Erasmus University Rotterdam

Thesis MSc Business Analytics and Quantitative Marketing

Capital Structure Arbitrage: A Factor Investing Perspective

Author
Jeroen de Groot (466 544)

University Supervisor
Dr. Andreas Alfons

University Second Assessor
Dr. Gustavo Freire

Robeco Supervisors
Dr. Thibault Lair
Dr. Patrick Houweling

November 16, 2023

Abstract

This paper contains an extensive and robust study investigating capital structure arbitrage (CSA) opportunities with a factor investing portfolio perspective. We evaluate regression, cointegration and Ornstein-Uhlenbeck models with various controls, including Random Forest and Boosting, to investigate possible interactions and non-linearities. On the US bond market for listed companies from 1994 to 2022, we can harvest CSA opportunities in both the credit and equities markets. The arbitrage opportunities are driven by both asset classes, equity and credit, and both asset classes contribute to the convergence of the opportunity. An investor may attractively combine individual models and obtain an ensemble model with a yearly turnover rate of 117% and break-even transaction cost of 1.12%. The CSA factor of the ensemble is significantly correlated with Equity and Credit Momentum and, empirically, does not obtain a significant 7-factor spanning alpha. However, to construct the 7-factor portfolio, an investor would have to short bonds, which is difficult and costly in practice.

Contents

1 Introduction

In recent years, there has been a growing interest in applying factor investing principles to credit markets, specifically corporate bonds. Factor investing entails allocating to quantitative investment strategies that historically have shown higher risk-adjusted returns than the market. In simpler terms, factor investing seeks to capitalize on specific characteristics or attributes of assets that have demonstrated a consistent impact on returns. Well-known factors such as Value (Basu, 1977) focus on investing in undervalued assets, Momentum (Jegadeesh and Titman, 1993) involves capitalizing on the trend-following behaviour observed in asset prices, whilst the observation that smaller companies, in terms of market capitalization, have historically outperformed larger ones forms the basis for Size (Banz, 1981), and Low-Risk (Haugen and Heins, 1972) prioritizes assets with lower volatility. These factors have been extensively studied in equity markets, leading to the development of successful factor investing strategies.

While factor investing in equity markets has gained significant attention, the academic literature exploring factor investing in credit markets is relatively new but rapidly expanding. Researchers and practitioners have recognized the potential for factor-based investment strategies to generate superior risk-adjusted returns in the corporate bond space (Correia et al., 2012; Jostova et al., 2013; Frazzini and Pedersen, 2014; Houweling and Van Zundert, 2017; Haesen et al., 2017). Consequently, the first factor investing strategies for corporate bonds have emerged, offering investors an alternative approach to traditional fixed-income investing.

Factor investing strategies are designed to harvest factors that earn high risk-adjusted returns while minimizing the exposure to unrewarded risks. While some factors are built on bond or issuer information, the inherent relationship between credit default risk and equity risk implies that other markets, such as equity or credit default swaps, also carry information for corporate bond investors. So-called "structural" credit risk models attempt to model this relationship, starting with Merton (1974). At the same time, bond and equity investors may have different opinions about the risk of a company defaulting and price default risk differently. This happens due to various factors like market segmentation, investors' risk tolerance, and information processing (Gebhardt et al., 2005). Capital structure arbitrage (CSA) is a strategy that seeks to exploit those mispricing opportunities arising from differences in the pricing of default risk by bond and equity investors. Investors employing CSA aim to capitalize on these divergences through long or short positions in bonds and equities, benefitting from subsequent price adjustments.

This paper aims to investigate opportunities that arise when stock and bond prices diverge, with a particular focus on the bond leg; the investment strategy has to be profitable without taking positions in the equity market. We are the first to investigate a CSA factor investing strategy in the bond market. A substantial difference with existing literature is that we 1) estimate a bond-equity relationship instead of a CDS-equity relationship, 2) use bond returns, 3) only trade the bond leg, and 4) do monthly portfolio construction based on the relative ranking of the opportunities, while typical CSA strategies do absolute ranking and only buy above a set threshold. The primary goal of this paper is to investigate what model (or combination of models) obtains the highest break-even transaction cost and models mean-reversion in credit returns. To answer this, we investigate:

1. Can capital structure arbitrage opportunities be harvested when focusing on the bond leg only?
2. What are the drivers of the formation and convergence of the arbitrage opportunities, and at what speed do the arbitrage opportunities disappear?
3. How robust are those findings to various modelling options?

To answer the research questions, we introduce models that use return level information, price level information, models based on an Ornstein-Uhlenbeck process (Ornstein and Uhlenbeck, 1930), models that estimate the correlation between debt and equity markets and models that fix the correlation. All models are evaluated on monthly data from 1994 to 2022, covering 29 years. We introduce a novel framework specific to mean reversion in the credit market that, among other evaluation criteria, uses autocorrelation in the signal series as a proxy for mean reversion.

The contribution of this paper is eight-fold. Firstly, we show that investors may harvest CSA opportunities focusing on the bond leg only. Secondly, a CSA factor may also harvest those opportunities in the equities market. Thirdly, we show that both the equity and the credit markets drive arbitrage opportunities. Fourthly, setting a prior to fix the credit-equity correlation improves returns, emphasizing the importance of correlation modelling. Fifthly, removing systematic risk improves the arbitrage modelling, but machine learning does not add value. Sixthly, price-level models show disappointing results. Seventhly, we show why models based on an Ornstein-Uhlenbeck process are valuable. Finally, we show that combining fixed and estimated correlation models in an ensemble outperforms any individual model in gross and net returns and a data-driven ensemble in net returns.

This paper is structured as follows: in Section 2, we discuss the literature; in Section 3, we introduce the data, and in Section 4, all methods and evaluation procedures are explored. We show results in Section 5, and we conclude and discuss the research in Section 6.

2 Literature Review

This section discusses three applicable streams of literature for this research: debt market-, equity market-, and factor investing literature.

Debt market From the 2000s till the Great Financial Crisis, there was a notable surge of interest in Credit Default Swap (CDS) contracts. Consequently, extensive research was conducted on CDS-equity trading and models that effectively capture the relationship between debt and equity risk. Studies by Yu (2006); Bajlum and Larsen (2008); Wojtowicz (2014) examine CSA using the CreditGrades model based on the structural model introduced by Merton (1974) and discussed in Bluhm et al. (2001) and Finkelstein et al. (2002). These studies find that CSA can yield attractive risk-adjusted returns, with Sharpe ratios, a measure for risk-adjusted returns, comparable to industry benchmarks. Several modifications of the structural models, such as CreditGrades, have been proposed in Leland and Toft

(1996); Byström (2006); Ozeki et al. (2011); Ju et al. (2015). All structural models aim to predict default risk and estimate a model implied credit spread (ICS) from the default risk. If the observed credit spread of the CDS is far from the ICS, there should be an arbitrage opportunity. Research shows that some structural models outperform CreditGrades in predicting default risk, but structural models still structurally underpredict default risk (Leland and Toft, 1996; Leland, 2012). Duarte et al. (2007) highlight the positively skewed returns associated with structural CSA and emphasizes the significant capital allocation required for this strategy as opportunities cluster. However, to enhance the predictability of default risk, it becomes imperative to establish a long-run relation between debt and equity. Lovreta and Mladenović (2018) show that such relations between the CDS and equity markets exist in cointegration relations, but only when allowing for structural breaks. They show that cointegrated CDS-ICS pairs yield higher returns with less risk. Kapadia and Pu (2012) further discuss the integration of equity and CDS markets and the implications for structural models. Schaefer and Strebulaev (2008); Friewald et al. (2014) show corporate bond/CDS and equity market integration. However, more recently Choi and Kim (2018) show disintegration between relative risk premia in bond and equity markets, where potential mispricing may break debt-equity market integration, a possible impediment to arbitrage.

There is a growing body of literature exploring the integration of CDS and equity markets using machine learning models (Bali et al., 2020; Chan et al., 2023; Mao et al., 2023), and more general in finance Gu et al. (2020); Kelly and Xiu (2023). Bali et al. (2020) and Mao et al. (2023) show imposing the dependence structure between debt and equity of Merton (1974) improves performance of multiple machine learning models for CDS-equity trading. The vast majority of literature discusses CDS-equity trading and not specifically bond-equity trading. Regarding the performance of structural models for bond-equity trading Eom et al. (2004) show that structural models do not accurately predict bond spreads. Some structural models overpredict, whilst other structural models underpredict. For CSA in the corporate bond market Velthuis (2007) shows that regression models that measure arbitrage opportunities based on historical credit and equity returns outperform the CreditGrades model with a long bond, short equity strategy. This paper extends Velthuis (2007) research by investigating whether we can use default risk information from structural models in regression models. We extend further with controls for systematic risk, and investigate machine learning for the regression models.

Equity Market In a CSA strategy, we aim to find the relation between debt and equity. Researchers have documented techniques in equity markets like statistical arbitrage, pairs trading and cointegration to establish the relationship between two equity instruments that garner considerable attention. This stream of literature may inspire methods that, with adjustment, may be implemented in debt-equity trading. Liew and Wu (2013) introduce the copula approach as an alternative to cointegration (Vidyamurthy, 2004) for pairs trading, subsequent studies by Rad et al. (2016); Krauss (2017) utilizing extensive samples and analyzing many academic papers suggest that cointegration and distance (Gatev et al., 2006) methods tend to generate superior returns. Contrary to Lovreta and Mladenović (2018), Rad et al. (2016) does not only establish a cointegration relation but defines a trading strategy based on the spread between two cointegrated securities. Krauss (2017) emphasizes that cointegration provides a more rigorous framework for pairs trading than the distance approach. Inspired by these results, we construct a cointegration framework in this research, in addition to the extended regression framework.

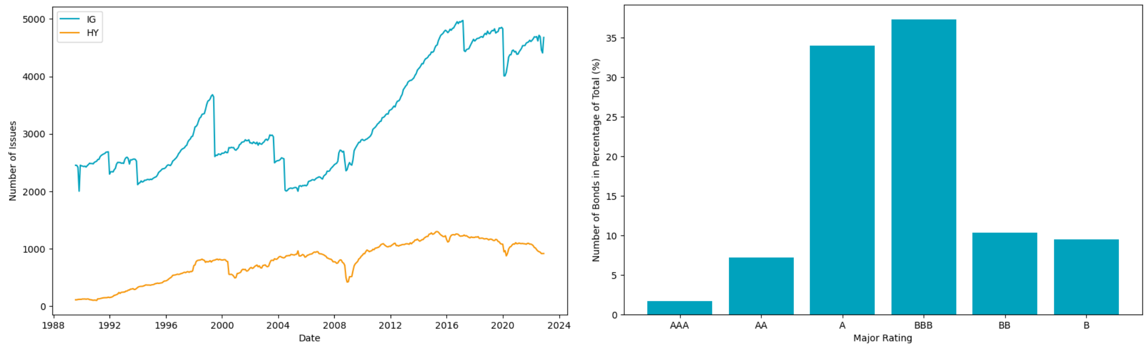
Avellaneda and Lee (2010) introduce an Ornstein-Uhlenbeck process (Ornstein and Uhlenbeck, 1930) for measuring mispricing between ETFs and individual stocks. Guijarro-Ordonez et al. (2021) extend this research toward statistical arbitrage by applying neural networks to frequency-transformed mispricing spreads that are input for the Ornstein-Uhlenbeck process. The papers show promising results. However, statistical arbitrage is based on short-term mean reversion (Pole, 2011). Therefore, methodologies that flourish in statistical arbitrage do not necessarily translate to good performance for the CSA strategy in this research, based on monthly data. We take inspiration from Avellaneda and Lee (2010) and create an Ornstein-Uhlenbeck process for a CSA factor investing strategy.

Factor Investing Factor investing in the corporate bond space is relatively new with the first strategies appearing after 2010 (Correia et al., 2012; Jostova et al., 2013; Frazzini and Pedersen, 2014; Haesen et al., 2017) compared to equities where factors have been reported earlier (Haugen and Heins, 1972; Basu, 1977; Banz, 1981; Jegadeesh and Titman, 1993). Factor investing in corporate bond markets is further discussed in Houweling and Van Zundert (2017) and Bektić et al. (2019) and factor investing shows higher risk-adjusted returns than the market. Researchers have yet to explore CSA in a factor investing strategy. We contribute to the factor investing literature by exploring a CSA factor for the corporate bond market.

3 Data

In this paper, we use the bond database from Robeco with monthly bond, stock, and issuer data (1989-2022). Before filtering, the data set consists of 3,200,834 bond-month observations of 61,267 unique bonds of 5,462 companies. Table 17 in Appendix A shows the filtering steps for the data. We use $1,352,676$ bond-month observations, which is 42.3% of the available sample. The primary decrease in the sample results from the constraint that the bond has to be a constituent of the US Investment Grade (IG) or High Yield (HY) index. Investment Grade bonds are bonds from companies with a low probability of default and are considered the most secure bonds. As credit rating drops, the default risk increases. The US IG/HY filter prevents exchange rate noise in the data from bonds issued in Europe or emerging markets (EM), and removes the influence of differences in taxation that account for approximately one-third of credit spreads (Schaefer and Strebulaev, 2004). Similarly, we remove bonds that do not have matched dollar-denoted equity returns.

We remove bonds rated CCC or lower and distressed bonds because the behaviour of those bonds is substantially different from the other bonds, for example, due to legal procedures concerning possible bankruptcy (Wang, 2011). We implement a final constraint of at least 36 months of observations to increase the speed of the programs as we require at least 36 months of observations for estimation in the rolling window that we apply. Figure 1a shows that there are more IG bonds in the investment universe than HY bonds, and that the available universe grows over time. Figure 1b shows that most bonds are rated A or BBB and that the highest rating is for ∼2% of bonds only.



(a) Number of bonds in the cross-section

(b) Rating Distribution

Figure 1: Growth of the Bond Universe and Distribution of Ratings

This paper uses equity and credit returns to estimate arbitrage opportunities. For equity returns, we use excess equity returns, constructed by subtracting the risk-free rate from equity returns (Haesen et al., 2017). Risk-free rates are the 1-month US treasury bill rates and are retrieved from the Dartmouth US returns library (Kenneth French Data Library)[1]. Following Houweling and Van Zundert (2017); Bessembinder et al. (2008) we employ credit returns, defined as the excess return of a corporate bond versus duration-matched Treasuries. Contrary to equities, bonds have a maturity date. Bonds with longer maturity carry an increased risk for which investors want to be rewarded. The compensation is the term premium. Additionally, there is a default premium, where companies with higher risk levels pay higher interest rates. The term premium can be earned efficiently by investing in government bonds, so the primary purpose of investing in corporate bonds is to earn the default premium. Therefore, we remove the term premium and model only the default premium. The Robeco data set contains major (BBB) and minor (BBB+/BBB/BBB-) ratings. This paper uses major ratings to obtain a larger peer universe when filtering rating, sector and maturity buckets. Otherwise, there are months with peer groups consisting of 2 bonds, and this is not a representative sample of the group. Table 1 shows that rating, sector, and maturity groups show different returns that differ substantially across time periods. Figure 15, 16, and 17 in Appendix A show the different credit return dynamics across groups over time, where one may observe the difference in volatility of the credit returns too. Ambastha et al. (2010) provide evidence for segmentation into HY and IG. More recently, Chen et al. (2014) show that the different market participants in IG and HY lead to different return dynamics. Therefore, we test models separately for the IG and the HY universe, where we prioritize IG as the IG universe is the largest, Figure 1a.

Integration of markets is crucial for profitable arbitrage (Chan et al., 2023). As a proxy for market integration, we use correlation. Table 2 shows a significant cross-correlation between credit and equity returns, evidence that we have information flow from equity to credits. However, only

[1]https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_factors.html

Table 1: Monthly credit returns (%) per rating, sector and maturity group

	Rating						Sector			Maturity		
Time Period	**AAA**	**AA**	**A**	**BBB**	**BB**	**B**	**Financial**	**Industrial**	**Utility**	**Short**	**Medium**	**Long**
Before 2000	0.016	0.020	0.035	0.027	0.187	0.187	0.045	0.053	0.058	0.069	0.063	0.023
Between 2000 and 2010	-0.053	-0.027	0.004	0.037	0.099	0.030	-0.050	0.055	0.052	0.059	0.006	0.009
After 2010	0.076	0.071	0.093	0.150	0.321	0.320	0.180	0.164	0.135	0.148	0.198	0.147

one-fifth of the correlation is significant. The table further shows that credit returns are significantly correlated with past credit and equity returns; we explore including past credit and equity information in this paper. Figure 18 in Appendix A shows that correlation is also time-varying, indicating that information spill-over during specific periods might be lower or more selective. This is challenging for factor investing because portfolios are constructed monthly, removing the possibility of buying more or fewer bonds in specific periods. Therefore, capital structure arbitrage opportunities may be especially challenging to harvest in a factor investing strategy.

Table 2: Auto- and cross-correlation for credit and credit with equity lags

Correlation	**Credit with credit lags**			**Credit with equity lags**		
Lag	**Mean**	**Median**	**Significance**	**Mean**	**Median**	**Significance**
0	1.00	1.00	100%	0.10	0.09	22.1%
1	-0.01	-0.01	22.1%	0.00	0.00	11.1%
2	-0.04	-0.04	13.8%	-0.02	-0.02	9.9%
3	-0.05	-0.06	13.7%	-0.02	-0.03	8.3%
4	-0.05	-0.06	9.4%	-0.01	-0.01	6.7%
5	-0.03	-0.02	8.8%	-0.04	-0.03	8.6%
6	-0.01	-0.01	8.2%	-0.03	-0.03	7.7%
7	-0.01	0.00	6.8%	-0.01	-0.02	5.4%
8	-0.04	-0.04	6.7%	-0.02	-0.03	5.4%
9	0.00	-0.01	6.5%	0.02	0.02	6.4%
10	0.01	0.02	6.9%	-0.04	-0.04	7.1%

4 Methodology

The aim is to find a combination of models that outperforms all other models and captures mean-reversion in credit returns. To evaluate all models, we first introduce portfolio construction and spanning regressions to assess the performance of one model relative to another model in Section 4.1. Secondly, we introduce three categories of models to evaluate in Section 4.2, Section 4.3 and Section 4.4. Thirdly, we introduce evaluation metrics in Section 4.5 to assess whether the models capture CSA opportunities, and lastly, in Section 4.6 we introduce an ensemble methodology to combine the models that capture CSA opportunities.

4.1 Portfolio Construction and Spanning Regressions

We construct equally weighted monthly portfolios from January 1994 to December 2022 because we require the first few years of data for bond-equity relation estimation. We hold the portfolio for one month and up to 12 months using the overlapping portfolio methodology of the influential paper of Jegadeesh and Titman (1993). The holding period of up to 12 months prevents extreme turnover and is realistic in practice (Houweling and Van Zundert, 2017). Portfolios consist of deciles, 10% of the cross-section, in line with Jegadeesh and Titman (1993) for equities and with Jostova et al. (2013); Houweling and Van Zundert (2017) for corporate bonds. Bonds are ranked monthly and sorted in the cross-section based on signals. An example of a signal is all bonds in the cross-section ranked by credit return. The bond with the highest return would have rank 1, and the top decile would contain the top 10% credit returns. Following Blitz (2012); Huij et al. (2014); Houweling and Van Zundert (2017) we construct long-only portfolios instead of long-short portfolios that are common in academic literature because shorting corporate bonds is difficult and costly in practice (Houweling and Van Zundert, 2017). Including the short positions would, therefore, inflate performance to unrealistic levels.

We compute performance statistics to evaluate portfolio return characteristics. Performance metrics for every portfolio are the mean and standard deviation of the annualized credit returns, Sharpe ratio, information ratio (IR) and tracking error. Sharpe on portfolio level is defined as

$$Sharpe = \frac{R_p - R_f}{\sigma_p},$$

where R_p is the portfolio's return, R_f is the risk-free rate and σ_p is the standard deviation of the portfolio credit returns. We define IR as,

$$IR = \frac{R_p - R_m}{\sigma_{pm}},$$

where R_m represents a benchmark to the bond set. In this research paper, the benchmark is the market return, where we define the market as all bonds in the US IG or US HY index. σ_{pm} is the tracking error defined as the standard deviation of the portfolio minus benchmark returns.

We take a three-step approach to infer an incremental improvement to a strategy, the influence of common factors on the strategy, and the significance of the unexplained returns considering factors based on spanning regressions (Fama and French, 1992). First, we regress the observed top-bottom credit returns of strategy A on the observed top-bottom credit returns of the second/previous strategy B to test incremental improvement

$$R_{TB_A,t} = \alpha + \beta R_{TB_B,t} + \varepsilon_t, \tag{1}$$

where α shows whether returns of strategy A are spanned by returns of strategy B, the sign of β the (direction of) correlation with the other strategy, and $1 \leq t \leq T$ (January 1994 to December 2022). Suppose estimated α is significantly positively different from zero. In that case, the returns of strategy A are significantly higher than the returns of strategy B, and the adjustment is a significant improvement for the strategy. When we refer to a spanning alpha or significant alpha in case of an improvement in the strategy, we refer to α, which shows the returns strategy A obtains over the returns of strategy B, common practice in financial literature (Houweling and Van Zundert, 2017). Secondly, we regress observed top-bottom portfolio returns

$$R_{TB,t} = \alpha + \beta_m R_{m,t} + \sum_f \beta_f R_{TB,f,t} + \varepsilon_t,$$

top portfolio returns,

$$R_{D_1,t} = \alpha + \beta_m R_{m,t} + \sum_f \beta_f R_{D_1,f,t} + \varepsilon_t,$$

and bottom portfolio returns,

$$R_{D_{10},t} = \alpha + \beta_m R_{m,t} + \sum_f \beta_f R_{D_{10},f,t} + \varepsilon_t,$$

where the sum over f is over the four well-known factors in the corporate bond market, Size, LowRisk, Value and Credit Momentum (Houweling and Van Zundert, 2017), we supplement the four well-known factors with Equity Momentum, Credit Reversal, and Equity Reversal as we construct a strategy based on data that those factors utilize too. When we observe a significant 7-factor spanning alpha, we obtain significant credit returns unexplained by the market and the four well-known factors plus the supplementary factors. Furthermore, we investigate the correlation of different strategies with the factors. Significant betas show that the CSA strategy shows significant common returns with one or more factors. If this is combined with high R^2, one may in theory approximately obtain the returns of the CSA strategy with an investment in a linear combination of the common plus supplementary factors. Now that we have introduced how we asses possible improvement of various strategies, we introduce all models in the next three sections. The models are split into three sections, as all models use a different 'level' of credit and equity information.

Section 4.2 contains regression models based on credit and equity returns, i.e. 'returns level'. Section 4.3 contains cointegration models that use information from the difference in the bond spread and share price, i.e. 'price level'. Lastly, Section 4.4, discusses Ornstein-Uhlenbeck models estimated on the cumulative residual series of the models from Section 4.2, i.e. 'cumulative returns level'.

4.2 Return Level Models

In this section, we introduce return level models in three subsections. First, we discuss a model-free approach in Section 4.2.1 to set a base model to compare all alternatives to. Secondly, we introduce regression models in Section 4.2.2 to explore 1) past dependencies, 2) raw mispricing versus return prediction, 3) influence of carry, and 4) structural information. Lastly, in Section 4.2.3, we introduce a two-pass approach to account for systematic risk and implement Random Forest and Gradient Boosting Regression Trees to explore non-linear and interaction effects.

4.2.1 Model-Free (Fixed Equity-Credit Correlation)

We consider a panel of credit and equity returns for bonds $i = 1, ..., N$ at time $t = 1, ..., T$ (July 1989 to December 2022). The simplest approach for deriving the relation between equity and credit returns of a company is via a model-free approach, the difference between a Z-score of credit and equity return

$$Z\text{-}score_i^c = \frac{r_{T,i}^c - \mu_i^c}{\sigma_i^c}, \quad Z\text{-}score_i^e = \frac{r_{T,i}^e - \mu_i^e}{\sigma_i^e},$$

$$Z\text{-diff}_i = Z\text{-}score_i^c - Z\text{-}score_i^e$$

where $r^c_{T,i}$ is credit return in month T for bond i, μ^c_i the mean of credit returns, and σ^c_i is the standard deviation of credit returns. Similarly, $r^e_{T,i}$ is equity return in month T for bond i, μ^e_i the mean of equity returns and σ^e_i is the standard deviation of equity returns. However, we require a monthly estimate of the arbitrage opportunity to construct portfolios. Therefore, we compute Z-diff in a rolling window of length L

$$Z\text{-}score^c_{t,i} = \frac{r^c_{t,i} - \mu^c_{t,i}}{\sigma^c_{t,i}}, \quad Z\text{-}score^e_{t,i} = \frac{r^e_{t,i} - \mu^e_{t,i}}{\sigma^e_{t,i}},$$
$$Z\text{-diff}_{t,i} = Z\text{-}score^c_{t,i} - Z\text{-}score^e_{t,i},$$

where the rolling window applies to the mean and standard deviation like $\mu^c_{t,i} = \frac{1}{L}\sum_{k=t-L+1}^{t} r^c_{k,i}$ and $\sigma^c_{t,i} = \sqrt{\frac{1}{L-1}\sum_{k=t-L+1}^{t}(r^c_{k,i} - \mu^c_{t,i})^2}$. t now starts at L, thus $t = L, ..., T$; we set $L = 36$, common in other research (Velthuis, 2007; Choi and Kim, 2018; Bai et al., 2019). For every period t, starting at $L = 36$, we have a cross-section of bonds with a 'score' for Z-diff. Bonds for which Z-diff > 0 credit returns have outperformed equity returns compared to their 3-year historical average. Bonds for which Z-diff < 0 credit returns have underperformed equity returns based on their 3-year historical average. Therefore, in the presence of arbitrageurs, we expect that bonds with the most negative Z-diff 'score' will most likely outperform in the following months. For those bonds, the equity credit dispersion is the largest, based on the historical average, and to revert to the historical average, credit should start outperforming, and equity should underperform. We introduce $Z\text{-diff}_{t,b}$, the Z-diff 'score' of bond b, the first bond in the second decile in ascending order of all bonds in the cross-section at time t. We buy all bonds for which $Z\text{-diff}_{t,i} < Z\text{-diff}_{t,b}$. We buy the bottom 10% of the cross-section.

Before introducing return level regression models, we note that one may express Z-diff in a regression form. To keep the notation simple, we ignore a rolling window and treat the model as if we only estimate it once as

$$r^c_{t,i} = \mu^c_i + \varepsilon^c_{t,i}, \quad r^e_{t,i} = \mu^e_i + \varepsilon^e_{t,i},$$
$$Z\text{-}score^c_i = \frac{r^c_{t,i} - \mu^c_i}{\sigma^c_i}, \quad Z\text{-}score^e_i = \frac{r^e_{t,i} - \mu^e_i}{\sigma^e_i},$$
$$Z\text{-diff}_i = Z\text{-}score^c_i - Z\text{-}score^e_i.$$

With the regression form, we may explain why the Z-diff model is a good benchmark. The model is essentially a return level regression model with correlation and volatility constraints, as shown in

$$\begin{aligned}
r^c_{t,i} &= \mu^c_i + \varepsilon^c_{t,i}, \quad r^e_{t,i} = \mu^e_i + \varepsilon^e_{t,i}, \\
r^c_{t,i} - 1r^e_{t,i} &= \mu^c_i - \mu^e_i + \varepsilon^c_{t,i} - \varepsilon^e_{t,i}, \qquad \text{imposing } \beta = 1 \\
r^c_{t,i} - r^e_{t,i} &= \alpha_i + \varepsilon_{t,i}, \qquad \text{with } \alpha_i = \mu^c_i - \mu^e_i, \quad \varepsilon_{t,i} = \varepsilon^c_{t,i} - \varepsilon^e_{t,i}, \\
Z\text{-}score^{c-e}_i &= \frac{\varepsilon_{t,i}}{\sigma_i} \\
&= \frac{\varepsilon^c_{t,i} - \varepsilon^e_{t,i}}{\sigma_i} \\
&= \frac{\varepsilon^c_{t,i}}{\sigma_i} - \frac{\varepsilon^e_{t,i}}{\sigma_i} \\
&= \frac{\varepsilon^c_{t,i}}{\sigma^c_i} - \frac{\varepsilon^e_{t,i}}{\sigma^e_i}, \qquad \text{if and only if } \sigma^c_i = \sigma^e_i
\end{aligned}$$

$$= Z\text{-}score_i^c - Z\text{-}score_i^e$$
$$= Z\text{-diff}_i.$$

The Z-diff model is equal to a regression of credit on equity returns if and only if the correlation between equity and credit is equal to 1 and the volatility of equity and credit is equal. Therefore, we define Z-diff in the category of fixed equity-credit correlation models.

4.2.2 Regression (Estimated Equity-Credit Correlation)

The next set of models is based on regressions, and the correlation between equity and credit returns is estimated. The simplest regression model is the following, where we regress credit on equity returns of a firm with a rolling window of length L

$$r^c_{(t-L):t,i} = \alpha_{t,i} + \beta_{t,i} r^e_{(t-L):t,i} + \beta_{x,t,i} X_{(t-L):t,i} + \varepsilon_{(t-L):t,i}, \tag{2}$$

where $X_{(t-L):t,i}$ can be any control variable. Without any controls, we refer to this model as the Returns Model. We set $L = 36$. We obtain a set of parameter estimates for every rolling window, resulting in a vector of parameter estimates of length $T - L$ starting at observation L for every bond i. To construct portfolios, we require a monthly arbitrage estimate. For every window starting at $t = L$ we have a vector of in-sample error terms $\varepsilon_{(t-L):t,i}$. To introduce the arbitrage estimate, we denote $\varepsilon_{t,l,i}$ as the in-sample error terms of the rolling window at time t, where $1 \leq l \leq L$. Therefore, we denote the last in-sample error at time t as $\varepsilon_{t,L,i}$. Following Velthuis (2007), we estimate the arbitrage opportunity at time t with a Z-score of $\varepsilon_{t,L,i}$ as

$$Z\text{-}score_{t,i} = \frac{\varepsilon_{t,L,i} - \mu_{t,i}}{\sigma_{t,i}} = \frac{\varepsilon_{t,L,i}}{\sigma_{t,i}}, \tag{3}$$

where $\mu_{t,i}$ is mean of the error process and $\sigma_{t,i}$ is the standard deviation of the error process. We do not cap the Z-scores because the Financial Crisis and COVID-19 result in many capped signals; they have the same value, and therefore, the inability to create decile portfolios, and we use equally weighted portfolios; thus, the magnitude of mispricing does not matter, only the order. A negative Z-score signals underperformance of the credit returns relative to equity returns. We expect the credit returns to catch up with the equity returns. Therefore, we buy sufficiently negative Z-scores

$$Z\text{-}score_{t,i} < Z\text{-}score_{t,b},$$

where Z-score$_{t,b}$ is the Z-score of bond b for which 10% of the Z-scores in the cross-section at time t, are lower. We extend upon this regression model to investigate 1) past dependencies, 2) raw mispricing versus return prediction, 3) influence of carry, and 4) structural model information.

Past Dependencies Kwan (1996); Gebhardt et al. (2005); Hilscher et al. (2015) find that information flows from equity to bond markets, and from Table 2 in Section 3, we infer this is the case for the data in this research as well. Therefore, lagged equity information might provide information for credit returns, and we extend the Returns Model with lagged equity returns

$$r^c_{(t-L):t,i} = \alpha_i + \beta_1 r^e_{(t-L):t,i} + \beta_2 r^e_{(t-1-L):t-1,i} + \beta_3 r^e_{(t-L):t-2,i} + \varepsilon^{PE}_{(t-L):t,i},$$

where $\varepsilon^{PE}_{(t-L):t,i}$ are the in-sample error terms from the model denoted with PE (past equity) to denote the in-sample errors are different from the Returns Model in-sample error terms. From Table 2, we infer that credit returns also contain significant autocorrelation. We extend the regression with lagged credit returns

$$r^c_{(t-L):t,i} = \alpha_i + \beta_{1,t} r^e_{(t-L):t,i} + \beta_{2,t} r^e_{(t-1-L):t-1,i} + \beta_{3,t} r^e_{(t-2-L):t-2,i} \\ + \beta_{4,t} r^c_{(t-1-L):t-1,i} + \beta_{5,t} r^c_{(t-2-L):t-2,i} + \varepsilon^{PD}_{(t-L):t,i},$$

where $\varepsilon^{PD}_{(t-L):t,i}$ are the in-sample error terms from the model denoted with PD (past dependencies) to emphasize that the error process is not equal to the Returns Model error process. By including the lags, we test whether the performance of the signal, accumulation of ε^{PD}, improves by including the effect of past dependencies. If we do not obtain significantly higher alpha over the entire sample period, the arbitrage opportunity is not better captured by including past dependencies. Furthermore, we test the similarity between $\varepsilon_{(t-L):t,i}$ and $\varepsilon^{PD}_{(t-L):t,i}$, $r^c_{(t-1-L):t-1,i}$, $r^c_{(t-2-L):t-2,i}$, $r^e_{(t-1-L):t-1,i}$, and $r^e_{(t-2-L):t-2,i}$ by assessing the rank correlation with Spearman's rank correlation coefficient Spearman (1961). We check the correlation of the ranked bonds in the cross-section between two signals monthly to asses whether the two signals rank the portfolio differently. Rank correlation further allows to assess the effect of incremental contribution on the signal. We can infer whether the ranked cross-section of $\varepsilon_{(t-L):t,i}$ is correlated with $\varepsilon^{PD}_{(t-L):t,i}$ but uncorrelated with the lagged equity and credit returns. If that is the case, we can infer whether lagged equity and credit returns drive the ranking in the cross-section. Exposure of credit returns to lagged credit and equity returns sheds light on omitted variable bias and identifies opportunities for more accurate modelling of CSA opportunities.

Raw Mispricing versus Return Prediction Raw mispricing refers to the estimation in the cross-section of returns and bond characteristics using all information available until time t, but estimating the relation between credit and equity contemporaneously, like the introduced models. Another approach is to predict credit returns for $t + 1$ and buy bonds for which the predicted credit return is highest. We test whether past errors contain information that can predict future returns. If predicting returns obtains significant alpha from a spanning regression, Equation (1), over a raw estimate of the mispricing approach, it is deemed more informative for arbitrage modelling. We estimate the return prediction model referred to as Predict Returns 1 (P1) as

$$r^c_{(t-L):t,i} = \alpha_{t,i} + \beta_{t,i}\varepsilon_{(t-1-L):t-1,i} + \eta^{P1}_{(t-L):t,i}, \tag{4}$$

where $\eta^{P1}_{(t-L):t,i}$ are the in-sample error terms. We investigate time series dependence for return prediction just as we did for raw mispricing estimates. To test whether including a time series dependence structure on past credit and equity returns adds value to forecast credit returns, we extend Equation (4) with ε_{t-2} and ε_{t-3} lags, and investigate spanning alpha.

Influence of Carry Before delving into the concept of carry, it is essential to introduce the concept of the yield curve. The yield curve serves as a vital tool in financial markets, providing insights into the relationship between interest rates and the maturities of bonds. The yield curve is typically shown with maturity on the x-axis and interest rate on the y-axis. An upward-sloping curve suggests anticipation of economic expansion, while an inverted or flat curve may signal economic concerns, potentially indicating an expectation of a recession (Campbell, 1995).

Consider carry as an asset's future returns, assuming prices stay the same (Koijen et al., 2018). For bonds specifically, that means that carry is the expected return when the yield curve does not change (Martens et al., 2019). In practice, the yield curve does change. By examining carry removal, we may isolate factors beyond expected returns that enhance our understanding of the underlying return drivers in an arbitrage setting. The first order Taylor expansion of credit returns shows credit returns are driven by carry and by duration times the change in spread

$$r^c_{t,i} \approx \frac{S_{t-1,i}}{10,000Y} - D_{t-1,i}\Delta\left(\frac{S_{t,i}}{10000}\right)$$
$$r^c_{t,i} = \frac{S_{t-1,i}}{10,000Y} - D_{t-1,i}\Delta\left(\frac{S_{t,i}}{10000}\right) + e_{t,i},$$

where $Y = 12$ to rescale spread on a 'yearly scale' to a 'monthly scale' for data in this research specifically. The first term is the carry, D_{t-1} is the last period's duration (a measure of interest rate sensitivity), and ΔS_t is the expected spread difference to the last period; by including e_t we equal both sides as the Taylor expansion is an approximation. Spread difference and duration are observed variables. Carry is the return an investor earns on the bond if the spreads do not change during the month. In this paper, we focus on arbitrage, for which mean reversion is an underlying assumption. Therefore, we are interested in modelling spread change. We investigate whether removing the carry component improves the arbitrage strategy. We introduce three 'carry free' models to extrapolate what factor beyond carry affects the arbitrage strategy. We again denote the in-sample error terms with a superscript to emphasize the difference from the error process of the Returns Model. To focus on spread change, we estimate a first model that removes the carry from the Taylor expansion

$$r^c_{(t-L):t,i} - \frac{S_{(t-1-L):t-1,i}}{10,000Y} = \alpha_{t,i} + \beta_{t,i} r^e_{(t-L):t,i} + \varepsilon^{MC}_{(t-L):t,i}.$$

This approach of removing carry still contains e_t. Therefore, to investigate whether we can improve further, we estimate the second model

$$D_{(t-1-L):t-1,i}\Delta\left(\frac{S_{(t-L):t,i}}{10000}\right) = \alpha_{t,i} + \beta_{t,i} r^e_{(t-L):t,i} + \varepsilon^{DSD}_{(t-L):t,i},$$

where we focus directly on the spread change times duration. Lastly, to complete the picture of the carry dynamics that influence the arbitrage dynamics, we regress only the change in spread on equity returns

$$\Delta\left(\frac{S_{(t-L):t,i}}{10000}\right) = \alpha_{t,i} + \beta_{t,i} r^e_{(t-L):t,i} + \varepsilon^{SD}_{(t-L):t,i}.$$

If duration drives the returns of the bond, including it in the regression may lead to maturity bias as duration effects are stronger for higher maturity bonds. We assess incremental improvements with spanning regressions from Section 4.1.

Structural Model Information Researchers base structural models on the theory of Merton (1974). The key idea presented in Merton (1974) is modelling corporate debt as an option on the value of the underlying firm's assets. Merton's framework builds on the Black-Scholes option pricing model (Black and Scholes, 1973). By viewing debt as an option, Merton provides insights into the relationship between a firm's leverage (debt-to-equity ratio), its underlying asset value, and the riskiness of the debt. An example of a structural model is the CreditGrades model, developed to profit from mispricing between a company's debt and equity.

CreditGrades is a practical extension of Merton's model (Finkelstein et al., 2002). Assumed is that the asset value follows a geometric Brownian motion. The recovery rate, a parameter that reflects the uncertainty around the exact value of the default barrier, follows a log-normal process. The default barrier level, therefore, varies over time; this results in more realistic short-term default probabilities. An improvement over Merton's model that underestimates default probabilities (Leland, 2002). A drawback of CreditGrades is that not all input variables are directly observable. Byström (2006) shows that including three assumptions improves the transparency of the calculation and obtains similar Distance-to-Default (DtD) results. From the structural model proposed by Byström (2006), we compute model implied credit spreads (ICS) based on Distance-to-Default

$$DtD_{t,i} = \frac{ln(Lev_{t,i})}{(Lev_{t,i} - 1) * \sigma^e_{t,i}},$$

where σ^e is equity volatility, we use 260-day historical volatility, and Lev is the leverage, computed as the value of liabilities over the value of liabilities plus the value of equity. Liabilities contain all liabilities of the company. For the computation of the probability of default, we accumulate time annual probabilities as

$$PD^M_{t,i} = 1 - (1 - PD^1_{t,i})^M,$$

where M is the time to maturity and PD^1 the annualized default probability from Byström (2006) as

$$PD^1_{t,i} = N(-DtD_{t,i}).$$

Following Byström (2006), we cap annualized default probabilities at a level of 20%, and we cap $< 0.5\%$ of all annualized default probabilities as a result. To relate the model implied credit spread to default probabilities, we may use

$$s_{t,i} = -\frac{ln(1-\lambda) * \exp -PD^M_{t,i} * M + \lambda}{M}, \tag{5}$$

from Manning (2004) where λ is the recovery rate on default, usually set at 0.4. A 'simpler' alternative is

$$s^{simple}_{t,i} = (1-\lambda) * PD^M_{t,i}, \tag{6}$$

where one states that spread is the risk premium an investor wants based on loss in case of default times the probability of default. Reverse engineering the recovery rate from the estimated model implied spreads shows that the model implied spread calculation from Equation (6) obtains recovery rates closer to 0.4 and is, therefore, the model of choice.

This paper investigates whether information from structural models can improve arbitrage modelling for the regression models. The first step to assess the potential is to identify whether the error of the Returns Model captures movement in the Distance-to-Default variable; we regress

$$\varepsilon_{t,i} = \alpha_i + \beta_{1,i}(DtD_{t,i} - DtD_{t-1,i}) + \varrho_{t,i} = \alpha_i + \beta_{1,i}\Delta DtD_{t,i} + \varrho_{t,i}.$$

The regression captures the exposure of the Returns Model error term to DtD. If present, we may improve the arbitrage modelling by controlling for DtD exposure, discussed in the next section.

4.2.3 Two-Pass Regression (Estimated Equity-Credit Correlation)

This section discusses several approaches for cleaning the credit and equity returns for arbitrage modelling. We aim to remove systematic risk. In the first paragraph, we motivate using controls for systematic risk. The second paragraph introduces linear specifications to control for systematic risk. In the third and last paragraph, we introduce Gradient Boosting and Random Forest Regression Trees to include interaction and non-linear effects.

Systematic Risk and Idiosyncratic Returns Lower-rated bonds and bonds with a longer maturity inherently bear more risk than higher-rated short-dated bonds; we call this systematic risk (Weinstein, 1981). We can correct for systematic risk by controlling for maturity and rating effects. Moreover, different sectors of the economy can exhibit varying risk and return characteristics due to the nature of their businesses, financial structures, and macroeconomic factors that affect them (Leary and Roberts, 2014). Controlling for the company's sector that issues the bond can control for this and further reduce systematic risk in the credit returns. Since models in Section 4.2.2 use credit returns, we expect the errors that we model the arbitrage opportunity from to be partially explained by maturity, sector and rating effects (Fama and French, 1989). Table 1 in Section 3 further shows that returns differ across rating, maturity and sector buckets. In the next paragraph, we introduce methods to capture systematic risk in a first-pass regression and model time-series arbitrage dynamics in a second pass on individual bond level, like the previously introduced models.

Linear Models The first approach to clean returns of systematic returns is to clean credit, and equity returns monthly by considering abnormal returns. Abnormal returns are the return of the bond minus the returns of the peer group of that bond. We construct the peer group on maturity, sector and rating buckets and compute abnormal credit and equity returns as

$$AR^c_{t,i} = r^c_{t,i} - r^c_{t,peers}, \quad AR^e_{t,i} = r^e_{t,i} - r^e_{t,peers},$$

where $r_{t,peers}$ are the peer group returns for month t for a peer group. For the return of the peer group of bond i, the return of bond i itself is excluded. We refer to the Returns Model with abnormal returns as the Returns Demeaned Model. For Z-diff, we refer to Z-diff Demeaned.

A second approach is a two-pass approach where, in the first pass, we estimate a pooled model to control for systematic effects and in the second pass, we estimate arbitrage on an individual bond level. First, we introduce the procedure for Z-diff. We monthly clean the credit and equity returns of the maturity, rating, and sector mean

$$r^c_{t,i} = \beta^M_t M_{t,i} + \beta^R_t R_{t,i} + \beta^S_t S_{t,i} + \varepsilon^c_{t,i},$$

$$r^e_{t,i} = \beta^M_t M_{t,i} + \beta^R_t R_{t,i} + \beta^S_t S_{t,i} + \varepsilon^e_{t,i},$$

where M is a categorical maturity dummy, including three dynamic maturity buckets. Dynamic entails that at every month t, we divide the cross-section in equally sized buckets across the maturity dimension. There are as many bonds in the bucket with the lowest maturity as in the middle and highest buckets by dynamically defining the barriers between the buckets. The barrier has to be dynamic as the distribution of maturity is not constant through time. R contains the company's rating, ranging from AAA to B, and S is a sector dummy for financial, utility or industrial companies. The cleaned credit and equity returns we use for Z-diff as

$$\varepsilon^c_{(t-L):t,i} = \mu^c_{t,i} + \zeta^c_{(t-L):t,i}, \quad \varepsilon^e_{(t-L):t,i} = \mu^e_{t,i} + \zeta^e_{(t-L):t,i}$$

$$Z\text{-}score^c_{t,i} = \frac{\zeta^c_{t,L,i}}{\sigma^c_{\zeta,i}}, \quad Z\text{-}score^e_{t,i} = \frac{\zeta^e_{t,L,i}}{\sigma^e_{\zeta,i}}$$

$$Z\text{-diff}_{t,i} = Z\text{-}score^c_{t,i} - Z\text{-}score^e_{t,i},$$

where $\zeta_{t,L,i}$ is the last in-sample error at time t, we refer to this model as Z-diff d. For pass 1, we test three linear specifications for return level regression models. The first specification is only the monthly subtraction of the maturity, rating, and sector mean with a cross-sectional regression

$$r^c_{t,i} = \beta^M_t M_{t,i} + \beta^R_t R_{t,i} + \beta^S_t S_{t,i} + \varepsilon^C_{t,i}.$$

In the first specification, we correct for group average credit returns. What would the credit returns be if a company has a short-maturity bond, is AAA-rated, and is a financial company? In the second specification, we add the equity returns of the specific rating, maturity or sector group, and estimate Returns RD as

$$\begin{aligned}\varepsilon^C_{(t-L):t,i} = {} & \gamma^M_{t,g} M_{(t-L):t,i} r^e_{(t-L):t,g} + \gamma^R_{t,g} R_{(t-L):t,i} r^e_{(t-L):t,g} \\ & + \beta^S_{t,g} S_{(t-L):t,i} r^e_{(t-L):t,g} + \varepsilon^1_{(t-L):t,i},\end{aligned}$$

where g is, for example, maturity below five years, AAA, or financial sector. This specification corrects credit returns for the equity return of the rating, maturity and sector group. If all AAA bonds have lower returns than lower-rated bonds, we should correct this if we want to focus on the idiosyncratic returns. Removing these systematic returns from the group leaves solely the idiosyncratic returns of the specific bond. We include interaction between the rating, maturity and sector groups in the third specification. Essentially, we remove the group average equity return effect of, for example (AAA, long maturity, and financials) for companies that belong to that group, and estimate Returns RD ID

$$\begin{aligned}\varepsilon^C_{(t-L):t,i} = {} & \gamma^M_{t,g} M_{(t-L):t,i} r^e_{(t-L):t,g} + \gamma^R_{t,g} R_{(t-L):t,i} r^e_{(t-L):t,g} + \beta^S_{t,g} S_{(t-L):t,i} r^e_{(t-L):t,g} \\ & + \sum_g \gamma_{t,g} M_{(t-L):t,i} R_{(t-L):t,i} S_{(t-L):t,i} r^e_{(t-L):t,g} + \varepsilon^1_{(t-L):t,i},\end{aligned} \tag{7}$$

where $\sum_g$ runs over all the possible groups by interactions of the six rating, three maturity and three sector buckets. To limit the number of models in the final evaluation procedure, we test which of the three specifications leads to the highest autocorrelation in the error series. We employ autocorrelation as a metric to target CSA opportunities, not necessarily portfolio performance. Lastly, we test what information ΔDtD from the structural model holds by

$$\varepsilon^C_{(t-L):t,i} = \gamma^M_{t,g} M_{(t-L):t,i} r^e_{(t-L):t,g} + \gamma^R_{t,g} R_{(t-L):t,i} r^e_{(t-L):t,g} + \beta^S_{t,g} S_{(t-L):t,i} r^e_{(t-L):t,g} \\ + \beta^D_{t,g} \Delta DtD_{(t-L):t,i} + \varepsilon^1_{(t-L):t,i}.$$

In the second pass, we run individual regressions on the error term of the first pass. In this pass, we focus on the bond-specific sensitivity to the equity return of the issuer

$$\varepsilon^1_{(t-L):t,i} = \alpha_{t,i} + \beta_{t,i} r^e_{(t-L):t,i} + \varepsilon^2_{(t-L):t,i}, \tag{8}$$

where, like the Returns Model, we construct a Z-score of the last in-sample residual of the rolling window at every month t as

$$Z\text{-}score^2_{t,i} = \frac{\varepsilon^2_{t,L,i}}{\sigma^2_{t,i}}. \tag{9}$$

For the first-pass estimation, we further employ (Gradient Boosting) Regression Trees and Random Forest (Friedman, 2001; Breiman, 2001) to assess whether the increased flexibility in non-linearities and interaction effects improves the in-sample fit and autocorrelation dynamics. One may model the interaction effects linearly, like Equation (7), and one can also add non-linear transformations of those variables in a 'linear' regression model. However, lacking prior knowledge about the actual interaction and non-linear function quickly leads to computational infeasibility because of the explosion of the number of features if many possibilities are included (Gu et al., 2020). In the next paragraph, we discuss the implementation of (Gradient Boosting) Regression Trees and Random Forest of the following models

$$r^c_{t,i} = f(M_t, R_t, S_t, M_t r^e_{t,g}, R_t r^e_{t,g}, S_t r^e_{t,g}) + \varepsilon^1_{t,i},$$
$$r^c_{t,i} = f(M_t, R_t, S_t, \Delta DtD, M_t r^e_{t,g}, R_t r^e_{t,g}, S_t r^e_{t,g}) + \varepsilon^1_{t,i},$$

The errors from these specifications are further regressed on individual bond level following the procedure from Equation (8), and standardized like Equation (9).

(Gradient Boosting and Random Forest) Regression Trees This section discusses the implementation of (Gradient Boosting) Regression Trees and Random Forest. We implement a naive single tree, a dynamic depth single tree and Gradient Boosting Regression Trees and Random Forest with dynamic hyperparameter tuning.

Regression Trees Trees flexibly capture interactions and are therefore attractive to test on the entire cross-section (Breiman, 2017). A first naive implementation is a single regression tree with a depth equal to five for the entire sample period. We extend this by tuning the depth parameter in the cross-section, dynamic depth. The depth parameter is re-estimated every month. Single regression trees are prone to overfitting and do not generalize well to unseen data (Kelly and Xiu, 2023). Therefore, we implement two ensemble methodologies: boosting and bagging. First, we discuss the model validation procedure, equivalent to single, boosting, and bagging trees.

Model Validation In pass one, the goal is to provide superior explanatory power in-sample and in the cross-section. In pass one, we do not focus on the time-series dynamics that drive the arbitrage opportunity. We focus on purifying the credit returns from cross-sectional rating, maturity and sector effects. Therefore, we employ K-fold cross-validation (Stone, 1974), not time-series validation (Kelly and Xiu, 2023), even though we use time-series data. Furthermore, as we aim to explain a contemporaneous relation with high in-sample explanatory power, we do not obtain error terms for estimation in pass two in a test set but retrieve them in the train set. K-fold cross-validation is visualized in Figure 2 iteration 1 to 5. After obtaining hyperparameters, we retrain the entire 36-month cross-section to retrieve in-sample errors, iteration 6, for a second pass regression. To prevent data leakage, we employ grouped K-fold cross-validation. We define groups on the company level. Herefore, for example, Coca-Cola can only appear in the train or validation set, not in both. This prevents data leakage as we employ company-level equity and bond-level credit returns. Bonds are a subset of the company, so we circumvent data leakage in credit returns. Furthermore, rating and sector are company-level characteristics too.

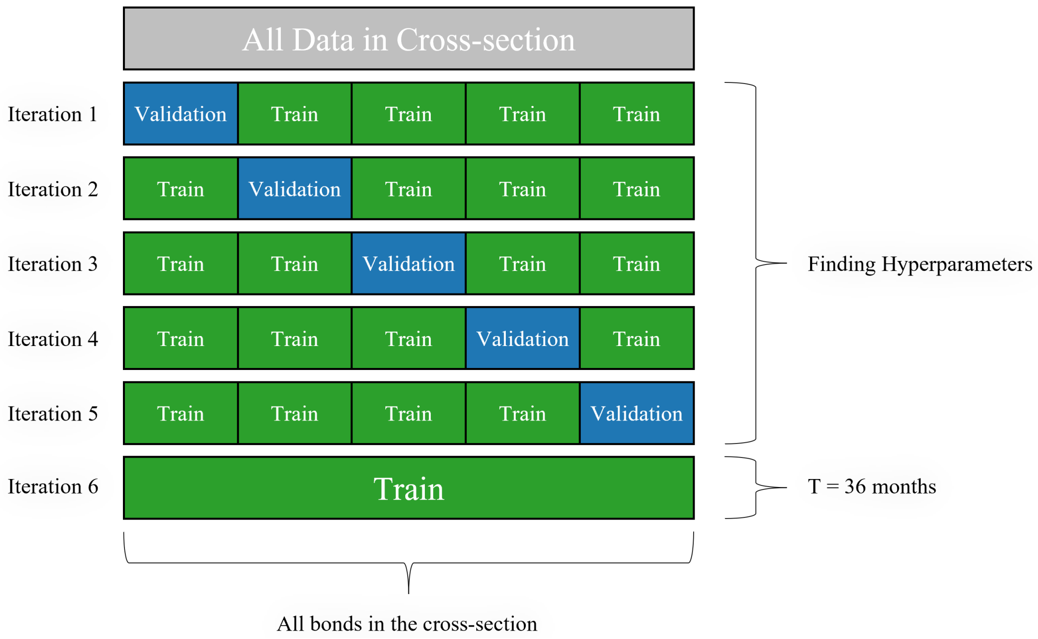


Figure 2: Estimation and training procedure for hyperparameters and in-sample error estimation

Ensemble Learning - Gradient Boosting and Random Forest In order to improve the predictive power of a single regression tree, one can employ boosting or bagging. By combining many different models, ensemble learning tends to be more flexible (less bias) and less data sensitive (less variance). Boosting and bagging combine a set of weak learners into a strong learner to minimize training and testing errors (Friedman, 2001; Breiman, 2001).

In boosting, a subsample of data is selected, fitted with a tree and then trained sequentially. That is, each new tree tries to compensate for the weaknesses of its preceding tree. With each iteration, the weak rules from each individual regression tree are combined to form one strong prediction rule. Individually, all trees are shallow and weak predictors, but combined, they form a strong predictor. Boosting methods differ in creating and aggregating weak learners based on the employed algorithm. Two popular boosting methods include Adaptive Boosting and Gradient Boosting (Freund and Schapire, 1997; Friedman, 2001). This paper uses Gradient Boosting because it is more commonly used in financial markets (Gu et al., 2020). Gradient Boosting works by sequentially adding predictors to an ensemble, with each one correcting for the errors of its predecessor by increasing the weight for data points with the highest errors. It trains on minimizing a loss function based on the residual errors of previous predictors (Friedman, 2001). In this research, we optimize over a quadratic loss function to align with the regression models. Important hyperparameters for Gradient Boosting are the depth, learning rate, number of trees, and percentage of the sample used for every tree. For an optimal subsample less than 100% of the sample we refer to Gradient Boosting as Stochastic Gradient Boosting (Friedman, 2002).

Another approach to regularize trees is bagging. In bagging, a bootstrapped subsample (Efron, 1992) with a random subset of all features is selected per tree. For every iteration (tree), we bootstrap a new subsample, and at every split, we construct a new random selection of features (Breiman, 2001). In bagging, all trees are trained in parallel, and a final prediction is made by averaging the predictions over all trees, contrary to boosting, where trees are trained sequentially. The depth of the trees, the number of trees, and the dropout for features and bootstrapping are hyperparameters.

Where boosting focuses on removing noise by focusing on the 'unexplainable' errors, bagging focuses on removing noise by ignoring large errors if they do not appear in many trees by taking the average over the trees. Boosting may, therefore, be more prone to overfitting. As we evaluate the in-sample fit of the boosting and bagging models, we must be wary of overfitting and test both approaches.

We expect hyperparameters to be stable over time, and for computational purposes, we re-estimate hyperparameters for Gradient Boosting Regression Trees and Random Forest every five years. To assess whether this assumption is valid, we investigate in-sample R^2 over time. We expect a slight drop in R^2 as we move away from the training window. We test standardized and rank standardized data as input variables, standard practice in the literature (Kelly and Xiu, 2023).

Besides model estimation on return level, the literature shows that cointegration methods using price-level information work well in equity markets and contain valuable information for structural models. We discuss such an approach in the next section of the paper.

4.3 Cointegration and the Error Correction Model

The profitability of a capital structure arbitrage strategy rests on the assumption that the prices will revert. Therefore, we investigate the degree of mean reversion by cointegration (Rad et al., 2016; Krauss, 2017). Furthermore, in the literature, it is clear that the profitability of strategies based on mean reversion for other asset classes depends on the degree of cointegration. Lovreta and Mladenović (2018) find that the profitability of CDS-equity trading crucially depends on the presence of cointegration and the stability of the cointegration vectors for a sample of European companies. Cointegration is a method that can measure the presence of a long-run equilibrium between debt and equity markets. We investigate whether the presence and strength of cointegration improve the returns of the capital structure arbitrage strategy. If that is the case, we can use it as a filter in our strategy. In practice, for example, one may only trade on a cointegrated subset of the universe.

We investigate double-sorted portfolios, where the first sort is in quintiles on cointegration and within every quintile, we sort on the degree of mispricing. If higher quintile portfolios obtain significant alpha over lower quintile portfolios, higher cointegration leads to superior returns. Moreover, we investigate whether a cointegrated subsample outperforms a non-cointegrated subsample.

The concept of cointegration is introduced in Granger (1981). Gregory and Hansen (1996) pioneer the test for cointegration whilst allowing for structural breaks in the time series. Cointegration is related to stationarity. A non-stationary time-series $x_{1,t}$ is $I(1)$ if the first difference of the time-series is stationary, $I(0)$. Consider a second $I(1)$ time-series $x_{2,t}$. Then, in general, it is true that a linear combination of the two time series is also $I(1)$. However, there may be a constant β such that $x_{1,t} - \beta x_{2,t}$ is $I(0)$. When this is true, the two time series are said to be cointegrated (Granger, 1981).

We examine Augmented Dickey-Fuller (ADF) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) statistics to test for stationarity (Dickey and Fuller, 1979; Kwiatkowski et al., 1992) of two time-series combinations: option-adjusted spread (OAS) and share price adjusted for dividends and splits (P^e) and P^e and bond price series constructed as

$$P^c_{t,i} = \sum_{k=1}^{t} (1 + r^t_{k,i}),$$

where r^t is the total credit return; the return including interest rate risk. We introduce the bond price series because we want the synthetic price of credit risk; a reflection of the market's assessment of the likelihood of default by bond issuers. The price series may reflect the price dynamics of the bond better than OAS; using the bond price series may increase the cointegration rate with share price. We test stationarity and cointegration on the entire sample. We are aware that estimating cointegration with information until time T, and trading with this information in month $t < T$ introduces forward-looking bias. However, consider a base case where OAS and share price are cointegrated for ten years. The cointegration of the series implies that if we find divergence in that period, it should mean revert. If the degree of cointegration does not improve credit returns in this scenario, it is improbable that it improves returns in a scenario without forward-looking bias. Nevertheless, we also study this assumption for stationarity by investigating rolling versus complete sample cointegration portfolio returns. We use ADF and KPSS statistics to test for stationarity. The null hypothesis of the ADF test states that the time series is non-stationary, implying $\gamma = 0$ in the following equation

$$\Delta x_{t,i} = \alpha_i + \beta_{0,i} t + \gamma x_{t,i} + \sum_{k=1}^{K} \beta_{k,i} \Delta x_{t-k,i} + \varepsilon_{t,i},$$

where $x_{t,i}$ is the level series, either $ln(OAS)$, $ln(P^c)$ or $ln(P^e)$, Δ the first difference of the respective series, α_i the constant and $\beta_{0,i} t$ the linear trend component. We choose AIC over BIC for lag selection as AIC emphasizes model fit. AIC penalizes model complexity less; thus, we can explore more lags (Burnham and Anderson, 2004). ADF has three configurations: no constant and no drift, constant without drift and constant with drift. We test for a constant with drift as a share price time series shows a trend. For KPSS statistics, the null hypothesis is the opposite of the ADF statistics; the time series is stationary, i.e. $d = 0$ in

$$x_{t,i} = \alpha_i + \beta_{0,i} t + d \sum_{j=1}^{t} u_j + \varepsilon_{t,i},$$

where u_j is a random walk. There are two configurations: stationary around a constant and stationary around a linear trend. We test stationarity around a trend for reasons similar to the ADF test. Because the null hypotheses are opposite and the tests different, the combination of the null hypotheses results in four unique cases (Schlitzer, 1995):

1. Both tests conclude that the series is not stationary - The series is not stationary
2. Both tests conclude that the series is stationary - The series is stationary
3. KPSS indicates stationarity, and ADF indicates non-stationarity - The series is trend stationary. One needs to remove the trend to make the series strictly stationary.
4. KPSS indicates non-stationarity and ADF indicates stationarity - The series is difference stationary. Differencing is to be used to make series stationary.

We are interested in cases 1 and 4 because when a differenced series is stationary, there is a basis to test for cointegration (Perman, 1991).

We test for cointegration on a subsample of non-stationary time series, cases 1 and 4. Two time series are cointegrated if the null hypothesis of no cointegration relationship is rejected. We employ Johansen's cointegration test (Johansen, 1991) to test for cointegration, most common in literature (Shrestha and Bhatta, 2018). We test for cointegration, including a linear trend. We choose a linear trend as we expect the relation between P^e and P^c to have a trend in most cases because the trend in share price is stronger than in OAS. If we find cointegration, the cointegration relation can be shown in an Error Correction Model (ECM) using Granger's theorem (Engle and Granger, 1987). Based on the ECM, the cointegrated time series shows a long-run equilibrium. The model specification allows for short-term deviations from the long-run equilibrium. The rolling ECM representation with a constant and a linear trend is

$$\begin{aligned} \Delta X_{(t-L):t,i} &= \alpha_{t,i}(\beta'_{t,i} X_{(t-1-L):t-1,i} + c_{t,i} + \gamma_{t,i} t) \\ &\quad + \sum_{k_{ar}=1}^{K-1} \Gamma_{k_{ar},t,i} \Delta X_{(t-k_{ar}-L):(t-k_{ar}),i} + \varepsilon_{(t-L):t,i} \end{aligned} \tag{10}$$

where $X = (ln(OAS), ln(P^e))'$, $\alpha_{t,i}(\beta'_{t,i} X_{(t-1-L):t-1,i} + c_{t,i} + \gamma_{t,i} t)$ is the error correction term around a linear trend, that reverts the time series to the long-run equilibrium for bond i, and the first term on the second line the autoregressive part of order k_{ar} that captures the short-term deviations from the long-run equilibrium. The long-run equilibrium is between option-adjusted spread and share price adjusted for dividends and share splits, or the share- and bond price series. Johansen (1991) describes two statistics to test the cointegration rank: the trace test and the maximum eigenvalue test. Following Lovreta and Mladenović (2018), we employ the trace test. We compute Johansen's trace test from eigenvalues $\lambda_{1,t,i}$ and $\lambda_{2,t,i}$, which we use to maximize the log-likelihood of equation (10). Following Johansen (1991), we define the null hypothesis of Johansen's trace test as

$$-T\sum_{j=1}^{2} log(1-\hat{\lambda}_{j,t,i}),$$

where $\hat{\lambda}_{j,t,i}$ are the solution of the problem

$$|\lambda S_{1,1} - S_{1,0}S_{0,0}^{-1}S_{0,1}| = 0,$$

where $S_{m,n}$ are defined as

$$S_{m,n} = T^{-1}\sum_{t=1}^{T} R_{m,t}R'_{n,t},$$

where $R_{0,t}$ is the matrix of residuals from regressing $\Delta X_{(t-L):t,i}$ on $(\beta_i' X_{(t-1-L):t-1,i} + c_i + \gamma_i t)$, and $R_{1,t}$ is the matrix of residuals from regressing X_{t-1} on the same set of regressors. When forming R, excluding the autoregressive part ensures that the residuals reflect deviations from the long-term equilibrium, providing a more accurate assessment of cointegration. Including the autoregressive part in R would mix short-term dynamics with long-term relationships, potentially leading to biased results. The Johansen trace test isolates long-term relationships by considering residuals, excluding the autoregressive part. We test the null hypothesis of $rk(\beta) = 0$ against the alternative that $rk(\beta) > 0$, where $rk(\beta)$ is the rank of the matrix β. If we reject the null hypothesis, we know that at least one cointegration relation exists. Moreover, we know that if we reject the null that $rk(\beta) = 1$ because we test for cointegration on a non-stationary subsample, implying that $rk(\beta) \neq 2$ because full rank would impose that the series are stationary (Johansen and Juselius, 1990).

Apart from testing the effect of the degree of cointegration on the portfolio credit returns derived from return level models introduced in Section 4.2.1 and 4.2.2, one may also construct portfolios on the divergence in the spread between OAS and P^e based on their historical relation. The time series that defines the arbitrage opportunity is

$$\delta_{(t-L):t,i} = ln(OAS)_{(t-1-L):t-1,i} - \beta_{0,t,i} ln(P^e)_{(t-1-L):t-1,i} + c_{t,i} + \gamma_{t,i}t,$$

where we estimate β_0 by regressing $ln(OAS)$ on a constant, the linear trend and $ln(P^e)$. We denote $\delta_{t,l,i}$ as the in-sample arbitrage spreads, where $1 \leq l \leq L$. Therefore, we denote the last in-sample arbitrage spread at month t as $\delta_{t,L,i}$. As different bonds have different arbitrage spread dynamics, we standardize all the spreads

$$\delta^*_{t,i} = \frac{\delta_{t,L,i} - \mu_{\delta,t,i}}{\sigma_{\delta,t,i}},$$

where $\mu_{\delta,t,i}$ is the mean of the arbitrage spread and $\sigma_{\delta,t,i}$ the standard deviation. We compute $\delta^*_{t,i}$ for every bond and rank accordingly every month. Essentially, $\delta^*_{t,i}$ is the signal. Following the reasoning for the return level models, we buy sufficiently negative spreads

$$\delta^*_{t,i} < \delta^*_{t,b},$$

where $\delta^*_{t,b}$ is the spread of bond b for which 10% of the spreads in the cross-section at time t, are lower.

Another approach to estimating arbitrage with a different level of information is to estimate arbitrage on a cumulative error series. In the next section, we introduce Ornstein-Uhlenbeck models tailored for such an estimation. Moreover, Ornstein-Uhlenbeck explicitly models mean reversion in the cumulative error series.

4.4 Ornstein-Uhlenbeck Models

Estimating a Z-score of the last in-sample residual as in Equation (3) does not consider the mean reversion of the series in the score. We model an Ornstein-Uhlenbeck (OU) process to consider an estimate of mean reversion (Ornstein and Uhlenbeck, 1930). Avellaneda and Lee (2010) introduce an Ornstein-Uhlenbeck process for modelling the outperformance of stocks relative to an ETF. We introduce the Ornstein-Uhlenbeck process in the credit market to estimate arbitrage opportunities between credits and equities. To estimate the Ornstein-Uhlenbeck process we construct a cumulative error series as

$$E_{t,l,i} = \sum_{k=1}^{l} \varepsilon_{t,k,i},$$

where ε is the error term from any return level model. Just as in Section 4.2.2 for ε consider now the cumulative in-sample cumulative error term $E_{t,l,i}$ for every window as the vector $E_{(t-L):t,i}$ for the entire sample, and model as follows

$$dE_{(t-L):t,i} = \kappa_{t,i}(\mu_{t,i} - E_{(t-L):t,i})dt + \sigma_{t,i} B_{(t-L):t,i},$$

where $B_{t,i}$ is a Brownian motion, $\kappa_{t,i}$ the speed of mean reversion, $\mu_{t,i}$ the unconditional mean, and $L = 36$. Because the last in-sample error is a cumulative sum including the first in-sample error of the rolling window, Ornstein-Uhlenbeck models use information of the entire rolling window length of the return level models. The Ornstein-Uhlenbeck process is a continuous process, but one can estimate the parameters in an AR(1) process without loss of generality as

$$E_{(t+1-L):t+1,i} = \alpha_{t,i} + \beta_i E_{(t-L):t,i} + \zeta_{(t-L):t,i}, \tag{11}$$

where $\zeta_{t,i}$ are normal, IID distributed random variables with mean 0 (Yeo and Papanicolaou, 2017). The parameter estimates for the Ornstein-Uhlenbeck process are

$$\kappa_{t,i} = -log(\beta_{t,i}), \quad \mu_{t,i} = \frac{\alpha_{t,i}}{1-\beta_{t,i}}, \quad \frac{\sigma_{t,i}}{\sqrt{2\kappa_{t,i}}} = \frac{\sigma^{\zeta}_{t,i}}{\sqrt{1-\beta^2_{t,i}}},$$

where $\sigma^{\zeta}_{t,i}$ is the square root of the variance of Equation (11). The parameter estimate of κ shows why the Ornstein-Uhlenbeck process is undefined on the non-cumulative error series. Specifically, for estimating the Ornstein-Uhlenbeck process on $\varepsilon_{t,i}$ directly, $\beta_{t,i}$ is mostly negative, rendering the OU process undefined in such cases. The estimated signal is

$$OU_{t,i} = \frac{E_{t,L,i} - \mu_{t,i}}{\sigma_{t,i}/\sqrt{2\kappa_{t,i}}} = \frac{-\mu_{t,i}}{\sigma_{t,i}/\sqrt{2\kappa_{t,i}}}, \tag{12}$$

as $E_{t,L,i} = \sum_{k=1}^{L} \varepsilon_{t,k,i} = 0$ by construction of the Ordinary Least Squares estimation of the residuals.

Next to a possible increase in portfolio returns, we investigate the drivers of the OU portfolio sorts by constructing portfolios on the separate components of the OU process. The top deciles contain bonds for which

$$OU_{t,i} < OU_{t,b}, \quad \mu_{t,i} > \mu_{t,b}, \quad \frac{\mu_{t,i}}{\sigma_{t,i}} > \frac{\mu_{t,b}}{\sigma_{t,b}}, \quad \kappa_{t,i} < \kappa_{t,b},$$

where b is the bond for which the cutoff value constructs 10% of the cross-section in the top decile. Because we investigate portfolio sorts on the kappa (κ) parameter, we estimate the OU process for the Z-diff not on the separate error series but on the Z-diff signal series as the difference between the equity and credit kappa has no meaning

$$E_{t,l,i} = \sum_{k=1}^{l} Z\text{-diff}_{t,k,i}.$$

One may not estimate the Ornstein-Uhlenbeck process on the individual credit and equity error terms due to the volatility difference in the two series; we need to consider the volatility difference between credit and equity. Furthermore, we investigate how OU portfolio sorts differ from a Z-score portfolio sort by investigating the in-sample error process to infer whether an OU process models a different mean-reversion pattern in the residual and credit returns. Finally, we investigate the performance improvement of modelling an Ornstein-Uhlenbeck process with spanning regression. We estimate the spanning alpha of every Ornstein-Uhlenbeck model over its return level matched model. For example, the Returns Model with Ornstein-Uhlenbeck portfolio returns regressed on the Returns Model from Equation (2) portfolio returns.

4.5 CSA Evaluation Framework

This section introduces an evaluation procedure to answer the research questions. We introduce a novel framework tailored specifically towards arbitrage strategies. We categorize all models into five categories: return level with fixed correlation, return level with estimated correlation, Ornstein-Uhlenbeck with fixed correlation, Ornstein-Uhlenbeck with estimated correlation, and cointegration models. For every category, we have introduced several models. Not all models are evaluated in the CSA framework as we pre-select promising models. Namely, we select categories with monotonic portfolio returns, reflecting a consistent return relationship (Haesen et al., 2017), and for the return level models we test the incremental improvement of the methods introduced in Section 4.2 including only models that demonstrate an improvement over the Returns Model. For the two-pass regression models we include the equity time effects that obtain highest autocorrelation in the signal series. For this selection we estimate Ornstein-Uhlenbeck models. Therefore, we do not retest the incremental improvement for Ornstein-Uhlenbeck models as we only estimate Ornstein-Uhlenbeck models for individual bond level controls of Section 4.2.2 that show an incremental improvement. For the final CSA evaluation framework, we introduce in the following paragraphs: (effective) lookback window, look-forward window, holding period, turnover, break-even transaction-cost, and autocorrelation.

(Effective) Lookback Window In this paragraph, we introduce the lookback window and then define an effective lookback window to compare returns level models and Ornstein-Uhlenbeck models on equal footing. Ornstein-Uhlenbeck models use information on the entire rolling window length of return level models because the last cumulative in-sample error contains information from all in-sample errors. We, therefore, investigate whether the information of multiple periods for return level models improves arbitrage modelling. The hypothesis is that if mispricing/divergence is persistent for multiple months, the probability of convergence should increase. Consider a lookback window of LB months; we transform the in-sample errors as

$$\varepsilon_{t,l,i}^{LB} = \frac{1}{LB}\sum_{k=1}^{LB} \varepsilon_{t,l-k+1,i},$$

essentially a moving average where $l = LB, ..., L$. For the return level models with estimated correlation, such as the Returns Model, we may define a signal for a lookback window of LB months as

$$S_{t,i}^{LB} = \frac{\frac{1}{LB}\sum_{k=1}^{LB} \varepsilon_{t,l-k+1,i}}{\sigma_{t,i}} = \frac{\varepsilon_{t,L,i}^{LB}}{\sigma_{t,i}}.$$

For a lookback window of 1 month, this is equal to the Z-score of Equation (3). For return level models with fixed correlation, such as Z-diff, the corresponding signal is defined as

$$S_{t,i}^{LB} = \frac{\frac{1}{LB}\sum_{k=1}^{LB} \varepsilon_{t,l-k+1,i}^{c}}{\sigma_{t,i}^{c}} - \frac{\frac{1}{LB}\sum_{k=1}^{LB} \varepsilon_{t,-k+1,i}^{e}}{\sigma_{t,i}^{e}}.$$

As for Ornstein-Uhlenbeck, the estimation window implicitly defines the lookback window. There is no further accumulation of the OU signal, and the signal is defined as

$$S_{t,i}^{LB} = OU_{t,i},$$

For the Ornstein-Uhlenbeck model, not every observation in the estimation window is weighted equally in the parameter estimates as the model employs a cumulative sum. The observations follow a linear decreasing weighting scheme. We define an effective lookback window to reflect the weighting scheme in the Ornstein-Uhlenbeck process. The effective lookback window neutralizes the weighing scheme and allows an apple-to-apple comparison for the effect of the lookback window. The effective lookback window equals the sum of the effective observation weights in the lookback window. Consider an example of the mean of an error series of length L

$$\mu_{\varepsilon} = \frac{1}{L}\sum_{l=1}^{L} \varepsilon_{l,i},$$

and a cumulative error series of length L

$$\mu_{E} = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{l} \varepsilon_{k,i} = \sum_{l=1}^{L} \frac{L+1-l}{L}\varepsilon_{l,i},$$

where the effective time weight for the error series is

$$ETW_{\varepsilon,l} = \frac{1}{LB},$$

and for the cumulative error series

$$ETW_{E,l} = \frac{LB+1-l}{LB} / \sum_{l}\left(\frac{LB+1-l}{LB}\right).$$

The different observation weights that are time dependent for the cumulative error series entail that effectively, some periods are weighted stronger than other periods; the ETW is illustrated in Figure 19 in Appendix B, and reflected in the effective lookback window by weighting the periods by their effective time weight

$$ELB = \sum_{l=1}^{LB} ETW_l(LB + 1 - l),$$

where one implicitly weighs the first effective time weight in a lookback window of 36 months 36 times, and the last observation (most recent) in the lookback window once to reflect the length of the lookback window.

Look-forward Window and Holding Period We introduce a look-forward window for an autocorrelation estimate for portfolios with a holding period of over one month. The holding period (HP) of a portfolio is how long an investor holds the bonds in the portfolio before selling them. The look-forward window for every lookback window is defined as

$$S_{t,i}^{LF,LB} = \frac{1}{LF} \sum_{k=1}^{LF} S_{t+k-1,i}^{LB},$$

where $LF = 1, ..., 5$. The holding period influences turnover in the strategies because bonds are held longer, influencing trading costs; therefore, we introduce turnover and break-even transaction cost as a metric that accounts for turnover in the next paragraph.

Turnover and Break-Even Transaction Cost Turnover (TO) is the percentage of the top decile that changes monthly multiplied by twelve to measure on a yearly scale. If the turnover of a portfolio is 1200%, that means that every month, all bonds in the portfolio are replaced. Because the turnover per model may differ and the holding period influences turnover, we define a metric that considers turnover. We define break-even transaction cost as

$$BETC = \frac{R_p}{2 * TO},$$

where we multiply by two because, for every bond we sell, we also have to buy a new one. The break-even transaction cost indicates how much every trade could cost before the strategy is loss-making. The power of break-even transaction cost is that we do not have to make assumptions about trading costs. Bond markets are relatively illiquid compared to equity markets, and liquidity in bond markets varies substantially between bonds, for example, across ratings (Chen et al., 2007; De Jong and Driessen, 2012). Therefore, assuming some level of trading cost will, by definition, not represent an investor's actual trading cost when applying the strategy. Break-even transaction cost are, therefore, more informative than net returns. We evaluate all signals for portfolio returns and break-even transaction cost. Nevertheless, we investigate autocorrelation to assess whether those portfolio returns and break-even transaction cost are actually driven by mean reversion.

Autocorrelation and Mean Reversion We employ autocorrelation as a proxy for mean reversion in the time series. Autocorrelation of every bond i at time t is defined by the Pearson correlation coefficient (Pearson, 1895) as

$$AC_{t,i}^{LF,LB} = Corr(S_{(t+1-35:t),i}^{LF,LB}, S_{(t-35:t),i}^{LB})$$
$$= \frac{\sum_{k=t-35}^{t}(S_{k+1,i}^{LF,LB} - \bar{S}_i^{LF,LB})(S_{k,i}^{LB} - \bar{S}_i^{LB})}{\sqrt{\sum_{k=t-35}^{t}(S_{k+1,i}^{LF,LB} - \bar{S}_i^{LF,LB})^2 \sum_{k=t-35}^{t}(S_{k,i}^{LB} - \bar{S}_i^{LB})^2}},$$

where the first estimation window observations, i.e. observations before observation L, are constructed from in-sample signals, which therefore contain forward-looking bias to compute the mean and standard deviation to obtain the Z-score for the signal. For a rolling window of 36 months, approximately 57% of the observations for bonds are before month 36. To some extent, this sample contains a forward-looking bias for the autocorrelation computation. We accept the bias to increase the available sample to compute autocorrelation from, with forward-looking bias 43% of the sample remains for autocorrelation estimation; without forward-looking bias, only 20%. We estimate autocorrelation with a 36-month rolling window. For the final autocorrelation of a model, we first take a cross-sectional average and then an average over time to prevent a higher weight to recent history as the universe of bonds grows with time. Note that we compute autocorrelation including $t+1$ for $S^{LF,LB}$ to include the observation of the convergence of the opportunity in the autocorrelation.

Due to computational limitations, we investigate a holding period of up to five months to test whether autocorrelation drives break-even transaction cost, but twelve months to assess possible convergence in break-even transaction cost independent of autocorrelation dynamics. We investigate a lookback window of 24, 30, 36, 42, and 48 months for the Ornstein-Uhlenbeck process for the CSA evaluation framework specifically. We investigate the Ornstein-Uhlenbeck process for several estimation windows to create enough variance in the sample compared to 36 months to infer the effect of autocorrelation on the return dynamics. We assess the influence of autocorrelation and the other variables on portfolio returns and break-even transaction cost as

$$R_{p,t} = c_t + \beta_{AC}AC_t + \beta_{ELB}ELB_t + \beta_{HP}HP_t + \vartheta_t,$$
$$BETC_t = c_t + \beta_{AC}AC_t + \beta_{ELB}ELB_t + \beta_{HP}HP_t + \beta_{TO}TO_t + \vartheta_t,$$

where we include turnover in the second equation because, by definition, break-even transaction cost is driven by turnover. We infer the significance and sign of all parameter estimates to investigate CSA and introduce an ensemble methodology for all models that capture CSA opportunities.

4.6 Ensemble

This section introduces two ensemble methodologies to combine signals. The first approach is a naive ensemble that averages all the signals, 1/N (Naive Ensemble). In the second approach, we do automatic signal selection via a Random Forest. We investigate whether autocorrelation is a good proxy of mean reversion for the Naive Ensemble. We do this for the ensemble instead of all models individually because the ensemble presents the average pattern of all models and is, therefore, representative of the mean reversion on average. If all models in the Naive Ensemble capture CSA opportunities following the autocorrelation proxy, we should observe divergence, before buying, and convergence, after buying, in credit returns for the top decile portfolio.

4.6.1 Naive Ensemble

For an ensemble, we have to choose between signal and portfolio blend. The signal blend takes the average of all the signals and buys one new portfolio. In a portfolio blend, one buys all the signals and takes the average of the returns of all portfolios. We use the signal blend as the signal blend performs better than the portfolio blend if the target is high alpha (Henke et al., 2020). Figure 20 in Appendix B shows an illustrative example showing that signal and portfolio blend buy different bonds, leading to different performances.

There is a different number of models for every model category (estimated/fixed correlation, return level/Ornstein-Uhlenbeck/cointegration). Therefore, the Naive Ensemble overweights category characteristics if we average all signals. We investigate the performance of the Naive Ensemble with equally weighted categories (Naive Ensemble Eqw) against the Naive Ensemble. For a correct implementation of the Naive Ensemble, it is important to cross-sectionally standardize every signal before averaging the signals because all signals are based on different models, and the distributions of those models are not necessarily equal. For a correct implementation of the Naive Ensemble Eqw, one additional cross-sectional standardization of the category averages is required before averaging over the categories. To compare the performance of the Naive Ensemble, we also investigate the hit ratio

$$\text{Hit Ratio} = \frac{\sum_{t=L}^{T} \begin{cases} 1, & \text{if } R_{p,t} > 0 \\ 0, & \text{if } R_{p,t} \leq 0 \end{cases}}{T},$$

where we essentially count the percentage of months with positive returns relative to the total number of evaluated months. A higher hit ratio may be interpreted as a higher probability of predicting future winners in the bond market (Korn et al., 2022). In the context of a Naive Ensemble, an elevated hit ratio indicates its favorability over the individual models it consists of, denoting more consistent and reliable performance.

4.6.2 Automatic Model Selection

As we have different model categories, the logical question is whether we can utilize the differences with automatic model selection. We use Random Forest Model Averaging (RFMA) for automatic model selection. We choose RFMA over Bayesian Model Averaging (BMA) because of 1) the computational intensity of BMA, 2) bagging of a Random Forest is asymptotically a particular case of BMA (Le and Clarke, 2022), 3) RFMA is more intuitive to benchmark against a Naive Ensemble, 4) RFMA would be a good benchmark for BMA because of 2). Wasserman (2000) further discuss BMA.

For RFMA, we test standardized and rank standardized features (signals), standard practice (Kelly and Xiu, 2023). The target is next month's credit return rank in the cross-section because we are only interested in the rank of every bond for next month, not necessarily the return itself. The rank target is more beneficial as we minimize the mean-squared error as a loss-function. Namely, it does not focus on predicting the extremes correctly since the target distribution is uniform.

We test two RFMA implementations. The first implementation uses a 1-month rolling window train set, no validation set, and a 1-month test set. We do not make hyperparameter selections on the test set to prevent forward-looking bias. We do not use a validation set for hyperparameter selection to prevent a gap between the train and test set. Without the gap, RFMA is closer to the Naive Ensemble, and the marginal improvement of applying RFMA can, therefore, be assessed more accurately. We do an initial test with various random training sets of the entire sample for hyperparameter selection to test if there is any sensitivity to the hyperparameters. The second implementation of RFMA is with expanding window with hyperparameter optimization. We prefer RFMA with a 1-month rolling window train set if the performance is similar to the RFMA with an expanding window since the 1-month training set allows to better assess the value add of RFMA against a Naive Ensemble. The expanding window setup is a training set of 3 years, validation of 1 year, and a test set of 1 month, with hyperparameter re-estimation every five years for computational purposes, discussed in Section 4.2.3. Lastly, as a proxy of the accuracy of the credit rank prediction independent of returns, we compute the Spearman Rank Correlation for the two RFMA implementations with the actual credit rank (Spearman, 1961).

Because we investigate multiple holding periods, we extend RFMA with a 2, and 3-month credit rank target. That is, we do not predict next month's rank but the average rank of a bond for the next 2 or 3 months. The hypothesis is that a multi-month credit rank target decreases turnover and that a multi-month credit rank target has higher portfolio returns for a holding period longer than one month than the 1-month credit rank target. It is important to note that this changes the training and test procedure. To prevent data leakage between the train and test set, we use a 1-month gap between the train and test set for the 2-month credit target and a 2-month gap between the train and test set for the 3-month credit rank target. We refer to the models as Random Forest 1, 2, and 3, depending on the credit rank target.

We assess the influence of the credit rank target on spanning alpha, turnover, break-even transaction cost, and compare this to the Naive Ensemble (Eqw).

Finally, we investigate what signals RFMA selects with feature importance scores as the features are the signals. For feature importance, we apply SHAP from Cooperative Game Theory (Shapley et al., 1953). SHAP is an explainer model that runs on the trained Random Forest model to explain the behaviour of the features. We define a simple explainer model to illustrate the basics of SHAP as follows

$$g(z') = \phi_0 + \sum_{j=1}^{J} \phi_j z'_j, \tag{13}$$

where $g(z')$ is the explainer function, ϕ_0 the base value, the model output without any input features, ϕ_j is the SHAP value of feature z'_j which is either zero (signal not present) or 1 (signal present) and $g(z')$ matches the output of RFMA assuring local accuracy. Together with two other properties, they guarantee that the SHAP values are unique (Lundberg and Lee, 2017). The explainer model from Equation (13) works on row-level data of features (signals) and, therefore, explains local predictions, i.e. the predicted credit rank in a certain month. We may extend Equation (13) to a global credit rank explainer as

$$\phi_j(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(J - |z'| - 1)!}{J!} \big(f_x(z') - f_x(z' \setminus j)\big), \tag{14}$$

where $|z'|$ is the number of non-zero entries in z', $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x', and $f_x(z')$ is the output of RFMA including feature z' for signal rank input x. x' are simplified inputs that map to the original input through mapping $x = h_x(x')$. In short, Equation (14) weighs all the differences between output with and without signal j and, therefore, constructs the effect for different credit ranks of signal j considering interaction effects with other signals. This procedure constructs the marginal contribution of every signal to the final predicted credit rank every month. The sum of absolute SHAP values for signal j ϕ_j is the total contribution of the signal to the predicted rank; it reflects to what degree the signal contributes to the final model output. If SHAP values for all signals are equal, RFMA is equal to the Naive Ensemble.

With the introduction of portfolio construction and spanning regressions in Section 4.1, the return level models in Section 4.2, cointegration models in Section 4.3, Ornstein-Uhlenbeck models in Section 4.4, evaluation criteria for CSA in Section 4.5, and finally the ensemble methodologies we answer all research questions in the next section.

5 Results

The results section is split into three main sections. The first section discusses the three subquestions in several subsections. The second section answers the main research question. The third section dives deeper into the incremental return level improvements, cointegration results, and Ornstein-Uhlenbeck results. We show that for return level models, we do not obtain an incremental improvement on individual-level modelling, and for cointegration models, returns do not show monotonicity. Furthermore, thirdly, a thorough investigation of how Ornstein-Uhlenbeck models capture CSA, and how the process behaves to explain the drivers behind Ornstein-Uhlenbeck models portfolio returns, and improvement over return level models.

5.1 Models that Capture Capital Structure Arbitrage Opportunities

We use a subset of the models introduced in Section 4.2, Section 4.3, and Section 4.4 for the CSA dynamics analysis, as discussed in Section 4.5. Namely, the cointegration models are omitted because the models do not provide monotonic returns across deciles, further discussed in Section 5.3.2. Secondly, in Section 5.3.1, we show that we do not obtain an incremental improvement on individual bond level for the return level models; thus we only include the Returns Model. The last paragraph of Section 5.3.1 explains the subselection of the two-pass models we use. The models in the CSA analysis are for return level Z-diff, Z-diff Demeaned, Z-diff d, Returns, Returns Demeaned, Returns RD, Random Forest, Boosting, and the latter three, including ΔDtD as a control variable. For every one of these models, we include the Ornstein-Uhlenbeck variant. Considering the five lookback windows, we have 5 times 11 times 2 = 110 models.

5.1.1 CSA on the Bond Leg Only

Whether the models capture CSA, Figure 3 shows break-even transaction cost for various levels of autocorrelation for return level and Ornstein-Uhlenbeck models. The figure shows that an increase in the autocorrelation magnitude leads to an increase in break-even transaction cost. Figure 21 in Appendix C shows a similar pattern for portfolio returns, although more masked for return level models. Table 3a and Table 3b show the drivers of break-even transaction cost and portfolio returns. For return level and Ornstein-Uhlenbeck models, portfolio returns are significantly driven by autocorrelation, shown by the t-statistic of -11.06 for return level models and 6.88 for Ornstein-Uhlenbeck models.

Table 3: Linear regression results for the relation between portfolio return, break-even transaction cost, and autocorrelation

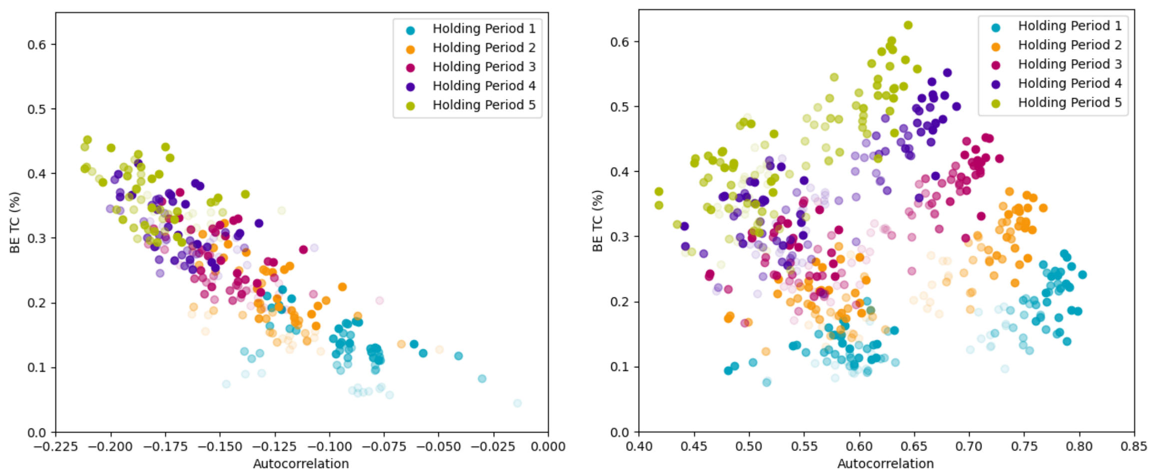
(a) Return level models

Variable	**Return**		**BE TC**	
Constant	1.47**	0.95**	-0.06**	0.00
	(20.76)	(14.35)	(-5.60)	(0.15)
Autocorrelation	-2.27**	-9.32**	-2.19**	-0.95**
	(-4.74)	(-14.88)	(-29.14)	(-11.06)
Effective Lookback Window		0.10**		0.03**
		(5.25)		(8.91)
Holding Period		-0.23**		0.03**
		(-14.00)		(9.12)
Turnover				0.00**
				(-3.54)
R-squared	0.07	0.48	0.74	0.89

(b) Ornstein-Uhlenbeck models

Variable	**Return**		**BE TC**	
Constant	1.00**	-0.20	0.92**	-0.55**
	(7.59)	(-0.34)	(16.74)	(-5.65)
Autocorrelation	1.55**	2.58**	-0.85**	0.97**
	(7.90)	(3.00)	(-10.48)	(6.88)
Effective Lookback Window		0.01		0.00
		(1.80)		(1.31)
Holding Period		0.08*		0.10**
		(2.12)		(12.86)
Turnover				0.00**
				(-4.66)
R-squared	0.19	0.36	0.29	0.91

Note. T-statistics are in parentheses; dependent variables are portfolio return and break-even transaction cost; * $p < 0.05$, ** $p < 0.01$.



(a) Returns Models

(b) Ornstein-Uhlenbeck models

Figure 3: Relation break-even transaction cost and autocorrelation for Returns and Ornstein-Uhlenbeck models, transparency reflects the lookback window; one month is most transparent to five months, least transparent

We conclude that for return level models in Table 3a 1) more substantial negative autocorrelation leads to higher BE TC. 2) A longer ELB, which means using more information, leads to higher BE TC. 3) A longer holding period leads to higher BE TC, contrary to the effect on portfolio returns, which means that the decrease in portfolio returns is not as substantial as the decrease in turnover. 4) Higher turnover leads to lower BE TC. Figure 22a in Appendix C shows more evidence for the hypothesis that a longer ELB leads to higher certainty for convergence as the increase in the magnitude of autocorrelation is mainly in the middle deciles. This signals that diverging patterns are pushed to the top and bottom deciles, which are the patterns that we want to capture.

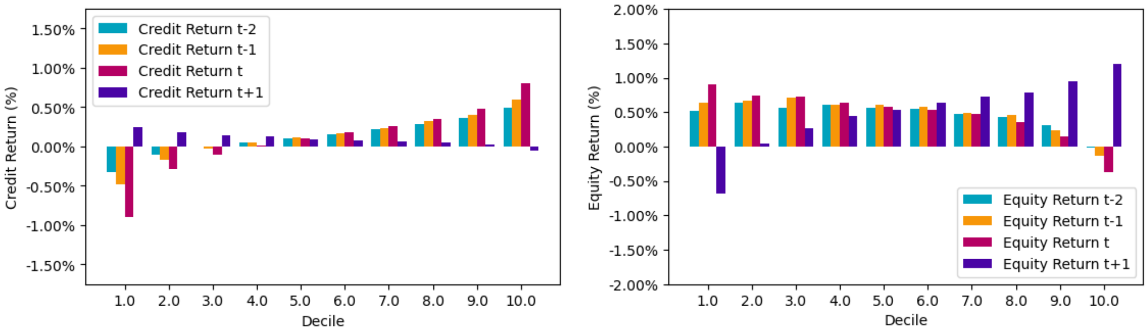
We conclude for Ornstein-Uhlenbeck models in Table 3b that 1) more substantial positive autocorrelation leads to higher BE TC, the slower the signal, the higher the BE TC. 2) A longer ELB has no significant effect, as every Ornstein-Uhlenbeck model uses a large part of history it is expected that the effect would not be as strong as for return level models (the relative difference between information used is smaller). 3) An increase in the holding period improves BE TC, as the holding period also increases portfolio returns for Ornstein-Uhlenbeck models this is expected. 4) Higher turnover leads to lower BE TC.

For both return level and Ornstein-Uhlenbeck models, approximately 90% of the variance is explained by the independent variables; the effect of autocorrelation is accurate. The difference in the sign between the effect of autocorrelation for return level and Ornstein-Uhlenbeck models is because Orstein-Uhlenbeck models are on a cumulative error series. Therefore, it is impossible to judge from the sign of the autocorrelation whether, for Ornstein-Uhlenbeck models, there is mean reversion in the arbitrage opportunity. The difference in the autocorrelation magnitude in absolute terms is due to overlapping data in the autocorrelation computation for Ornstein-Uhlenbeck models. If we use 48 months of cumulative data and move one month further, we only change one month of information. The cumulative data creates a moving average in the error term (Harri and Brorsen, 1998). Portfolio BE TC is better explained than portfolio returns reflected by a lower R^2 for explaining returns by autocorrelation, ELB, and holding period. This is reflected in Figure 21 in Appendix C as the relation between autocorrelation and portfolio returns is more masked than the relation between autocorrelation and BE TC in Figure 3.

Concluding, given the significance of autocorrelation and assuming that autocorrelation captures the dissapearance in the arbitrage opportunity, return level models capture CSA opportunities. To further investigate whether Ornstein-Uhlenbeck models capture CSA opportunities and whether autocorrelation is a good proxy for capturing CSA opportunities for return level models we investigate credit and equity returns prior and post arbitrage opportunities.

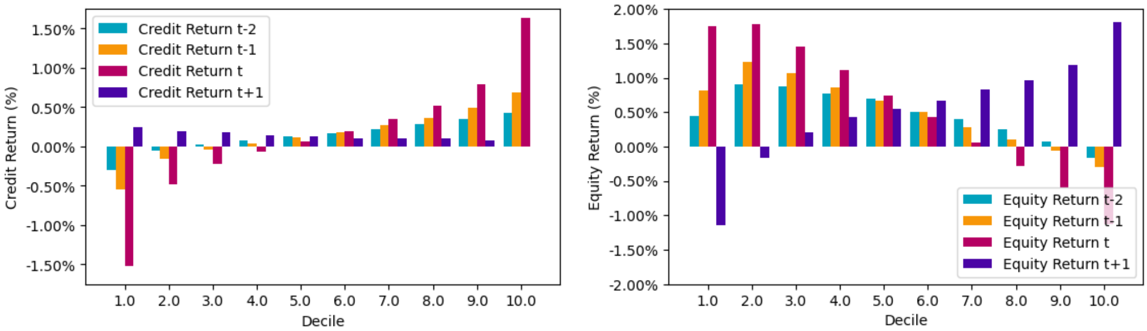
Figure 4 shows that up and until time t credit returns decrease and equity returns increase for Ornstein-Uhlenbeck models, and that at time $t+1$ this is reversed. Moreover, the pattern is monotonic across deciles. Therefore, Ornstein-Uhlenbeck models capture the formation and disappearance of the arbitrage opportunity; the Ornstein-Uhlenbeck models capture CSA opportunities. Figure 5 shows a similar pattern for return level models, therefore autocorrelation is a good proxy for evaluating CSA opportunities for return level models. In conclusion, all models capture CSA opportunities;

we can harvest CSA opportunities focusing on the bond leg only, with a portfolio factor investing implementation.



(a) Credit returns of Ornstein-Uhlenbeck models (b) Equity returns of Ornstein-Uhlenbeck models

Figure 4: Credit and equity returns across deciles before and after bonds are in the decile



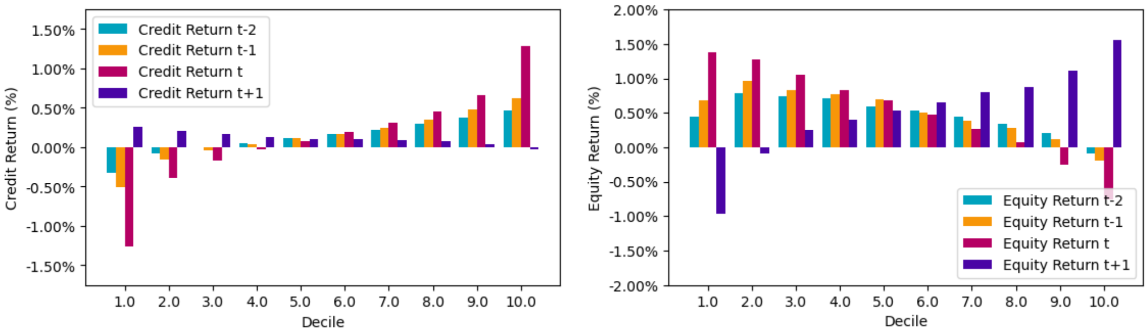
(a) Credit returns of return level models (b) Equity returns of return level models

Figure 5: Credit and equity returns across deciles before and after bonds are in the decile

5.1.2 Drivers of CSA Opportunities

We have shown all models capture CSA opportunities. To investigate what asset class drives the formation of the opportunity and drives the disappearance of the opportunity, we define a Naive Ensemble as in Section 4.6 and investigate again the development of credit and equity returns around time t. Figure 6 shows a similar pattern as Figure 4 and Figure 5 but approximately an average of the magnitude. This is expected as the return level and Ornstein-Uhlenbeck models are constituents of the Naive Ensemble. Most importantly, Figure 6 shows that for the Naive Ensemble, both asset classes drive the formation of the opportunity and the opportunity's convergence and disappearance.

Regarding the second part of the research question about the disappearance of arbitrage opportunities, we note that after a bond is a constituent of decile 1, the time it takes to return to middle deciles would signal that the opportunity has been arbitraged away. Figure 7 shows that most opportunities after three months are still in the top three deciles. The opportunities disappear relatively slowly for the Naive Ensemble. Note, however, from Figure 23 in Appendix D that 1) the speed of convergence, and 2) where the opportunity disappears to differ between return level and



(a) Credit returns of Naive Ensemble

(b) Equity returns of Naive Ensemble

Figure 6: Equity and credit returns across deciles before and after bonds are in the decile

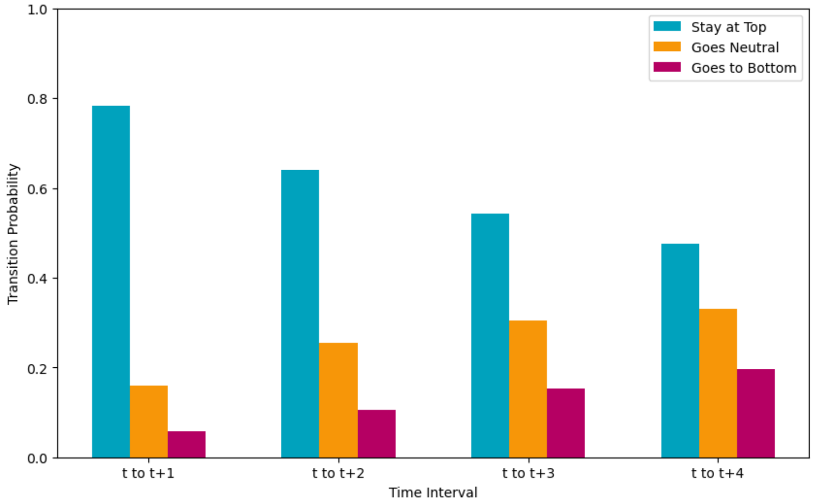


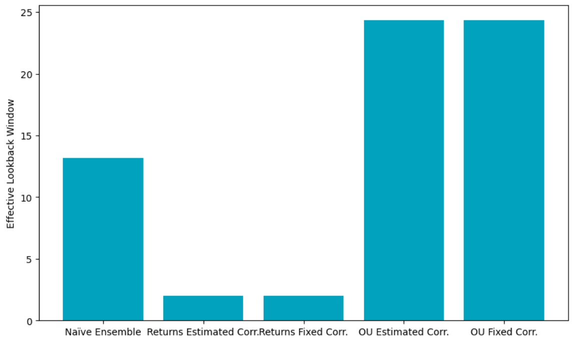
Figure 7: Transition probabilities Naive Ensemble

Note: Stay at top are deciles 1, 2, 3, Goes Neutral are deciles 4, 5, 6, 7, Goes to Bottom are deciles 8, 9, 10.

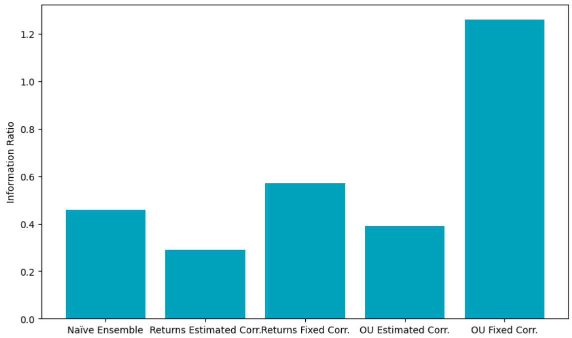
Ornstein-Uhlenbeck models. The quicker convergence in the return level models may be reflected in the larger magnitude of the credit returns at time t in Figure 5. The figure shows that return level models seem to capture more extreme cases than Ornstein-Uhlenbeck models do in Figure 4 because return level models have a larger difference in credit returns between t and $t+1$ across the top and bottom deciles. In conclusion, both asset classes contribute to the convergence and disappearance of the arbitrage opportunity, half of the opportunities have converged after approximately four months for the Naive Ensemble, and return level arbitrage opportunities converge quickest.

5.1.3 Robustness to Various Modelling Options

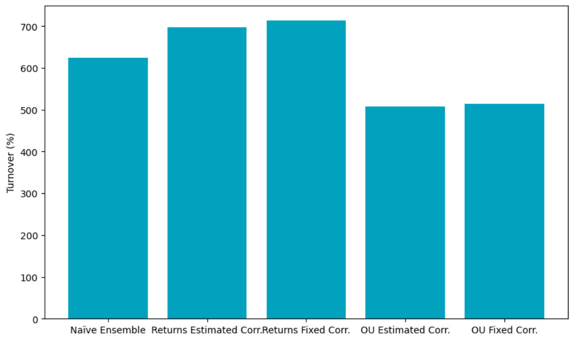
We know that all models capture CSA opportunities, that both asset classes drive the formation and convergence of the opportunities, and that the convergence speed varies between return level and the Ornstein-Uhlenbeck models. Figure 8 shows that between return level and Ornstein-Uhlenbeck models, a split in fixed- and estimated correlation reveals that the Naive Ensemble return dynamics vary substantially between its constituents. Namely, Figure 8d shows fixed correlation models, underrepresented in a Naive Ensemble, obtain higher BE TC than estimated correlation models. The fixed correlation models are underrepresented because Z-diff models only constitute 27% of all models. The varying performance between categories and underweight of fixed correlation models raise the question of whether the Naive Ensemble is optimal or cleverly selecting between the constituents may improve the model performance. Figure 24 in Appendix E shows that the first improvement is to weigh the categories equally as BE TC improves, which means that the signal average of fixed correlation OU models weighs as much as the signal average of estimated correlation OU models, even though there are more estimated correlation models.



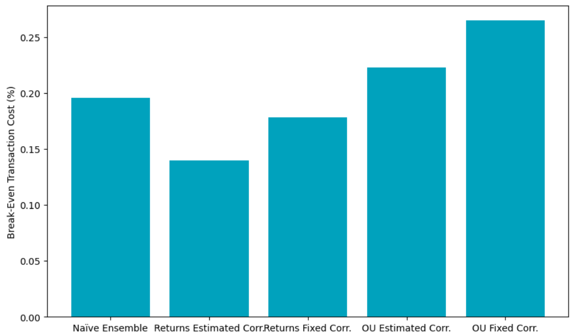
(a) Effective lookback window



(b) Information ratio



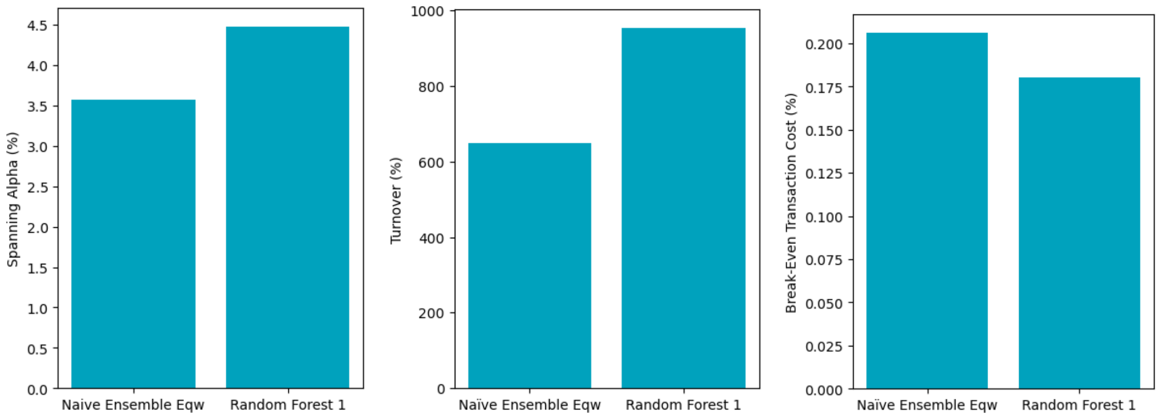
(c) Turnover



(d) Break-even transaction cost

Figure 8: Equity and credit returns across deciles before and after bonds are in the decile

The second ensemble methodology is automatic signal selection with RFMA. Figure 25 in Appendix E shows that Random Forest 1 selects models dynamically and that weights are not equal to 1/N. Figure 9a shows that model selection effectively improves the top decile returns for a 1-month holding period as Random Forest 1 spanning alpha is higher than Naive Ensemble Eqw spanning alpha. Figure 26 in Appendix E shows that performance between rolling and expanding window for Random Forest does not differ substantially, as a rolling window is closer to 1/N, and the performance is slightly better than expanding, Random Forest with a rolling window is the implementation of choice. That performance is better for a rolling window is somewhat surprising as Figure 27 in Appendix E shows that the Spearman rank correlation with next month's credit rank is higher for an expanding window than for a rolling window. That means that throughout the entire data sample, an expanding window predicts next month's rank closer to the actual rank than a rolling window, considering the entire cross-section. Therefore, a rolling window solely predicts better in the top as the top decile drives portfolio returns. Lastly, for robustness, Figure 28 in Appendix E shows that standardization or rank standardization and hyperparameters do not influence the loss; therefore, we choose fixed hyperparameters and standardized features for the rolling window. Rank standardized features contain less information than standardized features because they lose the scale difference. Figure 9c shows that the increase in performance for RFMA with rolling window comes at the cost of higher turnover and, therefore, lower BE TC than the Naive Ensemble Eqw.



(a) Spanning alpha (b) Turnover (c) Break-even transaction cost

Figure 9: Performance Naive Ensemble Eqw and Random Forest 1 (1-month holding period)

We extend the credit rank target. Figure 10 shows that all Random Forest model configurations have a more substantial alpha decay than the Naive Ensemble Eqw. Extending the Random Forest credit rank target decreases the difference in 1-, 2-, and 3-month holding period spanning alpha but does not improve the spanning alpha in absolute terms. Extending the Random Forest credit rank target is ineffective in improving spanning alpha. What does it mean for turnover and, therefore, for the break-even transaction cost?

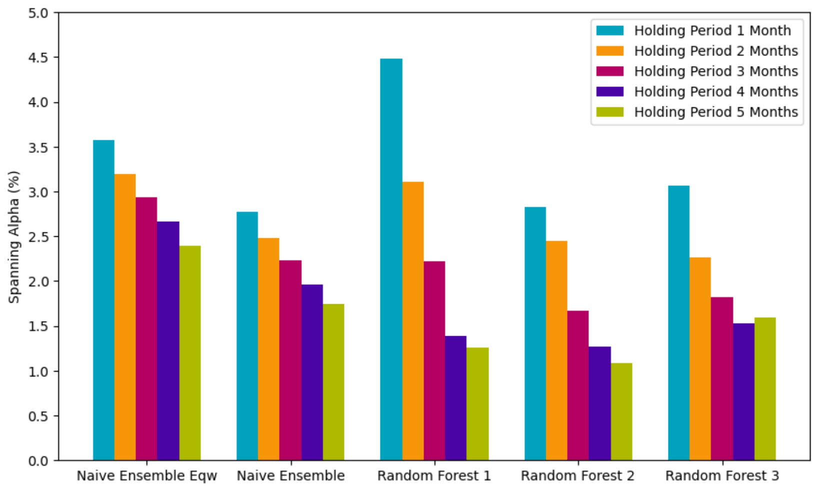
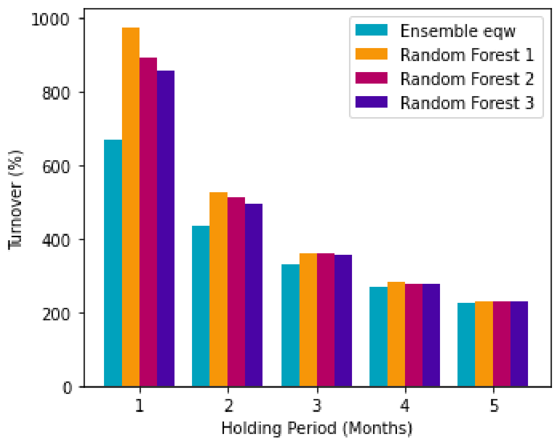


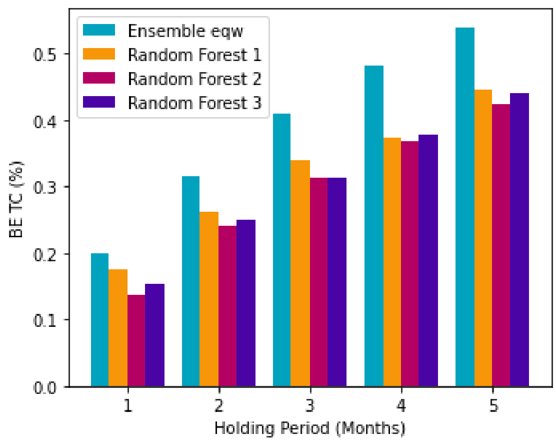
Figure 10: Top-bottom spanning alpha Random Forests and Naive Ensemble Eqw

Considering the effect of the multi-month credit rank target on turnover and, therefore, on the break-even transaction cost, Figure 11a shows that all RFMA credit targets lead to higher turnover and, therefore, lower break-even transaction cost for the same level of portfolio returns, Figure 11b. A 2- or 3-month target does reduce turnover, but only noticeably for a 1-month holding period. Again, extending the credit target is ineffective. A longer holding period is crucial for all methods to reduce the turnover rate and increase the break-even transaction cost. Considering transaction cost and performance, the longer the holding horizon, the more attractive the Naive Ensemble Eqw becomes, and the less attractive RFMA becomes, confirmed by Figure 12 that shows the difference expands to 20 basis points.

In conclusion, a Naive Ensemble Eqw obtains higher break-even transaction costs than a Naive Ensemble because of differences in the underlying categories and Random Forest Model Averaging because of lower turnover short-term and higher portfolio returns long-term.



(a) Turnover



(b) Break-even transaction cost

Figure 11: Naive Ensemble Eqw and Random Forests turnover and BE TC per holding period

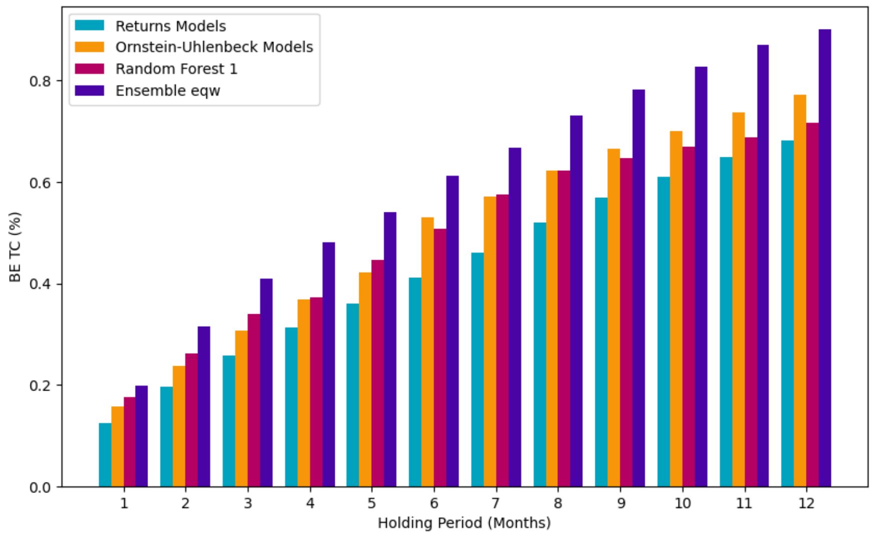


Figure 12: Convergence in break-even transaction cost

5.2 Best Capital Structure Arbitrage Model(s)

This section discusses the Ensemble Eqw specific results in Section 5.2.1. In the second section, we discuss which controls and whether including non-linear and interaction effects are effective in the two-pass approach.

5.2.1 The Ensemble

The Naive Ensemble Eqw obtains the highest break-even transaction cost, and models mean reversion in credit returns. The hit ratio is the final metric to show that an ensemble outperforms all other models. Figure 29 in Appendix F shows that an equally weighted ensemble obtains a higher hit ratio than all individual models and a higher hit ratio than RFMA.

We have shown that the Naive Ensemble Eqw outperforms the other models for the IG universe. For robustness purposes, we show in Table 22 in Appendix G that all ensemble models, except for returns estimated corr., also outperform the market for the HY universe. Furthermore, we test whether the strategy works in equity markets. Figure 31 in Appendix H shows that, if you switch the sign of the signal, CSA opportunities can also be harvested in the equity market because outperformance and IR have a monotonic pattern and IR is above the threshold of 0.5, which is considered good (Goodwin, 1998).

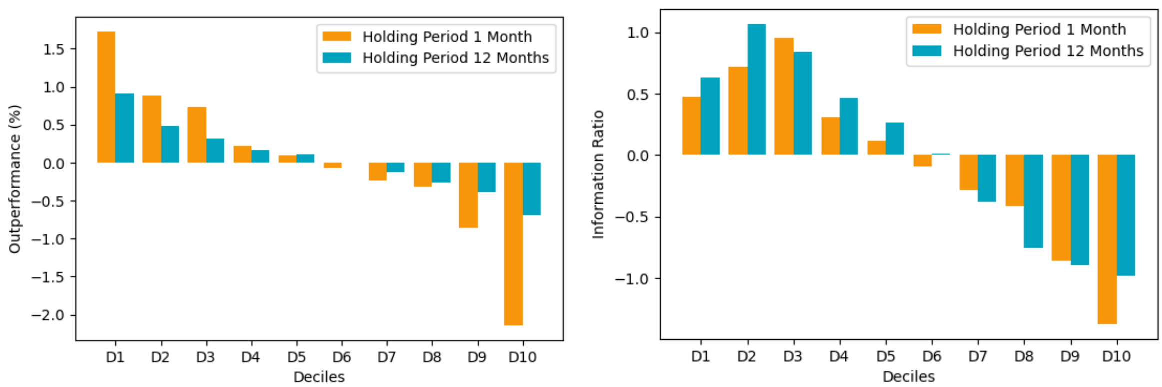
The Naive Ensemble Eqw models CSA and outperforms all individual models. A final performance evaluation is to analyze the 7-factor alpha from Section 4.1. Table 4 shows that the Naive Ensemble Eqw only obtains significant spanning alpha over the market. Considering the seven factors, a Naive Ensemble Eqw does not obtain a significant 7-factor spanning alpha. That means that, in theory, we may reconstruct the returns from the strategy by investing in a linear combination of the seven factors and the market. It does mean we would have to short Credit Momentum, which in practice is costly (Houweling and Van Zundert, 2017). In the top portfolio, we buy equity winners and credit losers, and we expect credit to outperform; the positive loading to Equity Momentum and the negative loading to Credit Momentum align with observed dynamics in Figure 6. The loading should be reversed for the bottom portfolio, but the table shows that the loadings are insignificant. For top-bottom, we therefore observe loadings similar to the top portfolio. Since we use credit and equity returns data in an Equity Momentum combined with Credit Reversal implementation, it is good to observe that the CSA strategy does not load significantly on Size, LowRisk, and Value for the top-bottom portfolio as we aim to model different dynamics. We, therefore, expect that we should model returns differently. We do not interpret the reversal factors since Table 21 in Appendix F shows that Equity- and Credit Reversal obtain significantly negative alpha to the market.

In conclusion, the Naive Ensemble Eqw obtains more periods with positive returns than any individual model and RFMA. Ensemble methodologies have positive returns over the market for the IG and HY universe. The Naive Ensemble Eqw CSA strategy works in the equity markets. The strategy shows monotonic outperformance and IR for a one- and 12-month holding period, Figure 13. Returns are not harvested in a small period, but over time, reflected by the hit ratio and visualized in Figure 30 in Appendix F. The CSA strategy significantly correlates with Equity- and Credit Momentum and, in theory, does not add anything above existing factors.

Table 4: Spanning regression results for the relation between Ensemble Eqw and factor top, bottom, and top-bottom portfolio returns and market returns

Model	**Ensemble Eqw**					
Variable	**Top**		**Bottom**		**Top-Bottom**	
Alpha	1.34%*	-1.45%	-2.11%**	0.02%	3.44%**	-2.37%
	(2.23)	(-1.64)	(-7.28)	(0.03)	(4.74)	(-1.93)
Market	1.42**	1.55**	0.96**	0.99**	0.45**	0.44**
	(33.10)	(42.71)	(46.55)	(48.24)	(8.74)	(9.17)
CreditReversal		-0.11**		-0.09**		-0.09**
		(-3.16)		(-4.62)		(-3.53)
EquityReversal		-0.01		0.27**		-0.10
		(-0.11)		(4.13)		(-1.22)
EquityMomentum6_1M		0.68**		0.06		0.25**
		(6.52)		(1.33)		(2.85)
CreditMomentum6_1M		-0.18**		0.01		-0.12*
		(-2.93)		(0.19)		(-2.42)
Size		-0.15*		-0.32**		-0.03
		(-2.05)		(-3.62)		(-0.41)
LowRisk		-1.83**		0.09		0.02
		(-10.91)		(1.92)		(0.50)
Value		0.11		-0.10		-0.04
		(1.80)		(-1.61)		(-0.51)
R-squared	0.76	0.86	0.86	0.88	0.18	0.36

Note. T-statistics are in parentheses; * $p < 0.05$, ** $p < 0.01$.



(a) Outperformance

(b) Information ratio

Figure 13: Naive Ensemble Eqw credit market performance

5.2.2 Controls - Systematic and Idiosyncratic Returns

In this section, we investigate the effect of the controls and non-linearities in machine learning, introduced with the two-pass approach. We draw four conclusions.

First, Table 5a shows that controlling for peer-group equity time effects does not significantly improve portfolio returns as there is no spanning alpha. Secondly, Table 5c shows that including ΔDtD does improve portfolio returns in some cases compared to the two-pass approach with only peer-group equity time effects. The pattern is, however, not strong, but if anything, there is economic value in the addition of ΔDtD. Thirdly, Table 5b shows that allowing for interaction effects and non-linearities with Boosting and Random Forest does not significantly improve portfolio returns consistently. The added value of ΔDtD is not in the interaction or non-linear interaction effects. Finally, Table 5d shows that monthly demeaning credit and equity returns with their peer group average leads to significant spanning alpha for 92% of the cases for Z-diff, but only 4% for Returns; however if anything demeaning is an improvement as there is positive spanning alpha for every case.

In short, estimating CSA opportunities with idiosyncratic returns via peer group demeaning is effective, especially for Z-diff. Controlling for peer-group equity time effects does not add value. Including ΔDtD adds, if anything, economic value. Non-linearities and interaction effects with Boosting and Random Forest do not improve model performance.

Table 5: Spanning regression alpha t-statistic results for two-pass controls and machine learning

(a) Equity time effects

Model		**Two-Pass**		
	t-stat	**Returns**	**Boosting**	**Random Forest**
Returns	above 2	8%	0%	0%
	below -2	0%	0%	0%

(b) Interaction and non-linearities

Model		**Two-Pass ΔDtD**	
Two-Pass ΔDtD	**t-stat**	**Boosting**	**Random Forest**
Returns	above 2	12%	12%
	below -2	0%	0%

(c) Including ΔDtD

Model		**Two-Pass ΔDtD**		
Two-Pass	**t-stat**	**Returns**	**Boosting**	**Random Forest**
Returns	above 2	16%	20%	24%
	below -2	0%	0%	0%
Boosting	above 2	36%	24%	44%
	below -2	0%	0%	0%
Random Forest	above 2	20%	16%	20%
	below -2	0%	0%	0%

(d) Peer-group demeaning

Model		**Demeaned**	
	t-stat	**Z-diff**	**Returns**
Z-diff	above 2	92%	
	above 0, below 2	8%	
	below 0	0%	
Returns	above 2		4%
	above 0, below 2		96%
	below 0		0%

Note. The percentage is of all spanning regression for lookback window and holding period one to five months; 8% is for 2 out of 25 combinations the t-value of spanning alpha is above 2 or below -2, significant at 5%.

5.3 Incremental Model Improvements

This section covers three topics. In the first section, we discuss return level models, in the second section, cointegration models, and in the third section, Ornstein-Uhlenbeck models.

5.3.1 Modelling of Arbitrage Opportunity on Return Level

In this section, we show that there are no incremental improvements on individual bond levels for return level models. Therefore, individual bond-level arbitrage estimation in CSA is only from the Returns Model.

Past Dependencies To model past dynamics that may drive arbitrage opportunities, we investigate past dependencies in the Returns Model with lagged equity and credit returns and the exposure of the signal from the Returns Model to past equity and credit information. The error from the Returns Model does not capture past dependencies. Table 6 shows that the rank correlation between the Returns Model and the Returns Model with equity and credit lags equals 0.81, indicating that both errors contain similar information. More importantly, the error from the Returns Model has a rank correlation of approximately 0 with 1 and 2 periods of lagged equity and credit returns, showing that the error does not contain the information that the lagged equity and credit returns do. Therefore, including past equity and credit information does not substantially influence the arbitrage signal, even though Figure 32 in Appendix I shows that R^2 of the Returns Model with equity and credit lags is consistently higher than the R^2 of the model without lags, the lags do contain information.

Table 6: Rank correlation portfolio return between Returns Model and Returns Model with equity and credit lags

Model	Returns CE_l1_l2	Equity_l1	Equity_l2	Credit_l1	Credit_l2
Returns Model	0.81	-0.02	-0.02	0.06	0.03

Note. Returns CE_l1_l2 is Returns Model with 2 equity and credit lags; Equity_l1 is last month's equity return; Credit_l1 is last month's credit return.

To conclude with significant evidence that including equity and credit lags does not improve the arbitrage modelling, Table 7 shows that the t-statistic of alpha is not significant. Moreover, the returns of both top-bottom portfolios are significantly correlated (47.80). The t-statistics of the credit and equity lags show that the Returns Model error contains less information about past credit returns than past equity returns, as both equity lags are significantly correlated. That both equity lags are correlated is in line with Kwan (1996); Gebhardt et al. (2005); Hilscher et al. (2015) that information flows from the equity to the bond markets. Including only the equity lags in the model, does not yield significantly better performance; alpha is not significant for Returns E_l1_l2 (-0.25).

Raw Mispricing versus Return Prediction We investigate whether the error of the Returns Model holds predictive power. We rank bonds on predicted next month's return instead of raw mispricing in the current month. Table 8 shows that predicting $t+1$ credit return with the mispricing in month t does not obtain significantly higher returns than constructing portfolios on the mispricing in month t. Moreover, increasing the mispricing information with the mispricing in month $t-1$ and month $t-2$ does not significantly outperform constructing portfolios on the mispricing in month t. The portfolio returns are also significantly correlated, even though the rank correlation between the Predict Returns and Returns Model is 0.01. The Predict Returns model, therefore, ranks bonds similarly in the tails, top and bottom deciles, but ranks differently in deciles two to eight. In conclusion, return prediction is not significantly more effective than raw mispricing.

Influence of Carry One may expect that an arbitrage strategy, which tries to harvest positive credit return movements, signals those opportunities better by removing the 'noise' of the carry component.

Table 7: Spanning regression results for the relation between Returns Model and Returns CE_l1_l2 and E_l1_l2 top-bottom portfolio returns

Variable	**Returns Model**		
Alpha	0.10%	1.25%	-0.05%
	(0.30)	(1.45)	(-0.25)
Returns CE_l1_l2	1.16**		
	(47.80)		
Credit_l1		0.17**	
		(3.43)	
Credit_l2		-0.04	
		(-0.76)	
Equity_l1		-0.15*	
		(-2.14)	
Equity_l2		0.49**	
		(8.14)	
Returns E_l1_l2			1.09**
			(88.57)
R-squared	0.87	0.23	0.96

Note. T-statistics are in parentheses; * $p < 0.05$, ** $p < 0.01$.

Table 8: Spanning regression results for the relation between Predict Returns Model with 1, 2, and 3 lags and Returns Model top-bottom portfolio returns

Variable	**Predict Returns**	**Predict Returns 2**	**Predict Returns 3**
Alpha	0.74%	0.34%	0.72%
	(1.24)	(0.59)	(1.21)
Returns Model	0.39**	0.23**	0.22**
	(11.57)	(7.06)	(6.70)
R-squared	0.28	0.13	0.11

Note. T-statistics are in parentheses; * $p < 0.05$, ** $p < 0.01$.

Table 9 shows that alpha increases as we focus more on spread change and remove carry; however, the monotonic behaviour is not reflected in significance, which does not increase. Furthermore, returns of the top-bottom portfolios are significantly correlated and R^2 approaches 1. Removing carry from the credit returns does not lead to a significant improvement. Therefore, carry does not significantly influence portfolio returns.

Table 9: Spanning regression results for the relation between Returns Model without Carry, Duration times Spread Difference, Spread Difference and Returns Model

Model	**Returns - Carry**	**Duration x dSpread**	**dSpread**
Alpha	0.15%	0.20%	0.36%
	(1.51)	(0.79)	(1.15)
Returns Model	1.01**	0.90**	0.81**
	(186.87)	(65.22)	(46.73)
R-squared	0.99	0.92	0.86

Note. T-statistics are in parentheses; * $p < 0.05$, ** $p < 0.01$.

Structural Model Information Industry practice for CSA strategies published in the early 2000s are structural models. Leland (2012) states that structural models underpredict default probabilities. Figure 33 in Appendix I shows that the default probabilities in the sample considered in this research are close to zero. Therefore, it is impossible to derive meaningful spreads with either Equation (5) or Equation (6). Default probability does, however, contain information that we may exploit in the two-pass approach as Table 23 in Appendix I shows that the error term of the Returns Model contains some exposure to ΔDtD, further explored in Section 5.2.2.

Issuer Level Signal An issuer-level signal allows for immediate arbitrage estimation upon issuance of a bond, unlike bond-level signals that require a specific estimation window of length L for each bond. An issue-level signal ranks 41.6% of the cross-section on average, while an issuer-level signal ranks 82.4% of the cross-section. Figure 34 in Appendix I shows that the median age of bonds is below the estimation window length of the issue-level signal; therefore, the issue-level signal does not rank many bonds compared to an issuer-level signal. Table 10 shows that an issuer-level signal performs poorly; the top decile underperforms the market. In conclusion, the arbitrage opportunities are bond-specific and cannot be harvested on the issuer level.

Table 10: Portfolio performance metrics for Returns Issue Level and Returns Issuer Level

Metric	**Returns**	**Returns Issuer**	**Returns Issuer***
Annualized Return	0.96%	-0.62%	-0.28%
Sharpe	0.15	-0.09	-0.05
Age	6.61	4.13	6.43
Outperformance	0.07%	-1.51%	-1.33%
IR	0.02	-0.41	-0.37

Note. *Returns Issuer in the last column shows results for issuer signal on subsample of Regr. Returns, bonds that are at least 36 months of age; IG investment universe.

Two-Pass Regression For two-pass regressions, we investigate whether we increase autocorrelation with interaction dummies and whether we should clean returns of systematic risk ex-post or ex-ante individual arbitrage estimation. Table 24 in Appendix I shows that the autocorrelation in the signal does not improve by including interaction dummies between the peer groups, and controlling for systematic effects followed by individual bond-level arbitrage estimation has more substantial negative autocorrelation than vice versa. Moreover, Table 25 in Appendix I shows that peer-group demeaning both equity and credit returns obtain stronger negative autocorrelation than demeaning only credit returns and that including a sector dummy increases autocorrelation further. Lastly, Figure 35 in Appendix I shows that Boosting and Random Forest hyperparameters are estimated effectively as there are no noticeable drops in R^2 for months that we do not re-estimate the hyperparameters.

Concluding on Return Level Incremental Effects Investigation of the Returns Model dynamics shows that 1) The Returns Model does not capture past dependencies, and including past dependencies does not significantly affect the portfolio returns. 2) Raw mispricing is a more effective signal than return prediction. 3) Removing carry does improve returns slightly, but not significantly. 4) The arbitrage opportunity is bond-specific; an issuer-level signal has lower returns than an issue-level signal. 5) Interaction effects between peer groups do not improve autocorrelation. 6) We estimate CSA opportunities more accurately by cleaning returns of systematic risk ex-ante than ex-post. 7) Peer-group demeaning has higher autocorrelation, including the sector and demeaning both equity and credit returns. In conclusion, there is little return level information to improve the Returns Model for signalling an arbitrage opportunity on the individual bond level. We investigate information on price level in the next section, where we discuss cointegration and error correction models.

5.3.2 Cointegration and the Error Correction Model

This section first discusses the influence of the cointegration requirement on the available sample for estimation. Secondly, we discuss the performance of an error correction model.

In Section 4.3 we discuss that we test for cointegration on two credit time series, test on the complete sample and a rolling window, and implement two different stationary tests. Whether we use OAS or P^c, apply a rolling window or static window on the complete sample, test on issue level or issuer level, or force a 10% significance level instead of 5%, the conclusion remains the same: cointegration does not occur frequently, and too few bonds remain available for a viable factor investing strategy. A cointegration requirement leaves around 20%-30% of the sample available for signal construction. That is 10%-15% of all bonds in the cross-section, as we cannot estimate arbitrage for bonds with an age below three years.

Table 11 shows on the issue level that 1) OAS has less non-stationary time series than share price. 2) The rate drops further because the two time-series are not stationary simultaneously (OAS & Share Price $<$ OAS or Share Price). 3) That the relatively low Case 1 + Case 4 rate drives the cointegration rate as $\sim$70% of non-stationary pairs are cointegrated. 4) Lastly, this holds for stationary and cointegration tests on the complete sample and on a 36-month rolling basis, Table 26 in Appendix J. Because OAS primarily drives the lower rate of non-stationarity, we also examine P^c, results are similar, and we draw the same conclusion, Table 27 in Appendix J. Table 28 in the same Appendix shows that few non-stationary pairs again drive the low cointegration rate for the issuer level. Due to the small cointegrated sample, an investment strategy solely on the cointegrated subsample is not feasible in practice. However, one may estimate an arbitrage strategy on the cointegrated subsample and combine the strategy with an investment strategy on the complete sample; the cointegration strategy may be a complement.

Table 12 shows portfolio performance for five ECMs. Spread is an ECM specification on all bonds in the cross-section, whilst CI Spread is an ECM specification on all cointegrated bonds in the cross-section. The table shows that ECM performance varies substantially between specifications. Moreover, a cointegration prerequisite does not consistently improve outperformance as top decile returns for OAS improve to (0.88% $>$ 0.25%), whilst for the bond price, the returns decrease (0.15% $<$ $-$0.20%). Similarly, IR results are not unambiguous. Table 29 in Appendix J shows that the ECMs generally do not show monotonicity in IR. In short, results are not unambiguous; if anything, the

Table 11: Non-stationary ratio according to ADF and KPSS statistics and cointegration ratio according to Johansen's trace test for OAS and share price on issue level for the complete sample

Issue Level Complete Sample	**Non-stationary ADF**	**Non-stationary KPSS**	**Case 1**	**Case 4**	**Case 1 & 4**	**Cointegrated with Prerequisite**	**Cointegrated without Prerequisite**
Alpha = 0.05							
OAS	73.9%	49.1%	45.4%	3.7%	49.1%		
Share Price	77.5%	56.3%	52.4%	4.0%	56.3%		
OAS & Share Price	64.0%	33.9%	29.5%	4.3%	33.9%	25.1%	51.1%
Alpha = 0.1							
OAS	66.6%	62.5%	52.8%	9.7%	62.5%		
Share Price	71.1%	69.6%	59.6%	10.0%	69.6%		
OAS & Share Price	53.5%	50.0%	37.2%	12.9%	50.0%	32.3%	61.6%

Note. Case 1 refers to non-stationary ADF and KPSS test statistic results. Case 4 refers to the ADF test result as stationary and the KPSS test result as non-stationary; The percentage is relative to the number of time series; The results shown are for tests around a linear trend; the results are similar for tests around a constant.

ECMs do not provide consistent returns. Cointegration may, however, contain information for the Returns Model. Therefore, we investigate the performance of the Returns Model on a cointegrated subsample and a non-cointegrated subsample.

Table 12: Portfolio performance metrics for cointegration models

	Option Adjusted Spread and Share Price			**Bond Price Return and Share Price**	
Metric	**Spread[1]**	**CI Spread[1]**	**CI Spread[2]**	**Spread[1]**	**CI Spread[1]**
Outperformance	0.25%	0.88%	0.24%	0.15%	-0.20%
IR	0.23	0.39	0.15	0.14	-0.04

Note. IG investment universe; [1]rolling window stationarity and cointegration test; [2]complete sample stationarity and cointegration test; alpha = 0.05 for tests.

Table 13 shows that the Returns Model on a cointegrated or non-cointegrated subset does not significantly out- or underperform the complete sample as both alpha t-statistics are insignificant. However, it is striking that the Returns Model shows higher returns on the non-cointegrated subsample than the cointegrated subsample. Within the cointegrated subsample, Table 30 in Appendix J shows there is no monotonic increase or decrease in the portfolio returns as cointegration increases for the Returns Model; there is no monotonic effect of cointegration on the Returns Model portfolio returns.

Table 13: Spanning regression results for the relation between Returns Model cointegrated and non-cointegrated subsample and Returns Model complete sample portfolio returns

Variable	**Cointegrated subsample**	**Non-cointegrated subsample**
Alpha	-0.69%	0.48%
	(-1.47)	(1.73)
Returns Model[1]	1.18**	0.84**
	(45.92)	(54.63)
R-squared	0.86	0.90

Note. T-statistics are in parentheses; [1]complete sample; cointegration alpha = 0.05; * $p < 0.05$, ** $p < 0.01$.

Considering all rating categories, we have shown that there is no significant effect of cointegration on portfolio returns for the ECMs and the Returns Model. However, Figure 36 in Appendix J shows that the rate of cointegrated time-series pairs is higher for lower-rated bonds. This is not surprising, as we show in Section 3 that bond dynamics differ across rating categories. We investigate whether larger rating buckets drive full sample results from Table 13 by splitting the cointegration effect across rating categories. Table 14 shows alpha significance across rating buckets for a cointegrated and not cointegrated subsample. The first clear pattern is that, in general, alpha increases for lower-rated bonds. The second and most important pattern is in the alpha difference column, the difference in alpha between a cointegrated and not cointegrated subsample; namely, there is no monotonic increase for alpha as ratings decrease. No evidence exists that a cointegrated subsample obtains higher returns for lower-rated bonds than higher-rated ones.

Table 14: Spanning regression alpha results across (non-)cointegrated subsample and rating categories

	Z-diff Cointegrated		Z-diff Not Cointegrated		
Rating	**Alpha (%)**	**Alpha - t-statistic**	**Alpha (%)**	**Alpha - t-statistic**	**Alpha Difference**
AAA-AA	1.25*	2.47	1.58**	5.11	-0.32
A	1.54**	3.39	1.46**	5.99	0.08
BBB	2.02**	3.47	1.95**	4.56	0.06
BB	2.56*	2.41	2.98**	3.98	-0.42
B	4.48**	2.95	4.09**	3.79	0.39
	Returns Model Cointegrated		**Returns Model Not Cointegrated**		
Rating	**Alpha (%)**	**Alpha - t-statistic**	**Alpha (%)**	**Alpha - t-statistic**	**Alpha Difference**
AAA-AA	1.34**	2.71	1.10**	3.89	0.24
A	1.55**	3.36	1.36**	5.27	0.19
BBB	2.35**	3.6	1.44**	2.88	0.91
BB	3.6**	3.36	3.39**	3.89	0.21
B	2.09	0.89	3.42*	2.59	-1.33

Note. There is a more idiosyncratic risk for BB and B-rated bonds. Therefore, we investigate the performance across all rating categories with quintiles instead of deciles; * $p < 0.05$, ** $p < 0.01$.

In conclusion, 1) a cointegration requirement reduces the available bond universe by 70 to 80%, driven by a lack of non-stationarity. 2) ECMs do not provide monotonic portfolio returns. 3) The Returns Model does not provide better returns on a cointegrated subsample. 4) A higher cointegration rate for lower-rated bonds does not lead to higher returns on the cointegrated subsample than the non-cointegrated subsample. Thus, no information on cointegration may improve arbitrage modelling.

5.3.3 Ornstein-Uhlenbeck and Mean Reversion

In this section, we discuss the spanning alpha of Ornstein-Uhlenbeck models over their matched return level models, the dynamics underlying the Ornstein-Uhlenbeck model, and investigate the value added of the mean-reversion parameter in the process. Compared to two-pass regression, we may interpret the Ornstein-Uhlenbeck process as a third pass where we accumulate the CSA residuals.

Table 15 shows that an Ornstein-Uhlenbeck (OU) process improves performance for all return level models because most spanning alpha is positive, 95% t-statistics > 0, and significant, 45% of t-statistics > 2. Table 31 in Appendix K shows that especially for Z-diff models portfolio returns substantially improve with the addition of an Ornstein-Uhlenbeck process as for 84%, 92%, and 52%, on average for 72% of lookback window and holding period combinations the model with Ornstein-Uhlenbeck process has significant spanning alpha over the model without. The improvement is possible because the OU process captures a different error pattern. Figure 37 in Appendix K shows modelling an OU process forces a more gradual mispricing pattern in the top and bottom deciles. The gradual pattern is reflected in the OU Ensemble versus the Returns Ensemble in Figure 4a and Figure 5a where the magnitude of the credit return at time t is lower for the OU Ensemble in the top and bottom deciles than the Returns Ensemble whilst credit returns at time $t+1$ is higher for the OU Ensemble. The credit return at time $t+1$ is the credit return that the method harvests, reflected by significant spanning alpha of OU relative to return level models in Table 15.

Table 15: Spanning regression average alpha t-statistic results for Model with OU and OU Kappa sort over matched return level model

	Ornstein-Uhlenbeck models		**Ornstein-Uhlenbeck models κ sorted**	
Alpha	**Average over all return level models**	**Returns Model**	**Average over all return level models**	**Returns Model**
t-stat >2	45%	48%	10%	72%
0 <t-stat <2	50%	52%	63%	28%
-2 <t-stat <0	5%	0%	27%	0%
t-stat <-2	0%	0%	0%	0%

Note. The percentage is of all spanning regressions for lookback window and holding period of one to five months; the t-value is the alpha of the spanning regression of portfolio returns of a model with Ornstein-Uhlenbeck process over the portfolio returns of that model without Ornstein-Uhlenbeck process; 48% is for 12 out of 25 combinations the t-value of spanning alpha is above 2, significant at 5%.

We have assessed that Ornstein-Uhlenbeck models add performance relative to return level models. However, the primary motivation for applying Ornstein-Uhlenbeck is the mean-reversion parameter κ in Equation (12) as the hypothesis is that explicitly modelling the mean reversion would drive higher returns. We show that this parameter does not drive the Ornstein-Uhlenbeck performance.

Table 16 shows that the OU process does significantly outperform a sort solely on μ (2.51) of the OU process, but by including just σ and not κ, OU actually underperforms (-1.19). Therefore, κ does not add value in Equation (12) and does not drive the performance. The table also shows that the OU process significantly outperforms a sort on κ (3.10), but that κ also obtains significant alpha over the OU process (3.14). Two-way significant spanning alpha is possible because the correlation between the two sorts is low, reflected in an R^2 of 0.03; the parameter κ is not significantly related to the portfolio construction of the OU process. Table 33 in Appendix K furthermore shows that the OU process significantly negatively loads on Credit Momentum whilst κ loads on different factors, further evidence that the two sorts are unrelated.

Figure 14 shows the effect of κ on the OU process; κ constrains the total mispricing in the OU process as the sum of the cumulative error decreases for higher quintiles of k. This is reflected in lower portfolio returns for mu portfolios with high kappa values, Table 34 in Appendix K. The value of the mean-reversion parameter κ is not visible in the portfolios constructed from the OU process. For the Returns Model, Table 15 shows that portfolios constructed on a sort of just κ from the OU

Table 16: Spanning regression results for Returns with OU process over Returns with OU but μ, μ/σ or κ sorted and results for Returns with OU κ sorted over Returns with OU

Variable	**Ornstein-Uhlenbeck**			**Kappa**
Alpha	1.32%*	-0.37%	2.34%**	2.07%**
	(2.51)	(-1.19)	(3.10)	(3.14)
μ	0.47**			
	(19.21)			
μ / σ		0.90**		
		(41.64)		
κ			-0.20**	
			(-3.35)	
Ornstein-Uhlenbeck				-0.15**
				(-3.35)
R-squared	0.52	0.83	0.03	0.03

Note. T-statistics are in parentheses; * $p < 0.05$, ** $p < 0.01$.

process obtain significant spanning alpha over the Returns Model without the OU process. The table also shows that this does not generalize to other return level models as only 10% of the t-statistics are positive and significant. Table 32 in Appendix K shows that for many models even 0% of t-statistics are significant. Therefore, we may not obtain consistent portfolio returns with k sorted portfolios.

Concluding, there is no robust value in the κ parameter of the OU process. It mainly functions as a constraint on the total sum of mispricing in the process. The performance improvement of Ornstein-Uhlenbeck is driven by something other than explicitly modelling mean reversion in the cumulative sum of the error process. The substantial incremental performance improvement over return level models is because the model estimates the total mispricing of the in-sample error process relative to only mispricing in recent months.

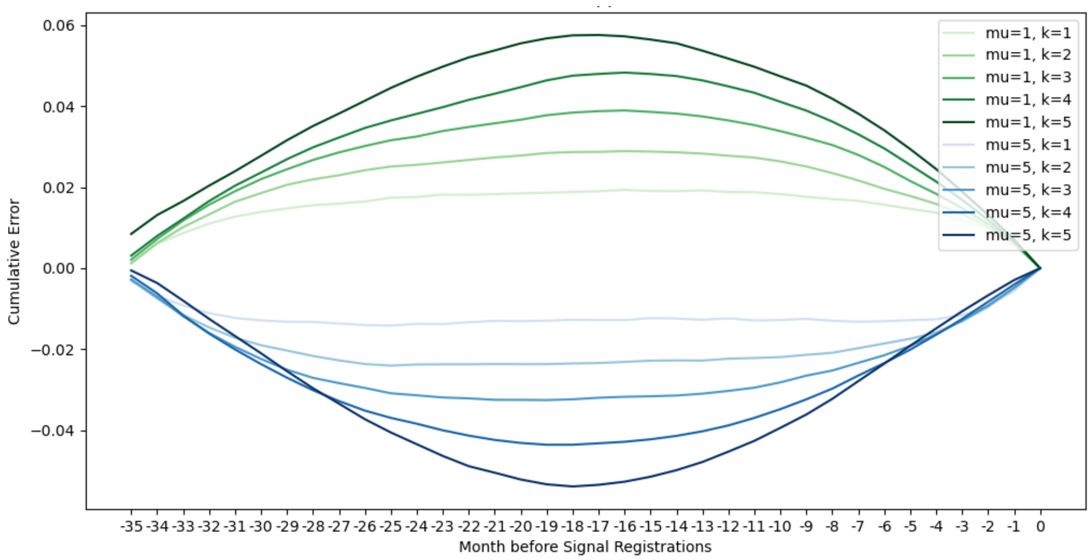


Figure 14: Influence of κ on cumulative in-sample error for the Ornstein-Uhlenbeck process, where quintile 1 contains the highest values for mu (μ) and k (κ)

6 Conclusion and Discussion

This research aims to find the best (combination of) model(s) that capture CSA opportunities with a factor investing portfolio perspective, where the best is reflected in break-even transaction cost.

We show that return level models and Ornstein-Uhlenbeck models can harvest CSA opportunities focusing on the bond leg only. Moreover, we may harvest the opportunities in the equity market, too. We show that both asset classes, equity and credit, drive the divergence that creates the arbitrage opportunity and that both asset classes contribute to the convergence and disappearance of the opportunity. Regarding speed, half of the opportunities have converged after approximately four months. We investigate robustness across fixed and estimated correlation categories and show that performance deviates between model categories but that opportunities may be harvested for every category independently. We show that a naive ensemble with equally weighted categories outperforms a naive ensemble where all signals are equally weighted because fixed correlation signals are underrepresented. We show that this equally weighted ensemble outperforms any individual model within the categories. A more sophisticated ensemble, Random Forest Model Averaging, obtains higher portfolio returns than a naive ensemble with equally weighted categories, but considering transaction cost, a naive ensemble with equally weighted categories outperforms Random Forest Model Averaging. We may attractively combine the individual models and obtain an ensemble model with a yearly turnover rate of 117% and break-even transaction cost of 1.12%. The CSA factor of the ensemble is significantly correlated with Equity and Credit Momentum and, empirically, does not obtain significant 7-factor spanning alpha. However, to construct the linear combination an investor would have to short bonds, which is difficult and costly in practice (Houweling and Van Zundert, 2017).

Within the models considered for the ensemble, we show that models based on idiosyncratic returns are a significant improvement over models based on systematic returns if we create the idiosyncratic returns by monthly demeaning credit and equity returns with peer group returns based on maturity, rating, and sector buckets. Controlling for peer-group equity time effects does not improve portfolio returns. Including ΔDtD as a credit risk control improves portfolio returns, although only sometimes significantly. Machine learning has no value added over linear controls.

For incremental model improvements we first conclude on individual bond level, then for the two-pass, followed by cointegration, and finally the Ornstein-Uhlenbeck models.

On individual bond level there are no incremental model improvements, we show that 1) the Returns Model does not capture past dependencies, and including equity and credit lags does not significantly improve portfolio returns. 2) Raw mispricing, like the Returns Model, is more effective than return prediction based on the raw mispricing. 3) Removing the carry component of credit returns in the Returns Model does not improve portfolio returns. 4) The arbitrage opportunity is bond-specific; an issuer-level signal underperforms a bond-level signal. The issuer-level signal even underperforms the market.

For the two-pass regression we investigate how to correct for systematic risk, we show that 1) ex-ante cleaning returns of systematic risk is more effective than ex-post. 2) Including interactions between rating, maturity, and sector buckets does not improve the effectiveness of equity time effects as a control. This is in line with results from the models considered for the Naive Ensemble, where peer-group equity time effects generally do not improve portfolio returns.

Regarding the cointegration model category, we show this category of models is not feasible in a CSA factor investing strategy. 1) A cointegration requirement reduces the available bond universe by 70 to 80%, driven by a lack of non-stationary time series. 2) Return level models do not provide better returns on a cointegrated subset, contrary to CDS-equity trading in Lovreta and Mladenović (2018). 3) A higher cointegration rate for lower-rated bonds does not lead to higher portfolio returns of the cointegrated subset compared to the non-cointegrated subset. 4) ECMs with or without cointegration requirements do not provide monotonic portfolio returns, are not superior to the more simple Returns Model and are therefore not included in the CSA framework. Contrary to pairs trading in the equity market (Rad et al., 2016), cointegration methods are not promising for a CSA factor investing strategy.

Finally, the motivation of applying Ornstein-Uhlenbeck is the possibility to explicitly model mean reversion with the k parameter of the process. We show for Ornstein-Uhlenbeck models that the mean-reversion parameter k in the process does not drive the significant improvement in the portfolio returns of the Ornstein-Uhlenbeck models relative to the return level models. Specifically, the mean-reversion parameter k constrains the total mispricing that the Ornstein-Uhlenbeck model captures. The performance improvement from Ornstein-Uhlenbeck is due to the focus of the process on the complete mispricing in the lookback window compared to only the most recent mispricing for the return level models. Modifying the Ornstein-Uhlenbeck process from Avellaneda and Lee (2010) we show promising results for CSA factor investing in corporate bonds.

This paper focuses on a factor investing strategy where bonds are ranked in the cross-section. We do not investigate the time-varying factor of arbitrage opportunities. Wojtowicz (2014) shows that for a CSA strategy with CDS-equity pair trading, most of the profits are made in a short time span because most CSA opportunities occur in a small time window. Further research may indicate whether a CSA strategy based on absolute signals instead of relative signals is more profitable than the factor investment strategy researched in this paper.

Within an ensemble, we show that fixed equity-credit correlation models outperform estimated equity-credit correlation models. The fact that there is a difference between fixed and estimated correlation is logical as the correlation between equity and credits is noisy, and a strong prior may counter some of the noise. However, a correlation equal to one is five to ten times higher than the observed correlation, which may be steep. Therefore, understanding why a fixed correlation ensemble shows higher returns than an estimated correlation ensemble may provide an opportunity to improve the portfolio returns shown in this paper. Another opportunity are frequency domain models. Recent developments for statistical arbitrage in equity markets have shown promising results in this direction (Guijarro-Ordonez et al., 2021). An initial investigation for corporate bond markets does not show promising results though.

Ultimately, this paper shows that we can harvest CSA opportunities using bond and equity returns with a factor investing portfolio investment strategy in credit and equity markets.

References

Ambastha, M., Dor, A. B., Dynkin, L., Hyman, J., and Konstantinovsky, V. (2010). Empirical duration of corporate bonds and credit market segmentation. *The Journal of Fixed Income*, 20(1):5–27.

Avellaneda, M. and Lee, J.-H. (2010). Statistical arbitrage in the us equities market. *Quantitative Finance*, 10(7):761–782.

Bai, J., Bali, T. G., and Wen, Q. (2019). Common risk factors in the cross-section of corporate bond returns. *Journal of Financial Economics*, 131(3):619–642.

Bajlum, C. and Larsen, P. T. (2008). Capital structure arbitrage: model choice and volatility calibration. *Available at SSRN 956839*.

Bali, T. G., Goyal, A., Huang, D., Jiang, F., and Wen, Q. (2020). Predicting corporate bond returns: Merton meets machine learning. *Georgetown McDonough School of Business Research Paper*, (3686164):20–110.

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

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.

Bektić, D., Wenzler, J.-S., Wegener, M., Schiereck, D., and Spielmann, T. (2019). Extending fama–french factors to corporate bond markets. *The Journal of Portfolio Management*, 45(3):141–158.

Bessembinder, H., Kahle, K. M., Maxwell, W. F., and Xu, D. (2008). Measuring abnormal bond performance. *The Review of Financial Studies*, 22(10):4219–4258.

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

Blitz, D. (2012). Strategic allocation to premiums in the equity market. *The Journal of Beta Investment Strategies*, 2(4):42–49.

Bluhm, C., Overbeck, L., and Wagner, C. (2001). The creditgrades model: A new framework for large portfolio default risk management. *Journal of Risk*, 4(2):1–25.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological Methods & Research*, 33(2):261–304.

Byström, H. N. (2006). Merton unraveled: A flexible way of modeling default risk. *The Journal of Alternative Investments*, 8(4):39.

Campbell, J. Y. (1995). Some lessons from the yield curve. *Journal of Economic Perspectives*, 9(3):129–152.

Chan, K. K., Kolokolova, O., Lin, M.-T., and Poon, S.-H. (2023). Price convergence between credit default swap and put option: New evidence. *Journal of Empirical Finance*.

Chen, L., Lesmond, D. A., and Wei, J. (2007). Corporate yield spreads and bond liquidity. *The Journal of Finance*, 62(1):119–149.

Chen, Z., Lookman, A. A., Schürhoff, N., and Seppi, D. J. (2014). Rating-based investment practices and bond market segmentation. *The Review of Asset Pricing Studies*, 4(2):162–205.

Choi, J. and Kim, Y. (2018). Anomalies and market (dis) integration. *Journal of Monetary Economics*, 100:16–34.

Correia, M., Richardson, S., and Tuna, İ. (2012). Value investing in credit markets. *Review of Accounting Studies*, 17:572–609.

De Jong, F. and Driessen, J. (2012). Liquidity risk premia in corporate bond markets. *The Quarterly Journal of Finance*, 2(02):1250006.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.

Duarte, J., Longstaff, F. A., and Yu, F. (2007). Risk and return in fixed-income arbitrage: Nickels in front of a steamroller? *The Review of Financial Studies*, 20(3):769–811.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 569–593. Springer.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, pages 251–276.

Eom, Y. H., Helwege, J., and Huang, J.-z. (2004). Structural models of corporate bond pricing: An empirical analysis. *The Review of Financial Studies*, 17(2):499–544.

Fama, E. F. and French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25(1):23–49.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Finkelstein, V., Pan, G., JPMorgan, G. S., Lardy, J.-P., Ta, T., Tierney, J., and Bank, D. (2002). Christopher c. finger riskmetrics group, inc. chris. finger@ riskmetrics. com.

Frazzini, A. and Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, 111(1):1–25.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Friewald, N., Wagner, C., and Zechner, J. (2014). The cross-section of credit risk premia and equity returns. *The Journal of Finance*, 69(6):2419–2469.

Gatev, E., Goetzmann, W. N., and Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3):797–827.

Gebhardt, W. R., Hvidkjaer, S., and Swaminathan, B. (2005). Stock and bond market interaction: Does momentum spill over? *Journal of Financial Economics*, 75(3):651–690.

Goodwin, T. H. (1998). The information ratio. *Financial Analysts Journal*, 54(4):34–43.

Granger, C. W. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1):121–130.

Gregory, A. W. and Hansen, B. E. (1996). Residual-based tests for cointegration in models with regime shifts. *Journal of Econometrics*, 70(1):99–126.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Guijarro-Ordonez, J., Pelger, M., and Zanotti, G. (2021). Deep learning statistical arbitrage. *arXiv preprint arXiv:2106.04028*.

Haesen, D., Houweling, P., and van Zundert, J. (2017). Momentum spillover from stocks to corporate bonds. *Journal of Banking & Finance*, 79:28–41.

Harri, A. and Brorsen, B. W. (1998). The overlapping data problem. *Available at SSRN 76460*.

Haugen, R. A. and Heins, A. J. (1972). On the evidence supporting the existence of risk premiums in the capital market. *Available at SSRN 1783797*.

Henke, H., Kaufmann, H., Messow, P., and Fang-Klingler, J. (2020). Factor investing in credit. *The Journal of Beta Investment Strategies*, 11(1):33–51.

Hilscher, J., Pollet, J. M., and Wilson, M. (2015). Are credit default swaps a sideshow? evidence that information flows from equity to cds markets. *Journal of Financial and Quantitative Analysis*, 50(3):543–567.

Houweling, P. and Van Zundert, J. (2017). Factor investing in the corporate bond market. *Financial Analysts Journal*, 73(2):100–115.

Huij, J., Lansdorp, S., Blitz, D., and van Vliet, P. (2014). Factor investing: Long-only versus long-short. *Available at SSRN 2417221*.

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

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, pages 1551–1580.

Johansen, S. and Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration—with appucations to the demand for money. *Oxford Bulletin of Economics and Statistics*, 52(2):169–210.

Jostova, G., Nikolova, S., Philipov, A., and Stahel, C. W. (2013). Momentum in corporate bond returns. *The Review of Financial Studies*, 26(7):1649–1693.

Ju, H.-S., Chen, R.-R., Yeh, S.-K., and Yang, T.-H. (2015). Evaluation of conducting capital structure arbitrage using the multi-period extended geske–johnson model. *Review of Quantitative Finance and Accounting*, 44:89–111.

Kapadia, N. and Pu, X. (2012). Limited arbitrage between equity and credit markets. *Journal of Financial Economics*, 105(3):542–564.

Kelly, B. T. and Xiu, D. (2023). Financial machine learning. Technical report, National Bureau of Economic Research.

Koijen, R. S., Moskowitz, T. J., Pedersen, L. H., and Vrugt, E. B. (2018). Carry. *Journal of Financial Economics*, 127(2):197–225.

Korn, O., Möller, P. M., and Schwehm, C. (2022). Drawdown measures: Are they all the same? *The Journal of Portfolio Management*, 48(5):104–120.

Krauss, C. (2017). Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2):513–545.

Kwan, S. H. (1996). Firm-specific information and the correlation between individual stocks and bonds. *Journal of Financial Economics*, 40(1):63–80.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3):159–178.

Le, T. M. and Clarke, B. (2022). Model averaging is asymptotically better than model selection for prediction. *The Journal of Machine Learning Research*, 23(1):1463–1516.

Leary, M. T. and Roberts, M. R. (2014). Do peer firms affect corporate financial policy? *The Journal of Finance*, 69(1):139–178.

Leland, H. (2002). Predictions of expected default frequencies in structural models of debt. In *Venice Conference on Credit Risk*.

Leland, H. E. (2012). Predictions of default probabilities in structural models of debt. *The Credit Market Handbook: Advanced Modeling Issues*, pages 39–64.

Leland, H. E. and Toft, K. B. (1996). Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads. *The Journal of Finance*, 51(3):987–1019.

Liew, R. Q. and Wu, Y. (2013). Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19:12–30.

Lovreta, L. and Mladenović, Z. (2018). Do the stock and cds markets price credit risk equally in the long-run? *The European Journal of Finance*, 24(17):1699–1726.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Manning, M. J. (2004). Exploring the relationship between credit spreads and default probabilities.

Mao, W., Zhu, H., Wu, H., Lu, Y., and Wang, H. (2023). Forecasting and trading credit default swap indices using a deep learning model integrating merton and lstms. *Expert Systems with Applications*, 213:119012.

Martens, M., Beekhuizen, P., Duyvesteyn, J., and Zomerdijk, C. (2019). Carry investing on the yield curve. *Financial Analysts Journal*, 75(4):51–63.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449–470.

Ornstein, L. and Uhlenbeck, G. E. (1930). On the theory of the brownian motion. *Physical Review*, 36(5):823–841.

Ozeki, T., Umezawa, Y., Yamazaki, A., and Yoshikawa, D. (2011). An extension of creditgrades model approach with lévy processes. *Quantitative Finance*, 11(12):1825–1836.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Perman, R. (1991). Cointegration: an introduction to the literature. *Journal of Economic Studies*, 18(3).

Pole, A. (2011). *Statistical arbitrage: algorithmic trading insights and techniques*. John Wiley & Sons.

Rad, H., Low, R. K. Y., and Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance*, 16(10):1541–1558.

Schaefer, S. M. and Strebulaev, I. A. (2004). Capital structure arbitrage: An empirical investigation of risk, return and hedging.

Schaefer, S. M. and Strebulaev, I. A. (2008). Structural models of credit risk are useful: Evidence from hedge ratios on corporate bonds. *Journal of Financial Economics*, 90(1):1–19.

Schlitzer, G. (1995). Testing the stationarity of economic time series: further monte carlo evidence. *Ricerche Economiche*, 49(2):125–144.

Shapley, L. S. et al. (1953). *A value for n-person games*. Princeton University Press.

Shrestha, M. B. and Bhatta, G. R. (2018). Selecting appropriate methodological framework for time series data analysis. *The Journal of Finance and Data Science*, 4(2):71–89.

Spearman, C. (1961). The proof and measurement of association between two things.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

Velthuis, R. (2007). Capital structure arbitrage: Can debt-equity trading be profitable? Master's thesis, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, Netherlands. Supervisors: Dr. Patrick Houweling (Robeco Asset Management), Arco van Oord MSc (Erasmus University Rotterdam); Co-reader: Prof. dr. Dick van Dijk (Erasmus University Rotterdam).

Vidyamurthy, G. (2004). *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons.

Wang, W. (2011). Recovery and returns of distressed bonds in bankruptcy. *The Journal of Fixed Income*, 21(1):21.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107.

Weinstein, M. (1981). The systematic risk of corporate bonds. *Journal of Financial and Quantitative Analysis*, 16(3):257–278.

Wojtowicz, M. (2014). Capital structure arbitrage revisited. *Available at SSRN 2497162*.

Yeo, J. and Papanicolaou, G. (2017). Risk control of mean-reversion time in statistical arbitrage. *Risk and Decision Analysis*, 6(4):263–290.

Yu, F. (2006). How profitable is capital structure arbitrage? *Financial Analysts Journal*, 62(5):47–62.

A Data

Table 17: Data filtering steps and effect on available investment universe

Step	Description	Observations	Bonds	Companies
0	Complete data set	3,200,834	61,267	5,462
1	Bond constituent of US Investment Grade or High Yield Index	2,280,666	42,783	4,200
3	Bond status Issued	2,034,892	38,315	3,614
4	Bonds with Dollar Denominated Equity Returns	1,677,802	31,687	3,239
5	Rating below CCC	1,624,082	31,044	3,090
6	Remove distressed bonds	1,606,068	30,951	3,085
7	At least 3 years of bond-month observations	1,352,676	17,743	2,158

Note. Step 4 is effectively a constraint to public companies only, with equity traded on a US exchange to prevent exchange rate noise. Step 6: Distressed bonds are bonds with a price below 70% of their issue price. Step 7: We require 36-months of observations for the estimation window, this is to speed up the estimation process.

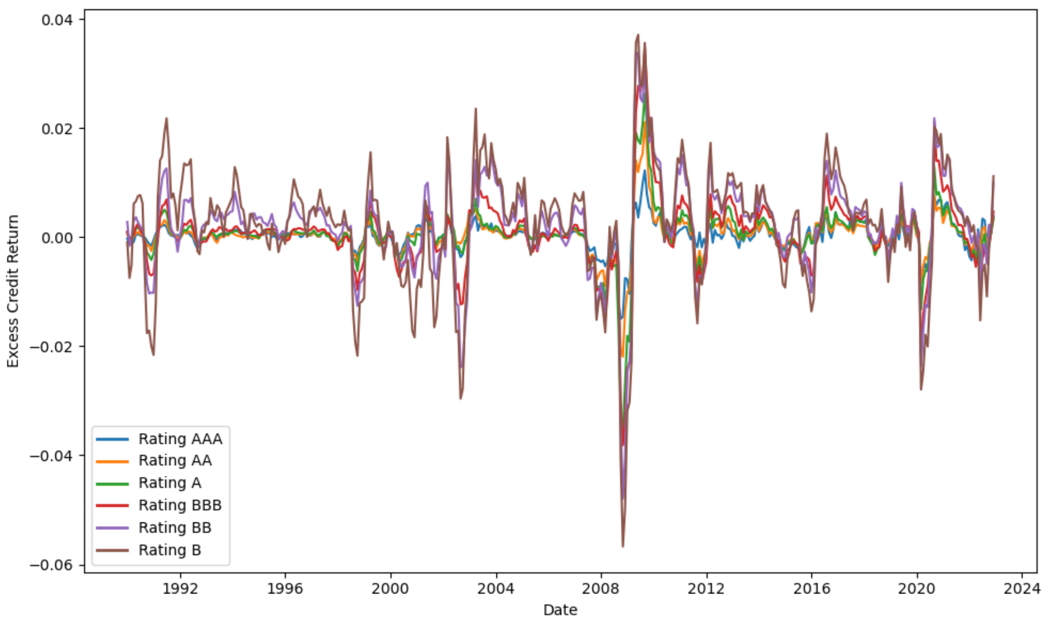


Figure 15: Credit returns per rating category over time

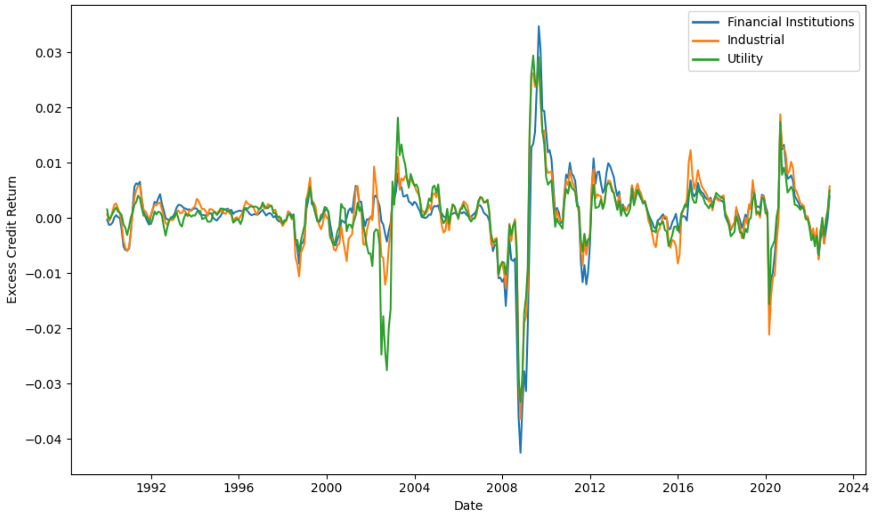


Figure 16: Credit returns per sector category over time

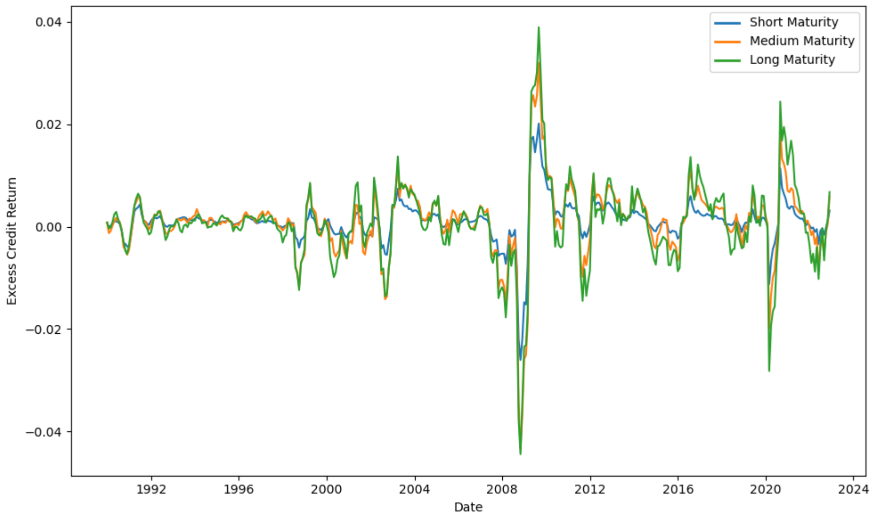


Figure 17: Credit returns per maturity category over time

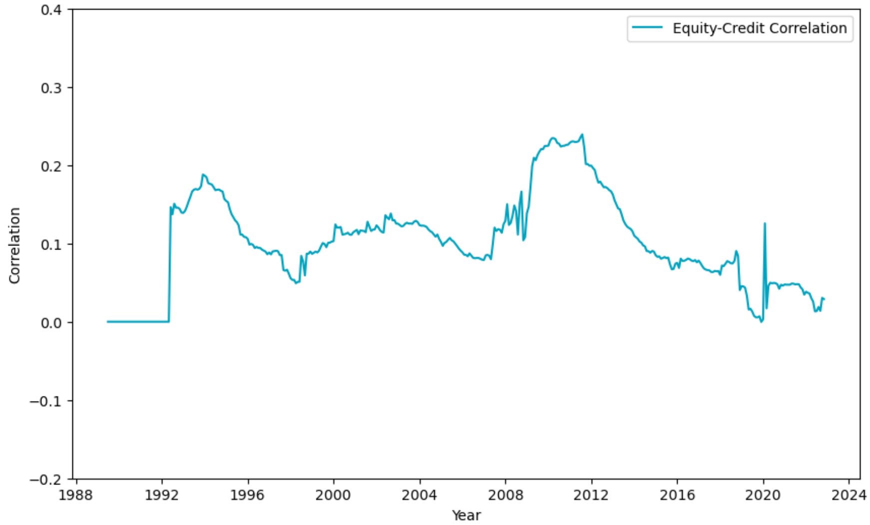


Figure 18: Bond value-weighted correlation between credit and equity returns

B Methodology

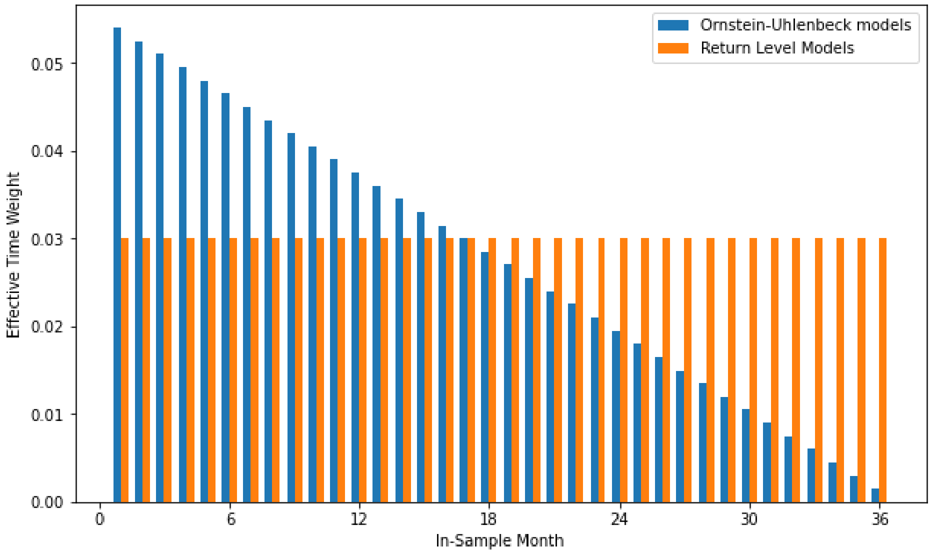
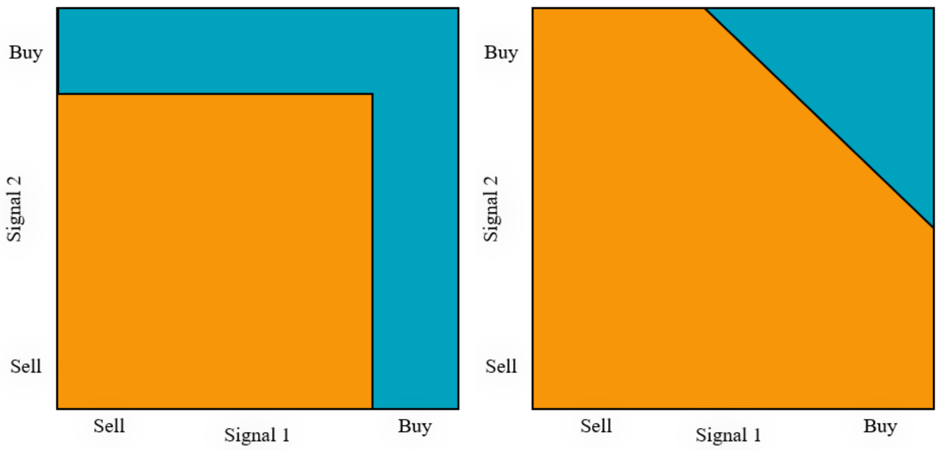


Figure 19: Effective time weight example for a return level and Ornstein-Uhlenbeck model with a 36-month estimation and lookback window



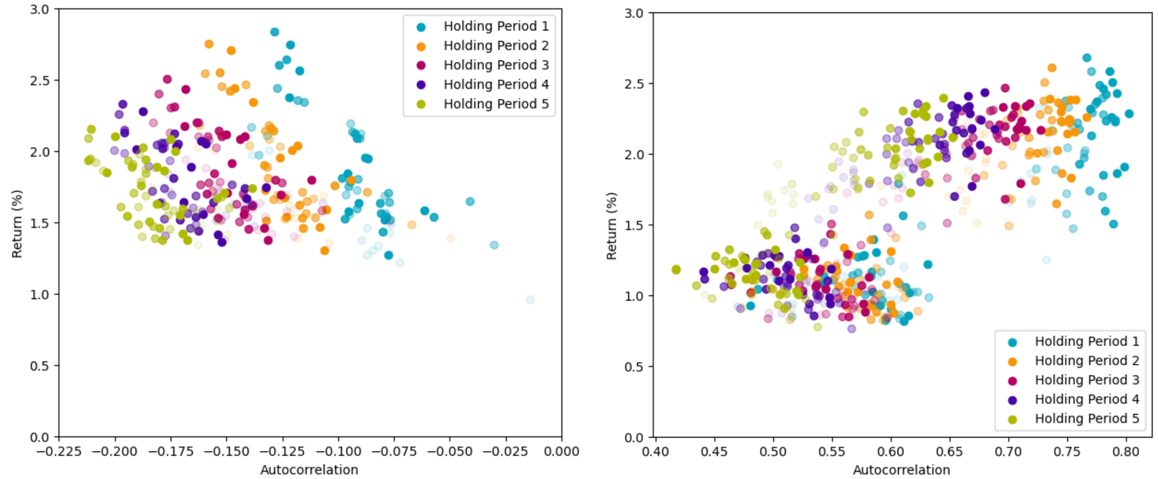
(a) Portfolio Blend

(b) Signal Blend

Figure 20: Illustrative example of portfolio and signal blend for two uncorrelated signals

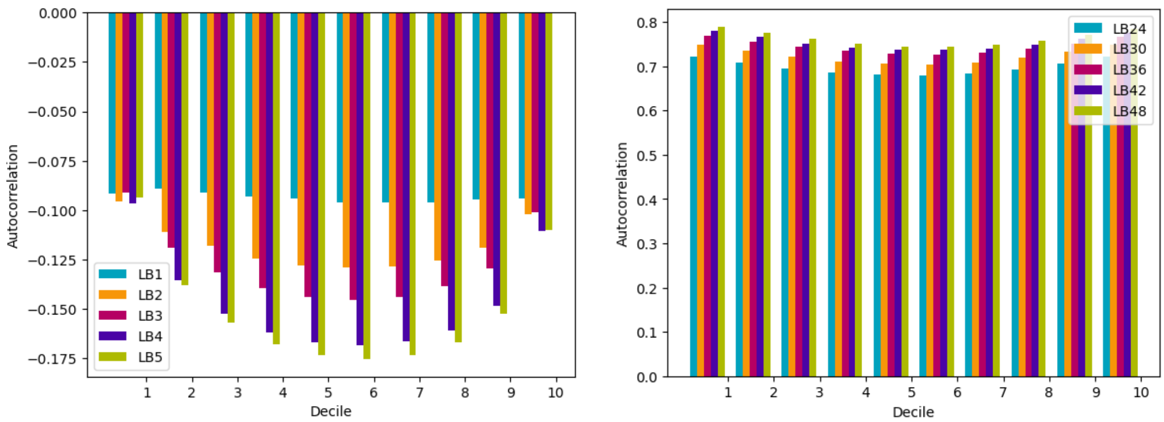
Adapted Source: Henke et al., 2020.

C Capital Structure Arbitrage Opportunities



(a) Return level models (b) Ornstein-Uhlenbeck models

Figure 21: Results on the relation between top decile returns and autocorrelation for return level and Ornstein-Uhlenbeck models, transparency reflects the lookback window; one month is most transparent to five months, least transparent



(a) Return level models (b) Ornstein-Uhlenbeck models

Figure 22: Autocorrelation across deciles with various lookback windows

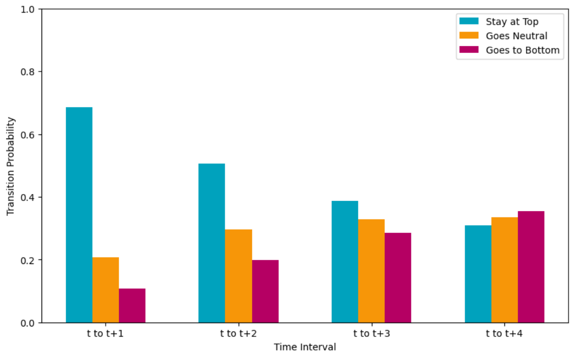
Table 18: Average spanning alpha over lookback windows with 1-month holding period

Alpha	Z-diff	Z-diff DM	Z-diff d	Returns	Returns DM	Returns RD	Boosting	Random Forest	RD dDtD	B dDtD	RF dDtD
Return Models	1.57%	1.60%	1.48%	0.93%	0.92%	0.55%	0.63%	0.77%	0.63%	1.00%	0.80%
OU Models	2.18%	1.76%	1.73%	1.69%	1.32%	1.11%	1.14%	1.35%	0.99%	1.09%	1.00%

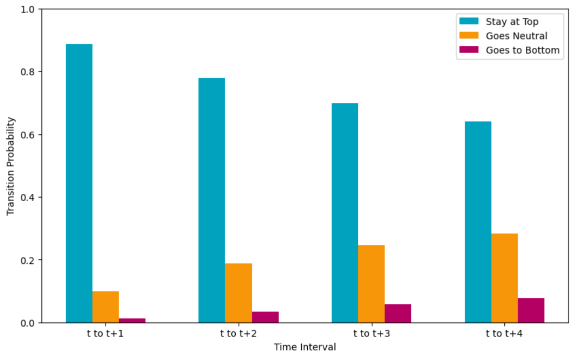
Table 19: Average spanning alpha over lookback windows and 1 to 5-month holding period

Alpha	Z-diff	Z-diff DM	Z-diff d	Returns	Returns DM	Returns RD	Boosting	Random Forest	RD dDtD	B dDtD	RF dDtD
Return Models	1.39%	1.38%	1.30%	0.91%	0.86%	0.54%	0.62%	0.78%	0.61%	0.90%	0.73%
OU Models	1.96%	1.53%	1.52%	1.58%	1.20%	1.04%	1.05%	1.28%	0.92%	0.99%	0.93%

D Drivers of Arbitrage



(a) Returns Ensemble



(b) Ornstein-Uhlenbeck Ensemble

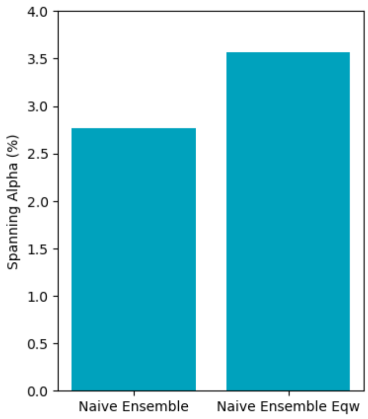
Figure 23: Transition probabilities

Note: Stay at top are deciles 1, 2, 3, Goes Neutral are deciles 4, 5, 6, 7, Goes to Bottom are deciles 8, 9, 10.

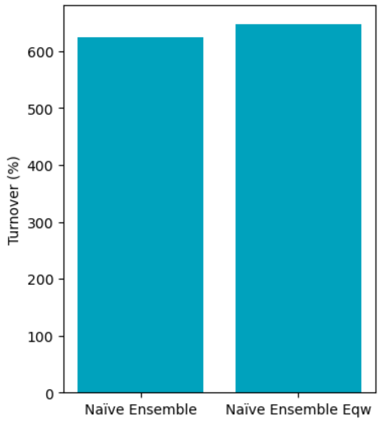
Table 20: Signal rank correlation return level models (1-month lookback window) and Ornstein-Uhlenbeck models (36-month lookback window)

Model	**Z-diff 1m**	**Returns 1m**	**Z-diff DM 1m**	**Returns DM 1m**	**Z-diff d 1m**	**RD 1m**	**B 1m**	**RF 1m**	**RD dDtD 1m**	**B dDtD 1m**	**RF dDtD 1m**	**OU Z-diff 36m**	**OU Returns 36m**	**OU Z-diff DM 36m**	**OU Returns DM 36m**	**OU Z-diff d 36m**	**OU RD 36m**	**OU B 36m**	**OU RF 36m**	**OU RD dDtD 36m**	**OU B dDtD 36m**	**OU RF dDtD 1m**
Z-diff 1m	1.00	0.61	0.84	0.45	0.84	0.39	0.44	0.46	0.39	0.45	0.46	0.30	0.16	0.25	0.11	0.25	0.10	0.11	0.12	0.10	0.12	0.12
Returns 1m	0.61	1.00	0.51	0.71	0.50	0.70	0.78	0.83	0.70	0.80	0.82	0.22	0.30	0.17	0.21	0.16	0.21	0.23	0.25	0.20	0.23	0.24
Z-diff DM 1m	0.84	0.51	1.00	0.60	0.94	0.51	0.52	0.50	0.51	0.49	0.48	0.25	0.13	0.31	0.16	0.29	0.14	0.14	0.13	0.14	0.13	0.13
Returns DM 1m	0.45	0.71	0.60	1.00	0.56	0.85	0.86	0.82	0.85	0.80	0.79	0.15	0.21	0.21	0.30	0.20	0.26	0.27	0.25	0.26	0.23	0.23
Z-diff d 1m	0.84	0.50	0.94	0.56	1.00	0.58	0.52	0.48	0.58	0.49	0.47	0.25	0.12	0.29	0.14	0.30	0.15	0.14	0.12	0.15	0.13	0.12
RD 1m	0.39	0.70	0.51	0.85	0.58	1.00	0.89	0.82	1.00	0.82	0.78	0.13	0.20	0.18	0.25	0.20	0.29	0.27	0.24	0.29	0.23	0.22
B 1m	0.44	0.78	0.52	0.86	0.52	0.89	1.00	0.92	0.89	0.91	0.87	0.15	0.23	0.18	0.26	0.18	0.27	0.30	0.27	0.27	0.26	0.25
RF 1m	0.46	0.83	0.50	0.82	0.48	0.82	0.92	1.00	0.82	0.88	0.91	0.16	0.25	0.17	0.25	0.16	0.25	0.28	0.30	0.25	0.26	0.27
RD dDtD 1m	0.39	0.70	0.51	0.85	0.58	1.00	0.89	0.82	1.00	0.82	0.78	0.13	0.20	0.18	0.25	0.20	0.29	0.26	0.24	0.29	0.23	0.22
B dDtD 1m	0.45	0.80	0.49	0.80	0.49	0.82	0.91	0.88	0.82	1.00	0.91	0.14	0.23	0.16	0.24	0.16	0.25	0.27	0.25	0.25	0.30	0.27
RF dDtD 1m	0.46	0.82	0.48	0.79	0.47	0.78	0.87	0.91	0.78	0.91	1.00	0.15	0.24	0.15	0.23	0.14	0.23	0.25	0.26	0.23	0.27	0.30
OU Z-diff 36m	0.30	0.22	0.25	0.15	0.25	0.13	0.15	0.16	0.13	0.14	0.15	1.00	0.50	0.74	0.29	0.73	0.25	0.29	0.33	0.25	0.31	0.32
OU Regr. Returns 36m	0.16	0.30	0.13	0.21	0.12	0.20	0.23	0.25	0.20	0.23	0.24	0.50	1.00	0.34	0.58	0.32	0.58	0.66	0.73	0.58	0.66	0.69
OU Z-diff DM 36m	0.25	0.17	0.31	0.21	0.29	0.18	0.18	0.17	0.18	0.16	0.15	0.74	0.34	1.00	0.49	0.92	0.40	0.41	0.37	0.40	0.36	0.34
OU Returns DM 36m	0.11	0.21	0.16	0.30	0.14	0.25	0.26	0.25	0.25	0.24	0.23	0.29	0.58	0.49	1.00	0.44	0.83	0.84	0.76	0.83	0.72	0.67
OU Z-diff d 36m	0.25	0.16	0.29	0.20	0.30	0.20	0.18	0.16	0.20	0.16	0.14	0.73	0.32	0.92	0.44	1.00	0.47	0.40	0.34	0.47	0.36	0.32
OU RD 36m	0.10	0.21	0.14	0.26	0.15	0.29	0.27	0.25	0.29	0.25	0.23	0.25	0.58	0.40	0.83	0.47	1.00	0.87	0.76	1.00	0.75	0.67
OU B 36m	0.11	0.23	0.14	0.27	0.14	0.27	0.30	0.28	0.26	0.27	0.25	0.29	0.66	0.41	0.84	0.40	0.87	1.00	0.86	0.87	0.84	0.77
OU RF 36m	0.12	0.25	0.13	0.25	0.12	0.24	0.27	0.30	0.24	0.25	0.26	0.33	0.73	0.37	0.76	0.34	0.76	0.86	1.00	0.75	0.79	0.83
OU RD dDtD 36m	0.10	0.20	0.14	0.26	0.15	0.29	0.27	0.25	0.29	0.25	0.23	0.25	0.58	0.40	0.83	0.47	1.00	0.87	0.75	1.00	0.75	0.67
OU B dDtD 36m	0.12	0.23	0.13	0.23	0.13	0.23	0.26	0.26	0.23	0.30	0.27	0.31	0.66	0.36	0.72	0.36	0.75	0.84	0.79	0.75	1.00	0.85
OU RF dDtD 36m	0.12	0.24	0.13	0.23	0.12	0.22	0.25	0.27	0.22	0.27	0.30	0.32	0.69	0.34	0.67	0.32	0.67	0.77	0.83	0.67	0.85	1.00

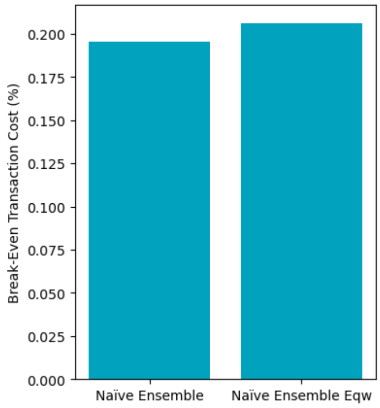
E Robustness of Findings



(a) Spanning alpha



(b) Turnover



(c) Break-even transaction cost

Figure 24: Performance Naive Ensemble and Naive Ensemble Eqw for a 1-month holding period

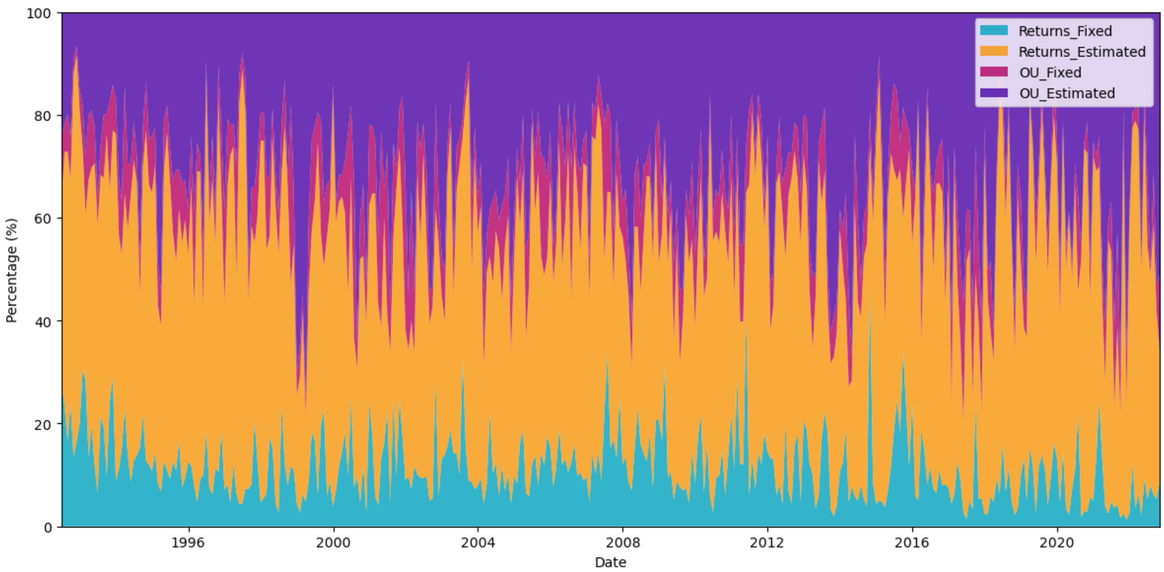
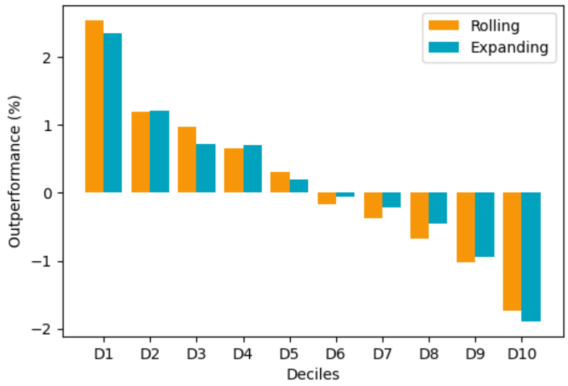
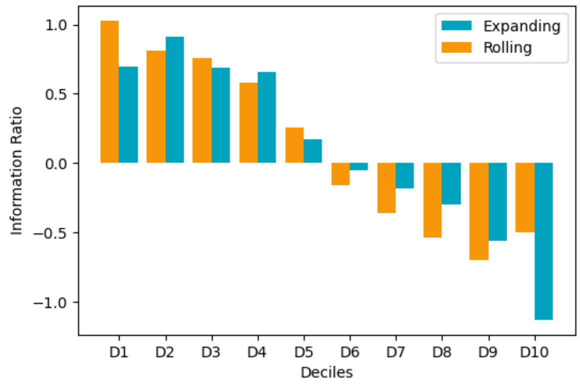


Figure 25: Random Forest model selection of return level and Ornstein-Uhlenbeck models with fixed and estimated correlation, based on SHAP values as fraction of total sum of SHAP values



(a) Outperformance



(b) Information ratio

Figure 26: Random Forest performance results for rolling and expanding window

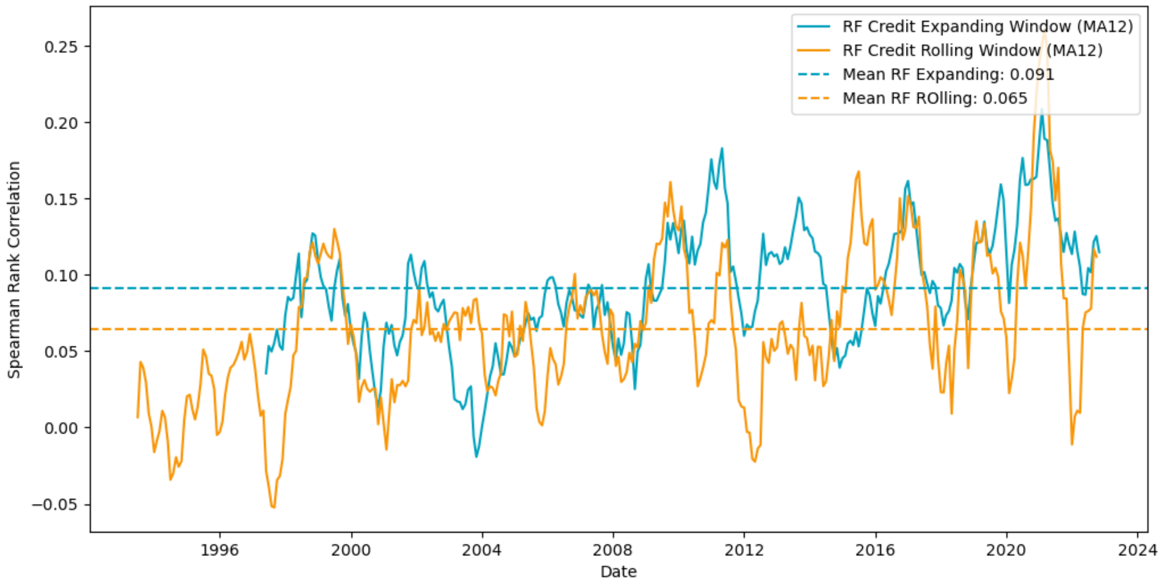
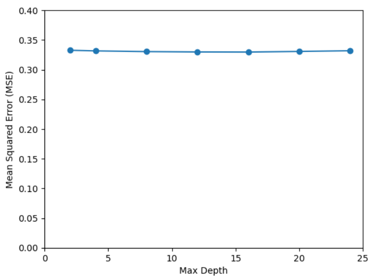
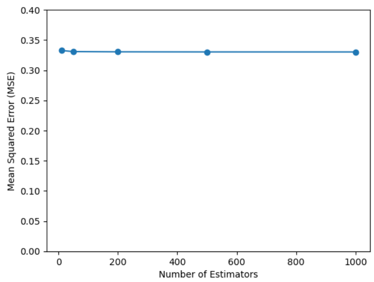


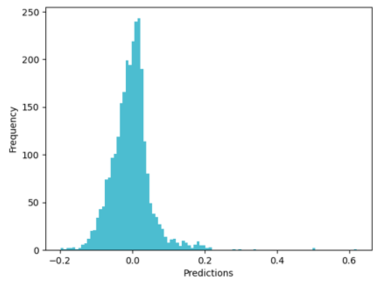
Figure 27: Random Forest Spearman rank results for a rolling and expanding window



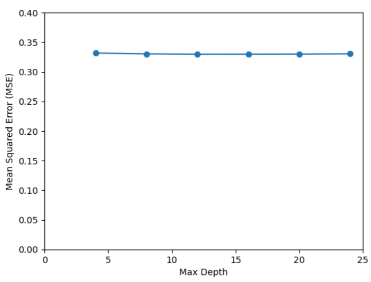
(a) Standardized depth



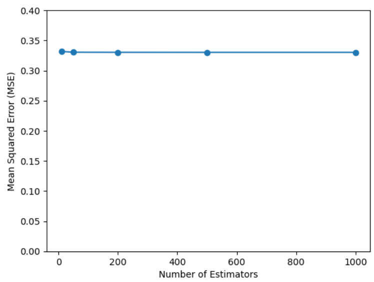
(b) Standardized n trees



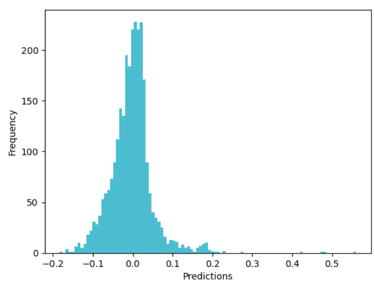
(c) Standardized prediction



(d) Rank standardized depth



(e) Rank standardized n trees



(f) Rank standardized prediction

Figure 28: Effect of (rank) standardization and hyperparameters on loss

F Naive Ensemble Eqw

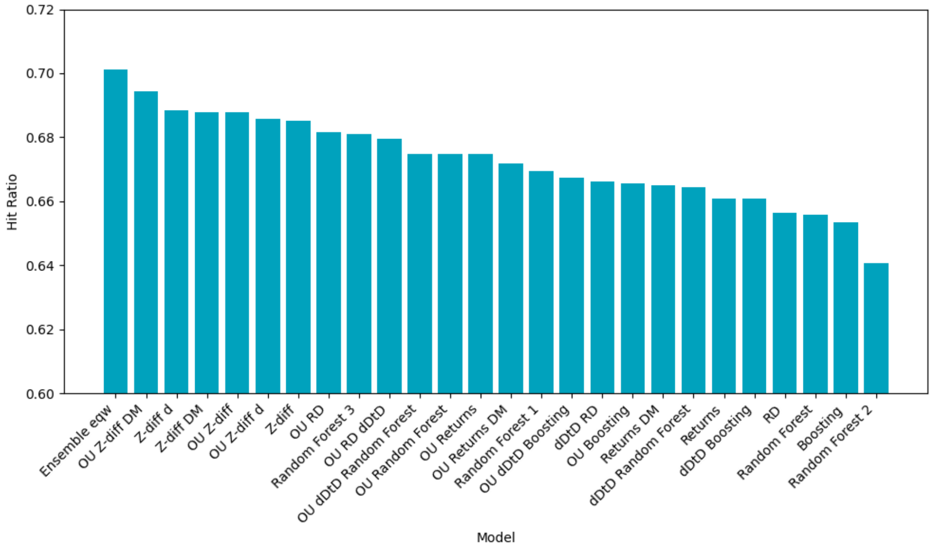


Figure 29: Hit ratio results for ensembles and individual models, averaged over lookback windows

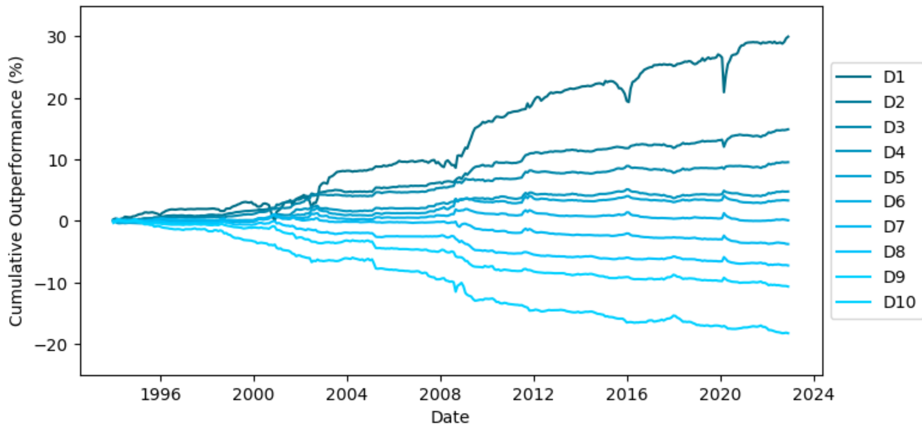


Figure 30: Cumulative outperformance Ensemble Eqw for 12-month holding period

Table 21: Spanning regression alpha results for top decile factor over the market

Top	Value	EquityMomentum	Size	CSA	LowRisk	CreditMomentum	EquityReversal	CreditReversal
Alpha	2.85%**	1.62%*	1.36%*	1.34%*	0.64%**	0.11%	-2.16%	-24.39%**
	(3.09)	(2.32)	(2.33)	(2.23)	(2.73)	(0.12)	(-1.90)	(-13.42)
Market	0.20**	0.15**	0.08	1.42**	0.11**	0.16*	0.25**	0.13
	(3.02)	(2.96)	(1.94)	(33.10)	(6.27)	(2.46)	(3.08)	(1.02)
R-squared	0.03	0.02	0.01	0.76	0.10	0.02	0.03	0.00

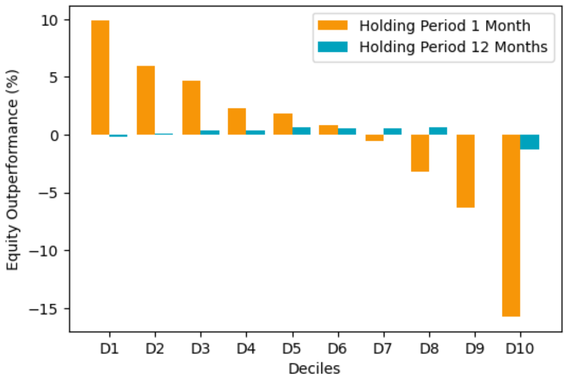
Note. T-statistics are in parentheses; * $p < 0.05$, ** $p < 0.01$.

G Model Performance High Yield

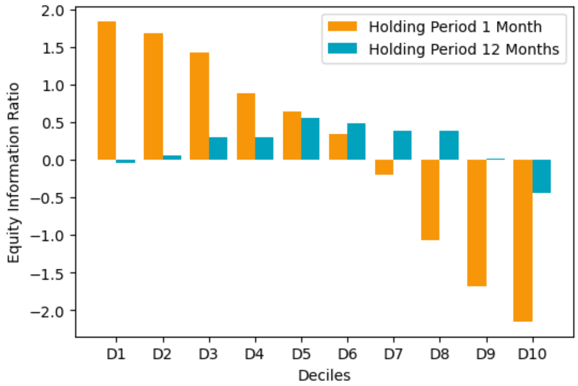
Table 22: Ensemble and Z-diff performance metric results for quintile portfolios and 1-month holding period in the High-Yield investment domain

	Reference Model	Ensemble Models										
		Naïve Model Selection		Random Forest Model Selection			Ornstein-Uhlenbeck		Returns		Ornstein-Uhlenbeck + Returns	
High Yield	**Z-diff**	**Naive Ensemble**	**Naive Ensemble Eqw**	**1-month Target**	**2-month Target**	**3-month Target**	**Fixed Corr.**	**Estimated Corr.**	**Fixed Corr.**	**Estimated Corr.**	**Fixed Corr.**	**Estimated Corr.**
Avg. Annualized Return	3.79%	4.16%	3.78%	5.29%	5.18%	4.61%	4.51%	3.97%	4.50%	2.98%	4.57%	3.24%
Std. Dev	7.94%	10.32%	10.05%	7.10%	7.76%	8.17%	7.43%	10.75%	8.81%	10.74%	8.11%	10.78%
Sharpe	0.48	0.40	0.38	0.74	0.67	0.56	0.61	0.37	0.51	0.28	0.56	0.30
Outperformance	0.71%	1.08%	0.70%	2.21%	2.10%	1.53%	1.43%	0.89%	1.42%	-0.10%	1.49%	0.16%
TE	5.34%	4.34%	5.13%	5.12%	5.13%	5.11%	4.93%	4.81%	5.34%	5.07%	5.36%	5.37%
IR	0.13	0.25	0.14	0.43	0.41	0.30	0.29	0.18	0.27	-0.02	0.28	0.03
BE TC	0.18%	0.34%	0.30%	0.30%	0.31%	0.29%	0.50%	0.44%	0.34%	0.21%	0.39%	0.27%

H Model Performance Equity Markets



(a) Outperformance



(b) Information ratio

Figure 31: Performance results across deciles for Naive Ensemble Eqw in the equity market

I Return Level

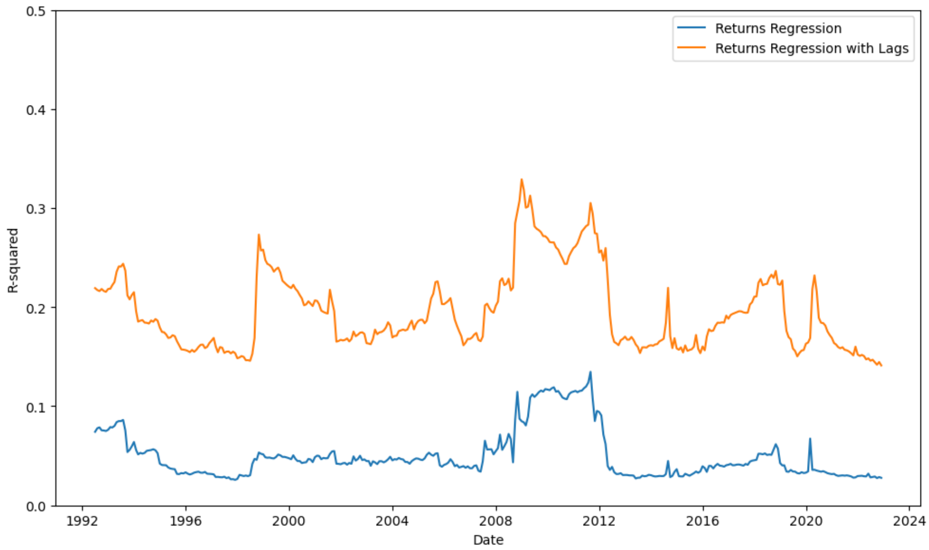


Figure 32: R-Squared for Returns Model and Returns Model with lags, 1992 - 2022

Table 23: Returns Model error exposure to ΔDtD

Model	**ΔDtD**
% Significant Beta	11%
Average R-squared:	0.03
Median R-squared:	0.01

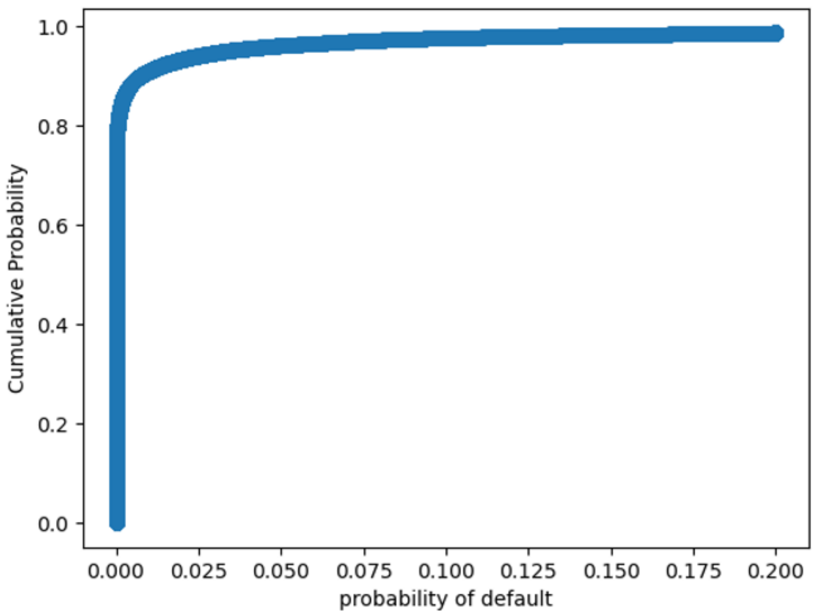


Figure 33: Cumulative default probability predicted by structural model for all bonds in the cross-section

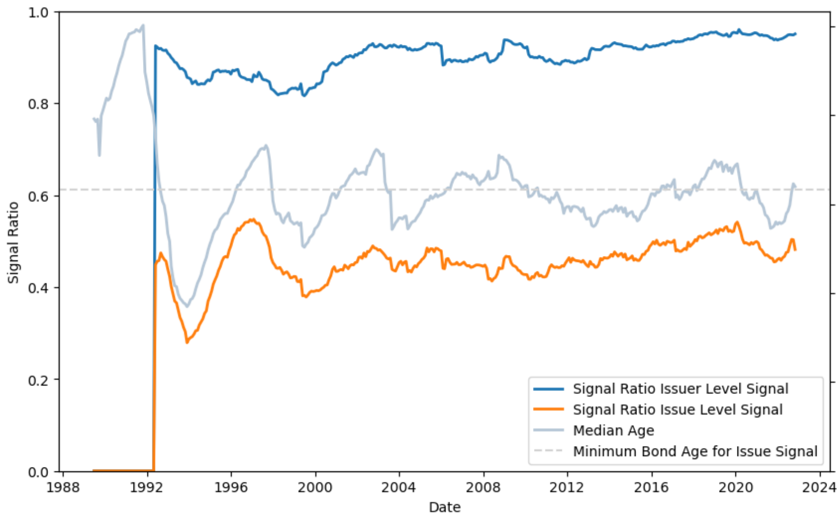


Figure 34: Ratio of non-NaN signal values for issuer- and issue-level signal with median bond age in the cross-section, 1992 - 2022

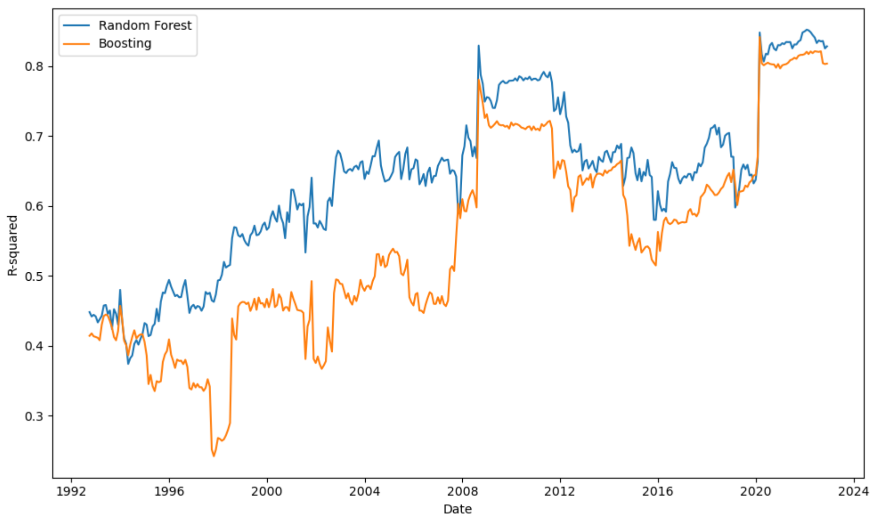


Figure 35: R-Squared for Boosting and Random Forest first pass with ΔDtD, 1992 - 2022

Table 24: First and second pass influence of interaction dummies and difference ex-ante and ex-post return cleaning of systematic risk on autocorrelation and R^2

Model	RD	RD ID	Returns	Returns RD	Returns RD ID	Returns RD ID Reversed
Autocorrelation	-0.070	-0.066	-0.014	-0.077	-0.073	-0.068
R-Squared	0.052	0.057	0.072	0.090	0.094	0.072

Table 25: First and second pass influence of demeaning credit and equity returns with and without sector bucket

Model	Credit	Credit and Equity	Credit with Sector	Credit and Equity with Sector
Autocorrelation	-0.073	-0.076	-0.082	-0.085
R-Squared	0.569	0.568	0.644	0.643

J Cointegration

Table 26: Non-stationary ratio according to ADF and KPSS statistics and cointegration ratio according to Johansen's trace test for OAS and share price on issue level and rolling basis

Issue Level Rolling	**Non-stationary ADF**	**Non-stationary KPSS**	**Case 1**	**Case 4**	**Case 1 & 4**	**Cointegrated with Prerequisite**	**Cointegrated without Prerequisite**
Alpha = 0.05							
OAS	82.6%	45.7%	42.5%	3.3%	45.7%		
Share Price	82.6%	48.1%	44.4%	3.8%	48.1%		
OAS & Share Price	68.7%	24.5%	21.2%	3.3%	24.5%	11.9%	52.0%
Alpha = 0.1							
OAS	75.1%	62.5%	53.8%	8.7%	62.5%		
Share Price	75.0%	65.4%	55.6%	9.8%	65.4%		
OAS & Share Price	57.1%	42.6%	31.8%	10.8%	42.6%	26.0%	64.1%

Note. Case 1 refers to non-stationary ADF and KPSS test statistic results. Case 4 refers to the ADF test result as stationary and the KPSS test result as non-stationary; The percentage is relative to the number of 36-month windows in the data; The results shown are for tests around a linear trend; the results are similar for tests around a constant.

Table 27: Non-stationary ratio according to ADF and KPSS statistics and cointegration ratio according to Johansen's trace test for price full return and share price on issue level for complete sample

Issue Level Complete Sample	**Non-stationary ADF**	**Non-stationary KPSS**	**Case 1**	**Case 4**	**Case 1 & 4**	**Cointegrated with Prerequisite**	**Cointegrated without Prerequisite**
Alpha = 0.05							
OAS	74.7%	58.4%	53.1%	5.3%	58.4%		
Share Price	77.5%	56.3%	52.4%	4.0%	56.3%		
OAS & Share Price	64.2%	36.6%	30.8%	5.7%	36.6%	14.8%	46.1%
Alpha = 0.1							
OAS	68.4%	70.7%	59.4%	11.3%	70.7%		
Share Price	71.1%	69.6%	59.6%	10.0%	69.6%		
OAS & Share Price	54.0%	54.6%	39.3%	15.3%	54.6%	24.4%	58.2%

Note. Case 1 refers to non-stationary ADF and KPSS test statistic results. Case 4 refers to the ADF test result as stationary and the KPSS test result as non-stationary; The percentage is relative to the number of time series; The results shown are for tests around a linear trend; the results are similar for tests around a constant.

Table 28: Non-stationary ratio according to ADF and KPSS statistics and cointegration ratio according to Johansen's trace test for OAS and share price on issuer level for complete sample and on rolling basis

Issuer Level Complete Sample	**Non-stationary ADF**	**Non-stationary KPSS**	**Case 1**	**Case 4**	**Case 1 & 4**	**Cointegrated with Prerequisite**	**Cointegrated without Prerequisite**
Alpha = 0.05							
OAS	74.8%	55.4%	47.4%	8.0%	55.4%		
Share Price	89.2%	72.8%	68.9%	3.9%	72.8%		
OAS & Share Price	67.3%	42.4%	34.2%	0.5%	34.7%	23.1%	70.3%
Alpha = 0.1							
OAS							
Share Price							
OAS & Share Price	54.9%	57.5%	38.0%	6.8%	44.8%	26.1%	78.6%
Issuer Level Rolling							
Alpha = 0.05							
OAS	0.0%	0.0%	0.0%	0.0%	0.0%		
Share Price	0.0%	0.0%	0.0%	0.0%	0.0%		
OAS & Share Price	68.9%	24.8%	21.4%	3.4%	24.8%	21.0%	50.0%
Alpha = 0.1							
OAS	0.0%	0.0%	0.0%	0.0%	0.0%		
Share Price	0.0%	0.0%	0.0%	0.0%	0.0%		
OAS & Share Price	57.5%	43.4%	32.3%	11.1%	43.4%	25.8%	62.0%

Note. Case 1 refers to non-stationary ADF and KPSS test statistic results. Case 4 refers to the ADF test result as stationary and the KPSS test result as non-stationary; The percentage for complete sample is relative to the number of time series; The percentage for rolling is relative to the number of rolling windows; The results shown are for tests around a linear trend; the results are similar for tests around a constant.

Table 29: Information ratio per quintile and decile for the Error Correction Models

Model	**Spread 1m**	**Spread 2m**	**Spread 3m**	**CI Spread 1m**	**CI Spread 2m**	**CI Spread 3m**	**CI Spread 1m Tot**	**CI Spread 2m Tot**	**CI Spread 3m Tot**
Q1	0.41	0.46	0.46	0.32	0.32	0.27	0.29	0.29	0.29
Q2	0.43	0.38	0.29	0.18	-0.09	0.09	0.26	0.26	0.01
Q3	0.14	0.08	-0.04	0.26	0.02	0.17	0.02	0.02	0.03
Q4	-0.02	-0.05	0.08	0.10	-0.02	0.12	-0.08	-0.08	-0.03
Q5	0.05	0.10	0.14	0.17	-0.01	0.18	0.03	0.03	0.13
D1	0.23	0.43	0.41	0.39	0.29	0.42	0.15	0.17	0.14
D2	0.45	0.34	0.40	0.18	0.25	0.06	0.26	0.16	0.37
D3	0.47	0.44	0.19	0.04	-0.02	-0.11	0.29	0.21	-0.12
D4	0.30	0.26	0.34	-0.05	-0.11	-0.06	0.15	0.02	0.13
D5	0.21	0.06	0.01	0.21	0.04	-0.08	0.05	-0.02	-0.14
D6	0.04	0.08	-0.08	0.13	-0.02	0.24	0.00	0.12	0.17
D7	0.00	0.02	0.12	-0.12	-0.07	0.14	-0.07	-0.03	-0.02
D8	-0.04	-0.11	0.02	-0.30	0.04	-0.20	-0.07	-0.04	-0.03
D9	0.11	0.15	0.14	0.10	0.07	-0.07	-0.05	0.06	0.11
D10	0.00	0.05	0.12	-0.03	-0.04	0.05	0.08	0.11	0.12

Table 30: Spanning regression alpha results for Returns Model over the market for different levels of cointegration

	Returns Model with LB 3		
Variable	Top	Middle	Bottom
Alpha	3.84%**	3.58%**	4.69%**
	(4.85)	(3.90)	(4.96)

Note. T-statistics are in parentheses; * p < 0.05, ** p < 0.01.

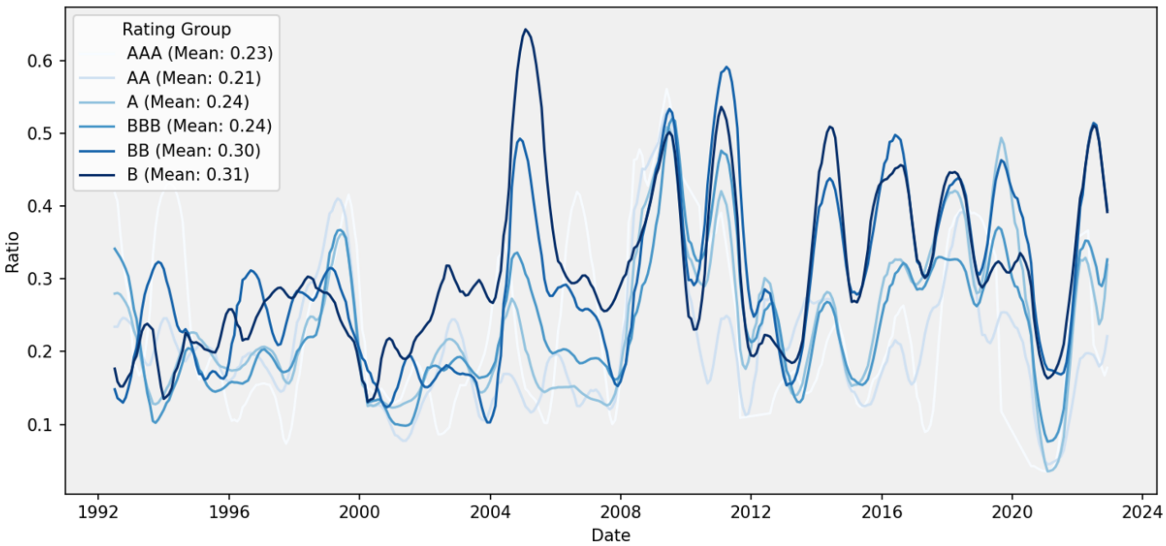


Figure 36: Cointegration ratio across rating groups for OAS and share price

Note. Cointegration ratio is on a rolling basis and issue level. The rating average is smoothed over a 6-month window, i.e. 6 observations.

K Ornstein-Uhlenbeck

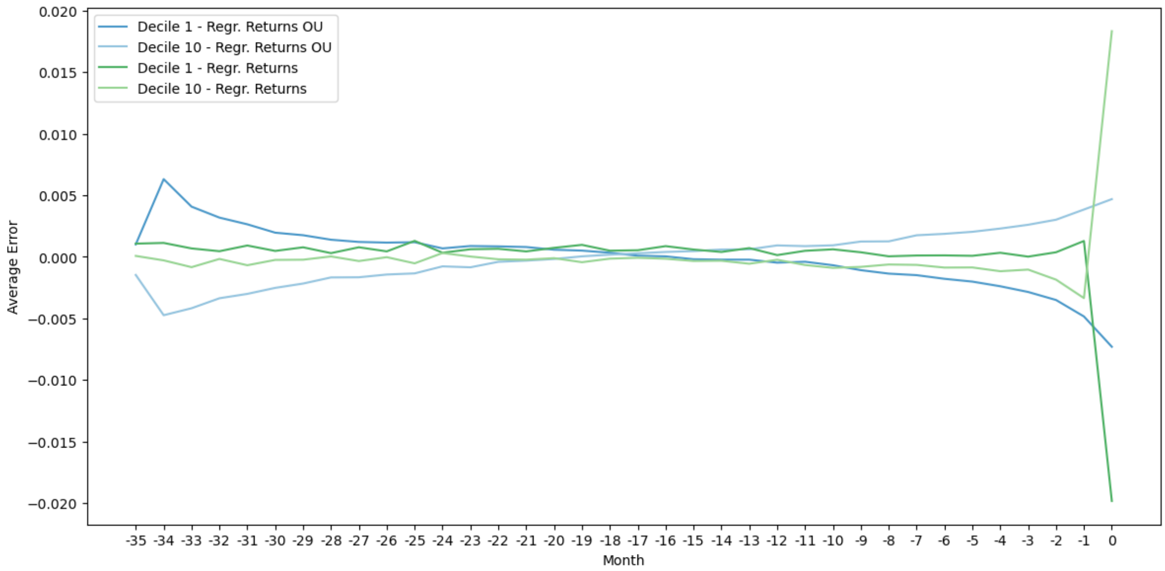


Figure 37: In-sample errors for top and bottom decile of Returns Model with and without OU process

Table 31: Spanning regression alpha t-statistic results for Model with OU process over Model without

	Raw		**Demeaned**		**Two-Pass**				**Two-Pass with ΔDtD**		
Model	**Z-diff**	**Returns**	**Z-diff**	**Returns**	**Z-diff**	**Returns**	**Boosting**	**Random Forest**	**Returns**	**Boosting**	**Random Forest**
2 < t-stat	84%	48%	92%	24%	52%	56%	12%	44%	30%	20%	32%
0 ≤ t-stat ≤ 2	16%	52%	8%	56%	48%	44%	88%	56%	70%	64%	52%
-2 ≤ t-stat < 0	0%	0%	0%	20%	0%	0%	0%	0%	0%	16%	16%
t-stat < -2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Note. The percentage is of all spanning regressions for lookback window and holding period of one to five months; the t-value is the alpha of the spanning regression of portfolio returns of a model with Ornstein-Uhlenbeck process over the portfolio returns of that model without Ornstein-Uhlenbeck process; 84% is for 21 out of 25 combinations the t-value of spanning alpha is above 2, significant at 5%.

Table 32: Spanning regression alpha t-statistic results for Model with OU Kappa sort over Model without

	Raw		**Demeaned**		**Two-Pass**				**Two-Pass with dDtD**		
Model	**Z-diff**	**Returns**	**Z-diff**	**Returns**	**Z-diff**	**Returns**	**Boosting**	**Random Forest**	**Returns**	**Boosting**	**Random Forest**
2 < t-stat	0%	72%	0%	0%	0%	0%	16%	8%	0%	0%	12%
0 ≤ t-stat ≤ 2	8%	28%	20%	80%	20%	72%	84%	92%	100%	100%	88%
-2 ≤ t-stat < 0	92%	0%	80%	20%	80%	28%	0%	0%	0%	0%	0%
t-stat < -2	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Note. The percentage is of all spanning regressions for lookback window and holding period of one to five months; the t-value is the alpha of the spanning regression of portfolio returns of a model with Ornstein-Uhlenbeck process, sorted by k, over the portfolio returns of that model without Ornstein-Uhlenbeck process; 84% is for 21 out of 25 combinations the t-value of spanning alpha is above 2, significant at 5%.

Table 33: Spanning regression results for Returns with OU process over Returns without OU, σ or κ, the market, and four common corporate bond factors

Variable	**Ornstein-Uhlenbeck**					**Kappa**		
Alpha	1.32%*	-0.37%	2.34%**	1.50%**	1.63%*	2.07%**	2.85%**	1.96%**
	(2.51)	-1.19	(3.10)	(2.67)	(2.52)	(3.14)	(5.04)	(3.16)
μ	0.47**							
	(19.21)							
μ / σ		0.90**						
		(41.64)						
κ			-0.20**					
			(-3.35)					
Ornstein-Uhlenbeck						-0.15**		
						(-3.35)		
Market				0.17	0.47**		-0.93**	-0.27**
				(1.14)	(11.55)		(-6.35)	(-6.80)
Size				-0.10			-0.10	
				(-1.54)			(-1.45)	
LowRisk				-0.01			-0.54**	
				(-0.22)			(-7.99)	
Value				-0.07			-0.09	
				(-0.78)			(-1.07)	
Credit Momentum				-0.49**			0.06	
				(-9.70)			(1.15)	
R-squared	0.52	0.83	0.03	0.48	0.28	0.03	0.31	0.12

Note. T-statistics are in parentheses; first five columns have OU portfolio returns as dependent variable; last three columns have kappa sorted portfolio returns as dependent variable; * p < 0.05, ** p < 0.01.

Table 34: Portfolio returns for kappa and mu double-sorted quintile portfolios

	Mu				
Kappa	1	2	3	4	5
1	0.10%	0.36%	0.27%	0.01%	0.18%
2	0.52%	0.55%	0.79%	0.89%	0.49%
3	1.04%	1.23%	1.04%	1.07%	0.67%
4	2.21%	1.49%	1.25%	1.10%	0.75%
5	2.81%	1.63%	1.48%	1.31%	0.86%