6
Panel A: IPCA												
Total R^2 (%)	3.06	5.61	6.19	6.81	7.53	8.16	3.91	6.81	7.40	7.79	8.49	9.24
Pred. R^2 (%)	0.90	1.04	1.02	1.01	1.15	1.28	1.23	1.29	1.29	1.33	1.50	1.70
Panel B: CA												
Total R^2 (%)	2.64	5.16	5.83	6.28	6.94	7.62	3.39	6.50	6.57	7.37	7.34	8.48
Pred. R^2 (%)	0.80	0.88	0.82	0.87	0.89	1.07	1.26	1.18	1.13	1.20	1.40	1.47



Note: Panel A and Panel B display results for a test sample split randomly into two sets consisting of the even and odd columns for IPCA and CA, respectively.

6 Conclusion and Further Research

This paper studies advanced unsupervised machine learning methods and their ability to explain the cross-section of delta-hedged equity option returns. I first define the evaluation framework used to analyze the performance of each of the modeling approaches. The evaluation framework assumes the no-arbitrage restriction and is, thus, solely concerned with evaluating the explanatory power derived from exposure to risk factors, as opposed to pure predictability. I consider two methods that include covariates in the empirical analysis: Instrumented Principal Components Analysis and the Conditional Autoencoder. In addition, I compare the results to two benchmark models: an observable factor model and Principal Components Analysis. I evaluate the performance of each model using the Total R^2, Predictive R^2, Sharpe ratio and variance. Besides analyzing the explanatory power of the modeling approaches, I determine the most important characteristics in terms of explaining the cross-section of equity option returns. Finally, I test the robustness of the models by splitting the test samples and recalculating the performance metrics.

The empirical results successfully show that the models that use covariates outperform the benchmark models. For the 5-factor model, the total R^2 of IPCA and CA is 8.0% and 7.7%, respectively, compared to only 4.4% for the benchmark PCA model. The observable factor model fails to provide any out-of-sample explanatory power. The Conditional Autoencoder provides the best overall results, especially from an economic perspective, yielding a Sharpe ratio of 2.61 under predictions from the 5-factor models. My findings show that from the extensive set of characteristics the volatility risk premium and option liquidity are most relevant in explaining the cross-section of returns, for both IPCA and CA. However, this study does not consider transaction costs, which might impact the importance of option liquidity. Finally, the robustness analysis reveals a weakness in the models, as their performance is highly dependent on the sample of option data.

6.1 Further Research

This paper encompasses a broad range of research areas. As such, there are multiple extensions for further research. First, the data could be expanded by including put options, more companies outside of the S&P 500 and options with different maturity or moneyness. This study only considers short-dated at-the-money options, which does not consider important elements in the option market such as the skew and convexity of the volatility surface. Even though the current set of characteristics is very extensive, the set of option-related characteristics could be extended in further research, for instance by including less well understood covariates. Second, more unsupervised machine learning methods could be considered and compared against. In addition, the current analysis could be extended by evaluating multiple variations of the Conditional Autoencoder. In an endeavor to improve the explanatory power, the current models can also be investigated from a pure predictability standpoint, without restricting the intercept. Finally, practical challenges such as transaction costs could be considered in further research, as they might help explain the importance of certain characteristics.

References

Abarbanell, J. S., & Bushee, B. J. (1998). Abnormal returns to a fundamental analysis strategy. *Accounting Review*, 19–45.

Ali, A., Hwang, L.-S., & Trombley, M. A. (2003). Arbitrage risk and the book-to-market anomaly. *Journal of Financial Economics*, *69*(2), 355–373.

Almeida, H., & Campello, M. (2007). Financial constraints, asset tangibility, and corporate investment. *The Review of Financial Studies*, *20*(5), 1429–1460.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of financial markets*, *5*(1), 31–56.

Amihud, Y., & Mendelson, H. (1989). The effects of beta, bid-ask spread, residual risk, and size on stock returns. *The Journal of Finance*, *44*(2), 479–486.

Anderson, C. W., & Garcia-Feijoo, L. (2006). Empirical evidence on capital investment, growth options, and security returns. *The Journal of Finance*, *61*(1), 171–194.

Asness, C. S., Porter, R. B., & Stevens, R. L. (2000). Predicting stock returns using industry-relative firm characteristics. *Available at SSRN 213872*.

Bakshi, G., & Kapadia, N. (2003). Delta-hedged gains and the negative market volatility risk premium. *The Review of Financial Studies*, *16*(2), 527–566.

Balakrishnan, K., Bartov, E., & Faurel, L. (2010). Post loss/profit announcement drift. *Journal of Accounting and Economics*, *50*(1), 20–41.

Baldi, P., & Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, *2*(1), 53–58.

Bali, T. G., Cakici, N., & Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, *99*(2), 427–446.

Bandyopadhyay, S. P., Huang, A. G., & Wirjanto, T. S. (2010). The accrual volatility anomaly.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of financial economics*, *9*(1), 3–18.

Barbee Jr, W. C., Mukherji, S., & Raines, G. A. (1996). Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal*, *52*(2), 56–60.

Barth, M. E., Elliott, J. A., & Finn, M. W. (1999). Market rewards associated with patterns of increasing earnings. *Journal of Accounting Research*, *37*(2), 387–413.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The journal of Finance*, *32*(3), 663–682.

Bauman, W. S., & Dowen, R. (1988). Growth projections and commin stock returns. *Financial*

Analysts Journal, *44*(4), 79.

Belo, F., Lin, X., & Bazdresch, S. (2014). Labor hiring, investment, and stock return predictability in the cross section. *Journal of Political Economy*, *122*(1), 129–177.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The journal of finance*, *43*(2), 507–528.

Bollen, N. P., & Whaley, R. E. (2004). Does net buying pressure affect the shape of implied volatility functions? *The Journal of Finance*, *59*(2), 711–753.

Boyer, B. H., & Vorkink, K. (2014). Stock options as lotteries. *The Journal of Finance*, *69*(4), 1485–1527.

Brandt, M. W., Kishore, R., Santa-Clara, P., & Venkatachalam, M. (2008). Earnings announcements are full of surprises. *SSRN eLibrary*.

Brooks, R., Chance, D. M., & Shafaati, M. (2018). The cross-section of individual equity option returns.

Brown, D. P., & Rowe, B. (2007). The productivity premium in equity returns. *Available at SSRN 993467*.

Cao, J., & Han, B. (2013). Cross section of option returns and idiosyncratic stock volatility. *Journal of Financial Economics*, *108*(1), 231–249.

Carr, P., & Wu, L. (2009). Variance risk premiums. *The Review of Financial Studies*, *22*(3), 1311–1341.

Chamberlain, G., & Rothschild, M. (1982). *Arbitrage, factor structure, and mean-variance analysis on large asset markets* (Tech. Rep.). National Bureau of Economic Research.

Chen, L., & Zhang, L. (2010). A better three-factor model that explains more anomalies. *Journal of Finance*, *65*(2), 563–595.

Chordia, T., Subrahmanyam, A., & Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of financial Economics*, *59*(1), 3–32.

Christoffersen, P., Goyenko, R., Jacobs, K., & Karoui, M. (2018). Illiquidity premia in the equity options market. *The Review of Financial Studies*, *31*(3), 811–851.

Cochrane, J. H. (2009). *Asset pricing: Revised edition.* Princeton university press.

Cooper, M. J., Gulen, H., & Schill, M. J. (2008). Asset growth and the cross-section of stock returns. *the Journal of Finance*, *63*(4), 1609–1651.

Cox, J. C., Ross, S. A., & Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of financial Economics*, *7*(3), 229–263.

Datar, V. T., Naik, N. Y., & Radcliffe, R. (1998). Liquidity and stock returns: An alternative test. *Journal of financial markets*, *1*(2), 203–219.

Desai, H., Rajgopal, S., & Venkatachalam, M. (2004). Value-glamour and accruals mispricing: One anomaly or two? *The Accounting Review*, *79*(2), 355–385.

Diether, K. B., Malloy, C. J., & Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, *57*(5), 2113–2141.

Duan, J.-C., & Wei, J. (2009). Systematic risk and the price structure of individual equity options. *The Review of Financial studies*, *22*(5), 1981–2006.

Eberhart, A. C., Maxwell, W. F., & Siddique, A. R. (2004). An examination of long-term abnormal stock returns and operating performance following r&d increases. *The Journal of Finance*, *59*(2), 623–650.

Eisfeldt, A. L., & Papanikolaou, D. (2013). Organization capital and the cross-section of expected returns. *The Journal of Finance*, *68*(4), 1365–1406.

Elgers, P. T., Lo, M. H., & Pfeiffer Jr, R. J. (2001). Delayed security price adjustments to financial analysts' forecasts of annual earnings. *The Accounting Review*, *76*(4), 613–632.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1–22.

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, *81*(3), 607–636.

Fodor, A., Krieger, K., & Doran, J. S. (2011). Do option open-interest changes foreshadow future equity returns? *Financial markets and portfolio management*, *25*(3), 265.

Francis, J., LaFond, R., Olsson, P. M., & Schipper, K. (2004). Costs of equity and earnings attributes. *The accounting review*, *79*(4), 967–1010.

Gettleman, E., & Marks, J. M. (2006). Acceleration strategies. *SSRN Electronic Journal*.

Goyal, A., & Saretto, A. (2009). Cross-section of option returns and volatility. *Journal of Financial Economics*, *94*(2), 310–326.

Green, J., Hand, J. R., & Zhang, X. F. (2017). The characteristics that provide independent information about average us monthly stock returns. *The Review of Financial Studies*, *30*(12), 4389–4436.

Gu, S., Kelly, B., & Xiu, D. (2019b). Empirical asset pricing via machine learning.

Gu, S., Kelly, B. T., & Xiu, D. (2019a). Autoencoder asset pricing models. *Available at SSRN*.

Guo, R.-J., Lev, B., & Shi, C. (2006). Explaining the short-and long-term ipo anomalies in the us by r&d. *Journal of Business Finance & Accounting*, *33*(3-4), 550–579.

Hafzalla, N., Lundholm, R., & Matthew Van Winkle, E. (2011). Percent accruals. *The Account-*

ing Review, *86*(1), 209–236.

Hawkins, E. H., Chamberlin, S. C., & Daniel, W. E. (1984). Earnings expectations and security prices. *Financial Analysts Journal*, *40*(5), 24–38.

Holthausen, R. W., & Larcker, D. F. (1992). The prediction of stock returns using financial statement information. *Journal of accounting and economics*, *15*(2-3), 373–411.

Horenstein, A. R., Vasquez, A., & Xiao, X. (2018). Common factors in equity option returns. *Available at SSRN 3290363*.

Hou, K., & Moskowitz, T. J. (2005). Market frictions, price delay, and the cross-section of expected returns. *The Review of Financial Studies*, *18*(3), 981–1020.

Hou, K., & Robinson, D. T. (2006). Industry concentration and average stock returns. *The Journal of Finance*, *61*(4), 1927–1956.

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, *28*(3), 650–705.

Huang, A. G. (2009). The cross section of cashflow volatility and expected stock returns. *Journal of Empirical Finance*, *16*(3), 409–429.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of finance*, *45*(3), 881–898.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, *48*(1), 65–91.

Jiang, G., Lee, C. M., & Zhang, Y. (2005). Information uncertainty and expected returns. *Review of Accounting Studies*, *10*(2-3), 185–221.

Jolliffe, I. T. (2002). Mathematical and statistical properties of population principal components. In *Principal component analysis* (pp. 10–28). Springer.

Kama, I. (2009). On the market reaction to revenue and earnings surprises. *Journal of Business Finance & Accounting*, *36*(1-2), 31–50.

Kelly, B. T., Pruitt, S., & Su, Y. (2017). Instrumented principal component analysis. *Available at SSRN 2983919*.

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

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *The journal of finance*, *49*(5), 1541–1578.

Lerman, A., Livnat, J., & Mendenhall, R. (2008). *The high-volume return premium and post-earnings announcement drift. available at ssrn 1122463.*

Lev, B., & Nissim, D. (2004). Taxable income, future earnings, and equity values. *The accounting review*, *79*(4), 1039–1074.

Lewellen, J., Nagel, S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial economics*, *96*(2), 175–194.

Litzenberger, R. H., & Ramaswamy, K. (1982). The effects of dividends on common stock prices tax effects or information effects? *The Journal of Finance*, *37*(2), 429–443.

Moskowitz, T. J., & Grinblatt, M. (1999). Do industries explain momentum? *The Journal of finance*, *54*(4), 1249–1290.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of financial economics*, *108*(1), 1–28.

Ou, J. A., & Penman, S. H. (1989). Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics*, *11*(4), 295–329.

Palazzo, B. (2012). Cash holdings, risk, and expected returns. *Journal of Financial Economics*, *104*(1), 162–185.

Pontiff, J., & Woodgate, A. (2008). Share issuance and cross-sectional returns. *The Journal of Finance*, *63*(2), 921–945.

Rao, R. K., Tang, H., & Chandrashekar, S. (2013). Do corporate cash holdings predictstock returns? *The Journal of Investing*, *22*(2), 29–39.

Rendleman Jr, R. J., Jones, C. P., & Latane, H. A. (1982). Empirical anomalies based on unexpected earnings and the importance of risk adjustments. *Journal of Financial economics*, *10*(3), 269–287.

Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, I. (2005). Accrual reliability, earnings persistence and stock prices. *Journal of accounting and economics*, *39*(3), 437–485.

Rosenberg, B., Reid, K., & Lanstein, R. (1985). Persuasive evidence of market inefficiency. *The Journal of Portfolio Management*, *11*(3), 9–16.

Scherbina, A. (2008). Suppressed negative information and future underperformance. *Review of Finance*, *12*(3), 533–565.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of business*, *39*(1), 119–138.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting review*, 289–315.

Soliman, M. T. (2008). The use of dupont analysis by market participants. *The Accounting*

Review, *83*(3), 823–853.

Thomas, J., & Zhang, F. X. (2011). Tax expense momentum. *Journal of Accounting Research*, *49*(3), 791–821.

Thomas, J. K., & Zhang, H. (2002). Inventory changes and future returns. *Review of Accounting Studies*, *7*(2), 163–187.

Titman, S., Wei, K. J., & Xie, F. (2004). Capital investments and stock returns. *Journal of financial and Quantitative Analysis*, *39*(4), 677–700.

Appendix

A Option and Stock Characteristics

Table 8: Extended overview of option and stock characteristics (grouped by literature source and sorted by ascending publication date)

Panel A: Option Characteristics

Literature Source	Acronym	Characteristic
Bollen and Whaley (2004)	*opt_demand*	Option demand pressure
Goyal and Saretto (2009)	*VRP*	Volatility risk premium
Fodor, Krieger, and Doran (2011)	*open_interest*	Option open interest
Brooks et al. (2018)	*opt_volume*	Option trading volume
Christoffersen et al. (2018)	*opt_liquidity*	Option bid-ask spread

Panel B: Stock Characteristics

Literature Source	Acronym	Characteristic
Fama and MacBeth (1973)	*beta*	Market beta
Basu (1977)	*ep*	Earnings to price
Banz (1981)	*mve*	Market capitalization
Litzenberger and Ramaswamy (1982)	*dy*	Dividend yield
Rendleman Jr, Jones, and Latane (1982)	*sue*	Unexpected earnings
Hawkins, Chamberlin, and Daniel (1984)	*chfeps*	Change in forecasted EPS (earnings per share)
Rosenberg, Reid, and Lanstein (1985)	*bm*	Book-to-market
Bauman and Dowen (1988)	*fgr5yr*	Forecasted 5-year EPS growth rate
Bhandari (1988)	*lev*	Leverage
Amihud and Mendelson (1989)	*baspread*	Stock bid-ask spread
	cashdebt	Cash flow to debt
	currat	Current ratio
	pchcurrat	Percent change in *currat*
	pchquick	Percent change in *quick*
Ou and Penman (1989)		

	pchsaleinv	Percent change in sales to inventory
	quick	Quick ratio
	salecash	Sales to cash
	saleinv	Sales to inventory
	salerec	Sales to receivables
Jegadeesh (1990)	*mom12m*	12-month momentum
Holthausen and Larcker (1992)	*depr*	Depreciation to PP&E
	pchdepr	Percent change in *depr*
Jegadeesh and Titman (1993)	*mom1m*	1-month momentum
	mom36m	36-month momentum
	mom6m	6-month momentum
Lakonishok, Shleifer, and Vishny (1994)	*sgr*	Sales growth
Barbee Jr, Mukherji, and Raines (1996)	*sp*	Sales to price
Sloan (1996)	*acc*	Working capital accruals
Abarbanell and Bushee (1998)	*pchcapx_ia*	Industry-adjusted percent change in CapEx
	pchgm_pchsale	Percent change in gross margin minus percent change in sales
	pchsale_pchinvt	Percent change in sales minus percent change in inventory
	pchsale_pchrect	Percent change in sales minus percent change in receivables
	pchsale_pchxsga	Percent change in sales minus percent change in SG&A
Datar, Naik, and Radcliffe (1998)	*turn*	Share turnover
Moskowitz and Grinblatt (1999)	*indmom*	Industry momentum
Barth, Elliott, and Finn (1999)	*nincr*	Number of earnings increases
Asness, Porter, and Stevens (2000)	*bm_ia*	Industry-adjusted *bm*
	cfp_ia	Industry-adjusted *cfp*
	chempia	Industry-adjusted change in number of employees
	mve_ia	Industry-adjusted *mve*
Chordia, Subrahmanyam, and Anshuman (2001)	*dolvol*	Stock dollar trading volume

	std_dolvol	Volatility of liquidity (dollar trading volume)
	std_turn	Volatility of liquidity (share turnover)
Elgers, Lo, and Pfeiffer Jr (2001)	*nanalyst*	Number of analysts covering the stock
	sfe	Scaled earnings forecasts
Amihud (2002)	*ill*	Illiquidity
Diether, Malloy, and Scherbina (2002)	*disp*	Forecasted EPS dispersion
J. K. Thomas and Zhang (2002)	*chinv*	Change in inventory
Ali, Hwang, and Trombley (2003)	*idiovol*	Idiosyncratic stock return volatility
Desai, Rajgopal, and Venkatachalam (2004)	*cfp*	Cash flow to price
Eberhart, Maxwell, and Siddique (2004)	*rating*	S&P debt credit rating
Francis, LaFond, Olsson, and Schipper (2004)	*roavol*	Earnings volatility
Titman, Wei, and Xie (2004)	*cinvest*	Corporate investment
Lev and Nissim (2004)	*tb*	Tax to book income
Hou and Moskowitz (2005)	*pricedelay*	Explained variation of stock return by lagged market returns
Jiang, Lee, and Zhang (2005)	*age*	Number of years of coverage on Compustat
Richardson, Sloan, Soliman, and Tuna (2005)	*egr*	Growth in common shareholder equity
	lgr	Growth in long-term debt
Anderson and Garcia-Feijoo (2006)	*grCAPX*	Growth in CapEx
Gettleman and Marks (2006)	*chmom*	Change in 6-month momentum
Guo, Lev, and Shi (2006)	*rd_mve*	R&D to market cap.
	rd_sale	R&D to sales
Hou and Robinson (2006)	*herf*	Industry sales concentration
Almeida and Campello (2007)	*tang*	Asset tangibility
Brown and Rowe (2007)	*roic*	Return on invested capital

Brandt, Kishore, Santa-Clara, and Venkatachalam (2008)	*ear*	Earnings announcement return
Cooper, Gulen, and Schill (2008)	*agr*	Asset growth
Lerman, Livnat, and Mendenhall (2008)	*aeavol*	Abnormal volume around earnings announcements
Pontiff and Woodgate (2008)	*chcsho*	Change in shares outstanding
Soliman (2008)	*chatoia*	Industry-adjusted change in asset turnover
	chpmia	Industry-adjusted change in net profit margin
Scherbina (2008)	*chnanalyst*	Change in *nanalyst*
Duan and Wei (2009)	*sratio*	Systematic volatility to total volatility (*tvol*)
Huang (2009)	*stdcf*	Cash flow volatility
Kama (2009)	*rsup*	Revenue surprise
Balakrishnan, Bartov, and Faurel (2010)	*roaq*	Return on assets
Bandyopadhyay, Huang, and Wirjanto (2010)	*absacc* *stdacc*	Absolute value of *acc* Volatility of accruals
Chen and Zhang (2010)	*invest*	Capital expenditures and inventory
Bali, Cakici, and Whitelaw (2011)	*maxret*	Maximum daily stock return
Hafzalla, Lundholm, and Matthew Van Winkle (2011)	*pctacc*	Percent accruals
J. Thomas and Zhang (2011)	*chtx*	Change in tax expense
Palazzo (2012)	*cash*	Cash to total assets
Eisfeldt and Papanikolaou (2013)	*orgcap*	Organization capital
Novy-Marx (2013)	*gma*	Gross profitability
Rao, Tang, and Chandrashekar (2013)	*cashpr*	Cash productivity
Belo, Lin, and Bazdresch (2014)	*hire*	Employee growth rate
Fama and French (2015)	*operprof*	Operating profitability
Hou, Xue, and Zhang (2015)	*roeq*	Return on equity
Brooks et al. (2018)	*tvol*	Total volatility

B Observable Benchmark Models

Table 9: Observable factor models

# factors	Observable factor model
1	$r_{i,t+1} = \alpha_i + \beta_{m,i} DH_{m,t} + \epsilon_{i,t+1}$
2	$r_{i,t+1} = \alpha_i + \beta_{m,i} DH_{m,t} + \beta_{size,i} LS_{size,t} + \epsilon_{i,t+1}$
3	$r_{i,t+1} = \alpha_i + \beta_{m,i} DH_{m,t} + \beta_{size,i} LS_{size,t} + \beta_{ivol,i} LS_{ivol,t} + \epsilon_{i,t+1}$
4	$r_{i,t+1} = \alpha_i + \beta_{m,i} DH_{m,t} + \beta_{size,i} LS_{size,t} + \beta_{ivol,i} LS_{ivol,t} + \beta_{voldev,i} LS_{voldev,t} + \epsilon_{i,t+1}$
5	$r_{i,t+1} = \alpha_i + \beta_{m,i} DH_{m,t} + \beta_{size,i} LS_{size,t} + \beta_{ivol,i} LS_{ivol,t} + \beta_{voldev,i} LS_{voldev,t} + \beta_{ch,i} LS_{ch,t} + \epsilon_{i,t+1}$
6	$r_{i,t+1} = \alpha_i + \beta_{m,i} DH_{m,t} + \beta_{size,i} LS_{size,t} + \beta_{ivol,i} LS_{ivol,t} + \beta_{voldev,i} LS_{voldev,t} + \beta_{ch,i} LS_{ch,t} + \beta_{disp,i} LS_{disp,t} + \epsilon_{i,t+1}$

Note: The variables in the factor models directly follow the definitions from Horenstein et al. (2018). That is, $DH_{m,t}$ is the delta-hedged return of the S&P 500 Index option at time t. The other variables are represented by long-short decile portfolios based on the respective characteristic. Specifically, $LS_{size,t}$ is the return of the long-short decile portfolio sorted based on market capitalization. $LS_{ivol,t}$ is the long-short portfolio return based on idiosyncratic volatility, defined as the standard deviation of the residuals of the Fama-French 3-factor model over the previous month's stock returns. $LS_{voldev,t}$ is based on the volatility risk premium, or the difference between implied volatility and historical volatility, where historical volatility is defined as the standard deviation of daily stock returns of the previous month. $LS_{ch,t}$ is the return of a portfolio sorted based on their cash-to-assets ratio. Finally, $LS_{disp,t}$ is based on the standard deviation of analyst forecasts divided by the absolute value of the mean forecast.

C Python Code

This section provides short descriptions of the programming files used in the empirical analysis of this study. All of the code is written in Python 3.7.0. In order to run the models, I suggest to use sufficient computing power.

1. Filter Option Data

Input	Raw option data from Option Metrics
Output	Filtered option data
Libraries	Numpy, Pandas
Length	194 lines
Description	Follows the filter procedure from Section 4.1 in order to generate a set of daily option return data.

2. Delta-Hedged Option Returns

Input	Filtered option data
Output	Monthly delta-hedged option returns
Libraries	Numpy, Pandas, Matplotlib
Length	146 lines
Description	Calculates the delta-hedged option return of each option in each month and provides summary statistics.

3. Characteristics

Input	Raw characteristics data from Option Metrics, CRSP, Compustat and I/B/E/S
Output	Combined and processed characteristics data matching the option data
Libraries	Numpy, Pandas, Statsmodels
Length	812 lines
Description	Calculates the delta-hedged option return of each option in each month and provides summary statistics.

4. Models

Input	Monthly option returns and monthly characteristics data
Output	Model (and benchmark) results for statistical and economic analysis
Libraries	Numpy, Pandas, Matplotlib, Scipy, Sklearn, Torch
Length	1048 lines
Description	Obtains all model results: Total R^2, Predictive R^2, Sharpe ratio, variance, important characteristics, no-arbitrage results and robustness tests.