ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Forecasting Credit Default Swap spreads

Erasmus University - Master Quantitative Finance - Thesis

Sjoerd van Driesum (345342) *

Erasmus University - Econometric Institute

Marcin Jaskowski

Dick van Dijk

May 31, 2016

Abstract

This paper studies the predictability of 5-year Credit Default Swaps by constructing multiple forecasting models that incorporate a wide range of explanatory variables. A model that combines forecasts shows to be able to outperform benchmark models in terms of multiple forecast evaluation measures. Two 'simple' trading strategies based on buy/sell signals of this combination of forecasts model turn out to be highly profitable. This model combines the forecasts of ten models that incorporate a wide range of explanatory variables. Six of these models are based on three new variable reduction techniques. A new proposed factor based on the cross-sectional idiosyncratic risk in Credit Default Swap spreads shows to be an addition to the most widely used explanatory variable in forecasting credit related financial instruments. The results show the usefulness of combining forecasts of models based on new variable reduction techniques that incorporate a wide range of explanatory variables in forecasting Credit Default Swap spreads.

Keywords: Credit Default Swaps, forecasting, variable reduction techniques

JEL Classifications: C55, G12, G17

*The author thanks Marcin Jaskowski and Dick van Dijk for all helpful comments.

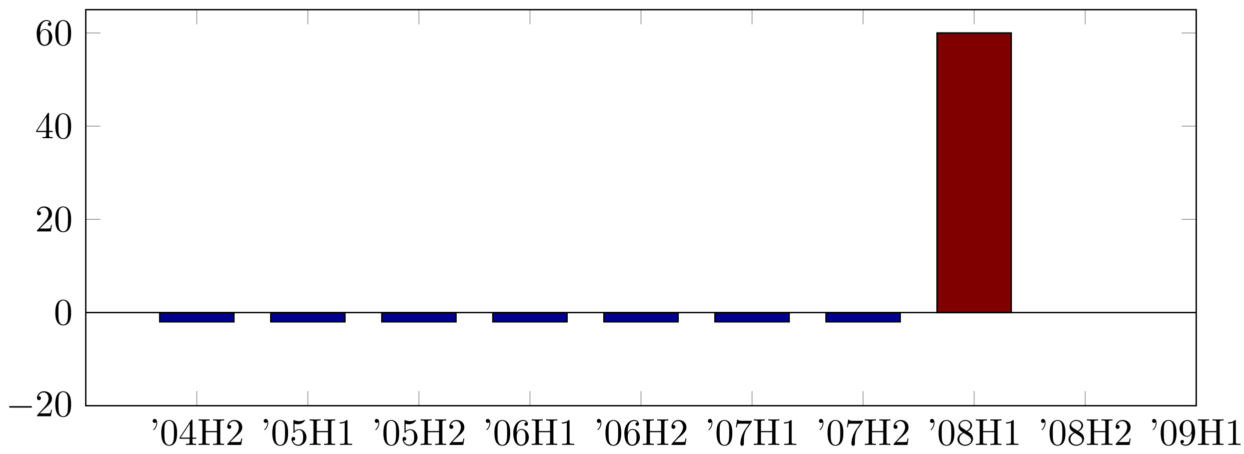
Contents

1 Introduction

Credit Default Swaps (CDSs) are credit derivative contracts designed to swap the credit exposure of fixed income products between two or more parties. This paper investigates the predictability of the CDS spread by means of thirteen different forecasting models over the period January 2002-February 2013. The first model uses an Autoregressive-Moving-Average ($ARMA$) approach and is the starting point for all other models. Nine models incorporate a large number of explanatory variables to forecast the CDS spread. Besides existing techniques to reduce the dimensionality of the explanatory variables, six of these nine models are based on three new techniques of variable reduction. The last three out of thirteen models are forecast combinations of the first ten models. One of the forecast combination models takes equally weights of the forecasts of those models that correctly predicted the direction of the CDS spread in the last period. This model outperforms the benchmark models in terms of all considered measures.

The CDS can be seen as an insurance against the default of the party that issued a fixed income product. Similarly to most insurance contracts, the buyer of the contract makes coupon payments to the seller until either the contract matures or the fixed income product issuer defaults. In return, the seller agrees to pay the buyer the premium of the fixed income product in case of a default. The premium is defined as the notional times one minus the recovery rate (RR). This rate is often equal to 40%. Figure 1 shows an example of the cash flows of the CDS buyer. In this 5-year contract the coupon is payed semi-annually until the issuer of the fixed income product defaults in the first half of 2008. The contract runs for another one and a half years, but because of the default the buyer of the contract stops making the coupon payments. The CDS spread represents the semi-annually coupon payments. In this example the CDS spread is equal to 2%: the buyer of the CDS contract has to pay 2% of the notional every 6 months to the seller of the CDS contract.

Figure 1: An example of the cash flows (in \$) of the buyer of a CDS contract. The blue bars represent payments from the buyer to the seller and the red bars vice versa. The notional of the contract is 100\$, starting date of the contract is '04H1 and matures at '09H1. Coupon payments of 2\$ are made semi-annually and the recovery rate (RR) equals 40%. The issuer of the fixed income product is defaulted in '08H1 resulting in the buyer receiving a payment of 60\$ (= notional × (1-RR)) in '08H1. No further coupon payments are made after the default.



A CDS is one of the most common credit derivatives and can involve many type of fixed income securities. The use of the swap gained much popularity in the years before the 2008 financial crisis. The volume of CDS based on notional value was about \$6 trillion at the end of 2004 and soared to \$58 trillion during the 2008 financial crisis. This volume fell to \$42 trillion one year later. The huge drop in volume indicates that the CDS was not only used to hedge credit risk, but investors did actively trade on these products with an incentive of making profits. Most CDS contracts are traded over the counter (OTC), rather than on an exchange. This makes it impossible to know how many CDS contracts a company has bought or sold. In case of a derivative traded OTC, the seller has to post collateral in case the default risk rises and in case its own credit rating worsens. One of the main examples during the financial crisis was American International Group (AIG). This company had sold roughly \$500 billion worth of CDSs OTC. The lack of transparency in the OTC market and default probabilities of the underlying fixed income product issuers going up made AIG facing a vicious cycle of massive losses, a potential downgrade of its credit rating and posting more collateral. Eventually, the US government had to intervene by making an \$85 billion loan to AIG to save it from bankruptcy.

After the crisis, there has been a large push by US regulatory and supervisory entities such as the Federal Reserve, the Securities and Exchange Commission (SEC) and the Commodity Futures Trading Commission (CFTC) to make the derivative market more transparent. The financial crisis and the resulting spate of regulations that followed made investors less interested in trading CDS contracts. The total notional amounts outstanding decreased to \$24 trillion by the first half of 2013.

The CDS spread represents a pure form of credit measure: when the probability of default goes up the seller of the contract wants to be compensated for the higher risk of paying out the premium. Therefore, the CDS spread increases, which translates in higher coupon payments to be made by the CDS buyer. The probability of default can be implied by multiple asset classes. Merton (1974) derives the implied probability of default using data on stock prices and company characteristics. Bharath and Shumway (2008) use this model to forecast default. Furthermore, another asset class that is often used to derive the probability of default is the corporate bond spread. This spread is often seen as a credit spread on top of a risk-free spread. Lin et al. (2015) find predictability of the bond spread by an array of 27 macro-economic, stock and bond predictors.

Most literature on CDS spreads is focused on constructing structural models and less on forecasting. Bai and Wu (2012) construct a structural model that explains the cross-sectional CDS variation and comes up with a way to forecast the CDS spread. They regress explanatory variables against the CDS spread in the cross-section and generate an impressive average R-squared of 77%. Therefore, I construct a new factor based on this cross-sectional approach and find it to be an addition to the already widely used explanatory variables in forecasting credit spreads.

Multiple research is based on the contribution of the price discovery of the CDSs and other asset classes. As mentioned before, the probability of default can be implied from the CDS, stock and bond market. Peña and Forte (2006) and Forte and Pena (2009) find the stock market is leading the CDS and bond market. This indicates that new information is incorporated in the stock market before the CDS and bond market. Du (2009) finds similar results and argues the CDS market is leading the bond market as well. Moreover, Fung et al. (2008) find that the lead-lag relationship between the stock market and CDS market is said to be dependent on the credit quality of the underlying reference entity. Forte and Lovreta (2009) find that at lower credit quality levels the relationship between the CDS market and stock market to be stronger.

The literature provides a wide range of explanatory variables that have been shown to be informative for the company's credit risk. As mentioned before, the CDS spread represents a form of the company's credit risk. Therefore, these explanatory variables should be helpful in forecasting the CDS spread. This list consists mainly of data on fundamental ratios (Altman (1968), Altman (1989)), stock returns (Duffie et al. (2007)) and options (Collin-Dufresne et al. (2001), Berndt and Ostrovnaya (2007), Cremers et al. (2008), Berndt and Obreja (2010) and Carr and Wu (2010)).

Incorporating a long list of explanatory often leads to overfitting, because of multicollinearity or non-identifiably of regression coefficients. Overfitting of the data ends up in great results in-sample, but poor results out-of-sample. Therefore, variable reduction methods are often used to avoid these problems. One of the most widely used variable reduction techniques is Principal Component Analysis (PCA), first proposed by Pearson (1901) and later independently developed and named by Hotelling (1933). The method uses an orthogonal transformation to convert a set of observations of possible correlated variables into a set of values of linearly uncorrelated variables. Another way of reducing the dimensionality is the use of shrinkage methods. The Least Absolute Shrinkage and Selection Operator ($LASSO$), first proposed by Tibshirani (1996), is one of these shrinkage methods. It shrinks the coefficients of the explanatory variables under a constraint until some of the coefficients hit zero. If the coefficient of an explanatory variable hits zero, this variable is deleted from the active dataset. Therefore, the $LASSO$ can be seen as a variable reduction method.

One of the shortcomings of PCA is that it can capture the variance of explanatory variables that do not have any predictive power. Kelly and Pruitt (2015) propose a new method that incorporates these shortcomings of the PCA. They construct factors that are driven by the explanatory variables and exclude information of the explanatory variables that do not have any predictive power. In my paper, I construct new methods that incorporate the shortcomings of the predictability as well. These method are based on the contribution to long- and short-term price discovery, first proposed by Hasbrouck (1995) and Granger (1969), respectively, and can be used to reduce the dimensionality by excluding the explanatory variables that do not contribute to the price discovery of the CDS spread.

Forecasting models do often not consistently outperform a benchmark, but can outperform a benchmark in sub-samples, as mentioned by Granger and Ramanathan (1984). Clemen (1989) and later on Armstrong (2001) compare multiple methods of combining forecasts. Both papers find that equal weighting of forecasts often performs pretty well compared to statistical optimal weights. Lawrence et al. (1986) combines judgemental and statistical forecasts and find that averaging forecasts to be the best approach.

The literature provides a wide range of forecast evaluation measures. Armstrong and Collopy (1992) compare multiple measures based on the forecast accuracy and find different outcomes for different measures. In forecasting financial time-series the directional accuracy of a forecast is often more important than the actual accuracy, because an investor can base his trading decisions on these outcomes. Pesaran and Timmermann (1992) are one of the first who constructed a test for the directional accuracy. Anatolyev and Gerko (2005) incorporate the profitability in their test, arguing that a model that forecasts large deviations more accurate than small deviations is of much more value. Greer (2003) compares multiple directional tests and finds different outcomes. One of the shortcomings for these tests is that investors have different utility functions. Some of the investor only buy a financial instrument when the forecast is larger than a certain threshold. This threshold can differ for different types of investors. Jordà and Taylor (2011) construct a measure where this problem does not occur. They construct their measure by comparing trading strategies with different buy/sell thresholds to a trading strategy based on a coin-flip with the same thresholds.

This paper is structured as follows. In section 2, I describe my dataset of daily CDS spread data on USD-denominated contracts of U.S.-based obligors over the period January 2002-February 2013 and I provide an overview of the used explanatory variables. In section 3, I construct thirteen different forecasting models, of which six models are based on three new variable reduction techniques. The last three models use combinations of forecast techniques. In the last part of this section, I describe the forecast evaluation measures. Section 4 starts with a comparison of the six models based on the three new variable reduction techniques and the model based on PCA in terms of number of selected variables and explained variance. Next, I discuss the results of all forecasting models using the forecast evaluation measures based on the entire sample. In the last part of section 4, I analyse the differences over time and between companies for one of the measures, after which I conclude in section 5.

2 Data

My dataset comprises daily 5-year CDS spread data on USD-denominated contracts of U.S.-based obligors from the Markit Database over the period February 2002-February 2013. Only contracts are included with the modified restructuring (MR) clause, which was the market convention before the introduction of the CDS Big Bang protocol in April 2009. After the protocol the no-restructuring (NR) clause has been the market standard. I only look at the MR clause, because I intend to have one time-series of CDS spreads for every company and during most of the time in my dataset the MR clause was the market standard.

The 5-year CDS contract is the most traded contract and therefore seen as the most liquid among all maturities. In financial markets the beginning and the end of a week consists of market noise. This research focuses on forecasting CDS spreads and to control for as much market noise as possible I only consider data on 5-year CDS spreads and explanatory variables of Wednesdays. However, multiple explanatory variables are not directly observable and are estimated according to different methods that require daily data. Therefore, these variables are constructed using daily data, but only data constructed for Wednesdays are taken into consideration for forecasting the one week ahead CDS spread.

A total of 32 explanatory variables are used to forecast the one week ahead CDS spread. These variables can be divided into eight blocks: implied probability of default measures, stock data, option data, fundamentals, rates, indices, idiosyncratic risk and an idiosyncratic risk factor. Five different databases are used to gather the data: CRSP, Option Metrics, Compustat, Bloomberg and data on the Fama and French (1993) factors is gathered from the Kenneth R. French Data Library [1]. Subsection 2.1 provides a detailed overview of these explanatory variables.

The data on CDS spreads is gathered from the Markit database, of which all available data is considered. Missing data on CDS spreads is handled carefully. The main incentive of this paper is to forecast the CDS spreads. Therefore, missing data on CDSs is only replaced by linear interpolation in case there is only a short period (no more than one month) of missing data. This data cleaning is based on daily data instead of weekly data, because by linearly interpolation between days the loss of information is minimised.

Matching between data on CDS spreads and the CRSP- and Markit Database is based on the tickers and names of companies. Companies that are not listed on any US stock exchange or do not have traded options are excluded from the dataset. Compared to handling missing data on the CDS spreads, companies with missing observations on stock- and option data are less easily excluded from the dataset. This data is solely used for constructing the explanatory variables. Therefore, only when more than 15% of data on stock and/or option prices are missing, a company is excluded from the dataset. Otherwise, linearly interpolation is used for the missing observations. Missing data at the beginning (end) of the period is replaced by the first (last) available observation.

The last matching is done with the Compustat and Bloomberg database. These databases are mainly used for data on fundamentals. Data from Compustat is preferred to data on Bloomberg. Therefore, the first matching is done with the Compustat database. Missing data of short periods is again replaced by linearly interpolation. The resulting missing data on fundamentals is gathered from Bloomberg. Bloomberg data is only used in case of longer periods of missing data in Compustat or when data on fundamentals for a specific company is not available in Compustat.

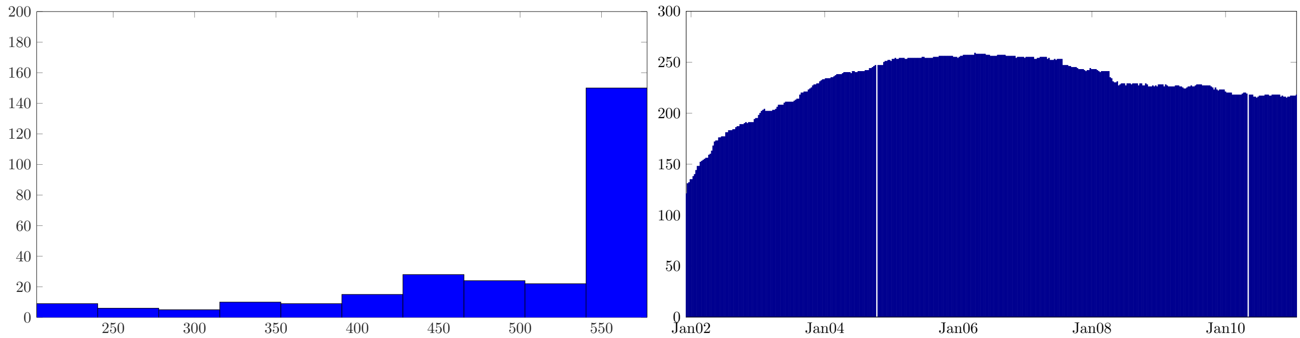
The last requirement of inclusion of a company in a dataset is that there are more than 200 weekly observations available. In section 3 I describe the use of a rolling window of 150 weekly observations to construct one period ahead forecasts. Therefore, the requirement results in at least 50 forecasts for every company in my dataset. Companies that do not have more than 150 weekly observations are excluded from the dataset.

The final dataset comprises of 278 companies over a period of 583 weeks. The left figure of Figure 2 shows the number of weekly observations for every company. It can be seen that not every company spans the entire period. However, the number of observation of most companies is larger than 500. This

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

indicates that most of companies span the entire sample. The right figure of Figure 2 shows that the number of companies increases until the beginning of the financial crisis at 2008 and starts declining after the start of the financial crisis. This could be expected by the fact that CDS contracts gained much more popularity in the years before the crisis and started to lose popularity during the crisis. Besides, I only look at CDS contracts with the MR clause, which was the standard until April 2009. Therefore, it could be that many companies chose to no longer issue CDS contracts with MR clauses. On average, every week there are more than 200 companies included in the dataset during the period Feb 2002 - Feb 2013.

Figure 2: (l) The histogram shows the number of weekly observations per company. (r) The barchart shows the number of companies that are included in the final dataset over time.



2.1 Explanatory variables

This section provides an overview of the explanatory variables used for forecasting CDS spreads. I consider 32 variables based on multiple asset classes. The literature provides informativeness of these credit characteristics.

Implied Probability of Default Measures

1. **IPD CDS** The implied probability of default (IPD) could be a good indicator for forecasting the actual CDS spread, because it represents a more pure and easier to compare measure of credit risk than the actual CDS spread. For the derivation of this implied probability of default assume a CDS contract with a unity notional amount. The protection seller is exposed to an expected loss L equal to:

$$L = IPD(1 - RR) \tag{1}$$

where IPD is the implied probability of default and RR the recovery rate. In absence of market frictions the risk neutral CDS spread should be equal to the present value of the expected loss:

$$CDS = \frac{IPD(1 - RR)}{1 + r} \tag{2}$$

Therefore, the implied probability of default can be calculated as:

$$IPD = \frac{CDS(1 + r)}{1 - RR} \tag{3}$$

Missing data on the recovery rate (RR) is replaced by 40%. For the risk-free rate (r) there are multiple choices. A natural choice for r is the Treasury Rate, which are yields on bonds that have no default risk. Besides, bond yield spreads are usually quoted in the market relative to a Treasury Bond. However, Hull et al. (2005) states that Treasury Rates tend to be lower than other rates that have a very low credit risk for a number of reasons:

(a) Treasury bills and Treasury bonds must be purchased by financial institutions to fulfil a variety of regulatory requirements. This increases demand for these Treasury instruments driving the price up and yield down.

(b) The amount of capital a bank is required to hold to support an investment in Treasury bills and bonds is substantially smaller than the capital required to support a similar investment in other very low-risk instruments.

(c) In the United States, Treasury instruments are given a favourable tax treatment compared with most other fixed-income investments, because they are not taxed at state level.

This leads many market participants to regard swap rates as a better proxy for the risk-free rate. Therefore, I use the 5-year swap-rate as a proxy for the risk-free rate.

2. **IPD stock** The stock market is seen as one of the most liquid asset class markets. In many cases, company specific news is first processed in the stock price before the price of other financial instruments is affected. Therefore, the implied probability of default (IPD) of the stock price could be a good indicator for forecasting CDS spreads. Nowadays, the KMV-merton model is one of the most widely used models to compute the implied probability of default. The KMV-merton model is based on Merton (1974), which states that equity can be seen as a call option on the assets of the company with strike price equal to the value of debt: when the asset value falls below the debt value a company defaults. This model first calculates the distance to default (DD) which measures the number of standard deviations the stock price is away from the default point: the point where the asset value equals the debt value. The probability of default implied by the stock price of a company is larger if the standard deviation of the stock price is higher, because the probability of substantial decreases in the stock price increases. To calculate the DD the assets are assumed to follow a Geometric Brownian Motion. The asset value and asset volatility are not direct observable. Therefore, two equations have to be solved simultaneously which requires a lot of computational time. Appendix A shows the full derivation of the KMV-merton model.

   Bharath and Shumway (2004) develop a naive predictor which does not have the problem of simultaneously solving the equations, but is still close to the KMV-merton model with rank correlation equal to 79%. Therefore, I follow this method to construct the probability of default implied by the stock price. The value of the assets is assumed to be the sum of equity and debt and volatility to be the weighted sum of the volatility of equity and debt. The equity value is direct observable as the product of the number of shares outstanding and the share price. The implied volatility of at-the-money put options is taken as the volatility of the equity. Li (2002) highlight that option implied volatility is seen as a more accurate measure of future volatility, which is exactly the kind of volatility I desire for this method. Bharath and Shumway (2004) approximate the market value of debt (D^*) with the face value of debt (F):

$$D^* = F \tag{4}$$

The face value of debt is equal to the short term liabilities (due in one year) plus half of the long term liabilities. The volatility of the debt of a company close to the default points is assumed to be very risky and correlated with the volatility of equity:

$$\sigma_D^* = 0.05 + 0.25 \times \sigma_E \tag{5}$$

This results in an easy to compute approximation of the volatility of the assets of the company:

$$\sigma_V^* = \frac{E}{E+D^*}\sigma_E + \frac{D^*}{E+D^*}\sigma_D^* = \frac{E}{E+F}\sigma_E + \frac{F}{E+F}(0.05 + 0.25 \times \sigma_E) \tag{6}$$

which does not require simultaneously solving of equations and results in the same equation as the KMV-merton model where the volatility of the company is replaced by the naive approach of equation (6):

$$DD^* = \frac{ln(\frac{E+F}{F}) + (r_{i,t-1} - 0.5\sigma_V^{*2})T}{\sigma_V^* \sqrt{T}} \tag{7}$$

where $r_{i,t-1}$ is the expected return of the assets of the company which is set equal to the one-year stock return. The DD represents the number of standard deviation the stock price is away from the default point. Therefore, the implied probability of default is given by the normal cumulative distribution function (CDF) of the negative value of the DD:

$$IPDStock = \mathcal{N}(-DD^*) \tag{8}$$

Data on the equity price and option implied volatility of the companies are gathered from the CRSP and Optionmetrics database, respectively, and have a daily frequency. However, data on the face value of debt is gathered from the Compustat database and is on a quarterly frequency. Therefore, the face value of debt is linearly interpolated between the quarterly measures of the face value of debt.

Stock price data

3. **Stock price,** computed as the midpoint of the closing bid-ask spread. A higher stock price would indicate that the probability of default decreases, because investors expect the company to be more profitable. This would lead to a decrease in the CDS spread. As mentioned before, the stock market is seen as the market which incorporates news the fastest.

4. **Traded volume,** the number of shares trades on a specific day for a specific company. Higher traded volume indicates more liquidity. When liquidity decreases, investor are less certain about

the direction of the stock price. Uncertainty often leads to more credit risk and therefore an increase in CDS spread.

5. **Stock Spread,** computed as the ask minus bid. Wider spreads indicate low liquidity. The same mechanism applies as for the Traded volume.

6. **SO stock,** the daily number of shares outstanding for a specific company. This number can change for several reasons: stock buyback programs, new issuance and stock splits. Although the motivation for these reasons can be positive or negative, a change in shares outstanding should increase the volatility of the stock price.

7. **RV stock,** the one year realised variance of the stock price. More variance means more uncertainty about the stock price and therefore one could expect an increase in the CDS spread.

8. **LT momentum,** the long-term momentum, measured as the one-year stock return. Higher stock returns should indicate more trust in the company. This leads to a decrease in CDS spread.

9. **ST momentum,** the short-term momentum, measured as the one-week stock return. The same applies for the ST momentum as for the LT momentum.

Option Data

10. **ATM put IV,** the implied volatility of an at-the-money put option with one-year maturity. The forward price is taken as strike price. Option implied volatility is often seen as an efficient predictor of future realised volatility and hence could be a good predictor for more risk. I look at the implied volatility for put options, because these capture the downside risk.

11. **OTM put IV,** the implied volatility of the most out-of-the-money put option with one-year maturity. Carr and Wu (2011) show the relationship between deep out-of-the-money put options and CDSs and should therefore be a better indicator for future realised volatility.

12. **RV ATM put,** the one-year realised volatility of an at-the-money put option. Again the forward price is taken as strike price and the volatility is computed as the one-year realised variance of the option price. More price movements in put option prices should indicate more risk for a company.

Fundamentals

13. **Leverage Ratio,** computed as the ratio of short-term and long-term liabilities. Short-term liabilities are due within one year. When this ratio is high, an outflow of capital is expected in the short-term. Therefore, a higher ratio leads to an increase in CDS spread.

14. **P/E ratio,** computed as the ratio of stock price to earnings. A high P/E ratio indicates investors expect high earnings growth in the future, which leads to a decrease in CDS spread.

15. **Dividend Index Yield,** the financial ratio of the amount of dividend a company pays out relative to its share price. High dividend paying companies are considered to be healthy companies, because it has excess cash which it can divide under his shareholders. Higher dividends lead to a decrease in CDS spread.

16. **P/B ratio,** computed as the ratio of stock price to book price. The stock price is forward looking, while the book price is backward looking. Therefore, a high P/B ratio indicates high expected future earnings, which leads to a decrease in CDS spread.

17. **P/S ratio,** computed as the ratio of stock price to sales. The same mechanism applies as for the P/B ratio.

18. **EV/EBITDA,** computed as the ratio of enterprise value (EV) to earnings before interest, taxes, depreciation and amortisation (EBITDA). A low ratio indicates an undervalued firm and is therefore expected to grow in the future, which makes a credit event less likely. An increase of this ratio leads to an increase in CDS spread.

19. **Size,** captured as the market capitalisation. Companies with a higher market capitalisation are seen as less risky. An increase leads to a decrease in CDS spread.

20. **P/C ratio,** computed as the ratio of stock price to cash flow. Cash flow is measured as EBITDA. The same mechanism applies as for the P/B ratio and the P/S ratio.

Rates

21. **Swap Rate,** the rate of a USD five-year fixed-for-floating swap. As mentioned before, this rate is often seen as the risk-free rate. An increase leads to an increase in CDS spread.

22. **Treasury,** the ten-year Treasury rate. Same mechanism as for the swap rate.

23. **T-Bill,** the three-months T-bill rate. Same mechanism as for the previous two rates.

24. **Libor,** the USD London-Interbanking-Overnight-Rate (LIBOR). This is the rate used for interbanking lending in the money markets. When banks charge eachother more for borrowing/lending, it often implies more uncertainty in the financial markets. This leads to an increase in the CDS spread.

25. **Slope Treasury,** computed as the difference between the ten-year treasury rate and the three-month T-Bill. A higher slope of the treasury curve indicates an expected period of economic expansion. In better times, the risk of credit events decreases. Therefore, an increase of the slope of the treasury curve leads to a decrease in CDS spreads.

Indices

26. **Return S&P,** daily return on the S&P 500 Index. Higher returns indicate more trust in the economy, which leads to a decrease in CDS spreads.

27. **RV S&P,** the realised volatility on the S&P 500 Index. The realised volatility is measured as the the one-year variance. More volatility indicates more uncertainty, which leads to a increase in CDS spreads.

28. **VIX,** represents the CBOE volatility index, which is a measure of the implied volatility of S&P 500 index options. More volatility indicates more risk, which leads to an increase in CDS spread.

Idiosyncratic Risk The following variables capture the idiosyncratic risk indicated by the stock price. Research often suggests that idiosyncratic risk often accounts for most of the variation in risk of an asset. Therefore, an increase in idiosyncratic risk leads to an increase in CDS spreads. Over time, multiple measures of idiosyncratic risk are developed. All these measures are still used to capture the idiosyncratic risk.

29. **FF3,** the idiosyncratic risk of the Fama-French three-factor model, which is the residual ($e_{i,t}$) of the following regression:

$$R_{i,t} - r_t = \alpha_i + \beta_{1,i}(R_{m,t} - r_t) + \beta_{2,i}SMB_t + \beta_{3,i}HML_t + e_{i,t} \tag{9}$$

with $R_{i,t}$ the stock return, r_t the risk-free rate, $R_{m,t}$ the market factor, SMB_t the Small-Minus-Big factor and HML_t the High-Minus-Low factor. The three factors and the risk-free rate are gathered from the Kenneth R. French Data Library.

30. **FF5,** the idiosyncratic risk of the Fama-French five-factor model, which is the residual ($e_{i,t}$) of the following regression:

$$R_{i,t} - r_t = \alpha_i + \beta_{1,i}(R_{m,t} - r_t) + \beta_{2,i}SMB_t + \beta_{3,i}HML_t + \beta_{4,i}RMW_t + \beta_{5,i}CMA_t + e_{i,t} \tag{10}$$

with RMW_t the Robust-Minus-Weak factor, CMA_t the Conservative-Minus-Aggressive Factor and all other factors are the same as in equation (10). Again, the five factors and the risk-free rate are gathered from the Kenneth R. French Data Library.

31. **CAPM,** the idiosyncratic risk of the Capital Asset Pricing Model (CAPM), which is the residual ($e_{i,t}$) of the following regression:

$$R_{i,t} - r_t = \alpha_i + \beta_{1,i}(R_{m,t} - r_t) + e_{i,t} \tag{11}$$

Idiosyncratic Risk Factor

32. **CSIR factor,** the Cross-Sectional-Idiosyncratic-Risk factor, which is based on Bai and Wu (2012) and has never been used in forecasting the CDS spread. Bai and Wu (2012) find that modelling the CDS spread in the cross-section returns high R-squares. By constructing a factor that captures the idiosyncratic risk in the cross-section, I tend to find an explanatory variable with high predictive power. In the first step of constructing this factor I follow Bai and Wu (2012) where I regress the IPD stock of item 2 on the CDS spreads in the cross-section for every week t:

$$CDS_{i,t} = \alpha_t + \beta_t IPDStock_{i,t} + e_{i,t} \tag{12}$$

This results in a $T \times N$ matrix of residuals $e_{i,t}$. In the second step I perform PCA on this $T \times N$ matrix to capture the cross-sectional variance over time. The first factor of the PCA analysis, the factor which explains the most variance, is used as the $CSIR$ factor. This one factor explains 64% of the cross-sectional variation of the idiosyncratic risk on average over time.

3 Methods

The first part of this section describes thirteen models to forecast the one week ahead CDS spread using a rolling window of 150 weekly observations. The explanatory variables are on a daily basis, as described in section 2. However, the aim of these models is to forecast the one week ahead CDS spread. Given the presence of the most market noise at the beginning and end of a week, I use end of day closing data of Wednesdays to forecast the one week ahead CDS spread.

I describe the first ten models, denoted by individual models, in section 3.1. The statistics behind the first four models have already been used in the literature of forecasting time-series. The last six individual models use a two stage approach of variable reduction techniques. The first stage makes use of three new approaches of variable reduction techniques. These three new approaches are based on existing methods that, originally, were not constructed for the reduction of variables. However, these methods tend to gain insights in the predictability of explanatory variables. Therefore, these methods are used in the first stage to determine the best possible selection of explanatory variables to use in forecasting the CDS spread. In the second stage, I construct forecasts with the subset of explanatory variables using two different techniques: PCA and equally weighting of forecasts made with individual explanatory variables.

The last three models are combinations of forecasts of the ten individual models. I make combinations of forecasts, because it is often seen in the literature that a single model does not outperform benchmarks for the entire sample, while it does for some sub-samples. In combining forecasts I tend to make use of the possibility that different models outperform in different sub-samples. The last part of this section describes eight measures of forecasting performance for the thirteen models. In the forecasting literature statistical measures of the fit of forecasts are widely used to evaluate the performance of models. However, more economic measures of directional accuracy and profitably of trading strategies based on forecasting models are often preferred. Therefore, I use both statistical and economic measures to evaluate the forecast performance of the thirteen models.

3.1 Individual models

One of the most frequent used approaches in modelling time-series data is to use auto-regressive-moving-average (ARMA) models, because it is able to capture a wide range of different time-series patterns. Besides, an ARMA model captures the (if any) autocorrelation of the time-series, which is often not captured by explanatory variables. Therefore, the first model I use is an ARMA model and is the starting point for all other models. To determine the number of lags of the ARMA model that fits the data best, I estimate multiple ARMA models for a small sub-sample of the data to determine the best fit for the ARMA model. I use the first 150 observations for the first five companies (in alphabetical order) to estimate the following regressions:

$$CDS_{i,t} = \alpha_{i,t} + \sum_{p=0}^{P} \beta_{p,i,t} CDS_{i,t-p} + \sum_{q=0}^{Q} \theta_{q,i,t} \varepsilon_{i,t-q} + \varepsilon_{i,t} \tag{13}$$

with i and t denoting the company and the time, respectively, and P,Q=$\{0,..,4\}$ the number of lags. To determine the number of lags of the ARMA model that fits the data best, I use the Schwarz et al. (1978) Bayesian Information Criterion (BIC). On average, $P = 1$ and $Q = 0$ gives the smallest BIC value. Therefore, the first model I use in forecasting the 5-year CDS spread is given by:

$$CDS_{i,t} = \alpha_{i,t} + \beta_{i,t} CDS_{i,t-1} + \varepsilon_{i,t} \tag{14}$$

The model described in equation (14) is the starting point for all other models I use to forecast the 5-year CDS spread. These other models incorporate the explanatory variables described in section 2. A natural choice to incorporate predictors into an ARMA model is to use an Autoregressive-Moving-Average with exogenous inputs (ARMAX) model. However, I base my approach on Bai and Wu (2012), who model the estimation errors by means of explanatory variables and find an average R-squared of 77%. Although Bai and Wu (2012) model the CDS spread in the cross-section and therefore did not use an ARMA model in their first step, this two stage approach shows the possibility of a gain in performance.

The first step in the two stage approach for the remaining nine individual models is to estimate equation (14). The second step is to model the errors terms of this equation with the use of the explanatory variables described in section 2. The second step is therefore given by the following equation:

$$e_{i,t} = f(X_{k,i,t-1}) + \eta_{i,t} \tag{15}$$

with $f(X_{k,i,t-1})$ defined as a function of the $k = 32$ explanatory variables X_k. The high dimensionality of the explanatory variables can lead to multicolliniarity and poor out-of-sample forecasts. Therefore, all nine models are (a combination of) variable reduction techniques. The one-week-ahead forecast for these models is given by:

$$\widehat{CDS_{i,t+1}} = a_{i,t} + b_{i,t} CDS_{i,t} + f(X_{k,i,t}) \tag{16}$$

3.1.1 CSIR

The second model models the error term of equation (14) by means of the $CSIR$-factor described in section 2, item 32. This factor captures the cross-sectional idiosyncratic risk of the CDS spread. For all other models this explanatory model is included in the variable reduction methods. However, I base one model entirely on this explanatory variable, because Bai and Wu (2012) found such impressive results by modelling the CDS spread in the cross-section. The $CSIR$ factor is not used in the literature before, but, according to results of Bai and Wu (2012), could be able to capture (most of) the cross-sectional behaviour of the CDS spreads. Formally, the model for the error term is given by:

$$e_{i,t} = \beta_{i,t} CSIR_{t-1} + \eta_{i,t} \tag{17}$$

with corresponding forecast:

$$f(X_{k,i,t}) = \beta_{i,t} CSIR_t \tag{18}$$

In this way the $CSIR$-factor can be seen as an idiosyncratic risk factor of the CDS spreads of the entire market. It captures the variance of the CDS spreads unexplained by the stock market. The exposure to this factor should therefore capture the idiosyncratic risk and should lead to better forecasts of the CDS spread.

3.1.2 Principal Component Analysis

One of the most widely used variable reduction methods is Principal Component Analysis (PCA). This method was first proposed by Pearson (1901) and later independently developed and named by Hotelling (1933). The method uses an orthogonal transformation to convert a set of observations of possible correlated variables into a set of values of linearly uncorrelated variables called principal components. By means of an eigenvalue decomposition the procedure starts by finding the first principal component that explains most of the variance of the dataset. The second step finds the principal that is orthogonal to the first factor and explains most of the remaining variance. The second step is repeated until the number of principal components is equal to the number of explanatory variables. This results in a new dataset of principal components with the same size as the dataset of explanatory variables. However, the first (last) principal component explains most (least) of the variance of the explanatory variables. Therefore, a small sub-set of principal components often explain most of the variance of the explanatory variables.

As explained earlier, I make use of a rolling window. Therefore, every period the model in equation (14) is estimated results in a new data-series of the 149 residuals $e_{i,t}$. PCA is performed on the corresponding 149 observations of the 32 explanatory variables $X_{k,i,t}$. I use the first three principal components to estimate the following regression for the 149 residuals:

$$e_{i,t} = \sum_{c=1}^{3} \gamma_{c,i,t} PC_{c,i,t} + \nu_{i,t} \tag{19}$$

Next, to construct $f(X_{k,i,t})$, which is equal to the forecast $\widehat{e_{i,t+1}}$, I assume the principal components follow an AR(1) process. Forecasts of these factors for period $t+1$ are then given by:

$$\widehat{PC_{c,i,t+1}} = \alpha_{c,i,t} + \beta_{c,i,t} PC_{c,i,t} \tag{20}$$

for $c = 1, 2, 3$. The function $f(X_{k,i,t})$ of equation (16) is then given by:

$$f(X_{i,t}) = \sum_{c=1}^{3} \gamma_{c,i,t} \widehat{PC_{c,i,t+1}} \tag{21}$$

3.1.3 3PRF

The next variable reduction method uses the Three-Pass Regression Filter ($3PRF$) of Kelly and Pruitt (2015). This $3PRF$ is a constrained least squares estimator and reduces to partial least squares as a special case. The idea behind this method is that the data can be described by an approximate factor model. One of the points of criticism on PCA is that it can capture the variance of explanatory variables that do not have any predictive power. The $3PRF$ is able to both find factors that are driven by the explanatory variables, for which the number of factors is much smaller than the number of explanatory variables, and handles the problem of pervasive irrelevant information among predictors.

Kelly and Pruitt (2015) establish an environment for the $3PRF$ wherein they make use of different variables. The *target* variable is wished to forecast with the use of the explanatory variables, the *predictors*. Forecasts of the target variable are constructed with use of *proxies*: factors driven by the predictors and the target in particular. These proxies are unobservable, but are always available from the target and predictors. Alternatively, the $3PRF$ gives the opportunity to supply these proxies if one has an idea of specific variables driving both the target and predictors. In this paper I use the first method, where I construct the unobservable proxies using the $3PRF$.

To construct the forecast with the desired number of proxies, the $3PRF$ makes use of a recursive algorithm of three steps. In this algorithm the proxies, predictors and target are defined by Z, X and $e_{i,t}$, respectively. The number of proxies is determined by the user of the algorithm and the first proxy, z_t, is set equal to $e_{i,t}$. The first step in the $3PRF$ is to regress all predictors x_k on Z over time:

$$x_{k,t} = \phi_{0,k} + z_t'\phi_k + \epsilon_{k,t} \tag{22}$$

for $k = 1, ..., 32$. In the second step the predictors x_t are regressed on the slope estimate $\hat{\phi}_i$ of equation (22) in the cross-section:

$$x_{k,t} = \phi_{0,t} + \hat{\phi}_i' F_t \tag{23}$$

The last step regresses the target variable $e_{i,t}$ on the slope estimate $\hat{F}_t$ of equation (23) and delivers forecast $\hat{e}_{t+1}$:

$$e_{i,t+1} = \beta_0 + \hat{F}_t'\beta + \eta_{t+1} \tag{24}$$

The next proxy, the second column of Z, is set by the in-sample errors of the forecasts of $e_{i,t}$: $z_t = e_{i,t} - \widehat{e_{i,t}}$. This new matrix of proxies Z is then used to go over the three steps described in equations (22)-(24). This algorithm will recursively go on until the required number of proxies is used to make an out-of-sample forecast $e_{i,t+1}$. In this paper the number of proxies is set equal to three.

This procedure assumes the data can be described by an approximate factor model. This factor is unobservable and is generated through the first two steps of the algorithm. In the first step, the exposure of the predictors to a small group of proxies over time is calculated. Predictors that do not have predictive power over time will have smaller exposures to these proxies. In the second and last step of constructing the factor, the exposures over time are regressed on the predictors in the cross-section. The factor used to forecast the target variables consists of the coefficients of this regression. Moreover, these coefficients capture the explanatory power of the predictors in explaining the exposures over time in the cross-section. In this way, the factor incorporates the predictive power of the predictors over time and the explanatory power of the predictors in the cross-section compared to each other. Predictors that have the most explanatory power over time and in the cross-section will therefore have the most weight in the factor.

3.1.4 Two stage variable reduction methods

The last six individual models are based on three different methods to reduce the number of predictors. The motivation of these methods is the same as for Kelly and Pruitt (2015): variable reduction methods such as PCA do not take into account whether the predictors have any predictive power for the independent variable. Moreover, when performing PCA on all predictors, the first principal components could explain most of the variance of the predictors that do not have predictive power for the 5-year CDS spread.

The following methods are all based on the same principle, where the first stage is to decide whether the predictors standalone have sufficient explanatory power for the independent variable $e_{i,t}$. These variables, denoted by $x^*_{j,i,t}$ for $j = 1, ..., J$ the number of predictors with sufficient explanatory power, are then used to construct a forecast for $e_{i,t}$. In the second step I construct two models with the variables $x^*_{j,i,t}$: the first one makes use of PCA and follows the procedure of section 3.1.2 to make one-step-ahead forecasts. The second method regresses the predictors solely on the error terms $e_{i,t}$ of equation (14) and constructs one-step-ahead forecasts for all these models:

$$\widehat{e_{j,i,t+1}} = \delta_{j,i,t} x^*_{j,i,t} \tag{25}$$

for $j = 1, ..., J$. These forecasts are then combined by equally weighting the predictions for the final one-step-ahead forecasts:

$$\widehat{e_{i,t+1}} = \frac{1}{J} \sum_{j=1}^{J} e_{k,i,t+1} \tag{26}$$

Price Discovery The first stage for this method is based on Hasbrouck (1995)'s information share. This method was first composed for a framework where a financial asset is traded in more than one market, for example the NYSE and the AEX. It assumes there is an unobserved component which drives the price of the financial asset in both/all markets. The information share represents the relative contribution of a market to the total variance of the unobserved component in the long-term.

For this method I assume the error term of equation (14) and the predictors described in section 2 are pure credit measures. Therefore, following Hasbrouck (1995), all these variables should be driven by an unobservable factor and can be used to measure the contribution of this variable to the price discovery. The first step in variable reduction is made by excluding predictors for which the contribution to the price discovery of the unobserved factor of the error term and the explanatory variable is considerably low.

Hasbrouck (1995) states that to determine the contribution of price discovery of two price series, the requirement of cointegration has to be met. This paper relaxes this condition, because it is not often seen that financial time-series fully cointegrate. For the remaining part of the procedure I follow Hasbrouck (1995), where I first determine the price discovery between the estimation error of equation (14) $e_{i,t}$ and $x_{k,i,t}$ the k-th explanatory variable. Because I follow the same procedure for every company and every explanatory variable I define $Y_t = Y_{k,i,t}$ from now on.

The differential of Y_t is defined as the error correction term, i.e., $z_t = \beta' Y_t = y_{1,t} - y_{2,t}$. Therefore, the cointegration vector $\beta = (1, -1)'$. Hasbrouck (1995) starts with the estimation of the following Vector Error Correction Model (VECM):

$$\Delta Y_t = \alpha\beta' Y_{t-1} + \sum_{j=1}^{k} A_j \Delta Y_{t-j} + \varsigma_t \tag{27}$$

where α is the error correction vector and ς is a zero-mean vector of serially uncorrelated innovations with covariance matrix Ω. The VECM has two portions: the first portion, $\alpha\beta' Y_{t-1}$, represents the long-run or equilibrium dynamics between the price series, and the second portion, $\sum A_j \Delta Y_{t-j}$, represents the short-run dynamics induced by market imperfections.

Hasbrouck (1995) transforms equation (27) into a Vector Moving Average (VMA):

$$\Delta Y_t = \Psi(L)\varsigma_t \tag{28}$$

and its integrated form:

$$Y_t = \Psi(1) \sum_{s=1}^{t} \varsigma_s + \Psi^*(L)\varsigma_t \tag{29}$$

where $\Psi(L)$ and $\Psi^*(L)$ are matrix polynomials in the lag operator L. The impact matrix, $\Psi(1)$, is the sum of the moving average coefficients, with $\Psi(1)\varsigma_t$ being the long-run impact of an innovation on each of the prices. If the rows of the impact matrix are identical, the long-run impact is the same for all prices. If I denote $\psi = (\psi_1, \psi_2)$ as the common row vector in $\Psi(L)$, equation (29) becomes:

$$Y_t = \iota\psi(\sum_{s=1}^{t} \varsigma_s) + \Psi^*(L)\varsigma_t \tag{30}$$

where $\iota = (1, 1)'$ is a column vector of ones. Hasbrouck (1995) states that the increment $\psi\varsigma_t$ in equation (30) is the component of the price change that is permanently impounded into the price and is presumably due to new information.

Hasbrouck (1995) argues that if the covariance matrix of the innovations Ω is diagonal then $\psi\Omega\psi'$ will consist of two terms, the first (second) represents the contribution to the common factor innovation from the first (second) market. The information share of market j is then defined as

$$S_j = \frac{\psi_j^2 \sigma_j^2}{\psi \Omega \psi'} \tag{31}$$

with σ_j the j-th diagonal element of Ω.

Equation (30) is a difficult equation to estimate and requires a lot of computational time. Baillie et al. (2002) argue the estimation of the VMA is not necessary to construct the information share. They state the VECM of equation (27) is sufficient by proving the following:

$$\frac{\psi_1}{\psi_2} = \frac{\gamma_1}{\gamma_2} \tag{32}$$

with $\alpha_\perp = (\gamma_1, \gamma_2)'$ the orthogonal of α of equation (27). This leaves a much easier model to estimate and speeds up computational time. Substituting equation (32) into equation (31) yields:

$$S_j = \frac{\gamma_j^2 \sigma_j^2}{\gamma_1^2 \sigma_1^2 + \gamma_2^2 \sigma_2^2} \tag{33}$$

However, if the price innovations are significantly correlated across markets, equation (33) does not hold. In this case, Hasbrouck (1995) uses Brezinski (2005)'s Cholesky factorisation of $\Omega = MM'$ to eliminate the contemporaneous correlation, where M is a lower triangular matrix. The information share is then given by:

$$S_j = \frac{([\psi M]_j)^2}{\psi \Omega \psi'} \tag{34}$$

where $[\psi M]_j$ is the jth element of the row of matrix ψM and M denoted by:

$$M = \begin{pmatrix} m_{11} & 0 \\ m_{12} & m_{12} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2(1-\rho^2)^{1/2} \end{pmatrix} \tag{35}$$

I construct the contributions to the information share relative to 150 estimation errors $e_{i,t}$ of equation (14) for every time t and explanatory variable k. Only the predictors for which the information share is larger than 50% are taken for the next and finale stage. As mentioned in section 3.1.4, in the second and final stage two methods are used to construct a forecast for $e_{i,t+1}$: PCA and equal weighting of forecast made with individual explanatory variables. I denote these models by $PD\ PCA$ and $PD\ EW$ whereby PCA and equal weighting of individual forecasts is used in the second stage, respectively.

Granger-Causality Another way of describing predictive power of an explanatory variable in forecasting $e_{i,t}$ is to find whether causality exists between $e_{i,t}$ and the predictors. In the first step of this method I test whether a causal relation exists between this residual and individual predictors and if this causation is in the right direction. The test I use is the Granger-Causality-test first proposed in Granger (1969). This method is comparable to the price discovery measure, because they both look at which market is leading and which market is lagging. However, the biggest difference between the two methods, is that the Granger-Causality-test looks at the short-term dynamics of the price development, while Hasbrouck (1995) looks at the long-term dynamics. The results of which market drives the other will therefore differ between the two methods. In particular, the outcomes of the Granger-Causality-test will differ much more over time than the Hasbrouck (1995)'s information share will.

The definition of causality used in Granger (1969) is based entirely on the predictability of the residual of equation (14). If an explanatory variable $X_{i,t}$ contains information in past terms that helps in the prediction of the residual and if this information is contained in no other series used in the explanatory variable, then the explanatory variable is said to cause the residual $e_{i,t}$. Explanatory variables that do not cause the residual are eliminated. The causal model is then given by:

$$e^*_{i,t} = \sum_{p=1}^{P} a_p e^*_{i,t-p} + \sum_{q=1}^{Q} b_q X^*_{i,t-q} + \varepsilon_{i,t} \tag{36}$$

The definition of causality given above implies that if $X^*_{i,t}$ is causing $e^*_{i,t}$ some b_q are not zero. To test for all b_q equal to zero I use a F-test with a 95% significance level. The input for an F-test requires two models. The first one is defined in equation (36). The second model excludes the Q lagged values of the explanatory variables $X^*_{i,t}$ of equation (36). The number of lags P and Q are determined using the BIC. As described in section 3.1.4, the explanatory variables that survived this first step are taken to the next stage where a forecast is made for $e_{i,t}$ using PCA and equal weighting of individual forecasts. I denote these models by $GC\ PCA$ and $GC\ EW$ whereby PCA and equal weighting of individual forecasts is used in the second stage, respectively.

LASSO The last method is based on the Least Absolute Shrinkage and Selection Operator ($LASSO$) first proposed by Tibshirani (1996). This method shrinks the regression coefficients by imposing a penalty on their size. The $LASSO$ coefficients minimise a penalised residual sum of square:

$$\begin{gathered} \hat{\beta}^{LASSO}_{i,t} = \underset{\beta}{\operatorname{argmin}} \sum_{p=0}^{P} \left(e^*_{i,t-p} - \beta_{0,i,t} - \sum_{k=1}^{K} x^*_{k,i,t-p}\beta_{k,i,t} \right)^2 \\ subject\ to \sum_{k=1}^{K} |\beta_{k,i,t}| \leq d \end{gathered} \tag{37}$$

which can be written in the equivalent $Lagrangian\ form$:

$$\hat{\beta}^{LASSO}_{i,t} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{p=0}^{P} \left(e^*_{i,t-p} - \beta_{0,i,t} - \sum_{k=1}^{K} x^*_{k,i,t-p}\beta_{k,i,t} \right)^2 + \lambda \sum_{k=1}^{K} |\beta_{k,i,t}| \right\} \tag{38}$$

The constraint $\sum_{k=1}^{K} |\beta_{k,i,t}| \leq d$ makes the solution non-linear in $e^*_{i,t}$. Therefore, there is no closed form solution and a quadratic programming algorithm is required to come up with a solution. This quadratic programming algorithm consists of five steps:

1. Set all coefficients $\beta_{k,i,t}$ equal to zero.
2. Find the explanatory variable $x^*_{k^1,i,t}$ most correlated with $e^*_{i,t}$.
3. Increase the coefficient $\beta_{k,i,t}$ in the direction of the sign of its correlation with $e^*_{i,t}$ until another explanatory variable $x^*_{k^2,i,t}$ has as much correlation with the residual, $r_t = e^*_{i,t} - \widehat{e^*_{i,t}}$, as $x^*_{k^1,i,t}$.

4. Increase $(x^*_{k^1,i,t}, x^*_{k^2,i,t})$ in their joint least squares direction until another explanatory $x^*_{k^3,i}$ has as much correlation with the residual r_t.

5. Continue until all explanatory variables are in the model.

Because of the nature of the constraint, making d sufficiently small will cause some of the coefficients to be exactly zero. When a coefficient hits zero during any step of the algorithm, this explanatory variable is removed from the active set of explanatory variables. Therefore, the $LASSO$ indirectly does a continuous subset selection. If d is chosen larger than the sum of the least squares estimate coefficients, then the $LASSO$ coefficients are exactly equal to these least squares estimates.

In this paper, d is not chosen upfront. I run the algorithm for decreasing values of d until there are fifteen (out of 32) explanatory variables with non-zero coefficients. This approach is new compared to the traditional way of using $LASSO$, because I do not use the coefficients constructed by the $LASSO$ technique and only take the fifteen explanatory variables that have a coefficient not equal to zero to the next stage. In the second stage I construct the models using the fifteen explanatory variables based on PCA and equal weighting of individual forecasts described in section 3.1.4 to forecast $e_{i,t}$. I denote these models by $LASSO$ PCA and $LASSO$ EW whereby PCA and equal weighting of individual forecasts is used in the second stage, respectively.

3.1.5 Stationarity, correlation and multicollinearity

One of the possible influential factors in time-series econometrics is that the time-series data is non-stationary. The statistical proportions of stationary processes do not change over time. In this research, especially explanatory variables that contain a trend, i.e. size, stock price, show time-dependent statistical proportions. Modelling non-stationary time-series can result in infinite persistence of shocks, spurious regressions, where a high R-squared can be obtained while the time-series are totally unrelated, and standard assumptions for asymptotic analyses are not valid. As mentioned before, multiple explanatory variables contain a trend which results in non-stationarity. The most obvious choice is to run test for a trend in the time-series, by for example a unit-root test, and then adjust for this possibly non-stationarity before constructing forecasts. However, this is not done in this research. Therefore, there is a possibility that the forecasting models base their forecasts on non-stationary processes.

Another issue that could arise in this research is the high correlation between some of the explanatory variables, i.e. stock price and size. Including explanatory variables that are highly correlated in a regression can lead to multicollinearity: one explanatory variable influences the other resulting in disturbance of the data and the statistical inferences made are not reliable. In this research multicollinearity is not an issue: none of the forecasting models includes two explanatory variables at the same time in a regression. However, almost all models make use of either PCA or an equal weighting of forecasts. PCA constructs factors that explain most of the variance of the dataset. Therefore, including explanatory variables that are highly correlated results in an over-presence of the same information in the first factors of the PCA. Fortunately, I take more than one principal component to construct forecasts. Therefore, not only the variance of highly correlated explanatory variables is used to forecast the CDS spread. For the equal weighting of forecasts the same mechanism applies. The forecast will contain more weight of the same information of the highly correlated explanatory variables.

3.2 Forecast Combination

The motivation behind combining forecasts is that models with the same predictive power often do not consistently outperform a benchmark over time, but can outperform these benchmarks in sub-samples. Therefore, it is often difficult to identify one single best model. However, by combining forecasts one can diversify the gains of the predictive power of different models for different sub-samples, resulting in a model structurally outperforming the benchmarks.

One of the most widely and perhaps simplest form of combining forecasts is to use equal weighting. As Clemen (1989) points out, equal weighting of forecasts often outperforms other weighting schemes. Equal weighting often underperforms in-sample, but outperforms out-of-sample. As Elliott and Timmermann (2013) point out, when all models are based on similar data and when they all perform roughly the same then equal weighting is optimal, because individual forecasts have identical variance and identical pair-wise correlation. Based on these theories, equal weighting could be a good addition to the individual models. Therefore, the first combination of forecasts is equal weighting (EW) and is defined by:

$$\widehat{CDS}_{i,t+1}^{EW} = \frac{1}{M}\sum_{m=1}^{M}\widehat{CDS}_{i,t+1}^{m} \tag{39}$$

with $M = 1, ..., 10$ denoting forecasts of the individual models.

As described before, it is often difficult to identify one best model over the entire sample. However, there are often models that outperform the benchmark in sub-samples. The last two combinations of forecasts are based on this phenomenon, where I only take an equal weighting of forecasts of the individual models that have been able to construct a ‘good’ forecast for the past week. I look at two measures to decide ‘good’ performance of a model for the past week: the Mean Square Prediction Error (MSPE) and the Correctly Predicted Sign (CPS). I denote these two models with *Lag MSPE EW* and *Lag Sign EW*. The former selects the models that outperformed the benchmark model in the last period in terms of MSPE and constructs an equally weighted forecasts of these models for the next period. In case non of the models outperformed the benchmark in the past weeks, it selects a single model that had the lowest MSPE in the last week. *Lag Sign EW* constructs an equally weighted forecast as well, but selects the models that correctly predicted the sign in the last period. In case none of the models was able to correctly predict the sign, the model with the forecast of the model with the lowest MSPE in the last period is selected.

3.3 Forecast evaluations

There is still a lot of discussion going on how to measure the performance of forecasting models. As explained in the last part of the previous section, there is a difference between the statistical and economic relevance of forecasts. In forecasting financial time series, forecasts based on a Random Walk (RW) or a coing-flip are often taken as benchmarks and are often seen as hard to beat. Therefore, I compare the performance of the thirteen models with forecasts based on these benchmarks for six statistical and economic measures and two 'simple' trading strategies. The benchmark is defined differently for all these measures and I describe them separately for all measures.

The first measure is based on the most widely used statistical measure: the Mean Squared Prediction Error ($MSPE$). This measure describes the goodness-of-fit of model m and is defined by the following:

$$MSPE_i^m = \frac{1}{T^i}\sum_{t=1}^{T^i}\Big(E[\widehat{CDS}_{i,t}^m] - CDS_{i,t+1}\Big)^2 \tag{40}$$

with T^i the number of forecasts for company i. The benchmark is composed by setting the forecast for the next period equal to the value of today.

The second measure is the Diebold and Mariano (2012) test. This test defines the null hypothesis of no difference of two forecasting model. I compare the thirteen forecasting models a benchmark model that is composed by setting the forecast for the next period equal to the value of today. The test is based on the difference of squared errors made by both models ($d = (CDS_{i,t} - \widehat{CDS}_{i,t}^{RW})^2 - (CDS_{i,t} - \widehat{CDS}_{i,t}^{m})^2$). The errors are assumed to be normally distributed. The statistic is then given by:

$$DM = \frac{mean(d)}{\sqrt{\frac{var(d)}{T^i}}} \sim \mathcal{N}\big(0,1\big) \tag{41}$$

A DM statistic larger than the on sided 95% critical value of 1.65 indicates that the model outperforms the benchmark model.

The following measures have more economic meaning than the first two measures described above. The first measure of economic forecast evaluation is the 'rather simple' Percentage Correctly Predicted Sign ($PCDS$). This measure counts the number of times a model has correctly predicted the sign of the weekly change for a specific company and is given by:

$$PCPS_i^m = \frac{1}{T^i}\sum_{t=1}^{T^i} I\big((\widehat{CDS}_{i,t}^m - CDS_{i,t-1}) \times (CDS_{i,t} - CDS_{i,t-1})\big) \tag{42}$$

with $I(\cdot)$ an indicator function which takes a value equal to one if the condition is true and zero otherwise. The benchmark for this measure is the coin flip scenario, where I assume the chance of predicting a positive change is equal to 50%. Therefore, outperforming the benchmark indicates a $PCPS$ higher than 50%.

The next measure is the distribution-free Pesaran and Timmermann (1992) test. This test focuses on the directional accuracy (DA) of the forecasts. The method compares the number of times model m predicts the right sign compared to the fraction of the number of positives and negatives of model m and the considered variable. Moreover, the test takes into account whether there are more weekly positive or negative changes in the time-series. Therefore, the difference between the DA of the benchmark and of the expected DA is a Hausman (1978) type statistic. For this measure I define PP^m (PN^m) and PP (PN) the fraction of positive (negative) changes for model m. The test then compares the $PCPS$ of equation (42) with the estimator of the expectation of the $PCPS$ under the null. The estimator of the expectation of the $PCPS$ for a company is then given by:

$$\hat{P}_*^m = PP^m \times PP + PN^m \times PN \tag{43}$$

the variances of the $PCDS$ and $\hat{P}_*$ are defined as follows:

$$\hat{V}(PCDS) = \frac{\hat{P}_* \times (1 - \hat{P}_*)}{T} \tag{44}$$

$$\hat{V}(\hat{P}_*) = \frac{(2PP-1)^2 PP^m(PP^m-1)}{T} + \frac{(2PP^m-1)^2 PP(1-PP)}{T} + \frac{4 \times PP \times PP^m \times PN \times PN^m}{T^2} \tag{45}$$

The test-statistic is then given equation (46) and follows a standard normal distribution:

$$PT = \frac{PCDS - \hat{P}_*}{\sqrt{\hat{V}(PCDS) - \hat{V}(\hat{P}_*)}} \sim \mathcal{N}(0,1) \tag{46}$$

where the 95% critical value of 1.65 indicates the model outperforms the benchmark model.

The next measure is based on Anatolyev and Gerko (2005). This market timing test is tied to a simple trading strategy where a virtual investor buys (sells) the CDS when a model predicts a positive (negative) change for the next period. This excess profitability (EP) takes the actual profit into consideration. The virtual investor invests a monetary unit value every period and unwinds it at the end of the period. The weekly return of this strategy based on buy/sell signals of model m for company i is then given by:

$$r_{i,t}^m = I(\widehat{CDS}_{i,t}^m) CDS_{i,t} \tag{47}$$

The average weekly return is then compared to a random walk trading strategy with the probability of a buy (sell) signal equal to the proportion of positive (negative) changes. Anatolyev and Gerko (2005) state that under the null hypothesis of conditional mean independence the forecasts $\widehat{CDS}_{i,t}$ are independent of $CDS_{i,t}$ for all lags and leads. Under this null hypothesis the estimators of the two trading strategies based on model m and the RW benchmark model are given by:

$$A_{i,T}^m = \frac{1}{T} \sum_{t=1}^{T} r_{t,i}^m \tag{48}$$

and

$$B_{i,T}^{RW} = \Big(\frac{1}{T}\sum_{t=1}^{T} I(\widehat{CDS}_{i,t}^{m})\Big)\Big(\frac{1}{T}\sum_{t=1}^{T} CDS_{i,t}\Big) \tag{49}$$

When a model has predictive power, the average return of the simple trading strategy $A_{i,T}^{m}$ will be higher than the average return of the random walk $B_{i,T}^{RW}$. Anatolyev and Gerko (2005) argue the most straightforward way to estimate the variance of the difference of average returns $A_{i,tT}^{m}$ and $B_{i,T}^{RW}$ is by:

$$\hat{V}_{i,EP}^{m} = \frac{4}{T^2}\hat{p}_{\widehat{CDS}}(1-\hat{p}_{\widehat{CDS}})\sum_{t=1}^{T}(CDS_{i,t}-\widehat{CDS}_{i,t}^{m})^2 \tag{50}$$

where

$$\hat{p}_{\widehat{CDS}} = \frac{1}{2}\Big(1+\frac{1}{T}\sum_{t=1}^{T} I(\widehat{CDS}_{i,t}^{m})\Big) \tag{51}$$

is the consistent estimator of the probability of a positive predicted sign. Moreover, the EP test is one sided with a normally distributed test statistic and given by:

$$EP_i = \frac{A_{i,T}-B_{i,T}}{\sqrt{\hat{V}_{i,EP}}} \sim \mathcal{N}\big(0,1\big) \tag{52}$$

As Anatolyev and Gerko (2005) mention, the EP test takes fuller advantage of possibly predictability than the Pesaran and Timmermann (1992) test, because the latter takes no profitability into account. Moreover, Christoffersen and Diebold (2006) show that absence of mean predictability dependency in the volatility can result in sign predictability. Therefore, the EP statistic is preferred to the PT statistic in terms of mean predictability considerations.

As described before, evaluation forecasts based on goodness of fit measures is maybe not the best way of measurement. Moreover, the first difficulty is to identify the right loss-function. Different loss-functions can give different results. Besides, an investor is often more interested weather he can base an actual trading strategy on the forecasts. Measures based on the sign an profitability like the PT and EP measures could therefore illustrate a more ideal picture of the performance of forecasting models. However, these type of measures are often constructed for one type of investor, because not every investor has the same signal of going long/short. Jordà and Taylor (2011) introduced a new statistical method which takes these points of criticism into account. The method assesses the correct classification (CC) ability of investment strategies.

Jordà and Taylor (2011) define the actual direction of the CDS spread as $d_{i,t}$ and an investor is long (short) when his forecast of the direction $\delta_{i,t}$ is larger (smaller) than a threshold parameter c. This results in the following classification rates for the binary decision problem a trader faces for a trading strategy:

		Prediction	
		Negative/Short	Positive/Long
Outcome	Negative/Short	$TN(c) = P(\hat{\delta}_{i,t} < c \mid d_{i,t} = -1)$	$FP(c) = P(\hat{\delta}_{i,t} > c \mid d_{i,t} = -1)$
	Positive/Long	$FN(c) = P(\hat{\delta}_{i,t} < c \mid d_{i,t} = +1)$	$TP(c) = P(\hat{\delta}_{i,t} > c \mid d_{i,t} = +1)$

where $TN(c)$ ($FN(c)$) and $TP(c)$ ($FP(c)$) represent the true (false) classification of negatives and positives, respectively. The curve that summarises all possible combinations $TN(c), TP(c)$ is called the *correct classification frontier* or CC frontier. This frontier summarises a production possibilities frontier for the classifier $\hat{\delta}_{i,t}$ $\forall c$. Moreover, the points on the frontier represent the maximum percentage of true positives achievable for a given value of true negatives. As can be seen from the definitions of the classification rates, when $c \to \infty$ then $TP(c) \to 1$ and $TN(c) \to 0$, and vice versa if $c \to -\infty$. Therefore, the CC frontier takes value in the unit square.

The test compares the area under the CC frontier of the forecasts made and that of a coin-toss. The area under the CC frontier of a coin-toss is equal to 0.5: in this case $TP(c) = 1 - TN(c)$ $\forall c$ which translates into a diagonal line from $[0, 1]$ to $[1, 0]$ as the CC frontier. The area under the curve (AUC) for a model is then the integral of the CC frontier:

$$AUC = \int_0^1 CC(r)dr \tag{53}$$

Following Hsieh et al. (1996), the test statistic of the AUC larger than 0.5 (the AUC of a coin-toss) is given by:

$$JT_{i,m} = \frac{AUC - 0.5}{\sigma} \sim \mathcal{N}(0, 1) \tag{54}$$

Following Hanley and McNeil (1982) and Obuchowski (1994) the approximation of the variance of AUC is given by:

$$\sigma^2 = AUC(1 - AUC) + (T_P - 1)(Q_1 - AUC^2) + (T_N - 1)(Q_2 - AUC^2) \tag{55}$$

where

$$Q_1 = \frac{AUC}{2 - AUC}, \quad Q_2 = \frac{2AUC^2}{1 + AUC} \tag{56}$$

The last two measures of the forecast performance of the models are the most practical. These measures are trading strategies based on the virtual possibility of trading the underlying price of the 5-year CDS contract. Both trading strategies are long (short) the CDS spread when a model forecasts a positive (negative) change in the CDS spread. The first trading strategy invests one dollar every week and unwinds his position the week after in every company. In other words, this measure represents the cumulative sum of weekly returns and is closely related to the EP test. Instead of testing whether this type of trading strategy is better than a coin-flip, it shows the actual profits that could have been obtained in the sample period. The second trading strategy invests one dollar at the beginning of the sample in every company according to the buy/sell signal of a model. Instead of unwinding his positions every week, it reinvests the full realised profit according to these buy/sell signals. This trading strategy is a measure of cumulative compounding of weekly returns.

The difference in results of the first trading strategy are more easy to interpret between different models: more profits are obtained if the model predicts the direction of large changes in the CDS spread more accurate. The second trading strategy is a more realistic scenario of an investor who invests according to buy/sell signals of the forecasting models. The initial investment of one dollar is kept in the trading strategy until the end of the sample period. Transactions costs are not taken into account for both trading strategies and there are not limitations on short selling. This means that when a model predicts a negative change (the investor goes short), but the price of the CDS spread more than doubles in value, the investor can lose more than his initial investment. Note again, these trading strategies trade in the underlying price of the 5-year CDS spread and are therefore not buying/selling the CDS contract itself.

An overview of the characteristics of the forecasting evaluation measures is given in Appendix B. The first two measures show similar characteristics: they both take the forecasting fit into consideration and their benchmark are both the random walk model where the forecast is set to the value of today. Key difference between the two measures is that the DM measure tests the difference in forecasting fit, while the MSPE does not involve statistical testing. The next four measures, $PCPS$, PT, EP and JT, all incorporate the correctly predicted sign of the weekly change in CDS spread. $PCPS$ is the most basic measure, which does not involve statistical testing, does not incorporate profitability and the actual fraction of positive and negative weekly changes in CDS spread. The PT and EP measures do involve statistical testing and incorporate the actual fraction of positive and negative weekly changes in CDS spread. Besides, the EP test takes profitability into account. The JT measure does have the same characteristics as the previous three measures. Furthermore, it is the only measure that looks at different type of investors: buy/sell signals can differ between different type of investors. The JT measure incorporates the entire space of buy/sell signals. The last two measures are the most practical and look at the profits/losses that would have been obtained if an investor would invest according to simple buy/sell signals of the different models.

4 Results

In this section I analyse the results of the thirteen forecasting models. First, I focus on the two stages of the three new methods (PD, GC, and $LASSO$). I highlight the differences between performing PCA on the sub-samples of explanatory variables of these three methods and on the full sample. Next, I present empirical results, including the confirmation of the motivation of the use of forecast combination and out of sample forecast evaluations using the measures described in section 3.3. Finally, I analyse one of the trading strategies over time and for the 278 companies.

4.1 Two stage variable reduction methods: Variable Selection and Explained Variance

The three new two stage variable reduction methods require more attention than the other more widely used methods. Note that in the first stage, these three methods select a sub-sample of explanatory variables that have sufficient predictive power. In the second stage, the selected sub-sample of explanatory variables is used to either construct forecasts by equally weighting forecasts made with these explanatory variables individually or by using PCA. Especially the number of selected variables and the PCA of the second stage reveal more interesting information on which variables are selected and if these variables show the same behaviour.

The left graph of Figure 3 shows the average percentage of number of selected explanatory variables of the companies over time. Note that PCA always includes all 32 explanatory for constructing the components which results in a horizontal line at 100\$. The number of selected variables for the PD method stays stable over time, which indicates that the contribution to the long-run price discovery of the explanatory variables is stable over time. Moreover, a variable is only selected if the contribution to the price discovery is above the threshold value of 50%. This implies that there are only a couple of variables for which the contribution is around 50% and the volatility of these contributions is high enough to move around the threshold. On average, the PD method selects around 70% of the explanatory variables, which indicates on average 10 variables do no contribute more than 50% to the price discovery.

The right graph of Figure 3 shows much more variation in the average number of selected variables for the GC model compared to the other models. Especially the large differences between PD and GC should be mentioned. Both methods select explanatory variables based on the contribution to the information share. This difference is pointed out by Rittler (2012), who argues the information share proposed by Hasbrouck (1995) reveals the long-run price discovery while the information share proposed by Granger (1969) reveals the price discovery dynamics in the short-term. Therefore, the average number of selected variables differs between the PD and GC method. Although both methods use the same rolling window with the same sample period, price discovery in the long run is apparently more stable than in the short run. Moreover, the PD method always selects on average more explanatory variables than GC.

$LASSO$ shrinks the coefficients of the explanatory variables until at least seventeen of them are shrunk to zero. Therefore, one would expect a horizontal line at $46\%(\approx \frac{15}{32})$. However, this is not the case, which indicates that for multiple companies the explanatory variables are highly correlated resulting in shrinking two variables to zero at the same time.

The right graph of Figure 3 shows the fraction of variance explained by the first three principal components and the total variance of the explanatory variables. Note that the total variance is different for all methods, because PCA is performed on different sub-samples for all methods. Therefore, the more unstable explained variance of GC is explained by the large variation in selected variables (see the left graph of Figure 3). The variance explained by PCA, PD and $LASSO$ show the same patterns over time. The increase in variance explained starting at the beginning of 2008 indicates that the explanatory variables that did have predictive power showed an increase in correlation when the financial credit crisis started.

The PD method shows most of the variance explained over time. This is remarkable, considering the selected number of explanatory variables is the largest after PCA. Only three factors explained on average between 85% and 90%, indicating the explanatory variables selected by PD have a high correlation. The difference in explained variance by PCA, GC and $LASSO$ looks more intuitive, because the method with the largest number of selected variables explains the least variance and vice versa. This relation occurs for the GC method as well: the explained variance increases (decreases) in periods of a decrease (increase) in the average number of selected variables.

Figure 3: (l) The left figure displays the average percentage of selected explanatory variables of the companies over time. Note that PCA always includes all 32 explanatory for constructing the components which results in a horizontal line at 100%. (r) The right figure shows the average fraction of explained variance of the first three principal components per method over time.

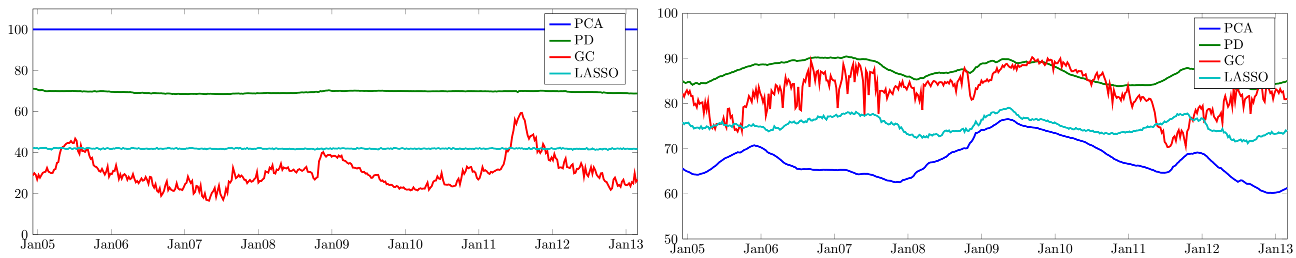


Table 1 shows the average number of times a variable is selected and the corresponding average loading to the first three principal components. There is a big difference in the average number of times a variable is selected between the four methods. The third column shows that the PD method almost always selects the same explanatory variables. On average, PD selects around 24 variables ($\approx 32 \times 74\%$) and the number of explanatory variables that are selected more than 84% of the time is equal to 24. All other variables are selected in less than 20% of the time. This result confirms the results shown in Figure 3, which show little variation in selected variables by the PD method.

The differences between the average number of times the explanatory variables are selected for the GC and $LASSO$ are less extreme. The average maximum (minimum) an explanatory variable is selected by GC is 47.4610% (19.4700%). For the $LASSO$ method the range is somewhat wider with 97.6676% and 14.9000%. However, for both methods the differences in percentages between the explanatory variables is much smaller than for the PD method.

The PD method selects especially the explanatory variables that are measured over a longer period: the long term variance, yearly returns, rates and fundamentals. As mentioned, the PD method reveals the long-run price discovery. Therefore, selecting these type of explanatory variables is not a surprise. On the other side, the GC method reveals the short-run price discovery, but seems to select explanatory variables that both are measured on the short- and long-term. Furthermore, $LASSO$ often selects explanatory variables that are not selected by the PD method or vice versa, e.q. 11 & 27.

The $CSIR$ factor (32) and both the IPD measures (1-2) have a substantial large loading to the first three principal components for all four methods. These variables, with especially the $CSIR$ factor, seem to be the explanatory variables that most drive the four models. The Slope Treasury (25) shows a high selection rate and loading compared to the other 'rates' variables (21-24). This explanatory variable, which is often seen as a measure of economic expansion/contraction, seems to be a leading indicator for the CDS spreads. The explanatory variables on stock data, option data, fundamentals, rates and indices are on average many times selected and have large loadings to the first three principal components. These explanatory variables tend to influence the performance of all models.

Table 1: The table shows the average number of times an explanatory variable is selected and the corresponding average loading to the first three principal components. Note that for PCA all explanatory variables are always selected resulting in all values of the first column equal to 100%. For the other three methods PCA is one of the two methods in the second stage. For these last three methods, the loading is set equal to zero in case an explanatory variable is not selected. The average is taken including these zero loadings. The last row represents the average percentage of selected variables and average percentage of variance explained by the first three principal components.

	PCA		PD		GC		LASSO	
	% Selected	Loading	% Selected	Loading	% Selected	Loading	% Selected	Loading
1.IPD CDS	100.0000	7.9157	99.7290	8.4904	42.5195	12.5653	94.6676	11.2884
2.IPD stock	100.0000	6.4064	84.3208	11.8828	47.4610	13.8528	58.9263	8.9837
3.Stock price	100.0000	9.9287	99.9797	12.1557	24.5580	5.0343	35.5813	6.3637
4.Traded volume	100.0000	4.0204	10.2226	1.4007	26.1343	8.1198	68.5719	15.7067
5.Stock spread	100.0000	2.4743	2.7723	0.8896	43.6232	15.0727	72.2127	15.0820
6.SO stock	100.0000	9.3006	99.4676	7.7419	19.4700	6.4689	46.8336	4.1966
7.RV stock	100.0000	9.3708	99.9992	10.8005	42.3296	10.9118	43.9894	7.0739
8.LT momentum	100.0000	9.3379	99.9727	21.3332	20.9822	6.5363	46.0176	12.2934
9.ST momentum	100.0000	18.8456	0.0009	0.0003	33.5120	11.0428	49.4256	21.0027
10.ATM put IV	100.0000	8.1648	97.2002	10.1611	35.2542	8.9285	55.8189	8.4727
11.OTM put IV	100.0000	2.1062	19.7340	6.0181	20.9271	8.4738	63.2273	24.6282
12.RV ATM put	100.0000	9.3530	99.9188	10.7293	33.3979	8.0621	28.6435	4.3126
13.Leverage ratio	100.0000	6.7982	99.3328	15.0140	17.5603	6.3962	58.9191	9.3771
14.P/E ratio	100.0000	8.3906	93.0350	14.2974	21.8913	5.6818	35.7571	8.1907
15.Div Ind Yld	100.0000	7.4779	84.9873	9.1244	32.6205	7.9570	36.0808	5.4764
16.P/B ratio	100.0000	8.5273	93.5888	11.1858	24.3494	5.6010	30.9526	5.0542
17.P/S ratio	100.0000	9.1099	94.3205	9.8333	22.7306	4.6429	15.4164	2.3788
18.EV/EBITDA	100.0000	8.1838	88.9609	12.9418	19.8412	5.2159	29.9719	6.3409
19.Size	100.0000	9.9199	99.6380	10.2073	23.9926	4.7253	14.9000	2.0451
20.P/C ratio	100.0000	7.4015	93.9297	18.3294	18.0571	5.0048	40.6922	9.5784
21.Swap rate	100.0000	9.0740	99.9990	6.5974	35.9174	9.5548	32.5131	2.9335
22.Treasury	100.0000	7.9140	99.9990	8.3798	34.5344	10.0879	44.1630	3.7828
23.T-Bill	100.0000	9.0320	99.9612	9.9895	36.1881	9.6751	27.8094	2.9820
24.Libor	100.0000	9.8357	99.9990	9.2880	45.6823	12.6779	34.1691	2.9109
25.Slope Treasury	100.0000	9.4934	100.0000	14.3279	34.8137	9.1360	35.8759	7.2926
26.Return S&P	100.0000	3.1657	0.0067	0.0032	32.4608	12.3579	62.4432	22.2189
27.RV S&P	100.0000	8.7214	99.9961	10.7606	51.8666	13.9101	38.3267	6.1909
28.VIX	100.0000	7.6722	99.9161	13.3385	35.2983	9.1032	67.7338	11.3075
29.FF3	100.0000	22.2883	0.0000	0.0000	33.5237	10.3842	28.1745	11.2938
30.FF5	100.0000	21.9544	0.0000	0.0000	33.5753	10.4142	29.9382	11.9061
31.CAPM	100.0000	22.3765	0.0000	0.0000	34.3050	10.7931	27.2025	10.2973
32.CSIR factor	100.0000	5.4388	99.8912	24.7543	39.6102	13.9646	68.9842	19.0377
Average	100.0000	67.7484	73.7775	86.7903	31.8434	82.9662	44.4981	74.7066

4.2 Confirmation of forecast combination motivation and forecast performance

Table 2 shows the average number of times the weekly forecasts of the specific model have the lowest squared error. This percentage is constructed by counting the number of weeks that a model has the lowest squared error and dividing this by the total number of weeks. The values in the table represent the averages taken over all companies. Models (3-5) and models (6,8,10), the two stage variable reduction methods with PCA as the second stage, outperform in more weeks than the $AR(1)$ model by means of squared errors. This indicates that modelling the error term of the AR(1) by means of various methods does improve the forecast performance.

The two stage methods that equally weight the forecasts of individual explanatory variables (7,9,11) seem not to be an improvement to the $AR(1)$ model. However, when PCA is performed on the same sub-sample of explanatory variables there seems to be an improvement. The first three factors seem to be able to capture the most important correlation of the selected explanatory variables in terms of forecasting performance. The best performing model based on the results of Table 2 is the $3PRF$. Similar to the results in Kelly and Pruitt (2015), the $3PRF$ demonstrates strong forecasting performance and is superior to the alternative methods.

Although multiple models are an improvement to the $AR(1)$ model, none of these models outperform the RW in terms of number of weeks of lowest squared error. As mentioned previously, this was expected, because single models do often not structurally outperform the benchmark model. However, the RW only has the lowest squared error for 21.2785% on average of the time. This means that in 78.7215% of the time at least one of the models is outperforming the RW by means of the fit of the predictions. This confirms the motivation of using combinations of forecasts in this research. The second and the third column show the motivation behind the short training period for computing the forecasting combinations based on the MSPE and Sign. The table shows that the individual models outperform for an average period of around two weeks. The forecasting combination models include forecasts based on the performance of last week and given that the individual models show good performance in the week after, the choice of only looking at the past period seems to be the right one.

Table 2: The second column shows the average number of times the forecasts of the specific model have the lowest weekly squared error. The benchmark model RW is constructed by setting the forecast for next week equal to the value of today. The second and the third column show the average number of weeks a model outperforms given that it outperforms in a certain week.

Model	% Lowest Squared Error	Length Outper MSPE	Length Outper Sign
1. RW	21.2785	-	-
2. AR(1)	4.9783	2.0125	2.2533
3. CSIR	7.8728	1.9686	2.2519
4. PCA	9.2055	1.8299	2.1890
5. 3PRF	15.424	1.7390	2.1725
6. PD PCA	10.8945	1.9147	2.2568
7. PD EW	2.3964	2.0108	2.2529
8. GC PCA	10.4752	1.8037	2.1478
9. GC EW	3.6113	1.9767	2.2369
10. LASSO PCA	10.9781	1.8272	2.1937
11. LASSO EW	2.8854	1.9958	2.2522

Table 3 shows the forecast evaluations of the thirteen models by means of the measures described in section 3.3. The first column indicates that none of the individual models outperforms the benchmark model in terms of $MSPE$ averaged over the companies. Furthermore, the second column, which shows the percentage of companies a model has outperformed the benchmark model in terms of the $MSPE$, shows that the best performing model only outperforms the benchmark for 5% of the companies. These findings of poor forecast performance in terms of fit of the individual models are confirmed by the DM statistic. The average Z-scores are negative indicating the benchmark model is on average better in forecasting the CDS spread than the individual models.

Table 2 showed that the $3PRF$ and the two stage models with PCA were outperforming the benchmark in 10%-15% of the time. However, these models perform poorly when looking at both measures of $MSPE$. The large $MSPE$ value of $3PRF$ in combination with the results of Table 2 indicate the $3PRF$ is often able to construct good forecasts, but often makes large errors as well. In terms of average $MSPE$ the two stage models with PCA outperform the two stage models with equal weighting of forecasts. However, the latter outperforms the benchmark for more companies than the former. This indicates that the two stage models with equal weighting are more often making large forecasting errors compared to the two stage models with PCA.

The individual models show better results for the more economic relevant measures, although the averages of the $PCPS$, PT, EP and JT do not seem promising. The directional accuracy tested by $PCPS$ and PT show similar results. None of the individual models outperforms the benchmark on average. However, the performance of the models is much better in terms of direction than in terms of fit. The EP test and the JT test incorporate the profitability of the buy/sell signals of the models by putting more weight on forecasts if there is big change in the CDS spread. The EP test considers a buy (sell) signal if the model predicts a positive (negative) change. This trading strategy is in line with the last two columns of Table 3. Therefore, these two measures show similar results: models with higher values in the two columns of the EP test are more profitable. Furthermore, if an investor would follow the second trading strategy (last column) by reinvesting his total capital every week it can outperform the benchmark strategy of *Alwayslong*. However, the JT test incorporates trading strategies with all different possibilities of buy/sell signals. This test shows that the models outperform the benchmark for more companies than the EP test. This indicates that an investor could make more money by adjusting the threshold for the buy/sell signals from zero to another value.

The results of the forecast combinations confirm the motivation of the use of them. Forecast combinations based on equal weighting (EW) outperforms the other models in terms of $PCPS$, EP and the first trading strategy. However, Table 2 shows that there should exist a combination of forecasts that could substantially improve the performance. Model 13 selects the models that outperformed in the last period in terms of $MSPE$ and constructs an equally weighted forecast combination with these models. This type of combining improves the results substantially in terms of most of the measures. Especially the EP test, JT test and the trading strategies indicate a big improvement. Even the number of companies for which it has a smaller $MSPE$ or DM statistic higher than 1.65 increases substantial. This confirms the motivation of the use of combining forecasts. The improvement indicates that there are some periods of subsequent weeks where some of the models outperform the benchmark. However, model 14 shows outstanding results indicating that it outperforms the benchmark model. This model selects the models that correctly predicted the direction of the CDS spread in the last period and constructs an equally weighted forecast of these models. The outstanding results indicate that the individual models correctly predict the sign in subsequent weeks. Not even in terms of the last six measures, which mainly focus on the directional accuracy, but even for the measures of fit the model shows huge improvements. Except for the DM-test, the percentage of companies for which model 14 outperforms the benchmark is higher than 75% for all other measures. The profits that could have been made of the two trading strategies confirms that model 14 significantly outperforms all other models.

Table 3: This table shows the forecast evaluation measures (described in Section 3.3) of the thirteen models and the benchmark model. The left column of the MSPE (Mean Squared Prediction Error) shows the average over all companies in basis points. The right column represents the percentage of companies the model has outperformed the benchmark model in terms of the MSPE. The forecasts of the benchmark model are set equal to today's value. The columns of DM (Diebold Mariano) represent the average Z-score and the percentage of companies for which the model outperformed the benchmark (Z-score larger than 95% critical value of 1.65). The benchmark is defined in the same way as for the MSPE and Z-score. The values for the % Outperforming are not defined, because the *DM* test compares the models to this benchmark. The *PCPS* (Percentage of Correctly Predicted Signs) averages the percentages over all companies. The % Outperforming indicates the percentage of companies for which the *PCPS* is larger than 50%. The benchmark model represents the coin-flip, with a theoretical probability of 50% correct predictions. The columns of *PT* (Pesaran-Timmermann), *EP* (Excess Profitability) and *JT* (Jorda & Taylor) represent the Z-scores averaged over the companies and the percentage of companies for which the Z-score is larger than the 95% critical value of 1.645. These three measures evaluate the forecast performance compared to the benchmark models. Therefore, the Z-score are not defined for the benchmark model. The % Outperforming is set equal to a theoretical value of 50% indicating an equal probability of outperforming. The last two columns are the two trading strategies that trade in the underlying price of the 5-year CDS contracts. The first trading strategy can be seen as the cumulative sum of returns after investing 1$ in every company every week. The second trading strategy cumulatively compounds the weekly returns after an initial 1 $ investment in every company. The values shown in the last two columns represent the sum of all profits/losses over all companies made at the end of the sample. Both trading strategies go long (short) the CDS if the model predicts a positive (negative) change. Transaction costs are not taken into account and there are no restrictions on short selling. The benchmark models for both trading strategies always predict a positive change in CDS (Always long strategy).

	MSPE		DM		PCPS		PT		EP		JT		Trade 1	Trade 2
Model	bps	% Outper	Z-score	% Outper	Mean	% Outper	Z-score	% Outper	Z-score	% Outper	Z-score	% Outper	$ profit	$ profit
1.Benchmark	2.0395	-	-	-	0.5000	50.0000	-	50.0000	-	50.0000	-	50.0000	680.47	653.35
2.AR(1)	2.2591	2.9091	-0.7115	0.3597	0.4952	53.5971	-0.3092	12.5899	0.1369	1.7986	0.3410	11.1913	-52.20	-94.11
3.CSIR	2.3496	1.0909	-0.7849	0.0000	0.4965	56.4748	-0.2998	12.2302	0.1416	6.1151	0.3910	14.4404	85.38	167.55
4.PCA	2.4961	1.8182	-0.7412	0.7194	0.4961	52.5180	-0.1768	12.2302	0.5487	11.8705	0.5400	17.6895	355.13	826.16
5.3PRF	124.2435	0.0000	-1.1257	0.0000	0.4981	57.5540	-0.2362	11.8705	0.5615	15.1079	0.5304	13.7184	302.29	414.00
6.PD PCA	2.3967	0.7273	-0.7296	0.7194	0.4984	57.1942	-0.0710	16.9065	0.5800	17.6259	0.5586	17.6895	384.08	1,667.02
7.PD EW	2.5602	3.6364	-0.6574	0.0000	0.4979	55.3957	-0.1099	12.9496	0.3086	5.3957	0.4778	14.4404	104.53	214.76
8.GC PCA	2.5628	2.1818	-0.7253	0.0000	0.4944	50.0000	-0.0956	13.6691	0.5233	15.1079	0.4796	17.3285	311.05	805.79
9.GC EW	3.9619	4.0000	-0.6230	0.3597	0.4988	56.8345	-0.0370	15.1079	0.3944	7.5540	0.5541	18.4116	140.86	388.98
10.LASSO PCA	2.4409	0.7273	-0.6276	0.3597	0.5024	59.3525	0.0989	19.7842	0.7337	19.4245	0.8224	24.1877	457.68	1,153.97
11.LASSO EW	3.5370	5.0909	-0.6596	0.0000	0.4999	58.9928	0.0067	15.4676	0.3765	6.4748	0.6069	17.3285	150.01	188.82
12.EW	3.5690	4.7273	-0.5795	0.0000	0.5063	66.5468	-0.0005	18.3453	0.7107	20.1439	0.8164	22.3827	394.73	1,361.89
13.Lag MSPE EW	2.4461	24.3636	-0.1652	2.1583	0.4264	9.3525	-3.6520	1.0791	1.2327	34.5324	1.5097	44.0433	696.44	6,641.88
14.Lag Sign EW	3.4925	75.6364	0.5042	9.3525	0.5760	88.8489	2.7441	79.1367	2.7870	81.2950	4.2643	90.6137	1,683.92	566,550.37

4.3 Trading Strategies: luck or intelligence?

This last subsection analyses the performance of the models in more detail. In the previous section the performance in terms of statistical and economic measures were discussed. These measures showed the averages over all companies and reported the percentage of companies for which the models outperformed. Especially the models based on combining the forecasts of the individual models seemed to perform well over the entire sample. However, by approaching the results in more detail I tend to answer the question whether these results are based on luck or whether the models are actually able to construct good forecasts. The most straightforward way to answer this forecast is to look at the the first trading strategy, which looks at the cumulative sum of weekly returns. This trading approach does not require any complicated methods for giving buy and sell signals and is easy to interpret (more money is better). Furthermore, this trading strategy is less influenced by past returns compared to the second trading strategy of cumulative compounding of returns.

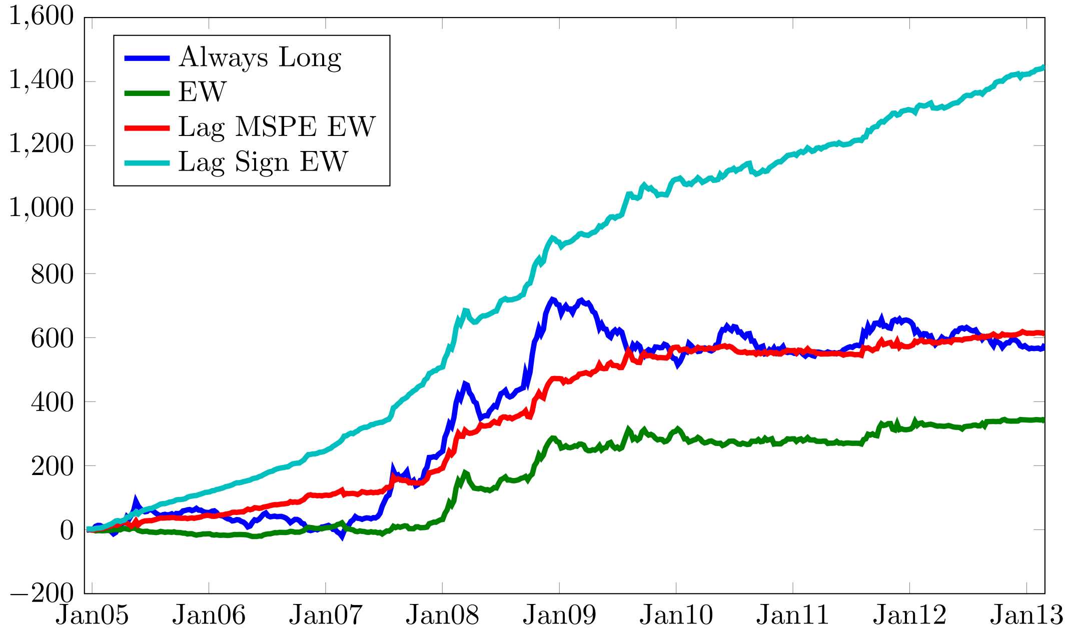
In the previous section, the combination of forecasts showed the best performance over the entire sample. With the last model, which selects the individual models that correctly predicted the sign in the last period and equally weights the forecasts of these model for the next period, as clear winner in terms of outperforming. Therefore, Figure 4 shows the cumulative sum of weekly returns for these models and the benchmark model *Always Long*. The figure shows that the profitability at the end of the sample is almost equal for *Always Long* (dark blue) and *Lag MSPE EW* (red). However, the volatility of cumulative profits is much higher for *Always Long*. Thus, most investor would prefer the trading strategy based on *Lag MSPE EW*, because it shows the same profits with less risk.

The dark blue (*Always Long*), green (*EW*) and red (*Lag MSPE EW*) lines show similar behaviour. In the period 2007 to mid 2009 the trading strategies based on the first three models seem to be profitable. After this period the profitability decreases. The two main reasons for this decrease are already mentioned in section 1. The first reason is that the lag of transparency and large taken risk in CDS contracts were the main drivers of the financial credit crisis. As a result, there has been a large push by US regulatory and supervisory entities to the make the derivative market more transparent. Therefore, investor who had an incentive of making profit on CDS contracts before the crisis were less interested in trading these CDS contracts after this large push. Second, the market standard changed from the MR clause to the NR clause in April 2009. Because of this event, there was much less trading in CDS contracts with the MR clause, which resulted in less volatility in these CDS contracts after April 2009. In this research I focus on the CDS contracts with the MR clause. Thus the volatility of the CDS spreads in my dataset decreased after the introduction of the new standard.

In contrast to the decreasing profits of *Always Long*, *EW*, and *Lag MSPE EW*, the light blue line (*Lag Sign EW*) shows a stable increase of profits over time. In the last section, this combination of forecasts showed to be able to outperform the benchmark over the entire period. Figure 4 shows that it structurally outperforms the benchmark as well. Compared to the other models, this model is not influenced by the financial crisis and the change of the new market standard. These results confirm what is shown in Table 3 of the last section. Moreover, this model shows less volatility in profits than the *Always Long* strategy over time. Based on these results, the luck component of constructing forecasts for this strategy looks to be eliminated.

Appendix C shows the profitability of the other models over time. These figures confirm the results of the previous findings. The profitability of the AR(1), *CSIR*, *PCA* and 3*PRF* models do not show much volatility, but do not seem to be able to outperform the *Always Long* strategy. Of the six new methods, the methods that use *PCA* in the second stage outperform the methods that use EW. Furthermore, all six new methods show similar behaviour: the peaks are at the same time, but the methods that use *PCA* in the second stage correctly predict the direction for more companies than the methods that use EW.

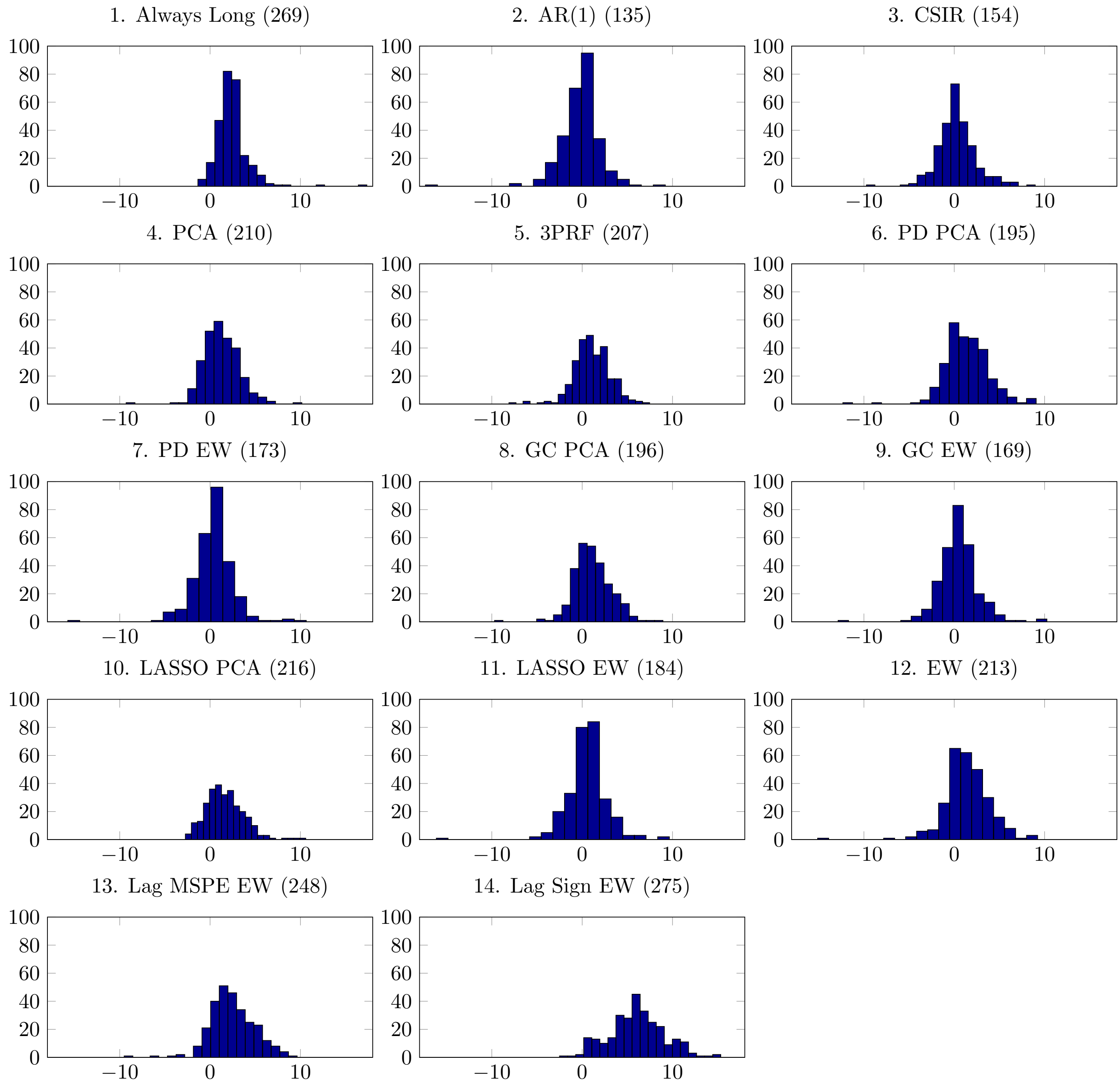
Figure 4: The graph shows the performance of the first trading strategy over time. This trading strategy goes long (short) 1$ in every company when a model forecasts a positive (negative) change in the CDS spread. In other words, the graph shows the cumulative sum of weekly returns over all companies. The dark blue line (Always Long) is the benchmark model which always predicts a positive change. The other three models are the combination of forecast models. On average, these models showed the best performance over the entire sample (see Table 3).



So far, the performance of the models is shown for the entire sample and for the profitability over time. These results could be biased given that my sample consists out of 278 companies and the results shown previously are all summed over the companies. Therefore, Figure 5 shows a histogram of the profitability of all companies over the entire sample period. First thing to highlight is that the *AlwaysLong* strategy is profitable for almost all companies (269 out of 278). Only the last model is profitable for more companies (275). Models 1, 4, 6, 8, 10, 12, 13 show a clear right-skewed distribution. All other models seem to be more centered around the mean. All individual models have a mean close to zero, but most of them make huge losses for a small amount of companies and struggle to make up for these losses with other companies. Therefore, the profits made by these models shown in Table 3 of the previous section are somewhat disappointing.

The profits of the combination of forecasts show on average good results for all companies. Both *EW* (12) and *Lag MSPE EW* (13) do make huge losses for a couple of companies which looks like the reason why these models do not structurally outperform the benchmark model. This is in contrast to the last model, which already showed great results in the previous sections, but now confirms its great ability to forecast the CDS spread. For only three companies this model was not able to generate profitable trading strategies. The mean of the histogram centres around a 7$ dollar profit, which is a huge improvement compared to the 2$ of the *Alwayslong* strategy. This indicates that on average a ‘simple’ trading strategy based on model 14 would double an investor’s money seven times in the period end of 2004 - beginning of 2013.

Figure 5: The histograms show the profitability of the first trading strategy in dollars of all companies over the entire sample period. The number between brackets is the number of companies for which a profit is earned at the end of the sample. Note the total number of companies is equal to 278.



5 Conclusion

In this paper, I provide a comprehensive study on the predictability of the 5-year CDS spreads of USD-denominated contracts of U.S.-based obligors. The dataset comprises 278 companies over a period of 583 weeks. The predictability is analysed by the use of thirteen models. The first model is an ARMA model and is the starting point for all other models. Nine of these models model the error term of the ARMA by incorporating a large number of explanatory variables. One of the explanatory variables is a new factor which captures the cross-sectional-idiosyncratic-risk of CDS spread, which shows to have predictive power for the CDS spread. The other explanatory variables are constructed using data on different asset classes, fundamentals and macro-economic variables. Three of the thirteen models are combinations of forecasts of these nine models that incorporate explanatory variables.

Multiple variable reduction techniques are used for the forecasting models. Three new approaches of variable reduction techniques are proposed. These techniques make use of a two stage approach. The first stage selects a number of explanatory variables that have sufficient explanatory power. The second stage either uses PCA to construct a forecast with these variables or equally weights the forecasts made with models consisting of one of the selected variables. These models show an improvement to the more original PCA.

The forecasting performance is evaluated using multiple statistical tests. These tests compare the fit of the forecasts, the directional accuracy and the profitability of the models compared to a benchmark model. On average, the models show better performance in terms of directional accuracy than in terms of forecasting fit. I find that only a combination of forecasts is able to outperform benchmark models. This indicates that the nine models that use explanatory variables do show periods of outperforming, but are not able to structurally outperform the benchmark models. I find that a model that combines forecasts of models that correctly predicted the sign in the previous period outperforms all other models. Moreover, trading strategies based on this model were able to generate huge profits. Investing one dollar in every company at the end of 2004 and reinvesting the total capital made according to the buy/sell signals of this model would yield a profit of $566,550.37 at the beginning of 2013. Following this strategy, money is only lost for 3 out of 278 companies.

The impressive results show that this combination of forecasts approach can be useful in practice. Investor can base profitable trading strategies on this approach and financial institutions can get an advantage in credit hedging costs by potentially delaying the purchase of the CDS contract. However, only the 5-year CDS contracts with the MR clause are considered. This clause was the market standard until April 2009. Therefore, the decision to only include these contracts could have biased the results. However, the decrease in number of CDS contracts did not decrease substantially after April 2009. This research could be extended by including the new market standard, the NR clause, and incorporate CDS spreads of different maturities, although liquidity will cause issues given that the 5-year CDS contract is the most liquid among all maturities.

The trading strategies give an indication of the performance of the models. These strategies do not incorporate transaction costs and restrictions on short selling. More important, the trading strategies trade in the underlying price of 5-year CDS contracts. To my knowledge, these type of price trackers are not commonly offered by most brokers. Therefore, the profits made by these trading strategies do not replicate a realistic scenario. However, in financial markets there is (almost) always a counterparty who wants to take on the other side of a trade. Furthermore, these trading strategies are more realistic when trading in derivatives with the 5-year CDS contracts as the underlying.

This paper makes three contributions to the literature. The first is the introduction of the Cross-Sectional-Idiosyncratic-Risk factor, which shows to have predictive power for the CDS spread. The second contribution consists of new techniques of variable reduction. These techniques tend to outperform the widely used PCA. Lastly, I show that these techniques can be used for models that forecast the CDS spreads. By incorporating a large number of explanatory variables on different asset classes, fundamentals and macro-economic variables the models are able to outperform benchmark models in sub-samples of the dataset. A model that constructs the next period forecast by combining forecasts of these models that correctly predicted the sign in the last period is able to detect the sub-samples in which the models outperformed the benchmark. This model is able to outperform the benchmark for the entire sample in terms of statistical and economic measures and two ‘simple’ trading strategies based on the forecast direction are able to generate huge profits.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

Altman, E. I. (1989). Measuring corporate bond mortality and performance. *Journal of finance*, pages 909–922.

Anatolyev, S. and Gerko, A. (2005). A trading approach to testing for predictability. *Journal of Business & Economic Statistics*, 23(4):455–461.

Armstrong, J. S. (2001). Combining forecasts. In *Principles of forecasting*, pages 417–439. Springer.

Armstrong, J. S. and Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, 8(1):69–80.

Bai, J. and Wu, L. (2012). Anchoring credit default swap spreads to firm fundamentals. *Available at SSRN 2020841*.

Baillie, R. T., Booth, G. G., Tse, Y., and Zabotina, T. (2002). Price discovery and common factor models. *Journal of financial markets*, 5(3):309–321.

Berndt, A. and Obreja, I. (2010). Decomposing european cds returns*. *Review of Finance*, 14(2):189–233.

Berndt, A. and Ostrovnaya, A. (2007). Information flow between credit default swap, option and equity markets. *Tepper School of Business, Carnegie-Mellon Working Paper*.

Bharath, S. T. and Shumway, T. (2004). Forecasting default with the kmv-merton model. In *AFA 2006 Boston Meetings Paper*.

Bharath, S. T. and Shumway, T. (2008). Forecasting default with the merton distance to default model. *Review of Financial Studies*, 21(3):1339–1369.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654.

Brezinski, C. (2005). La méthode de cholesky. *Revue dhistoire des mathématiques*, 11(2):205–238.

Carr, P. and Wu, L. (2010). Stock options and credit default swaps: A joint framework for valuation and estimation. *Journal of Financial Econometrics*, 8(4):409–449.

Carr, P. and Wu, L. (2011). A simple robust link between american puts and credit protection. *Review of Financial Studies*, 24(2):473–505.

Christoffersen, P. F. and Diebold, F. X. (2006). Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, 52(8):1273–1287.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

Collin-Dufresne, P., Goldstein, R. S., and Martin, J. S. (2001). The determinants of credit spread changes. *The Journal of Finance*, 56(6):2177–2207.

Cremers, M., Driessen, J., Maenhout, P., and Weinbaum, D. (2008). Individual stock-option prices and credit spreads. *Journal of Banking & Finance*, 32(12):2706–2715.

Diebold, F. X. and Mariano, R. S. (2012). Comparing predictive accuracy. *Journal of Business & economic statistics*.

Du, Y. (2009). Price discovery of credit risk.

Duffie, D., Saita, L., and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3):635–665.

Elliott, G. and Timmermann, A. (2013). *Handbook of Economic Forecasting SET 2A-2B*. Elsevier.

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

Forte, S. and Lovreta, L. (2009). Credit risk discovery in the stock and cds markets: Who leads, when, and why . *When, and Why (December 17, 2009)*.

Forte, S. and Pena, J. I. (2009). Credit spreads: An empirical analysis on the informational content of stocks, bonds, and cds. *Journal of Banking & Finance*, 33(11):2013–2025.

Fung, H.-G., Sierra, G. E., Yau, J., and Zhang, G. (2008). Are the us stock market and credit default swap market related? evidence from the cdx indices. *Journal of Alternative Investments, Summer*.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Granger, C. W. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204.

Greer, M. (2003). Directional accuracy tests of long-term interest rate forecasts. *International Journal of Forecasting*, 19(2):291–298.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *Journal of Finance*, pages 1175–1199.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, pages 1251–1271.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hsieh, F., Turnbull, B. W., et al. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics*, 24(1):25–40.

Hull, J. C., Predescu, M., and White, A. (2005). Bond prices, default probabilities and risk premiums. *Default Probabilities and Risk Premiums (March 9, 2005).*

Jordà, Ò. and Taylor, A. M. (2011). Performance evaluation of zero net-investment strategies. Technical report, National Bureau of Economic Research.

Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316.

Lawrence, M. J., Edmundson, R. H., and O'Connor, M. J. (1986). The accuracy of combining judgemental and statistical forecasts. *Management Science*, 32(12):1521–1532.

Li, K. (2002). Long-memory versus option-implied volatility predictions. *The Journal of Derivatives*, 9(3):9–25.

Lin, H., Wu, C., and Zhou, G. (2015). Forecasting corporate bond returns: A regressed combination approach. *Available at SSRN 2346299.*

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

Obuchowski, N. A. (1994). Computing sample size for receiver operating characteristic studies. *Investigative Radiology*, 29(2):238–243.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Peña, J. I. and Forte, S. (2006). Credit spreads: Theory and evidence about the information content of stocks, bonds and cdss. *Business Economics Working Papers, Universidad Carlos III, Departamento de Economía de la Empresa.*

Pesaran, M. H. and Timmermann, A. (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics*, 10(4):461–465.

Rittler, D. (2012). Price discovery and volatility spillovers in the european union emissions trading scheme: A high-frequency analysis. *Journal of Banking & Finance*, 36(3):774–785.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

A Derivation KMV-merton model

The KMV-merton uses the proposed method of Merton (1974) by looking at the distance to default (DD). The DD means the number of standard deviations the stock price is away from the point where it defaults. After deriving this DD, one can approximate the implied probability of default by assuming stock prices follow a normal distribution.

Merton (1974) uses a Black and Scholes (1973) framework where he states the value of a company follows a geometric. This model is later developed and noways broadly used by credit rating agencies such as Moodies and S&P and is often referred to as the KMV-model. In this model the capital structure of a company is defined as follows:

$$V_A(t) = D(t) + V_E(t) \tag{57}$$

Which in other words can be seen as the value of the assets of a company is the sum of the debt and the equity value. In this framework the equity can be seen as a call option on the assets of the company with strike price equal to the debt value. The value of the assets are assumed to follow a geometric Brownian motion such that

$$dV = \mu V dt + \sigma_v V dZ \tag{58}$$

Given the equity is a call option on this value of assets with strike price equal to the debt value, the equity price is obtained as

$$V_E(t) = V_A(t)\Phi(d1) - D(t)exp(-r(T-t))\Phi(d2) \tag{59}$$

with:

$$\begin{gathered} d1 = \frac{log\left(\frac{V_A(t)}{D(t)}\right) + (r - 0.5\sigma_A^2)(T-t)}{\sigma_A\sqrt{T-t}} \\ d2 = d1 - \sigma_A\sqrt{T-t} \end{gathered} \tag{60}$$

Using Ito's formula once can show that

$$\sigma_E = \frac{V_A}{V_E}\frac{\partial V_E}{\partial V_A}\sigma_A \tag{61}$$

Then the Distance-to-Default is given by:

$$DD(t) = \frac{log\left(\frac{V_A(t)}{D(t)}\right) + (r - 0.5\sigma_A^2)(T-t}{\sigma_A\sqrt{T-t}} \tag{62}$$

And implied probability of default:

$$IPD_s(t) = P[V_A \leq D] = ... = \Phi(-DD(t)) \tag{63}$$

The value of the assets ($V_A(t)$) and the volatility of the assets (σ_A) are obtained by simultaneously solving Equations 59 and 61. This requires a lot of computational time for large datasets.

B Forecast evaluation measures

Table B.1: The table shows the characteristics of the different forecasting evaluation measures described in section 3.3. The second column indicates whether a measure takes the forecasting fit into consideration. The second and the third column indicate whether the sign of the predicted change and the profitability are taken into consideration, respectively. The fourth column shows the benchmarks of the forecasting measures. RW is defined as the random walk, where the forecast for the next period is set to the value of today. The coin-flip benchmark forecasts a 50% chance of correctly predicted the sign of the change in CDS spread. The weighted coin-flip takes the realised number of positive and negative changes into account and Always Long benchmark gives always buy signals to a trader. The last column is especially for the *JT* measure, because that is the only measure that takes different kind of buy/sell signals into consideration. The other measures which take profitability into account consider a buy (sell) signal when the predicted change is positive (negative).

Measure	Statistical test	Fit	Sign	Profitability	Benchmark	Different type of investors
MSPE	-	+	-	-	RW	-
DM	+	+	-	-	RW	-
PCPS	-	-	+	-	Coin-flip	-
PT	+	-	+	-	Weighted Coin-flip	-
EP	+	-	+	+	Weighted Coin-flip	-
JT	+	-	+	+	Weighted Coin-flip	+
Trade 1	-	-	+	+	Always Long	-
Trade 2	-	-	+	+	Always Long	-

C Figures trading strategies

Figure C.1: The graph shows the performance of the first trading strategy over time. This trading strategy goes long (short) 1$ in every company when a model forecasts a positive (negative) change in the CDS spread. In other words, the graph shows the cumulative sum of weekly returns over all companies. The dark blue line (Always Long) is the benchmark model which always predicts a positive change. The other models are the first four models described in section 3.

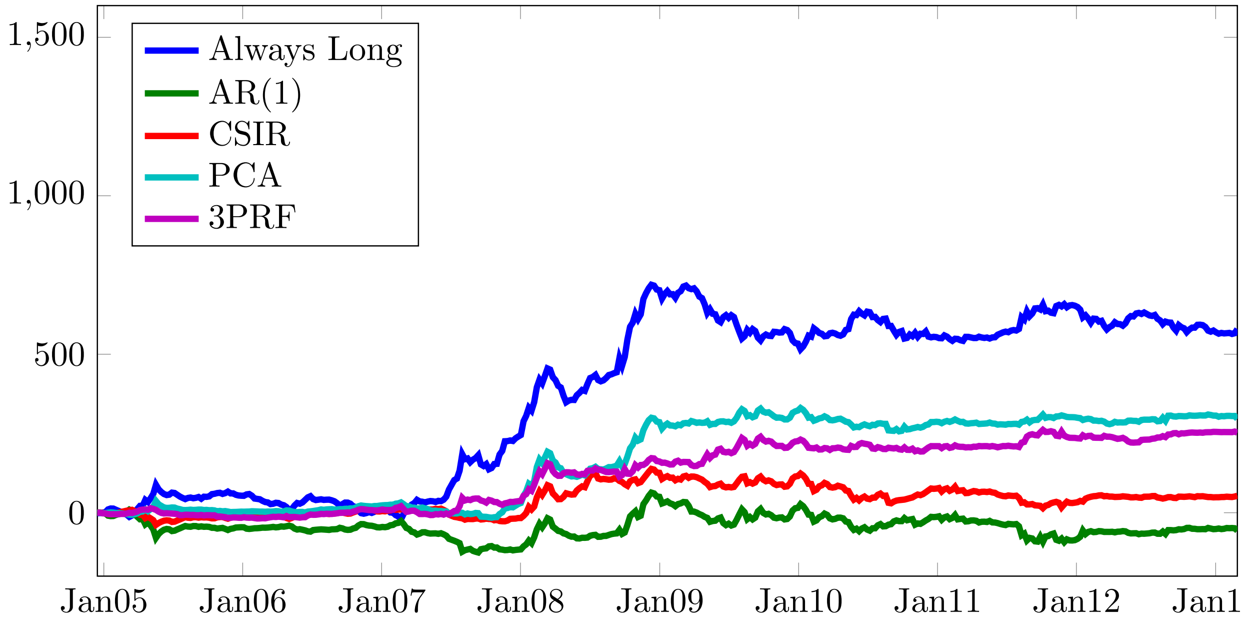


Figure C.2: The graph shows the performance of the first trading strategy over time. This trading strategy goes long (short) 1$ in every company when a model forecasts a positive (negative) change in the CDS spread. In other words, the graph shows the cumulative sum of weekly returns over all companies. The dark blue line (Always Long) is the benchmark model which always predicts a positive change. The other models are the six models that use the three new variable reduction techniques.