UNIVERSITY OF SOUTHAMPTON

FACULTY OF BUSINESS, LAW AND ART

Southampton Business School

Forecasting Financial Markets with Online Information

by

Paul Vincent Gaskell

Thesis for the degree of Doctor of Philosophy

September 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF BUSINESS, LAW AND ART

SOUTHAMPTON BUSINESS SCHOOL

Forecasting Financial Markets with Online Information

Doctor of Philosophy

By Paul Vincent Gaskell

This thesis explores the relationship between what investors say on online social media and price movements in financial markets. Recent studies have applied techniques from the natural language processing literature to distil the content of blogs, micro-blogs and user generated content sites to a 'sentiment' measure, pertaining to whether the content is good or bad for a given stock. Sentiment is then measured over a time series and compared to stock price returns. There is general agreement in the literature that a relationship between online sentiment and returns exists, but the strength, sign and timing of this relationship vary across studies.

In this thesis I argue that this type of sign-strength-timing variability is an inherent part of the sentiment-price relationship. My rationale for this is that existing sentiment metrics miss important contextual information that can significantly alter the interpretation of a piece of text. The fact that sentiment measures lack this type of contextual awareness means that the relationship between sentiment and price will vary based on factors that are latent from the sentiment measure.

Based on this argument I make three key contributions in this thesis: *firstly*, I document significant evidence that sign, strength and timing variability are a characteristic feature of the online textual sentiment-price relationship. *Secondly*, I develop a novel time series analysis methodology, signal diffusion mapping (SDM), that is capable of modelling and forecasting effectively based on relationships that are characterised by this type of variability. *Third*, I show that when appropriately modelled using SDM, it is possible to use the sentiment signal to forecast prices. Using this methodology I document that the sentiment-price relationship is much stronger than has previously been assumed in the literature. I go on to show it is possible to develop trading strategies based on SDM that generate excess returns once reasonable costs have been accounted for.

I conclude that there is economically meaningful financial information in online social media, and that a characteristic of this information with respect to prices is variability. Modelling variability more accurately using SDM opens the possibility for using online information directly in asset pricing models or trading strategies.

Table of Contents

List of Tables

List of Figures

Academic Thesis: Declaration of Authorship

I **Paul Vincent Gaskell** declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

FORECASTING FINANCIAL MARKETS WTH ONLINE INFORMATION

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Part of this work has been published as: Gaskell, P., McGroarty, F. and Tiropanis, T., 2015. Signal Diffusion Mapping: Optimal Forecasting with Time-Varying Lags. *Journal of Forecasting.* Published Online September 2015

Signed: …….

Date: ……

Acknowledgements

First and foremost I'd like to thank Jennifer, Amy and Evan for putting up with me for the last couple of years, without your love and support writing this thesis would not have been possible. Also Jennifer, my parents and my sister for providing ad hoc proof reading services throughout.

I wish to thank my supervisors, Frank McGroarty and Thanassis Tiropanis, whose advice has been invaluable, and the web science DTC, funded through the Digital Economy Theme[1], for giving me the opportunity to purse this research.

[1] The Digital Economy Theme is a Research Councils UK cross council initiative led by EPSRC and contributed to by AHRC, ESRC, and MRC. This work was supported by the EPSRC, grant number EP/G036926/1.

Notational Conventions

Papers two and three contain a number of equations that require the manipulation of vectors and matrices. To unify the style of the thesis, I have used standard vector notation throughout, so r indicates a scalar variable, $\mathbf{r}$ indicates a vector and $\mathbf{R}$ a matrix. The only time I have deviated from this convention is in the factors of well-known asset pricing models, where the factors are typically denoted in capital letters. These cases only occur in the introduction to the thesis where the context of the equation should make the usage obvious.

I have also reserved a number of letters for recurring variables, p always indicates price, r returns, q a measure of textual sentiment and t is reserved for time. As is common in the financial literature, I primarily use the natural logarithm of a stock price return series in my analyses. This means the r mostly indicates the log return, but can also designate the raw return. The exact definition of r is always made clear in the text. As p is reserved, I designate probabilities with Pr**.** I have also reserved f to indicate a function and E to designate the expected value of a random variable. On a number of occasions I find it necessary to state the length of several vectors. To limit the number of characters used for this purpose, I use the special notation N_q to indicate the length of vector $\mathbf{q}$, intuitively this notation mean 'number of elements of type q'. This is the only occasion outside of the introduction where I deviate from standard vector notation.

Chapter 1: Introduction

Recent years have seen a rapid growth in online services, such as StockTwits[2] and Seeking Alpha[3], facilitating the sharing of financial information. McIntyre (2009) reports that these services now rival established financial media monopolies in the scale and quality of the content they provide. Goldman Sachs (2015) considers this to be part of a wider trend in investment practice, reporting that younger investors, born between the years 1980 and 2000, are beginning to eschew traditional ways of investing based on relationships with institutional equity managers. Instead, this capital is being allocated on the basis of conversations and reviews from online community resources. In monetary terms this change is significant, as Goldman Sachs report that the estimated size of the addressable opportunity created by online social investment services is $4 trillion.

Besides direct influence on investors capital allocation, there are other ways the web is being considered to influence financial services practice. Recent research reports on how consumers are increasingly likely to use peer review in purchasing decisions. Deloitte (2007), reports that 82% of US internet consumers have been directly influenced by peer reviews when choosing which products and services to buy. Similarly, Datamonitor (2010) describes how consumers are moving away from traditional forms of print reviews to using online consumer review services. Since ultimately consumer purchasing choices affect company performance, these reviews have implications for company earnings and the pricing of stocks.

To give some idea of the scale of the content that now exists online, the sample of data I gathered for this thesis covers the 100 stocks in the Standard and Poor's 100 stock index for the period running January 2014 to March 2015. This sample includes over 10.2 million messages and over 6 billion individual words from 45,516 different content sources. These sources range from micro-blogging services like Twitter, to social media sites like Facebook, to dedicated financial content aggregator sites like

[2] StockTwits is available at: http://stocktwits.com/
[3] Seeking Alpha is available at: http://seekingalpha.com/

financialcontent.com. There is no good way of telling just how much information now exists online, but I certainly haven't collected every relevant piece of content, meaning that the total volume of content is likely to be extremely large indeed. Financial theory considers information to be the primary, and only important, driver of asset prices, so even though the reports mentioned so far do not constitute peer reviewed findings, the sheer scale of online activity warrants closer examination.

The financial literature in this area is sparse. At the time of writing there are four main papers published in mainstream financial journals addressing the relationship between the content of online services and prices in financial markets (Antweiler and Frank 2004; Das and Chen 2007; Chen et al. 2014; Sprenger et al. 2014). The aim of these studies is to evidence whether there is any information in these services that can be used to model or forecast asset prices. To do this, researchers distil online content into a 'textual sentiment' metric, measuring whether the general tone of the messages implies a good or bad outlook for a stock. For example, phrases like "bull market" or "SP500 rises by 5pts", would indicate positive news about a market prices, whereas, "bear market" or "SP500 falls by 5pts" would be indicative of negative news. A feature of these approaches is that sentiment is considered to be a simple function of the words in the text, so that the word 'good' would be expected to hold positive sentiment regardless of the context in which it is used. Studies then apply econometric techniques to test whether this variable is related to prices.

There is agreement across studies that a relationship exists, but studies give contradictory accounts of the sign, strength and timing of the relationship; where, by sign I refer to whether the correlation between sentiment and price is positive or negative in sign, the strength is how closely the two variables are correlated and the timing is the lag time between information occurring in the sentiment series and it being reflected in prices. For example, Antweiler and Frank (2004) and Das and Chen (2007) both report contemporaneous correlations between sentiment measures and asset price returns but find no evidence of lagged relationships. By contrast, Sprenger et al. (2014) report a short-run lagged between sentiment and price, where sentiment is shown to negative influence future returns. The situation is further complicated by the findings of Chen et al. (2014), who document a positive correlation between sentiment and the return of stocks over the following two- to three-months following the date a piece of content is published. Further to this, Chen et al. also document that the cumulative return of a stock is positively

correlated to sentiment up to three years after a piece of content is published.

Perhaps unsurprisingly, given the variation in results across studies, authors have given different theoretical interpretation to their findings. For example, Antweiler and Frank (2004) consider their results to be the product of a small pricing inefficiency so that online sentiment is reflected in prices almost instantly, but there are occasional delays that lead to some predictability in price features. On the other hand, Chen et al. (2014) argue that investors may be drawing utility directly from community membership, this motivates them to post valuable information online and creates significant price predictability.

While they differ in their interpretation, the focus of these differing results is variability; variability in the sign, strength and timing of the relationship. As sentiment measures are constructed from word frequencies, what this means in practice is that returns are reported to respond differently to changes in the frequency of similar groups of sentiment carrying words over different studies. There are at least three plausible explanations for what might be causing this:

Firstly, investors could interpret the same words differently due to changes in their state of mind, or heuristics they are using to process the information they receive. There is a large behavioural literature documenting how investor cognitive biases affect their trading decisions, including behavioural asset pricing theories, such as Daniel et al. (1998), Daniel et al. (2002), Grinblatt and Han (2005), Barberis et al. (1998), Harrison Hong and Stein (1999) and Frazzini (2006), that consider how price predictability can emerge as a product of these biases. A key theme within this literature is that investor cognitive biases prevent them from responding immediately to new information, creating a lagged relationship between information and price. This relationship takes the form of a wave or oscillation where returns are sometimes positively correlated to information and sometimes negatively. As a result, this body of theory provides a potential explanation to account for the sign-strength-timing variability reported in the online sentiment studies.

Secondly, sign-strength-timing variability could emerge from the relationship between investors and the authors of online content. A number of studies, such as Davis and Tama-Sweet (2011), Huang et al. (2012), Loughran and Mcdonald (2011), Larcker and Zakolyukina (2012) and Rogers et al. (2011), have documented how managers of stock exchange listed companies may use deceptive language to misinform investors about the state of conditions in the company during earnings announcements and press releases. In a

less Machiavellian example, Tetlock et al. (2008) discusses how different people express themselves with different words, for example, some authors will generally be more hyperbolic than others when describing an event. In these cases investors have to make a judgement about the authors meaning based on a prior experience of the authors natural level of hyperbole. In either of these examples there is a natural feedback relationship between an investor's interpretation of the content, trading decision, and revised interpretation of future content from the same author. For example, an investor who is misled by a particular manager and loses money is likely to revise their assessment of future content produced by the same manager. Sign-strength-timing variability arises in this case because, although investors may be misled or mistaken in the short term, over time they move to correct their mistakes, which may create a lagged price response to information.

Third, there is often a level of assumed understanding an author will consider their reader to possess. For example, an active participant in financial markets would be expected to have a different understanding of words like 'tax', 'liability', 'bull' and 'bear', than a member of the general population. Without an understanding of how language is used in a specific context some of the meaning of the text is lost. A number of studies have shown evidence this is an important factor in classifying financial content, including Henry and Leone (2009) and Loughran and McDonald (2011b). This type of context can also change quite rapidly, 'Libor', for example, would have been a neutral word to most financial researchers before the recent rate fixing scandal. Now, associating a financial institution with Libor is likely to hold significant negative connotations. As with the other examples, the sentiment-price relationship would be expected to vary as the context this content is published in varies. There are a number of similar contextual features of language that are not measured by counting the frequencies of sentiment carrying words. A good example of this is how financial content often gives a commentary of expected future valuations of stocks, for example, a piece of text may make a statement like 'Dow Jones expected to top 17,000 points this week'. Read the week before, this statement appears positive and suggests investing in a fund that tracks the Dow may provide positive returns, read a week after when the target has been missed and the statement takes on a different semantic meaning.

Current sentiment analysis methodologies are not equipped to model this type of variability as they are based entirely on the frequency of words in a piece of text. The type

of factors I list above will often not be explicitly referenced in the text, making any variability they cause not directly contained in – or latent from – the sentient measure. A further complication is that in general some of these factors are not measurable. For example, it is difficult to see how investors trust in a given author could be explicitly measured in most cases. The result is that some sign-strength-timing variability is almost certainly going to persist irrespective of the sentiment analysis methodology being used. Given these difficulties, the fact that significant statistical relationships have been reported in the literature, by authors such as Antweiler and Frank (2004), Das and Chen (2007), Chen et al. (2014) and Sprenger et al. (2014), suggests the presence of a much stronger relationship between online textual sentiment and price exists, it just may require significant methodological innovations to gain a deeper understanding of it.

There are two gaps I identify in this literature that I will address in this thesis: *firstly*, based on the available theory and evidence there is good reason to suspect there will be variability in the strength, sign and timing of the sentiment-price relationship. There is currently no literature which studies this aspect of the relationship in detail. *Secondly*, current methodologies are limited in their ability to model relationships that exhibit the type of strength-sign-timing variability I describe.

To bridge these gaps, the substantive part of this contribution is developed in three papers; the conceptual link between the three is an investigation into how online textual sentiment can be used to forecast prices in financial markets. More specifically, this link focuses on how the variability in the online textual sentiment-price relationship can be modelled in order to forecast prices with textual sentiment measures.

In *chapter three*; I address the first gap in the literature, and study in detail the time-evolution of the sentiment-price relationship. To do this I have assembled the largest and most diverse corpus of online financial content studies in the literature to date. The corpus contains over 10.2 million pieces of content from 45,516 different online sources, ranging from micro-blogs, to articles, to message-boards. To put this in perspective, Antweiler and Frank (2004) use the second largest corpus of 1.5 million message-board posts, these are drawn from two online services, Yahoo Finance and Raging Bull. This matters because different online platforms are likely to attract different user bases, and place different restrictions on how information is accessed and shared. For example, Twitter places a 140 character limit on messages, whereas, a blog post may have no character limit at all. The

size of the corpus I use in this thesis reduces the chance of bias being introduced as a result of these factors.

I then adapt econometric techniques pioneered in the momentum literature by authors, such as Moskowitz et al. (2012), to study lagged relationship between sentiment and price at a time horizon of up to 5 months. I find evidence of significant long-run relationship between sentiment and price. I show using cross-correlation analysis, that this is characterised by an oscillation or waveform over the length of the lag period, suggesting that sign-strength-timing variability is a feature of the sentiment-price relationship. I note that this is similar in form to the type of relationship studies in the momentum literature, where researchers have documented an oscillatory relationship between return series and their own lagged values (see for example; Jegadeesh and Titman 1993; Asness et al. 2014; Asness et al. 2013). Using multivariate regressions, I show that some of the sentiment-price relationship can be explained by price momentum, but that there is also some incremental information about future returns contained in the lagged values of the sentiment series that cannot be explained by prior prices.

The importance of this paper to the overall thesis is that I provide empirical evidence of an economically meaningful relationship between sentiment and price that also shows significant sign-strength-timing variability. As this variability cannot be fully explained by price momentum, the evidence suggests that one or more of the other sources of variability I have discussed play a role in the sentiment-price relationship. Yet these sources of variability are not realistically measurable, leading to the conclusion that to more effectively model the sentiment-price relationship requires developing modelling techniques for capturing variability in the sign, strength and timing of time series relationships.

In *chapter four*; I address the methodological challenge of modelling sign-strength-timing variability, suggested by the findings of earlier studies and evidenced explicitly in the findings of chapter three. To do this I introduce the signal diffusion mapping (SDM) algorithm which has been published in the Journal of Forecasting[4]. SDM is the optimal Bayesian estimator for time varying relationships that exhibit sign-strength-timing variability of the type I describe. I develop SDM based on techniques from the computer science and statistical physics literatures for finding the optimal temporal alignments

[4] Gaskell, P., McGroarty, F. and Tiropanis, T., 2015. Signal Diffusion Mapping: Optimal Forecasting with Time-Varying Lags. *Journal of Forecasting.* Published Online September 2015.

between time series. I adapt these techniques for forecasting by incorporating them into a recursive Bayes estimation framework. The result is the optimal Bayesian forecast of time varying lagged relationships. I show the effectiveness of the SDM algorithm with tests on simulated data.

In *chapter five*; I return to the same modelling task I attempted in chapter three, this time armed with the SDM algorithm. If the sentiment price is significantly variable as I have suggested, then SDM should generate a better forecast than the modelling strategy I applied in chapter three. To test this I employ a similar approach to the Diebold-Mariano test (Diebold and Mariano 1995) to compare the forecast produced by SDM against the forecast produced by the simpler techniques I applied in chapter three, I show that SDM greatly improves the forecast, I go on to show how it is possible to generate large excess returns via a simple trading strategy based on the output of the SDM estimator.

I then introduce a novel hypothesis testing framework based on the output of the SDM algorithm to analyse where the extra predictive power comes from. I find that there are stock specific lag distributions which characterise the sentiment-price relationship. I also show that these distributions are significantly variable over time. I conclude that this is evidence in support of the fact that there are contextual factors, specific to stocks and particular time periods, that play an important role in determining how the words in text are interpreted by investors. I conclude that sign-strength-timing variability is a significant aspect of the sentiment-price relationship, and that SDM can be used to model it in a much more effective manner than existing approaches.

In developing SDM I draw on statistical methods from outside the field of finance and econometrics. The SDM method I develop in chapter four is based on mathematical concepts from the computer science and statistical physics literatures. This not only fulfils the interdisciplinary requirement for my research program, but also provides significant improvement on existing econometric methods. As well as providing an innovative modelling strategy for research into the online textual sentiment-price relationship, the general issue of variability in modelling semantic meaning in text touches a number of other areas and disciplines interested in the study of online content. As a result, the methodology I develop for this thesis is generalizable to a number of other contexts in the web science, computer science and internet science literatures.

The following chapter contains a literature review highlighting some of the most

relevant literature in this area; I first introduce the major bodies of theory that inform this research. In section 2.2 I consider how the relationship between information, investors and prices is viewed in mainstream financial theory. In section 2.3 I review what literature exists documenting the relationship between the information of textual content and prices. In section 2.4 I describe what methodologies have been applied to measure textual sentiment, both in the financial and computer science literatures. In section 2.5 I describe the time series modelling techniques that have been used to model the relationship between textual sentiment and other variables like asset prices. Finally, in section 2.6 I synthesise the evidence from the literature into an argument about how variability is an inherent feature of the online textual sentiment-price relationship.

Chapter 2: Literature Review

2.2. Information, investors and prices in financial theory

In section 1 I describe how investors may interpret textual sentiment information differently based on a number of factors latent to the content of the text. In this section, I introduce how, in financial theory the relationship between prices and explanatory variables is often thought to vary based on factors not captured in financial models. How this variability is captured and modelled depends on the underlying assumptions researchers make about investor behaviour. The importance of this section to the narrative of my thesis is to show where the idea of sign-strength-timing variability in the information-price relationship might be located within financial theory.

There are a number of assumptions about investor behaviour that are relatively uncontroversial in mainstream financial literature. These are summarised by Lo (2008) as:

Individuals prefer …

(i) … more money to less

(ii) … money now over money later

(iii) … to avoid risk

(iv) All agents act to maximise their own self-interest

Based on these assumptions, investors act in ways to maximise the value of their asset holdings whilst attempting not to take unnecessary risks. To do this an investor needs to make probabilistic decisions about the future prices of stocks. The information investors use to make these decisions is the main, and only important, driver of asset prices in financial theory. There are two categories of information the mainstream literature considers: *objective information* – material information about the conditions inside a firm or institution that affect the fundamental valuation of a stock, for example reports of company earnings, changes in management or macro-economic news; and *investor*

sentiment – defined by Baker and Wurgler (2007) as the thoughts, attitudes and opinions investors have about stocks that are not justified by the facts at hand. The role each of these types of information plays in the price formation process is a key differentiating factor between the two major bodies of asset pricing theory: efficient markets theory; and behavioural theories and momentum.

2.2.1. Efficient markets theory

Efficient markets theory is based on the assumption that investors are probabilistically rational. Asset pricing models based on this assumption can be traced back to Markowitz's (1952) modern portfolio theory (MPT) . Markowitz proposes that the concepts of probabilistic return and risk can be captured mathematically by the mean and variance of a stock's price. It follows from this assumption that the expected return and risk of holding a portfolio of stocks is a function of their means, variances and covariance. Markowitz (1999) discusses how the assumptions of MPT serve as both a hypothesis about investor behaviour, but also as a maxim for how a good, or rational, investor *ought* to act in a given situation. On the one hand, if the mean and variance of prices is an adequate description of the way stock prices move then the hypothesis is that investors behave in a way that allows these quantities to be defined. On the other hand, if these assumptions are true then there are mathematically optimal decisions an investor can make.

Under the assumptions of MPT investors have two levers they can use to build portfolios; the first is to hold portfolios of stocks with higher expected returns. This is both intuitively obvious and follows from the fact that investors prefer more money to less. If it is not possible for investors to estimate the return of stocks directly, the second lever is to diversify the risk of holding the portfolio. Diversification is important because, given the mean and variance assumptions, it can be shown mathematically that the ratio of the mean of a portfolio to its variance will generally increase the more stocks are held in the portfolio. The rate of increase is also a function of how correlated the constituent stocks in the portfolio are, the less positively correlated, the greater the increase in the mean-variance ratio of the portfolio. This allows for mathematically optimal portfolios to be created without the necessity for an estimate of future returns.

Building on MPT, Fama (1965a; 1965b; 1970) effectively removes the idea of

forecasting expected returns from the mainstream literature in proposing the efficient market hypothesis (EMH). The EMH is based on the synthesis of a large body of evidence showing that it is very difficult to systematically 'beat the market'. Fama (1970) reviews a large range of proposed trading systems and shows that none of them return profits in excess of reasonable transaction costs. This work has been corroborated by authors, such as Brock et al. (1992), who similarly find that technical trading rules, reported to be used by investors, do not generate returns in excess of reasonable costs. The conclusion Fama draws from this is that prices must be an accurate reflection of all available information. If they were not, there should be a period of time over which prices moved in a stable trajectory as investors adjust their holdings. This should cause prices to trend. Fama (1965b) shows that, although prices are serially correlated, this correlation is too small for investors to profit from it in excess of costs. The EMH itself comes in three forms; in its weak form it prescribes that asset price returns cannot be predicted by their own prior return or other available market data, the semi-strong form asserts that prices cannot be predicted by any publicly available information, whilst the strong form asserts that prices cannot be predicted at all as they reflect all information both public and private (Fama 1970).

The EMH is one of the most influential concepts in asset pricing theory. By removing the ability to forecast expected returns, the only remaining lever for investors is diversification. Following this to its logical conclusion, it means that all investors should attempt to hold the portfolio with the most widely diversified risk, which corresponds to holding every stock in the market. Based on these assumptions the price of an asset should be a function of the difference in the risk profile of an asset relative to other potential investments. This insight forms the basis of the capital asset pricing model (CAPM), where the expected return of a stock is given based on its risk profile set against the risk of holding the market portfolio. This is expressed mathematically as

$$E[r_{i,t}] = r_{free,t} + \beta_{i,mkt}\left(E[r_{mkt,t}] - r_{free,t}\right) \quad (1)$$

where $r_{i,t}$ is the return of the ith stock at time t, r_{free} is the return from a risk free asset, for example U.S. Treasury Bonds, and r_{mkt} is the return of the market portfolio. The parameter $\beta_{i,mkt}$ represents the sensitivity of fluctuations in the stock's price to fluctuations in the market return. The expected return of the asset in CAPM is a premium on the excess variance the investor takes on by holding the stock.

CAPM has been extremely influential in shaping the way financial researchers think about risk. The model implies there is a portion of the stock's variance that cannot be reduced by diversifying over the market portfolio. This systematic risk represents the effect of exogenous shocks on the market as a whole. The risk of a world war, for example, affects all stocks in the market and so cannot be diversified away. Idiosyncratic risk, features of a company that are independent of other companies, can be diversified away by constructing portfolios.

In the CAPM model the only parameter is the relationship between the return of the stock and the market portfolio provided by $\beta_{i,mkt}$. As specified, equation (1) implies that this relationship is a fixed property; a company has a pre-defined systematic risk profile. CAPM in this form is known to perform quite poorly in modelling asset price returns. One of the ways it has been extended is to make $\beta_{i,mkt}$ time varying in a conditional CAPM model, this takes the same form as equation (1) but $\beta_{i,mkt}$ is fitted based on the rolling lagged values of the market portfolio (Jagannathan and Wang 1996). Conditional CAPM implies that systematic risk is affected by changing context, that there are factors affecting investor perceptions of risk that alter the premium demanded for an incremental unit risk at different points in time.

An assumption of CAPM is that investors are only motivated by the mean and variance profiles of stocks. An issue with this, raised by Merton (1973) and Fama and French (1993; 2004), is that the market does not exist in a vacuum from other factors in the wider economy. For example, an investor is likely to be motivated by the goods and services that can be consumed using profits from trading. In this case there may be a covariance between returns and consumer prices. Another argument against CAPM is that there is empirical evidence that other variables explain the cross section of returns.

Fama and French (1993) find that two factors significantly influence the cross section of returns, the size of companies and the ratio of their fundamental accounting valuation to current stock price. These factors are added to CAPM to form the Fama French three-factor-model, given as

$$E[r_{i,t}] = r_{free,t} + \beta_{i,mkt}\left(E[r_{mkt,t}] - r_{free,t}\right) + \beta_{i,SMB}E[SMB_t] + \beta_{i,HML}E[HML_t] \quad (2)$$

where CAPM forms the first part of the model as before, SMB_t stands for 'small (market

capitalisation) minus big', and HML_t stands for 'high (book-to-market ratio) minus low. The *SMB* factor captures the risk premium for holding small capitalisation stocks, whilst the *HML* factor captures the risk premium for holding stocks whose current price is relatively large compared to their accounting valuation.

Similar to conditional CAPM, the parameters associated with the extra factors in equation (2) are allowed to be time varying, they are also allowed to take both positive and negative values. Again, the interpretation of these changes must be that an investor will price an incremental unit risk differently at different points in time. There are contextual aspects to investor interpretations of the factors in the model that change over time, but these are not expressed explicitly in equation (2).

This represents the state of the art in asset pricing theory based on the assumptions of rational expectations. It is not possible to forecast future returns directly because of the EMH. Instead, asset pricing models are tools to assess the risk profile of companies or to assess the performance of investments against a benchmark of how much an asset should have returned based on its risk profile. Effectively, the EMH removes variability in the time axis from the modelling of the information-price relationship. This is justified by the assumption that investors can quickly and rationally interpret new information they receive into their trading strategies. What is allowed to vary is investor interpretations of the risk profile of companies based on the explanatory factors in the models. Factors like the return of the market portfolio or book-to-market ratio of a company act as yardsticks for assessing how risky an asset is, but these relationships are able to vary over time based on other un-modelled variables.

2.2.2. Behavioural theories and momentum

The second major branch of asset pricing theory concerns the observation by Jegadeesh and Titman (1993) that prices exhibit momentum. Momentum is the phenomenon where stocks which have greater than average returns over a six- to twelve-month period will tend to have higher than average returns over the following three month period. This phenomenon has been observed over a large range of stocks and asset classes, and over a large range of different time periods (Asness et al. 2013; Asness et al. 2014). More recently, Moskowitz et al. (2012) document that momentum is not only present in the

cross-section of returns, but that the one month return of a stock can be forecast with its own twelve month lagged return. The momentum literature is supplemented by the long run reversal in prices documented by Thaler and De Bondt (1985), where excess profits can be earned by betting against stocks with high three- to five-year returns. There is also a short term reversal effect documented by Jegadeesh (1990) and Gutierrez and Kelley (2008), although the effect is generally not significant enough to trade for profits in excess of costs.

This evidence contradicts the assumptions of the EMH; the serial correlations show prices do not fully reflect all available information. In turn, this casts doubt on the hypothesis that investors are mean-variance optimising agents in the sense of Markowitz. A major strand of financial theory now deals with the expected behaviour of prices when investors are not considered to be probabilistically rational in this sense. The genesis of this research can be traced to Kahneman and Tversky (1979), who document how people make systematic mistakes in probabilistic reasoning tasks. From this foundation, a number of reported biases in investor decision-making have been proposed in asset pricing models purporting to yield price momentum as a consequence of investor psychology.

Daniel et al. (1998) and Daniel et al. (2002) consider investors to be overconfident in their own abilities. Due to this overconfidence they will over-react to their own private information. They also have self-attribution bias so they are more likely to act on new information that confirms their current thinking than on information that contradicts it. This exacerbates the initial overreaction creating price momentum. Over the long term the effects of overconfidence subside and investors alter their holdings accordingly, this generates long run reversal.

Barberis et al. (1998) consider momentum to be generated by representativeness. Representativeness describes the way in which investors assign specific properties to different stocks. An investor may mentally label some stocks ‘growth stocks’, for example. Investors will tend to favour keeping this mental model over changing it to suit new information. This creates a lag between new information occurring and investors altering their mental models to reflect it.

Grinblatt and Han (2005) and Frazzini (2006) model momentum as the product of the disposition effect, which is the tendency of investors to sell winners and hold losers. Investors are more likely to sell a stock that rises in value by one unit than a stock that falls

in value by the same amount. Similar to the theory of Barberis et al. (1998), this causes investors to initially under-react to new information creating price momentum.

Hong and Stein (1999) consider a different approach to modelling momentum. They propose that there are two types of traders present in the market, *newswatchers* and *momentum traders*. Newswatchers only trade based on their private information, this private information is considered to diffuse slowly through the market. Momentum traders adopt a naïve trend following strategy based only on prior returns. As the momentum traders observe these changes they adjust their positions, thus creating serial correlation.

The running theme of these papers is that momentum is the product of investor psychology. These psychological traits cause prices to be positively serially correlated over time horizons between three and twelve months, then negatively correlated in the longer term. Despite momentum being a recognised effect across the financial literature, there is no single accepted theory of what drives the momentum effect. There are also features of empirical momentum that are unexplained by existing theories. For example, Daniel and Moskowitz (2011) document how momentum portfolios experience crashes, short periods during market downturns where the stocks with lower than average prior returns greatly outperform stocks with higher than average prior returns. As existing behavioural theories exclusively consider the pricing of a single risky asset, market-wide features such as this current remain outside of their explanatory power.

One way momentum has been reconciled with existing asset pricing theory is as a further explanatory variable in the Fama French model. Carhart (1997) proposes this in a four-factor-model

$$E[r_{i,t}] = r_{free,t} + \beta_{i,mkt}\left(E[r_{mkt,t}] - r_{free,t}\right) + \beta_{i,SMB}E[SMB_t] + \beta_{i,HML}E[HML_t] + \beta_{i,UMD}E[UMD_t] \quad (3)$$

where UMD_t stands for up-minus-down, which is a factor representing the risk premium for stocks' one year momentum. This casts investor psychology as a conditional risk factor in a similar manner to the market efficiency models described in section 2.2.1.

What this model doesn't capture is the fact that, unlike the other explanatory variables, momentum can be shown to *forecast* future prices. In existing behavioural theories the effect of investors' cognitive biases is to create a time-distortion in the way

objective information is included in prices. The specifics of how this relationship works remains a very active area of investigation. Given the lack of agreement in the literature, possibly the strongest mathematical statement that can be made regarding momentum at present is that expected returns over some holding period *h*, are some function of a stocks prior returns, that is

$$E\left[\sum_{s=1}^{s=h} r_{i,t+s}\right] = f\left(r_{i,t-m:t}\right) \quad (4)$$

where the literature suggests values for *h* between one month and three months, and *m* is the length of the look-back period, typically taking values of up to one year.

A feature of the difference between behavioural theories and empirical evidence from the momentum literature is that empirical work typically considers *f* to be the cumulative return of the stock over the period *t – m*:*m*, whereas, theoretical work typically considers *f* to be a more complex waveform, or oscillation in prices over the look-back period. Note also that equation (4) does not include any description of the relationship of a stocks momentum to other stocks, or to the market as a whole. The evidence suggests these relationships are an important part of the momentum effect but there is no universally accepted explanation for how or why this occurs.

Equations (2) and (4) serve to provide a simple mathematical distinction between EMH based models and behavioural models. Under the EMH there is no future predictability in returns, it is possible to find explanatory variables concerning the cross-section of returns but there is no variability permitted on the time axis. Behavioural theories put the timing of the information-price relationship at the centre of research. Returns are not only forecast-able based on serial correlation, but this serial correlation may itself be time varying or complex.

2.2.3. The adaptive markets hypothesis

The two bodies of theory I have discussed thus far both view investors interpretation of information as a homogenous feature of all investors in the market. Under the efficient markets interpretation investors are entirely homogenous mean, variance optimising

agents. In the case of behavioural theories investors have psychological biases but these biases are the same for all investors. In either case there may be latent contextual factors which change prices reaction to information or explanatory variables, but these are exogenous to the thought process of the investor which is assumed to be fixed. There is another body of theory which attempts to unify empirical evidence in favour of both market efficiency and behavioural models by proposing a market made up of heterogeneous agents.

Lo (2004; 2005) argues that it is possible to reconcile these bodies of theory by framing financial markets in evolutionary terms. Lo's adaptive markets hypothesis (AMH) asserts that investors are not necessarily probabilistically rational, but they are capable of adapting and learning. The market then has a type of 'ecology', where different 'species' – for example, investment banks, hedge funds and pension funds – are all heterogeneous agents acting to further their own interests. Market efficiency in the AMH sense is then a special case of a price formation process governed by a particular form of market ecology. Under the assumptions of the AMH, investor risk preferences are a function of their historical experiences. For example, the technology bubble in the early 2000's would have bankrupted many risk seeking investors, the remaining population of investors would then be expected to have an altered set of risk preferences (Lo 2007).

There is some empirical evidence in support of the AMH. Researchers, such as Urquhart and Hudson (2013), have shown that returns sometimes show serial correlation and other times do not. A particularly striking example of adaptive behaviour the authors present is a 'runs analysis' of the FT30 stock index, showing that before Fama's work on market efficiency there was significant serial correlation in the index, this ends abruptly after the EMH become widely accepted in the 1970s. Another form of evidence comes in the form of surveys, such as Menkhoff and Schmidt (2005), who ask equity managers to disclose the trading strategies they use. Most respondents report that they use several different strategies to help them trade. Some of these will be 'value' strategies based on ideas consistent with the EMH, others will be strategies based on the assumption of price momentum. The authors note that it is rare for an investor to report relying exclusively on one type of strategy.

Importantly for the arguments I make in this thesis the AMH allows for the way investors react to new information to be variable over time. Since investors can learn and

adapt, it is possible that an investor may respond differently to the same piece of information at different points in time. It is also possible that changes in market ecology, for example a fund withdrawing from trading in a particular category of stock will cause changes to the price formation process. The consequence of this is that prices response to new information is not just a feature of exogenous context, but also the context of what type of investors are currently trading in a given asset, and what their current states of mind are.

2.3. Financial studies of textual sentiment

In practice a large amount of the financial information discussed in the previous sections comes published in some form of text document, for example, companies quarterly earnings announcements will be published in a document containing both the numerical quantities of the company's fundamental accounting values, and a qualitative description of the figures. The earnings figures in the document are directly referenced in the HML_t factor of the Fama French three-factor-model (equation (2)). As the numerical quantities in the document are known to drive prices, a natural question is whether the qualitative content does as well. In order to answer this question financial researchers have used a range of text processing methodologies, mostly derived from the computer science and machine learning literatures. These methodologies aim to measure the polarity or affect in the text so this can be used as an explanatory variable in econometric models.

A feature of textual content is that the type of information under study does not fit neatly into the existing categories of financial information I introduced at the beginning of section 2.2. Textual content can contain both objective information and traces of investor sentiment. In order to capture this distinction, Kearney and Liu (2014) categorise measures of affect in text as *textual sentiment.* The subject of this thesis is specifically the relationship between textual sentiment expressed on online social media and prices. There are only a small number of papers in the financial literature on this topic. A much larger literature, however, deals with other forms of content. Kearney and Liu separate these different forms of text content into three categories; *corporation-expressed sentiment*, including the content of corporate annual or interim reports and company earnings announcements; *media-expressed sentiment*, for example the content of articles in the

mainstream financial press of analysts' reports; and *sentiment expressed online*, for example blog posts and the content of micro-blogging platforms like Twitter. In the rest of this section I review the main contributions to the literature for each of these categories to make the link between the theories presented in section 2.2 and the actual content of text documents more explicit.

2.3.1. Corporation-expressed sentiment

The most widely studied form of textual content in the financial literature is the public disclosures of stock market listed companies. These come in the form of quarterly earnings announcements and annual or interim reports. Many of these reports are legally required by regulatory authorities like the Securities and Exchange Commission. As these documents are authored by company managers, they have access to inside information about the conditions inside the firm that may not be fully reflected in the accounting values in the company's financial statements. The aim of research in this area is to test whether this information is picked up by investors.

There is broad consensus in the literature that the tone of announcements has some short term forecasting ability for either returns or quarterly earnings. This has been documented by a large number of authors including Feldman et al. (2008), Davis et al. (2012), Doran et al. (2012), Mangen and Durnev (2010), Ferris et al. (2013) and Price et al. (2012). The most widely accepted interpretation of these findings is that managers convey some objective, inside information about the conditions in a company through the text of these disclosures. Loughran and McDonald (2013) also report that textual sentiment may reflect other concepts linked to prices like uncertainty. Within this literature textual sentiment is conceptually another form of objective information, but one that may be difficult for investors to access.

Other authors have suggested that investors may be psychologically manipulated by managers' use of language in disclosures. Henry (2006) documents how the written style of disclosures can alter investors' reactions to content. Subsequently, Henry (2008) hypothesises that investors may be psychologically influenced by a document's tone. The rationale Henry provides is that investors use mental framing as a heuristic in their decision making, the tone of disclosures influences this framing and subsequently alters investors'

heuristics with respect to the company. Other researchers go further, suggesting managers actively attempt to influence investors' impressions of earnings information. Davis and Tama-Sweet (2011) document how managers use alternative forms of media communication to frame the content of earnings announcements. Huang et al. (2012), Loughran and McDonald (2011a) and Larcker and Zakolyukina (2012) all document how managers may use deceptive language to mislead investors about future earnings. Finally, Rogers et al. (2011) documents how overly optimistic language, coupled with insider selling of stock, can be used as a predictor of prosecutions for fraud.

There is an important distinction between the ways textual sentiment is interpreted theoretically in this literature. One interpretation is that the information is simply another form of objective information. Under this interpretation investors may still make probabilistically rational judgements based on the information they receive. The short lagged predictability documented in the literature can be explained by the difficulty investors have accessing the objective elements of the content.

The second interpretation implies that investors are to a certain extent programmable based on the textual content they read. This makes the relationship between managers and investors potentially far more complex since, if managers can understand the value of attempting to manipulate investors' mental models; it is likely that investors would be aware of this attempted manipulation. There is some indication that this is the case, for example, Loughran and McDonald (2011a) have documented media services that claim to identify managers' use of deceptive language. The actual content of the text in these cases is a product of the game of cat and mouse investors and managers are playing over the sharing of information. The strict rules on sharing information in major financial markets mean that there is only a limited space for this type of manipulation to occur. Still, managers only have to disclose earnings quarterly so may use the content of filings to buy themselves time to correct failings in the company. In more extreme cases, such as those documented by Rogers et al. (2011), this may lead to managers committing fraud.

The suggestion in these papers is that the honesty of the author of the document is a latent contextual factor which significantly modifies the way the content is interpreted by investors. The amount of money investors make is a function of the quality of information they receive. If they are misled, or misinterpret information, this will cause a drop in their wealth which is likely to alter their future decisions when faced with information from the

same source in future. This type of relationship is similar to Soros' reflexivity theory, where prices move based on a symbiotic feedback relationship between information and price (Soros 2009). Under these conditions, the relationship between sentiment and prices is a function of the trust investors have in the author of a piece of content, which is a function of previous relationships between sentiment and price. In other words, the current sentiment-price relationship varies based on prior sentiment-price relationships. This type of relationship could drive the strength-sign-timing variability I discuss in chapter one.

2.3.2. Media-expressed sentiment

The second most widely studied form of text is the content of the mainstream media. In a financial context this includes newspapers, the financial press and reports by stock analysts. A difference between this literature and the corporate-expressed sentiment literature is that these studies often do not model stock specific sentiment. New articles will generally report on market or sector wide issues and so are not suitable for modelling the return of stocks directly. Instead, studies look at the effect of textual sentiment on market indices or portfolios of stocks.

Similarly to the corporate-expressed sentiment literature, there is broad agreement that the content of mainstream news articles can be used to forecast future returns or earnings at short lag lengths. Examples of studies that reach such conclusions include Ferguson et al. (2014), Engelberg et al. (2012), Carretta et al. (2011) and Tetlock (2007). There are also studies that have looked at other forms of market behaviour, for example, Liu and McConnell (2013) and Buehlmaier (2013), both of which concluded that the tone of press reporting on corporate acquisitions can affect the outcome of a takeover. The predictive power of analysts' reports is less clear. Huang et al. (2012) document that analysts' reports can predict earnings up to three years in advance, whereas, Twedt and Rees (2012) report there is a contemporary relationship between sentiment and prices but no significant lagged relationship.

There are again two separate theoretical interpretations of these findings. Tetlock et al. (2008) conclude that the media encodes hard to quantify aspects of firms performance that investors might judge rationally but with a small time delay. On the other hand García (2013) shows that the predictive power of the mainstream news is accentuated during

recessions. Garcia hypothesises that this is a reflection of investors' altered cognitive state during these times changing their perceptions of the information they read. The second interpretation links directly to the behavioural asset pricing literature. This link is drawn by Sinha (2010), who compares the predictive power of media expressed sentiment and the momentum effect. Sinha concludes that although the two effects are similar there is information in textual sentiment that cannot be explained by momentum. As with the corporate-expressed sentiment literature, the implication of these studies is that there are contextual factors, such as content being created during a recession, that are latent from the words in the text but affect how the information in the text is incorporated into prices.

As an example of the type of contextual feature that may alter investor interpretations of content is the time a piece of text refers to. Often, financial content makes some forecast of the future value of a company, for example, 'AAPL to suffer losses in Q4', clearly references a period in time. Existing sentiment measures do not take timing into account as they only measure the polarity of the content. If this statement was read in the third quarter of the financial year then the semantic meaning is clearly negative, but read in the first quarter of the following year and investors may see it as a sign AAPL will bounce back in the new-year, making it a good time to buy stock. With financial forecasts of this type the time period referenced in the forecast has a massive impact on the reading of the content, yet it is exactly the sort of contextual information sentiment measures do not detect. In terms of sign-strength-timing variability, the sign of the relationship in this case changes with the timing.

2.3.3. Online textual sentiment

The online textual sentiment literature concerns the measurement of sentiment expressed in blog posts, message-boards and micro-blogs. Kearney and Liu (2014) note that this is by far the least regulated of the forms of textual sentiment studied in the literature. Company filings have set formal structures and laws which govern their content. To a lesser degree press reports are also bound by structural norms and regulation. By contrast, different online platforms provide different structural constraints to the sharing of content, for example, micro-blogs like Twitter have character limits on messages restricting their length whilst blogs generally do not, but the authors are not bound by the same laws or codes of

conduct as with other forms of content. There is also a certain level of anonymity provided by the use of online services that makes the authors less likely to suffer reputational damage from the content they post.

There is consensus in the literature that a contemporary relationship exists between online textual sentiment and prices. Some studies, such as Sprenger et al. (2014) and Chen et al. (2014) document a lagged relationship, where sentiment can be used to forecast future prices. Other studies, such as Antweiler and Frank (2004) and Das and Chen (2007) report a contemporaneous relationship between sentiment and prices but no lagged relationship. Where a lagged relationship has been documented the sign and timing of the relationship vary, Sprenger et al. document a short-run, negative correlation between sentiment and price, whereas, Chen et al. document a positive correlation between sentiment and price up to three years in the future.

The online medium studied varies; Antweiler and Frank and Das and Chen study online message-board services, Chen et al. study the content of Seeking Alpha, a website dedicated to financial articles with an active comments section, Sprenger et al. study Twitter. Each of these platforms comes with their own structural constraints for information sharing. Twitter allows users to share messages of no more than 140 characters, whereas, Seeking Alpha is an articles and comments site allowing members to express themselves in lengthier prose. Each of these services is also likely to engage slightly different types of investor, as different cultures of users emerge around different services.

Another feature of online services that distinguishes them from other forms of media is that they contain content that may be created by investors as well as content that investors consume. This distinction has been noted in the theoretical interpretations authors give to their findings. Antweiler and Frank (2004) discuss how investors must think before they place a trade. Messages about a stock in this case might indicate that a particular stock has grabbed an investor’s attention. An issue with this argument conceptually is that it suggests investors post messages somewhat unconsciously, that their posts are a sort of exhaust left behind during the decision making process. This type of reasoning does not provide any insight into what motivates investors to do this.

Motivation is another aspect of the online sentiment literature that distinguishes it from the wider literature. There is a clear legal or monetary motivation for managers to file

earnings announcements and journalists to write financial news stories. There is no such imperative for investors to post information online. The only study that has attempted to address the issue of motivation is Chen et al. (2014); the authors argue that investors may derive utility from the act of community participation, or may reap tangible economic benefits from active community membership.

In support of this hypothesis Chen et al. cite a number of articles documenting cases where social relationships played a role in investors' decisions. Hong et al. (2004) document how social relationships play a role in whether households choose to invest in the stock market. The authors argue that observational learning and utility derived from shared social experiences increase the attractiveness of investing. Advice from friends may also lower barriers to entry, for example, by providing advice on the best brokerage service or strategy to pursue. Ivković and Weisbenner (2007) find that co-located households are more likely to invest in the same stocks; the authors consider this evidence of a type of 'information diffusion' effect that occurs as a result of neighbours sharing information with each other.

2.4. Methods for measuring textual sentiment

In this section I discuss the methodologies that have been used to model the textual sentiment-price relationship in the financial literature. I describe the ones that have been used most frequently in detail and provide an overview of other methodologies. Many of the sentiment analysis methods present in the financial literature were pioneered by computer scientist and machine learning researchers. To give a full account of the technology that is available for research in this area I also describe some of the more recent innovations in the wider sentiment analysis and natural language processing literatures. One of the features of these methodologies I wish to draw attention to is that they present a static interpretation of the semantic meaning in text. I discuss the implications this has in light of the arguments presented above about variability in the context the text is written in being a key feature of the sentiment-price relationship.

There are two main methodologies for measuring textual sentiment that have been applied in financial studies: dictionary based classifiers (DC) and naïve Bayes classifiers (BC). In the financial literature the DC has been applied much more often than the BC

(Kearney and Liu 2014). The basic anatomy of either approach is the same. A set of messages needs to be selected that the researcher believes references a stock's price or other financial variable. This is typically done by some simple keyword selection process, for example, Sprenger et al. (2014) select messages based on specific tags given to messages about stock prices on Twitter. Messages then undergo some form of pre-processing. Common types of pre-processing include removing punctuation and in some cases common words like 'and' and 'the'. The content of these messages is then classified and a measure of sentiment taken over different time periods. These measurements are then compared to the price of the asset using some econometric model. In the following subsections I describe the main procedures applied at each stage in the process, I also summarise some of the other innovations that have been reported in the computer science literature but have not yet made it into financial journals. This is to give an overview of not just what has been done, but also what is possible with existing technology.

2.4.1. Naïve Bayes classifiers

The BC is based on the assumption that the semantic content of text can be adequately described by a histogram of word frequencies. This is the 'bag-of-words' assumption that underpins many of the content analysis and sentiment analysis algorithms in the machine learning literature. The bag-of-words assumption implies that a message is equivalent to the set of frequencies it contains. For example, a message **me**, containing N_{wd} words, could be described as the vector

$$\mathbf{me} = \left[wd_1, wd_2, wd_3 \ldots wd_{N_{wd}}\right] \tag{5}$$

where wd_i is the frequency of the *i*th word. The basic task of a BC is to assign a probability score to a message based on its likely membership to a set of different classes. A typical example in the sentiment analysis literature is to define a class representing positive tone, cl_{pos}, and then calculate the conditional probability of this class given a message, that is, $Pr\left(cl_{pos}\middle|\mathbf{me}\right)$. Due to the bag of words assumption, **me** is simply a vector of word frequencies, so cl_{pos} can also be defined based on the word frequencies directly. That is

$$Pr\left(cl_{pos} \mid wd_1, wd_2 \ldots wd_{N_{wd}}\right) \tag{6}$$

which implies the word frequencies that define the class in equation (5) are independent of the messages they are contained in.

The BC then assumes it is possible to assign probabilities to words based on their likely class membership. The way this is done in practice is for researchers to hand classify a training sample of messages as belonging to different classes. Words are then scored based on how frequently they occur in each class. A simple way of doing this is to assign probabilities based on the percentage of words in the training set that occurred in messages assigned to a given class, that is

$$Pr(wd_i \mid cl_{pos}) = wd_{i|cl_{pos}}/wd_{i|train} \tag{7}$$

where $w_{i|cl}$ is the sum of the frequencies of the *i*th word that occurred in messages assigned to a given class, $w_{i|train}$ is the frequency of the *i*th word in the training set.

There are a number of other common approaches to assigning these probabilities. Sparck Jones (1972) introduced the term frequency inverse document frequency (tf-idf) approach; one of the most commonly used methods in the computer science literature. The purpose of tf-idf is to normalise the raw frequencies of words by the frequency of their occurrence across messages. This is done by dividing the word count by the number of messages with at least one occurrence of the word and taking logs. Another issue with assigning probabilities to words is that not all words that occur in the full sample will necessarily occur in the training set. To counter this it is common to use Laplace smoothing, so that each word is assigned a minimum frequency of 1 for each message.

Under the 'naïve' assumption that word frequencies are independent variables, the conditional probability of a message belonging to a class can be written as

$$Pr(\mathbf{me} \mid cl_{pos}) = \prod_i Pr(wd_i \mid cl_{pos}) \tag{8}$$

It is then possible to calculate the value of equation (6) using Bayes theorem;

$$Pr(cl_{pos} \mid \mathbf{me}) = \frac{Pr(cl_{pos})}{Pr(\mathbf{me})} Pr(\mathbf{me} \mid cl_{pos}) \tag{9}$$

where equation (9) describes the probability model the BC uses. The classifier itself is the algorithm that assigns probabilities to different classes. A typical example in the sentiment

analysis literature is a case where there are two classes, positive sentiment cl_{pos} and negative sentiment cl_{neg}. It is also typical to see a neutral classification but the logic is the same as in the two class case.

In the two-class case the classifier faces a straight choice between two options; the conditional probability of negative sentiment is defined similarly to equation (9). That is

$$Pr(cl_{neg} \mid \mathbf{me}) = \frac{Pr(cl_{neg})}{Pr(\mathbf{me})} Pr(\mathbf{me} \mid cl_{neg}) \tag{10}$$

where we have simply substituted cl_{pos} for cl_{neg} in equation (9). Relative class membership can then be given as an odds ratio by dividing equation (9) by equation (10). That is

$$\frac{Pr(cl_{pos} \mid \mathbf{me})}{Pr(cl_{neg} \mid \mathbf{me})} = \frac{Pr(cl_{pos})}{Pr(cl_{neg})} \frac{Pr(\mathbf{me} \mid cl_{pos})}{Pr(\mathbf{me} \mid cl_{neg})} = \frac{Pr(cl_{pos})}{Pr(cl_{neg})} \prod_i \frac{Pr(wd_i \mid cl_{pos})}{Pr(wd_i \mid cl_{neg})} \tag{11}$$

where a value of greater > 1 indicates the message belongs to the positive sentiment class and a value of < 1 indicates the negative class.

The BC has been used much less frequently in the financial literature than the DC, although the BC has been used more frequently in the analysis of online textual content. Of the four published studies on online textual sentiment, three of these have used a BC (Antweiler and Frank 2004; Das and Chen 2007; Sprenger et al. 2014) and only one, Chen et al. (2014), has used a DC.

The advantage of this approach over the DC is that it allows for the researcher to input some domain specific knowledge into the classifiers probability model of how words are related to prices. This allows the classifier to potentially pick up on features of language, like the name of a prominent manager or CEO, which would not be included in a generic sentiment dictionary. The downside of this approach is that, in terms of the likely variability I argue characterises the online textual sentiment-price relationship, it implies there is a fixed probability distribution that accurately describes the relationship between words and prices. Further, it assumes that enough concepts are present in the training sample for the measured probabilities to be contextually relevant across the rest of the study period. This is quite problematic in a financial context because generally information is more valuable the less common it is. Domain specific knowledge researchers input into the classifier at the training stage is already public, so of questionable value. For example,

before the recent scandal the word ‘Libor’ would probably not have been considered to have significant price implications. After the news broke the semantic meaning of the word changed significantly.

2.4.2. Dictionary classifiers

The DC approach is very similar to the BC approach. The main difference is that typically DCs use predefined word lists, rather than having researchers train the classifier themselves. A further difference is that often DCs will either assign a word to a class or not, rather than assigning them a probability score. A typical example of this is Chen et al. (2014), who define the relative negative sentiment of a message as the count of negative sentiment carrying words divided by the message length. Although the use of a list of pre-defined terms means the classification is not domain specific, it does mean that terms can be included from a longer timeline of historical events. For example, if there are particular terms for a stock market crash then clearly a crash does not happen every day but it would be useful to know if people were using similar language to when the last one happened.

There are a number of general purpose dictionaries available. Two that have been used often in the financial literature are DICTION and the General Inquirer dictionaries, studies employing these dictionaries include Tetlock (2007), Engelberg et al. (2008), Feldman et al. (2008), Tetlock et al. (2008), Kothari et al. (2009), Doran et al. (2012), Carretta et al. (2011), Demers and Vega (2011), Loughran and McDonald (2011b), Engelberg et al. (2012), Ferris et al. (2013), Price et al. (2012) and Twedt and Rees (2012), all using General Inquirer, and Davis et al. (2012), Davis and Tama-Sweet (2011), Demers and Vega (2011), Mangen and Durnev (2010), Rogers et al. (2011) and Ferris et al. (2013), who all use DICTION. An issue with using these dictionaries is that they can often miss words which carry a different meaning in financial terminology to standard usage. Kearney and Liu (2014) give the example to ‘tax’ and ‘liability’, both of which are classified as carrying negative sentiment in the General Inquirer dictionary but are not negative in financial context. Studies, such as Henry and Leone (2009) and Loughran and McDonald (2011b), find that the use of finance specific word lists provide a better gauge of sentiment than general word lists. The words list developed by Loughran and Mcdonald (2011b) has become increasingly popular in the financial literature. Studies using this dictionary

include Doran et al. (2012), Huang et al. (2012), Ferguson et al. (2014), García (2013), Jegadeesh and Wu (2013), Chen et al. (2014), Liu and McConnell (2013) and Loughran and McDonald (2013).

A key aspect of both the DC and BC approaches I wish to highlight is that they aim to classify content based on word frequencies alone. There is no way of picking up changes in the time period referenced in a piece of content, nor is there any way of detecting factors like the trust relationship between investors and authors. The result is that the type of sign-strength-timing variability these factors are likely to cause is not picked up at the sentiment measurement stage, and so is latent from the measure.

2.4.3. Other methods

The overwhelming majority of financial research articles use either the BC or DC approach. There are, however, a number of other methodologies for sentiment extraction that have not been popularised in the financial literature. In this section I describe a number of these approaches.

Several finance papers have explored the use of term weighting schemas on classifier performance. These approaches can be seen as a cross between the BC and the DC approaches. Researchers first obtain a dictionary of sentiment carrying words then weight the terms based on some measure of their relative sentiment. Loughran and McDonald (2011b) employ a term weighting schema similar to tf-idf, Jegadeesh and Wu (2013), employ a term weighting schema based on the correlation between the words frequency and time series of stock price returns. Jegadeesh and Wu report that their term weighting schema significantly increases the strength of the sentiment-price relationship. The authors conclude that term weighting may be just as important as word selection when defining a sentiment measure.

There are a number of techniques that are popular in the computer science literature but conspicuously absent from the financial literature. Support vector machines are a commonly used alternative to the BC, these appear to have been popularised in a sentiment analysis context by Mullen and Collier (2004). Antweiler and Frank (2004) use both a support vector machine and BC but find the results are very similar, in the paper the

author's only report results from using the BC.

Often researchers in the computer science literature will make use of lexical semantics and other linguistic structures to help identify when words are sentiment carrying. This can include methodologies like part-of-speech tagging[5], where sentence structure is used to identify the role words are playing in a sentence. This can be particularly useful in a sentiment context because of the difficulty with dealing with negation. Bag-of-words based approaches have difficulty with negation because they do not recognise the interrelationship between words or sentence structure.

Another more recent innovation is the use of sentiment ontologies; these are large knowledge-bases of semantic concepts. The aim of the ontology is to provide a more contextual description of a word or set of words in a given context (Cambria et al. 2013; Grassi et al. 2011). This idea is familiar to the financial literature as finance specific dictionaries have been introduced for a similar reason. The major difference between ontologies and domain specific word lists is that the network structure of the ontology allows for the presence of inter-relationships between words. This allows for the creation of probability models that are more consistent with a human understanding of language.

A unifying theme throughout each of these approaches is that they aim to improve classifier performance by introducing more contextual information to the classification algorithm than is present in the initial group of words. This could be based on an understanding of sentence structure, linkages to other semantic concepts, or the introduction of domain specific lists of sentiment carrying terms. A feature of the type of variability I've described throughout this introduction, however, is that even if aspects of investors trust in an author, or the time period referenced in the text could be extracted using more sophisticated language processing techniques, these factors are also variable over time. Current methodologies are simply not capable of modelling this aspect of the sentiment-price relationship.

[5] See, for example, the Stanford NLP project part-of-speech tagger http://nlp.stanford.edu/software/tagger.shtml

2.4.4. Classifier performance

A number of studies have reported some descriptive statistics based on how well different classifiers perform. Typically, authors will train the classifier then apply is to a different sample of content and which has been hand classified in advance by the researcher. How well the classifier performs is then a function of how similar to the human classification the algorithm achieves.

Huang et al. (2012) document the performance of different approaches in classifying the content of analyst reports. The BC performs best with an out of sample classification accuracy of 76.91%. Finance specific word lists are the next best performing; the list supplied by Loughran and McDonald (2011b) achieves an accuracy of 62.02%. Dictionary based approaches based on general word lists are reported to perform no better than random. Kearney and Liu (2014) note, however, that there are significant differences in the form and style of text from different content sources. Company filings, for example, have defined formats they are expected to follow. Similarly, analyst reports are likely to be structured in a fixed format. In contrast, online textual content is not bound by regulation of formal rules which makes the classification of online sentiment potentially more difficult than content from other sources.

Illustrating this point, Rosenthal et al. (2015) document the result of a competition to find the sentiment classification algorithm that most accurately classifies the tone of messages from Twitter. The winner of the competition achieved an accuracy of 64.84%, well below the 76.91% reported by Huang et al. Rosenthal et al. (2015) also test the accuracy of human classification. The authors gave a set of messages to some human participants then scored each participant based on how often their classification matched the majority classification. They found that the best *human* classifier will only agree with the majority view between 71.2% and 86.4% of the time, the lower estimate being for messages that contained sarcasm. This result raises some questions about how well investors could be expected to interpret price signals in online forums of this type.

The findings of these studies suggest there is some noise expected in the classification of text because it simply is difficult to ascertain the meaning the author is trying to convey. This could take the form of random error, but Rosenthal et al.'s result suggests there is also systematic error introduced by the authors writing style. Authors that are more sarcastic will see up to a 15% reduction in the ability of the reader to understand

the meaning of the text. This echoes the argument of Tetlock et al. (2008), that there may need to be some normalisation of an author's style of writing to adjust for the fact investors will be used to information being presented in a particular way. Subsequently, they would be expected to adjust their trading behaviour not on the words in the text per se, but on a combination of the words and an understanding of how the author historically presents information.

2.5. Econometric methods for comparing sentiment metrics to prices

Once a textual sentiment metric is defined researchers will attempt to model and hypothesis test aspects of the sentiment-price relationship using some form of time series analysis. By far the most common approach is to use some form of linear regression model. Another popular methodology is to examine the returns from applying some simple trading rule based on the sentiment variable. In this section I describe these two methodologies in detail. There are a number of less common approaches that have been applied in the financial literature, such as vector auto-regressions and volatility models. I summarise some of the features of these less common approaches at the end of this subsection. Finally, I discuss some other methodologies that have been applied outside of the financial literature to provide a full overview of the field. The aim of this section is to discuss how time series modelling strategies that have been applied in the literature do not account for the type of sign-strength-timing variability I describe as characteristic of the sentiment-price relationship.

2.5.1. Linear regressions

Most studies in this area will apply some form of linear regression model in their analysis. Kearney and Liu (2014) summarise the basic linear regression model that has been applied in the literature to date as

$$y_t = \alpha + \sum_{s=1}^{s=N_s} \beta_s^0 \, y_{t-s} + \sum_{s=0}^{s=N_s} \sum_{i=1}^{i=N_x} \beta_{i,s}^1 \, x_{i,t-s} + \sum_{s=0}^{s=N_s} \sum_{j=1}^{j=N_q} \beta_{i,s}^2 \, q_{i,t-s} + \varepsilon_t \qquad (12)$$

where y is typically a time series of asset price returns but may be some other price or market feature, x_i represents the ith control variable out of a total of N_x different control variables, q_i is similarly the ith sentiment variable out of a total of N_q different sentiment variables included in the model. The intercept is given by α, whilst β^0, β^1 and β^2 are vectors or matrices of parameter values. In practice different researchers may use different numbers of lags of different variables we have set all of these to N_s for notational convenience. Studies into the online textual sentiment-price relationship also tend to use simpler models, for example omitting control variables and autoregressive terms from regressions.

In the literature to date equation (12) has been fitted with the ordinary least squares algorithm in most cases, and in all of the online textual sentiment papers we have cited. The assumption is that the error term ε_t follows a normal distribution so that price is a set of linear functions of explanatory variables with normally distributed errors. The issue with this formulation, in light of the sign-strength-timing variability I have discussed, is that the assumption of the model is that price is related to sentiment via a set of linear functions. I argue that there are aspects of the sentiment-price relationship that are variable due to latent factors, like trust, that are not reasonably measurable. Since equation (12) requires ε_t to follow a normal distribution, the only way of introducing this type of variability is through further explanatory variables, but this is not possible unless they can be measured.

2.5.2. Trading strategies

A more direct way of illustrating the strength of the sentiment-price relationship is to show returns based on some sentiment trading strategy. The sense in this approach is that the results of the strategy can be assessed independently of the strategy used, and mimic more closely how sentiment may be applied in real world trading scenarios. In the online textual sentiment literature this has been done in two ways; Sprenger et al. (2014) use a threshold filter to select days where there is significant excess sentiment around a stock. They then buy this stock and sell it again quickly to realise the return. Chen et al. (2014) take a different approach, they form portfolios of stocks based on which have the highest sentiment over some look-back period. They then hold these stocks for several months. A number of variations on these strategies have been tried in the wider textual sentiment

literature, for example, Tetlock et al. (2008) trade mainstream media news stories by each day forming long-short portfolios by buying stocks with positive sentiment and short selling stocks with negative sentiment.

These strategies do not make any attempt to model time varying latent factors explicitly; rather, they attempt to counter this type of complexity by introducing simple heuristics that should hold true in most cases. As a result, they are similar to the linear regression approach in that they assess the strength of the statistical relationship between sentiment and price in spite of sign-strength-timing variability as opposed to with it.

2.5.3. Other approaches

Other less common approaches that have been tried in the financial literature are vector auto-regressions (VAR; Tetlock 2007), logistic regressions (Huang et al. 2012; Loughran and McDonald 2011a; Loughran and McDonald 2011b; Rogers et al. 2011; Buehlmaier 2013) and GARCH models (Antweiler and Frank 2004). Each of these brings some extra insight, but in terms of the sign-strength-timing variability I have discussed, they suffer from all of the same issues as linear regression models or trading strategies.

To expand on this point, consider the logistic regression approach as used in the literature to-date. Logistic regression is an intuitively useful model in this case because the sentiment variable is typically measured as a compound of two categorical outcomes, the frequency of positive sentiment words and negative sentiment words. Logistic regression can be seen as a generalisation of the linear regression approach to use with this type of categorical data. This is done by first modelling the probability of a given outcome using the logistic function. Given as

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{13}$$

where σ is the logistic function and x is the value of the categorical variable in question. After these probabilities have been calculated the logistic regression model can be fitted with ordinary least squares in the same manner as linear regression. Lags of the categorical variable can also be included in the same manner as linear regression, so the model can be specified as

$$r_t = \alpha + \sum_{i=1}^{i=p} \beta_i \sigma(x_{t-i}) + \varepsilon_t \tag{14}$$

so the model is a simple linear model of the relationship between the return at *t*, and *p* lags of the predictor variable, each transformed to probabilities using the logistic function.

The main advantage of logistic regression is that the model assumes sentiment is a categorical variable rather than a continuous one. What is clear from equation (14), however, is that the model still assumes that returns are a simple linear function of lagged sentiment, each parameter in **β** models an independent relationship between a single lagged value of *x* and the returns series. What I have argued throughout this introduction is that this relationship is likely to vary over time so this type of model is likely to miss important features of the relationship and subsequently underestimate its strength.

Another common econometric approach applied to the financial sentiment literature by Antweiler and Frank (2004) are GARCH models. GARCH models are an extension of the classic autoregressive model to cases where the variance of the time series process is time varying. The standard autoregressive model is given as

$$x_t = \alpha + \sum_{i=1}^{i=5} \beta_i x_{t-i} + \varepsilon_t \tag{15}$$

where α is the intercept, **β** is a vector of parameters and ε is the error term, assumed to be a normally distributed random variable so that

$$\varepsilon_t = N(0, \sigma_\varepsilon) \tag{16}$$

where σ_ε is the standard deviation of the error term, typically set to 1.

The autoregressive model can be extended to an autoregressive moving average model by including further moving average parameters, that is

$$x_t = \alpha + \varepsilon_t + \sum_{i=1}^{i=5} \beta_i x_{t-i} + \sum_{i=1}^{i=5} \theta_i \varepsilon_{t-1} \tag{17}$$

where again the error is assumed to be normally distributed with consistent variance.

GARCH extends this model by assuming that, rather than being constant, the error

term also follows and autoregressive process, so that

$$\sigma_{\varepsilon,t}^2 = \omega + \sum_{i=1}^{i=q} \alpha_i \varepsilon_{t-i} + \sum_{i=1}^{i=p} \beta_i \sigma_{\varepsilon,t}^2 \quad (18)$$

The autoregressive moving average model is than as stated in equation (17) but where the error term variance is now controlled by the model in equation (18).

For bi-variate regressions, GARCH models are typically applied as a way of adjusting for time-varying variance in a time-series before the series are compared using linear regression. This is important as linear regression assumes that the variances of the processes being compared are constant. Some returns data is known to have time-varying variance and so fitting GARCH models in advance of regression analysis makes sense in this case. Also, since there is little available work on the structure of sentiment time series it may be the case that sentiment series also have time-varying variances.

When using GARCH models, however, the series are still compared using linear regression. As a result, the comparison of the returns series is still based on an independent assessment of the relationship of each lag of the sentiment series with the leading value of the returns series. Although the model may capture important aspects of the frequency structure of the time-series under study, it assumes there is no variability in the lagged relationship between the series of the type I have discussed being a likely feature of this relationship.

Outside of the mainstream financial literature studies, such as Bollen et al. (2011), have applied neural networks to fit time series models between sentient and price. The sense in this approach is that the neural network may detect non-linear relationships between the series that would not be detected by standard econometric techniques. In theory, the fact that the neural network can detect complex non-linear relationships suggests this approach may be useful in modelling sign-strength-timing variability in the sentiment-price relationship. There are, however, two major drawbacks with their use in this regard: *firstly*, neural networks require large training datasets to be effective, and out of sample validation for hypothesis testing. This makes their usage impractical with time series of limited length. As gathering large corpuses of textual content is practically challenging constructing very long sentiment time series is difficult. *Secondly*, because the network is trained in advance it will not change its model based on new input. Contextual

changes, like changes in the semantic meaning of the word Libor, cannot be accounted for once the network is trained.

2.6. Synthesising the evidence for sign-strength-timing variability

In this section I synthesise the theoretical and empirical evidence for why variability would be expected to be a central aspect of the online textual sentiment-price relationship. This is important for understanding why I focus this thesis on the methodological challenges of modelling statistical relationships that exhibit sign-strength-timing variability. All of the sources of variability I describe in this section have been discussed in previous sections, I will repeat them here to make more explicit some of the assumptions about the sentiment-price relationship that underpin the rest of this thesis. I will also discuss how existing state of the art technologies from both the financial and natural language processing literatures do not capture this property of time series relationships.

My argument is justified around three possible sources of variability: *variability in interpretation, variability in authorship* and *variability in context.* The nature of, and evidence for, each of these is described in the following three subsections. Please note that I do not intend this to represent a complete list, simply a set of enough examples to suggest variability is likely to be a defining aspect of the online textual sentiment-price relationship.

2.6.1. Variability in interpretation

The first source of variability describes the way investors interpret information may vary based on their current state of mind. Under the assumptions of behavioural theory investors use heuristics to make trading decisions. New information drives price, but how this information is interpreted is a product of an investor's state of mind at the time the information is received. This may introduce variability into the textual sentiment price relationship for the same reasons prices exhibit momentum. In this simple case, behavioural theories suggest that investors will under-react to new information. So we might expect to see an oscillating, lagged relationship between sentiment and price in the same way prices exhibit lagged serial correlation.

A more complex situation arises if, as assumed in the AMH, different species of investor use different heuristics when interpreting text or adapt the heuristics they are using over time. In this case we would expect a lagged relationship but also that the strength and timing of this relationship will vary due to changes in market ecology. As a result, there may be two relationships between online textual sentient and price, one where sentiment measures objective information but is the incorporated slowly into prices due to investors cognitive biases. The second could be that sentiment measures the extent to which different cognitive biases are represented in the current market ecology.

In either of these cases sign-strength-timing variability arises as a consequence of investors current cognitive states. These cognitive states are not likely to be fully reflected in the words used in text, leaving the investors cognitive state latent from the online textual sentiment-price relationship.

2.6.2. Variability in authorship

Another potential source of variability is in the trust, or contextual understanding investors have about the authorship of a piece of content. Trustworthiness of content is an issue that has been raised by a number of authors in the company filings and disclosures literature, for example, Huang et al. (2012), Loughran and McDonald (2011a), Larcker and Zakolyukina (2012) and Rogers et al. (2011). Each of these authors argues that managers may manipulate investors interpretations of earnings announcemnets with deceptive us of language. A less malicious type version of the argument above is discussed in Tetlock et al. (2008) who note that some authors have different writing styles to others. Some authors may use many sentiment carrying words and phrases but mean to convey the same message. Understanding the message requires investors to make judgements on how subjective they believe the author intended to be.

In either of these cases, investors must make a judgement about the effect of content on prices. Ultimately, the proof of the accuracy of this judgement is whether prices move in the way the investor believed they would. This means that there is a feedback relationship between information and trading behaviours, where investors judge future sentiment-price relationships as a function of observed senitment-price relationships that occurred in the past. For example, an investor may read a report by the manager of a

company which implies that the companies earnings are likely to be strong for the quarter. At face value, this is a clear indication that buying the stock will lead to a profit for the investor. If at the end of the quarter the stock price has plummeted, when the manager again uses language implying strong earnings for the company the same investor may take this as a signal to sell due the experience of having suffered a previous loss from trusting the managers account. Sign-strength-timing variability in this case results from the process of continually reviewing the accuracy of prior accounts from different authors.

2.6.3. Variability in context

This third source of variability refers to aspects of the context in which a piece of content was published that modify the semantic meaning in the text, but that are not explicitly referenced. This has been addressed in the financial literature by authors like Henry and Leone (2009) and Loughran and McDonald (2011b), who have shown that dictionaries of specifically financial sentiment carrying words are superior classifiers of sentiment in a financial context than general dictionaries of terms. In this case, the fact that the message contains a type of semantic information particular to the financial community is not explicit in the text itself, but is assumed by association. This is a widely accepted issue in the sentiment analysis literature where authors like Cambria et al. (2013) and Grassi et al. (2011) have pioneered methodologies for enhancing the context of sentiment analysis measures through the use of massive ontologies.

There are aspects of the context of a piece of content that can cause sign-strength-timing variability. An example of this is the point in time a piece of content references, for example, making a forecast that a stock will rise to a particular price by June 3rd 2015, will be interpreted very differently on the 2nd of June to the 4th of June. This type of context does not indicate polarity explicitly, so measuring whether the message is generically good or bad will not account for occasions when timing is a factor in whether the statement would be considered good or bad.

2.6.4. The issue with modern technology

The issue with the technologies for measuring textual sentiment I describe being used in the financial literature in section 2.4, is that they rely on there being a set of fixed sentiment carrying words. Even when more complex methodologies are used, such as the sentiment ontologies I describe in section 2.4.3, the assumption is that there are concepts that convey a fixed semantic meaning. These can be captured and recorded and then used to model future sentiment as function of prior sentiment. The Libor scandal provides a good example to illustrate a number of issues with this approach, the Libor scandal broke in the mainstream media in 2012, in summary, a number of employees at major financial institutions colluded to fix the London interbank offered rate (Libor). This is the mechanism that fixes the cost of banks borrowing from other banks in the UK market, and underpins the price of a large number of other financial instruments. Pre-2012, Libor would take a specific meaning in financial vocabulary. To most the term would not convey any particular semantic meaning. Post-2012, being associated with Libor would be seen as negative for an institution drastically altering the semantic meaning of the word.

How would this be modelled using current techniques? Neither the BC nor the DC approaches allow for words changing their semantic meaning mid-sample, so pre- and post-2012 samples would need to be classified differently. This leads to the sense in sentiment ontologies, as Libor could be linked to good sentiment pre-2012 and bad sentiment post-2012. Digging deeper, however, there were reports of issues with Libor as early as 2008 (Mollenkamp and Whitehouse 2008) and some suggestion that Libor fixing dates back as early as 1991 (Keenan 2012). What we are trying to ascertain in a study into online textual sentiment is what sort of price response we expect given the occurrence of a set of words, which implies making a judgement about what we expect a population of investors to do given after reading a piece of text. What this example shows is that for a term like Libor, there is a web of overlapping, time-varying factors that affect its interpretation. Given the different accounts, we have three dates for when Libor should be considered negative by investors, 1991, 2008 and 2012. Which of these represents the shift in investor perceptions from neutral to negative depends on which author you consider to have influence over the opinions of most investors. Assessing sign-strength-timing variability in this case requires making a judgement about investors reading of a text that requires information that is difficult or impossible to measure.

In light of this difficulty, some issues with the time series analysis methodologies I describe in section 2.5 become apparent. Linear regression models assume that factors can be found with linear relationship to asset price returns, up to the point where the residual fluctuations in the return series are normally distributed Gaussian noise. What the Libor example demonstrates is that there are latent factors affecting the interpretation of words that cannot reasonably be measured in this way. Even with more advanced modelling strategies, such as neural networks, there are 'known unknowns' that obfuscate the relationship. The types of variability I describe in this section suggest that we know there is going to be variability in the relationship in future we just don't know what it's going to look like. Neural networks require training on a sample of the data that contains all relevant inter-relationships between variables, yet the situation I describe is one where I know there are likely to be relationships between variable in the future that have not occurred in the past. To model the sentiment-price relationship requires a strategy for dealing with these 'known unknowns'.

This forms the backdrop to the three papers I present as the substantive part of this thesis, in chapters two, three and four. Up to this point, the arguments I have made about sign-strength-timing variability are not well evidenced in the literature. In the literature to-date this type of variability is either assumed not to exist at all, for example in the assumptions of the EMH, or is intentionally left out of text classification models like the dictionary and naïve Bayes classifiers. In chapter three I use a standard econometric approach to look for evidence of this type of behaviour in the sentiment-price relationship. The main result I report in chapter three is that there is significant evidence of sign-strength-timing variability in the relationship. In chapter four, I then address the issues of modelling time series relationships that have these properties. Chapter five then applies the methodology I develop in chapter four to analyse the sentiment-price relationship I observed in chapter three in more detail.

For the next three chapters I am going to switch to using the plural 'we' rather than 'I' as I have up to this point. This reflects that fact that the following three chapters are intended to be standalone papers, and I have included my supervisory team as co-authors for each.

Chapter 3: Returns to Buying Online Sentiment Winners

3.1. Introduction

A recent Goldman Sachs (2015) report discusses the ‘socialisation’ of financial investing, described by the authors as the “increased social nature of personal investing driving consumer empowerment”. In practical terms, this describes the trend for investors to eschew traditional investment channels and sources of investment advice in favour of community lending and advice platforms. These platforms largely live online, tying the trend for increased personal empowerment closely to the rise of online investment communities. To set this trend in context, Cogent Research (2008) report that one in four US adults had gone online looking for investment advice, whilst (Deloitte 2007) report that 82% of US Internet customers were directly influenced by peer review services in their investment decisions. Although these figures are from the industry press rather than peer-reviewed journals, they suggest that when people go online they tend to be influenced by communities in how they spend their money, and they go online to look for investment advice. In monetary terms the scale of this activity is economically significant, Goldman Sachs (2015) estimate that there is a $4 trillion addressable market in investment services that utilise online social platforms.

These trends have not escaped financial researchers who have in recent years become increasingly interested in measuring sentiment or affect in text. A large number of studies now exist detailing the relationship between ‘textual sentiment metrics’, derived from news articles and company filings, and stock prices (Kearney and Liu 2014). In these studies, researchers define a quantitative measure of the tone of a document based on features of language used in the text. For example, the word ‘bullish’ would be considered to carry positive sentiment in a financial context whilst ‘bearish’ would be expected to carry negative sentiment. Metrics are based on the frequency of these words in the text, so that by ‘textual sentiment’, what is really being measured is some function of the number of times sentiment carrying words appear.

In contrast to this larger literature, the literature concerning *online* sentiment and prices is still quite sparse. There are four main finance papers describing the relationship between online content textual sentiment and prices:

Antweiler and Frank (2004) study the relationship between the content of over 1.5 million messages from the Yahoo Finance[6] and Raging Bull[7] message boards, and the price of stocks in the Dow Jones Industrial Average and Dow Jones Internet Commerce Index. They find that textual sentiment measures predict price features, such as trading volume and volatility, at short lag lengths. They also find that sentiment is contemporaneously correlated to asset price returns, but that there is no predictive relationship between the sentiment and returns.

Das and Chen (2007) study the relationship between the content of 145,110 messages from a number of financial message boards and the prices of 24 stocks in the Morgan Stanley High-Tech Index. The authors report that sentiment predicts next day prices, but not next day returns. Similar to the findings of Antweiler and Frank, the authors document a contemporaneous relationship between sentiment and returns.

Sprenger et al. (2014) document a predictive relationship between messages from the micro-blogging platform Twitter[8] and returns to stocks referenced in the messages. The study examines 249,533, 140 character messages containing the ticker symbol of a stock exchange listed company. The authors show evidence of a negative correlation between sentiment and prices at short lag lengths, this relationship is shown to be strong enough to make small profits in excess of costs from a simple trading strategy.

Chen et al. (2014) also document a predictive relationship between textual sentiment and returns. Using 97,070 articles and 459,679 comments on these articles from the financial content site Seeking Alpha[9]. The authors document a positive correlation between sentiment and returns over the following three months from the date the content was posted. The authors go on to show that this relationship persists for up to three years.

The results of Sprenger et al. and Chen et al. suggest a predictive relationship exists between sentiment and prices, but there are contradictory accounts of the sign, strength and

[6] Yahoo Finance, available at http://finance.yahoo.com/
[7] Raging Bull, available at http://ragingbull.com/
[8] Twitter, available at https://twitter.com/
[9] Seeking Alpha, available at http://seekingalpha.com/

timing of this relationship. For example, Antweiler and Frank and Das and Chen both report contemporaneous correlations between sentiment and returns but no lagged relationship. Sprenger et al. report short run negative lagged correlations between sentiment and price, and Chen et al. report long run positive relationships between sentiment and price of up to three years.

A further issue with this body of evidence is that studies to-date, have used relatively small, platform specific samples. Antweiler and Frank and Das and Chen study message-board postings, for example, although there are a range of other content sharing platforms that exist. Each platform comes with its own structural constraints regarding how content is shared, so it is likely different platforms will attract slightly different demographics, or cultures of people to use the services. Twitter, as studied by Sprenger et al., is a good example of this; messages on Twitter are constrained to be no more than 140 characters, placing a structural constraint on the information that is being conveyed. Another consideration is how different platform may host different types of investment communities, the comments section on Seeking Alpha, for example, may be home to discussions between different types of investors to those found posting on Yahoo Finance.

As a result we see two gaps in the current literature: *firstly*, there is a lack of empirical evidence of the time evolution of the sentiment-price relationship. Studies have focused mainly on short lag lengths, whilst Chen at al. show there may be a long run relationship as well. The sign-strength-timing variability reported across studies also suggests there may be a more complex interaction between sentiment and price than has usually been considered. *Secondly*, studies to-date have used one or two sources of content to form their sentiment measures, and have used fairly small corpus sizes, this risks biasing the results towards the behaviour of particular types of investors or investment communities associated with specific online platforms.

To address these issues, in this paper we consider the lagged relationships between the component stocks of the Standard and Poor's 100 (S&P100) stock index and the content of over 10.2 million messages from 45,516 separate content sources referencing these stocks over a 15-month period running from January 2014 to March 2015. As a result, our corpus is an order of magnitude larger than the number of messages that have been used in previous studies, and the range of content sources is far less platform specific than previous studies.

Our main finding is that the sentiment signal oscillates around prices in a predictable manner. We see evidence of short-run (20- to 40-trading days) negative correlation between sentiment and price, followed by a midterm (80- to 100-trading days) positive correlation, ending in longer-term (100- to 120-trading days) negative correlation. We see that these correlations strengthen significantly the larger the number of lagged sentiment days we consider in the regression. We go on to show that using a simple trading rule of buying 'sentiment winners', stocks which have on average greater negative sentiment attached to them over the past five months via our measure and holding them for the following two months, returns 9.766% APR in excess of the market rate over the sample period once reasonable transaction costs have been accounted for. This corresponds to approximately double the return of holding an equal weighted portfolio of the underlying assets over the same period.

Our results suggest that the same words that make up our sentiment measure lead to different trading behaviours at different time scales from the date the content was published. This somewhat confirms the joint findings of Sprenger et al. (2014) and Chen et al. (2014), since our results show that both the short-run negative correlation documented by Sprenger et al., and the longer-run positive relationship documented by Chen et al., could be consistent with price responding to sentiment in an oscillatory fashion. This evidence points to there being a latent factor influencing the way word frequencies are interpreted by investors that is not captured by current sentiment analysis methodologies which influences the sign, strength and timing of the sentiment-price relationship.

Looking to the wider literature, there are a number of possible sources of latent variability that have been documented in relationships between prices and information. These include:

Variability in investor interpretations of new information – there is a large body of research showing that prices exhibit momentum, that is, stocks with greater than average 6- to 12-month returns tend to continue to produce excess returns for the following 3- to 6-month period (Jegadeesh and Titman 1993; Asness et al. 2014; Asness et al. 2013). A number of behavioural asset pricing theories, such as Daniel et al. (1998), Daniel et al. (2002), Barberis et al. (1998), Hong and Stein (1999), Grinblatt and Han (2005) and Frazzini (2006), purport to yield price momentum as a consequence of the cognitive biases and heuristics investors use to make trading decisions. For example, Daniel et al. (1998)

consider investors to be overconfident about their private information and likely to over attribute their successes to their own skill and their failures to bad luck. As new information comes to light, investors' irrational belief in the value of their information creates a delayed reaction the new evidence. The presence of self-attribution bias also causes a build-up in overconfidence during times when the market is performing well, magnifying existing biases. Over the long term, the effect of these biases reduces, leading to a long-term price reversal following midterm momentum.

In another example, Hong and Stein (1999) reach approximately the same conclusions via a slightly different route. They consider two types of investor; *newswatchers* and *momentum traders*. Newswatchers rely entirely on their private information, whilst momentum traders rely solely on past prices to inform their decisions. The model assumes that private information diffuses slowly into the marketplace, so there is an initial under reaction to new information. Once momentum traders pick up the price signal from observing the effects of the newswatchers' trades they follow suit, creating price momentum. As in the Daniel et al. (1998) theory, over time the effects are reduced and prices are reversed in the long run. Hong and Stein also propose a mathematical model of their theory illustrating the interplay between newswatchers and momentum traders as oscillating waveform following an information event.

Based on the assumptions of these theories it is possible that textual sentiment holds some price relevant information, but that investors under-react to this information. This leads to an oscillatory effect in the sentiment price relationship due to investor cognitive biases. Clearly some of the information required to form a full model of the state of mind of investors in the market will be missing from the words in the text, making this a potential candidate for the oscillatory pattern we observe in the textual sentiment-price relationship.

Variability in the authorship of online content – there is a trust relationship between the author and reader of online content. In other areas of the financial literature authors, such as Huang et al. (2012), Loughran and McDonald (2011a), Larcker and Zakolyukina (2012) and Rogers et al. (2011), have noted that managers of companies may use deceptive language to mislead or manipulate investor perceptions of earning announcements. As online content is much less regulated than company filings it is natural to assume that some manipulation or misleading may take place. What authors like Rogers et al. (2011) suggest,

however, is that the fact the information presented to investors is intended to be misleading may itself be material information about the likely conditions in a company. The relative trust investors have in the author of the content acts as a semantic marker altering the interpretation of the text.

A similar possibility, noted by Tetlock et al. (2008), is that different authors have different writing styles, so the same information about a stock may be conveyed with different language by different authors. In this case rational investors would be expected to adjust their interpretation of the language used to match an understanding of the author's style. Another possibility is that investors do not pick up on this nuance and act literally on the content of the text, subsequently altering their trading decisions based on a change of authorship style. Since the dataset we consider covers many different online communities it is reasonable to assume there will be varied styles across different content sources, but also that the main contributors, and thus dominant writing styles, across these services will be variable over time.

The result of these factors is that, given the same wording, different semantic meaning will be placed on the text due to an understanding of its provenance or authorship. In the case of financial investing, ultimately, the quality of information from a given author is assessed based on how well it matches to the future price of stocks. The trust relationship between reader and author in the investors mind is likely to be constantly reassessed based as new information about prices occurs. In this case the oscillatory relationship could be due to investors re-assessing their opinion of prior content in light of new price movements.

Variability in the context the content was created in – text content does not contain a full list of all of the contextual information an investor is expected to know on reading the text. In a financial context, authors like Henry and Leone (2009) and Loughran and McDonald (2011b), have noted that dictionaries of specifically financial words are better at classifying financial sentiment than general dictionaries. For example, words such as 'liability' or 'tax' would generally be considered to be negative in general parlance, but in a financial context are not necessarily. The reason is that in financial content online there is an assumed level of contextual knowledge the creator of the content assumes the reader of the content has about financial investing. Adding in this context improves the accuracy of sentiment classification. Jegadeesh and Wu (2013) expand on this by showing that, even

when financially specific words are considered, fitting a stock, and time specific model to the derive term weightings for sentiment carrying words improves their ability to classify sentiment. This suggests that even words chosen for their supposedly unambiguous financial meaning can be interpreted in different ways by investors at different times and in different contexts.

The fact that this type of contextual information is an important aspect of measuring the semantic content of text is also widely accepted outside the financial literature. The techniques researchers currently use for measuring textual sentiment are largely drawn from the natural language processing literature, of which sentiment classification is a sub-discipline. Major avenues of research include the creation of massive online data-stores of linked semantic context, in recognition of the fact that even simple concepts like ‘good’ and ‘bad’ can have vastly different meanings if referenced in different contexts (Cambria et al. 2013; Grassi et al. 2011). There is evidence documented in the financial literature that time varying context may play a role in the textual sentiment-price relationship. García (2013) shows evidence that investor interpretations of sentiment varies during recessions. Clearly recessions do not happen all of the time, so as the present context changes over time the suggestion is so will investors interpretations of content.

Applying this same logic, it is plausible that pieces of content reference uncertain or changing points in the future. For example, the news that ‘the Dow Jones Industrial average will top 17,000 points next week’, would presumably constitute good news in most cases, however, if you have just placed a bet that this would occur this week then the semantic meaning of the statement changes significantly. Much of the existing financial content makes some assessment of the likely future return of stocks, so investors may buy on the expectation of a statement being true only to sell when the prediction window for the statement expires. This type of temporal information is latent from our sentiment measure as we only consider the date the message was posted.

Of the three latent factors we suggest, only the first has a natural proxy in the literature. Behavioural asset pricing theories are largely based on the empirical observation of price momentum, first documented by Jegadeesh and Titman (1993). Momentum describes the tendency for prices to continue going in the same direction. Empirically, this means that it is possible to systematically profit from a strategy of buying winners and selling losers, where winners are stocks that have better than average performance over the

previous 3- to 12-months, and losers are stocks with lower than average performance over the same period. Holding this portfolio will earn excess returns over the next three months. Recent work by Asness et al. (2013) and Asness et al. (2014) has shown this effect to be remarkably persistent across asset classes and time periods.

Moskowitz et al. (2012) document a similar ‘time series momentum’, whereby the future one-month return of a stock is shown to be predictable based on its prior 12-month return. The difference between the two effects being that the first refers to the predictability in the relative price of an asset to the wider market and the second to the predictability of the return of a single stock from its own prior price. In either case, the picture is one where a mid-term (up to one year) sample of past prices can be used to forecast a shorter-term (up to three months) return period, so there is an asynchronous relationship between the predictive period and the predicted future period.

The momentum literature is augmented by the long-run reversal phenomenon documented by Thaler and De Bondt (1985), where over the longer term (three- to five-years), returns on average reverse so that forming portfolios of average long-term losers outperforms the market. Together, these bodies of evidence describe how monthly returns are likely to be positive following 3- to 12-months of past positive returns and negative in the longer term, forming an oscillatory effect on prices.

In order to test to what extent the sentiment-price relationship can be explained by momentum, we use multivariate regression analysis to test for whether the sentiment-price relationship can be explained by serial correlations in price. We show that whilst the two effects are similar, the predictive power of the model is significantly increased by including both factors. Our results suggest that online textual sentiment contains economically significant, incremental information above what is already recorded in serial correlation in price.

In conclusion; as there is predictability over and above serial correlation in price, we suggest that one of the other forms of viability in the lagged relationship is playing a role in forming the oscillations we observe. If this is the case then it may be very difficult to find suitable proxy variables for some forms of variability in the relationship. For example, it seems unlikely that there exists a simple measure of investors trust in a piece of contents author that we could use in forecasting models. As a result, it is likely that variability is a feature of the relationship that needs to be modelled in its own right. The

crux of this point relates to what we consider a piece of text to be with respect to price. Current methodologies assume there are features of a piece of text that have a consistent interpretation, that is, there are words that always hold incremental positive or negative information about prices. What we show is evidence that the relationship between words and prices is not that simple, echoing the conclusions of Jegadeesh and Wu (2013), we argue that the semantic meaning of a word is determined largely by contextual factors.

Our findings should then speak to a number of different areas of the literature. Firstly, we contribute to the nascent online textual sentiment literature by conducting a sentiment study over the largest and most diverse range of content yet considered. We confirm some of the contemporary and short-run effects observed by Antweiler and Frank (2004), Das and Chen (2007), Sprenger et al. (2014) and Chen et al. (2014) but also highlight a new and significant aspect of the longer-term sentiment-price relationship. We provide methodological innovations supported by interesting empirical results that could inform future work in this area. Secondly, we draw a link to the under reaction and momentum literature, highlighting an empirical feature of the sentiment-price relationship that could shed new light on existing behavioural theories.

The rest of this paper is laid out as follows: section 3.2 describes our data and variable construction, section 3.3 introduces the lagged regression model we used in our analysis, section 3.4 shows our empirical results, section 3.5 shows the returns from applying some simple trading rules based on our results, and section 3.6 concludes and discusses future work.

3.2. Data and variable construction

In this section we introduce our data and variable construction. For price data, we follow Sprenger et al. (2014) in using the component stocks of S&P100. We find that for 93 of these stocks, we have messages every day in the sample period, so we omit the 7 stocks without full coverage from the analysis (a full list is given in the Appendix). We download daily adjusted closing prices from the Yahoo Finance website.

For text data, we have access to a huge range of social media content crawled between the beginning of January 2014 and the end of March 2015. To gather this content,

we searched sources for company-specific content using query strings given in the Appendix. These queries are designed to return messages where the content references the company by name or the company's stock ticker directly. Some companies, however, have ambiguous names or ticker symbols. In these cases, the query strings are edited to remove any ambiguity. In total, we have more than 10.2 million messages containing over six billion separate words covering 45,516 different content sources for the 93 component stocks of our reduced S&P100 index. A breakdown of the messages by stock is given in the Appendix with corresponding descriptive statistics. To put this dataset in context, the second largest corpus considered in a study of online sentiment was 1.5 million message board posts from the Yahoo Finance and Raging Bull message-boards by Antweiler and Frank (2004). As a result, our dataset is an order of magnitude larger and drawn from a far more diverse range of sources than previous studies have considered.

3.2.1. Price variables

We convert the stock price data to log returns in the usual manner, so that $r_{i,t}$ denotes the logarithmic return of the *i*th stock at time *t*, that is

$$r_{i,t} = log\left(\frac{p_{i,t}}{p_{i,t-1}}\right) \tag{19}$$

where $p_{i,t}$ is the adjusted closing price of the *i*th stock at time *t*, and *log* denotes the natural logarithm.

There is a risk in defining returns in this manner that any relationship we reveal might be due to the influence of sentiment on the market as a whole rather than a stock-specific relationship. To control for this, we follow Sprenger et al. (2014) in defining a simple measure of abnormal returns as the return of the stock minus the average return of the basket of stocks over the same time period, that is

$$ar_{i,t} = r_{i,t} - r_{mkt,t} \tag{20}$$

where the $r_{mkt,t}$ indicates the return of the market portfolio at time *t*, in our case this is an un-weighted portfolio of all the stocks in the sample.

3.2.2. Measuring textual sentiment

The size and diversity of our corpus of messages gives us some unique challenges that have not been addressed in the online textual sentiment literature before. Antweiler and Frank (2004), Das and Chen (2007) and Sprenger et al. (2014) all use machine learning techniques to measure the tone of documents. For these techniques to work requires the hand classification of a subset of messages, but given the diversity of the content sources, to hand code a representative sample of the content to train a classifier on would be too laborious. The other option, used by Chen et al. (2014), for online textual sentiment and, for example Doran et al. (2012), Huang et al. (2012), García (2013), Jegadeesh and Wu (2013) and Liu and McConnell (2013) for other forms of textual sentiment, is to use the dictionary of positive and negative financial sentiment-carrying words provided by Loughran and McDonald (2011b). Message tone can then be assessed by counting the number of words in a message that occur on these lists. Following this second approach, we count the number of positive and negative words as listed in the Loughran and McDonald dictionary for each message for each day. We further omit messages that occur on days where the market does not open and count days in Eastern time, so as to align with the time zone of the New York Stock Exchange.

A second consideration we face is that the length of messages in our corpus and the writing styles of the authors are varied. A typical construction for a sentiment metric used by, for example, Kothari et al. (2009), Ferguson et al. (2014) and Chen et al. (2014), is to count the number of negative sentiment-carrying words and divide them by the number of words in the text. A consideration we face because of the varied content, however, comes from the fact that the words are not evenly distributed in texts but follow an approximate Zipf-Mandelbrot distribution (Mandelbrot 1966). What this means in practice is that non-sentiment-carrying words in the text such as ‘and’ or ‘the’ will increase as a multiple of the number of times a sentiment-carrying word appears, so we would expect that the sentiment content of longer texts would be underweighted by this approach. Conversely, we would also expect that positive and negative words would be playing approximately the same role in the structure of a document, so their relative relationship to each other would be a less-biased indication of document tone.

In light of this, we classify the tone of a message as the natural logarithm of the ratio of positive to negative words in a document. The log transform is similar to that

employed by Antweiler and Frank (2004) and Sprenger et al. (2014) in their metrics and is useful as it evenly distributes the ratio around the mean. A consequence of doing this is that both sides of the ratio are sum-able over messages, so we can define sentiment as the log of the total amount of positive words in all messages at time *t* divided by the total number of negative words for the same time period, that is

$$q_{i,t} = log\left\{\frac{\left(1 + lmp_{i,t}\right)}{\left(1 + lmn_{i,t}\right)}\right\} \tag{21}$$

where $q_{i,t}$ is the total sentiment expressed towards the *i*th stock at time *t* via our measure, $lmp_{i,t}$ is the count of the number of positive sentiment-carrying words in messages referencing the *i*th stock at time *t*, and $lmn_{i,t}$ is the count of negative sentiment-carrying words for the *i*th stock at time *t*. We add one to these counts to control for situations where there may be no positive or negative word over a particular time period.

3.3. Methodology

In this section, we introduce the model we are going to use to analyse our data based on the axes of the sentiment-price relationships we wish to consider. The area of the financial literature where empirical results most closely resemble the type of variable time series relationships we describe in section 3.1 is the momentum literature. This literature has some well-established techniques for studying long-run price predictability and accounting for complexities in the time-evolution of time series relationships. A typical approach employed by authors such as Moskowitz et al. (2012) and Jegadeesh and Titman (1993) is to perform multiple regression tests with different averaging periods. For example, Moskowitz et al. (2012) construct a model where the return of a stock over the previous 12 months is used to forecast the following one-month return. The predictor and holding period windows are not always adjacent in time. Jegadeesh and Titman (1993), for example, skip a week between the lagged return period and forecast period. Including all these features in a model requires considering three axes of the data: the predictor period, holding period, and lag length.

Considering these axes in the context of the online information price-relationship, the predictor period is suggestive of the fact that one day might not always fully measure a

single information event. Similarly, variation in holding periods suggests that it may take several time periods for investors to fully respond to an information event. The third factor, lag length, suggests it might take some time before investors react at all to a sentiment event.

3.3.1. Modelling lag lengths

To model the lag length between sentiment events and prices, we are going to use a standard time-sequencing approach in the financial literature and hypothesise a simple linear relationship between the returns series and the sentiment series lagged by τ days. One way of doing this is for each individual stock, but more interesting in this context is to know if there are global features describing how investors generally interpret online information about a range of stocks.

To do this, we first normalise the sentiment series to z-scores following Tetlock et al. (2008). We do this because there may be features of the sentiment series that are distinct to individual stocks; messages may be dominated by a small number of authors with writing styles that are distinctly different from the general population. Denoting the z-score for the *i*th stock at time t as $q'_{i,t}$, we then specify this model as

$$r_{i,t} = \alpha + \beta_q q'_{i,t-\tau} + \varepsilon_t \tag{22}$$

where α and β_q are the intercept and slope we will fit in the corresponding regression, τ is the lag length being considered, and ε is the error term. We will use r in the descriptions of these models, but in our analysis, we will consider both returns and abnormal returns in the same manner. The interpretation of β_q in equation (22) is then the average size over the 93 stocks of the movement in a single stock's return given a movement in the same stock's sentiment series.

3.3.2. Modelling prediction windows

To model the average time it takes for a sentiment event to occur, we are going to use simple moving averages over different numbers of days on the sentiment series. Denoting

the moving average function as *MA*, we define this function as

$$MA(q_{i,t}, m) = \frac{1}{(m+1)} \sum_{k=0}^{k=m} q_{i,t-k} \tag{23}$$

where m is the window over which the average runs. Combing equations (22) and (23), we have a model capable of assessing a variety of different lag lengths and prediction windows, that is

$$r_{i,t} = \alpha + \beta_q + MA(q_{i,t-\tau}, m)' + \varepsilon_t \tag{24}$$

3.3.3. Modelling holding periods

To model holding periods, we use the same approach as for the sentiment windows. Denoting the holding period function as *H*, similar to the moving average function, we define this fully as

$$H(r_{i,t}, h) = \frac{1}{(h+1)} \sum_{k=0}^{k=h} r_{i,t+k} \tag{25}$$

where the only difference between the moving average function and this, is that the parameter of the function *h* moves the window forward in time rather than backwards.

3.3.4. Full model specification

Aggregating these further features into one model gives a functional form flexible enough to model all three axes of the relationship, that is

$$H(r_{i,t}, h) = \alpha + \beta_q MA(q_{i,t-\tau}, m)' + \varepsilon_t \tag{26}$$

where the dash indicates that we convert to z-scores after applying the average to adjust for variations in the writing styles of authors in relation to different stocks. By varying *m* in equation (26), we vary the average, where intuitively *m* = 0 means that we apply no average to the series. By varying *τ*, we vary the time differential between the series, where

$\tau = 0$, we consider the contemporary relationship. And by varying h, we vary the holding period, where $h = 0$ corresponds to a holding period of a single day.

As we are using moving averages in fitting the unknown parameters of the model α and β_q with linear regression, we would expect the presence of auto-correlated errors. To control this, we fit the model using the generalised least squares algorithm (Aitken, 1934) as implemented in the statsmodels library of the python programming language.

3.4. Empirical Results

In this section, we present findings from fitting various parameterisations of equation (26) to our data.

3.4.1. Lagged regressions

We start by considering simple lagged regressions, where the holding period $h = 0$, but the lag length τ and the prediction window m are allowed to vary. We show values for $m = 1$, $m = 20$, and $m = 40$ and for the value of τ ranging from 0 to 120, so in effect calculating the cross correlation function between sentiment and price of over a six-month period in trading days. We have limited the number of averaging windows to make the presentation of the results easier to follow but have run the same analysis on a larger range of values of m with very similar results.

Following the way Moskowitz et al. (2012) present their results, figure 1 shows the plot of the t-statistics for fitting these regressions to the returns series r. The x-axis is the lag length τ, and the y-axis is the value of the t-statistic for the regression. The coloured lines represent the different averaging windows, where blue corresponds to $m = 1$, green to $m = 20$, and red to $m = 40$. We see that there appears to be oscillations in the value of the regression parameters, where the most significant of these fall between $\tau = 20$-40, $\tau = 80$-100, and $\tau = 100$-120, with the first oscillation being negative, then positive, then negative again. For each of these oscillations, the t-statistics would indicate the parameters are significant at the $p < 0.1$ level.

The evidence from figure 1 indicates that the sign of the relationship between a given piece of content and price changes as we move away in time from when the content was posted. This oscillation in the strength and sign of the relationship shown in figure 1 is very similar to the time series momentum effect documented by Moskowitz et al. (2012), which draws a natural link to behavioural asset pricing literature and theories, such as Hong and Stein (1999), that yield oscillatory price movements as the product of investors under-reacting to new information.

Figure 2 shows the corresponding plot to figure 1, but reporting on the abnormal returns series rather than raw returns. We see immediately from the lower t-statistics that the relationship between sentiment and abnormal returns is much weaker. This suggests that the sentiment effect measured in figure 1 is influenced by a market factor rather than being entirely discrete to individual stocks. One interpretation of this is that there must be overlapping information contained in messages about certain stocks, so that a message that specifically references Apple's stock price would also be read by investors as news influencing the price of other similar shares that are not explicitly referenced in the message content. The general shape of the results is the same as figure 1, suggesting that it is the same effect being measured in either case, but the result for abnormal returns is noisier, again supporting the idea of the leaking of information across contexts where it is difficult from the content of the message to understand the implication of the content of the message in all possible contexts it could be applicable.

Moskowitz et al. (2012) also documents some cross correlation in the time series momentum effect between stocks, drawing another parallel between the momentum literature and our findings. Interestingly, an aspect of time series momentum documented by Moskowitz et al. is that it is correlated across stocks, in much the same way we observe that the sentiment-price relationship is stronger when market factors are also included in price. Behavioural asset pricing theories, such as Daniel et al. (1998), Daniel et al. (2002), Barberis et al. (1998), Hong and Stein (1999), Grinblatt and Han (2005) and Frazzini (2006), all concern the pricing of a single risky asset. As a result they are not suitable for describing the market factor we see in the sentiment-price relationship.

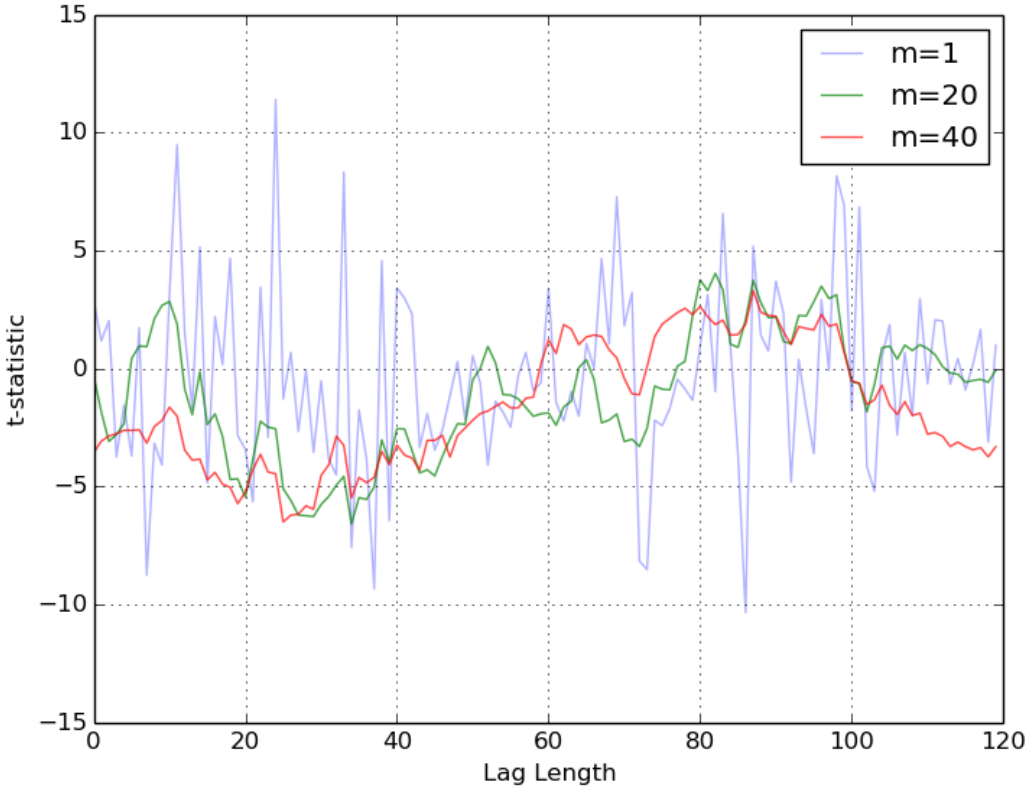

Figure 1: Cross correlation analysis of raw returns: The x-axis shows the lag length τ, and the y-axis shows the value of the t-statistic for the fitted parameter of equation (8). The key in the upper right-hand corner shows the different values of the prediction period m displayed on the plots. For the $m = 1$ model, we see there is a large number of significant t-statistics, but the lag length where these occur is unpredictable. Where we have applied the prediction window, an oscillating pattern emerges, where the sign of the correlation is initially negative though lags 20–40, then turns positive through lags 80–100 before finally turning negative again through lags 100–120.

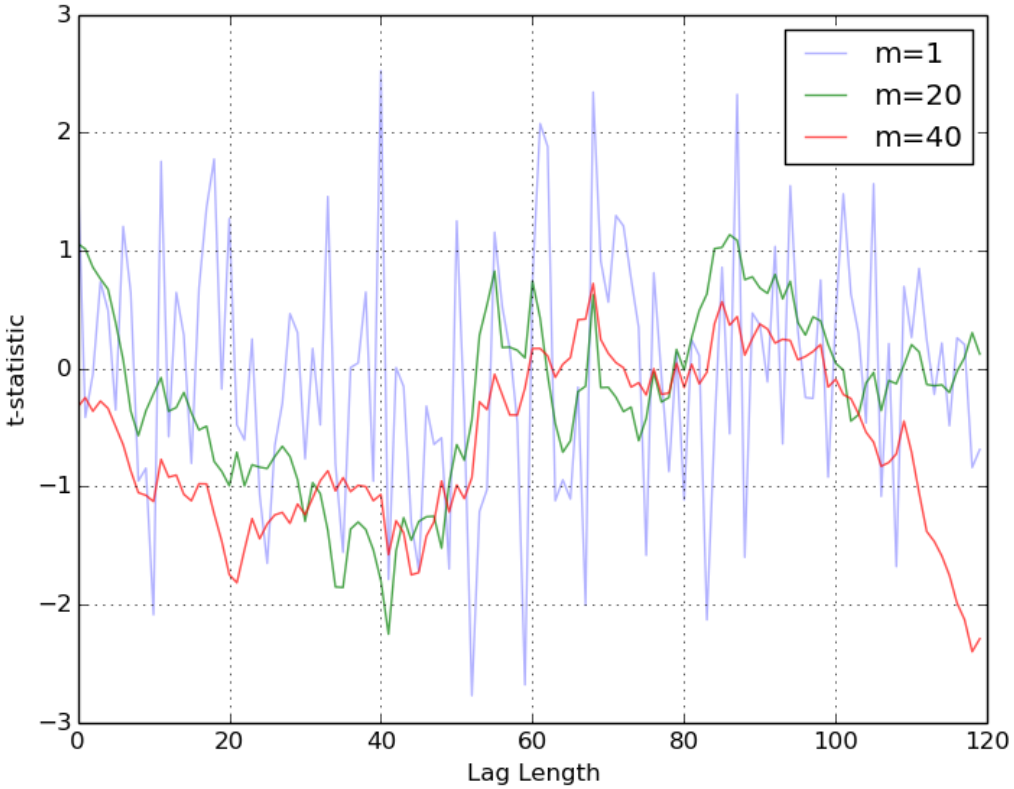

Figure 2: Cross correlation analysis of abnormal returns: Similarly to figure 1, the x-axis shows the lag length τ, and the y-axis shows the value of the t-statistic for the fitted parameter of equation (8), the difference being that the model is now fitted to abnormal returns rather than raw returns. The key in the upper right-hand corner shows the different values of the prediction period m displayed on the plot. The t-statistics are calculated using the generalised least squares algorithm (Aitken 1934) to control heteroskedasticity. We use the version provided in the statsmodels library of the python programming language. The key difference between figures 1 and 2 is that now the t-statistics are much smaller than for raw returns. The oscillations remain similar and take positive and negative signs at approximately the same lag lengths as raw returns.

3.4.2. Lagged regressions with different holding periods

Figures 1 and 2 only consider one-day holding periods. To expand our analysis, figures 3 and 4 consider holding periods of $h = 20$ and $h = 40$ days. Again, we considered a range of other holding periods in our initial analysis, but the results are similar, so we omit further results for ease of presentation. Figure 3 shows the results of this analysis for raw returns, thus the corresponding plot to figure 1 with the holding periods included. The key in the top left-hand corner of the table shows the model specification. The x-axis and y-axis are the lag length and t-statistics, respectively, as before.

We see from figure 3 that considering longer holding periods strengthens the relationship very significantly in all cases. The different oscillations we observed in figures 1 and 2 are still present and occur at approximately the same lag lengths. We also see that there is a slight upwards trend in the sentiment-price relationship as the lag length increases. This supports the finding of Chen et al. (2014) who reported a similar increase at greater lag lengths, however, this finding did not include evidence of the same type of oscillatory effect we observe here.

Figure 4 shows the corresponding plot to figure 2 for abnormal returns and different holding periods. Again, we see a similar pattern in the shape of the oscillations. We also see the strength of the relationship increase significantly once the holding periods have been included. An interesting aspect of figure 4 that we do not observe from the other plots is that the sign of the correlation between sentiment and price is almost always negative. Also, there is a general negative trend in the t-statistics as the lag length increases.

The general story from figures 1-4 is that there is sign-strength-timing variability of the type we describe in section 3.1. At different timings from the date a piece of content is created there are different sign and strength responses from the returns series. These responses are strong enough to have significant coefficient values.

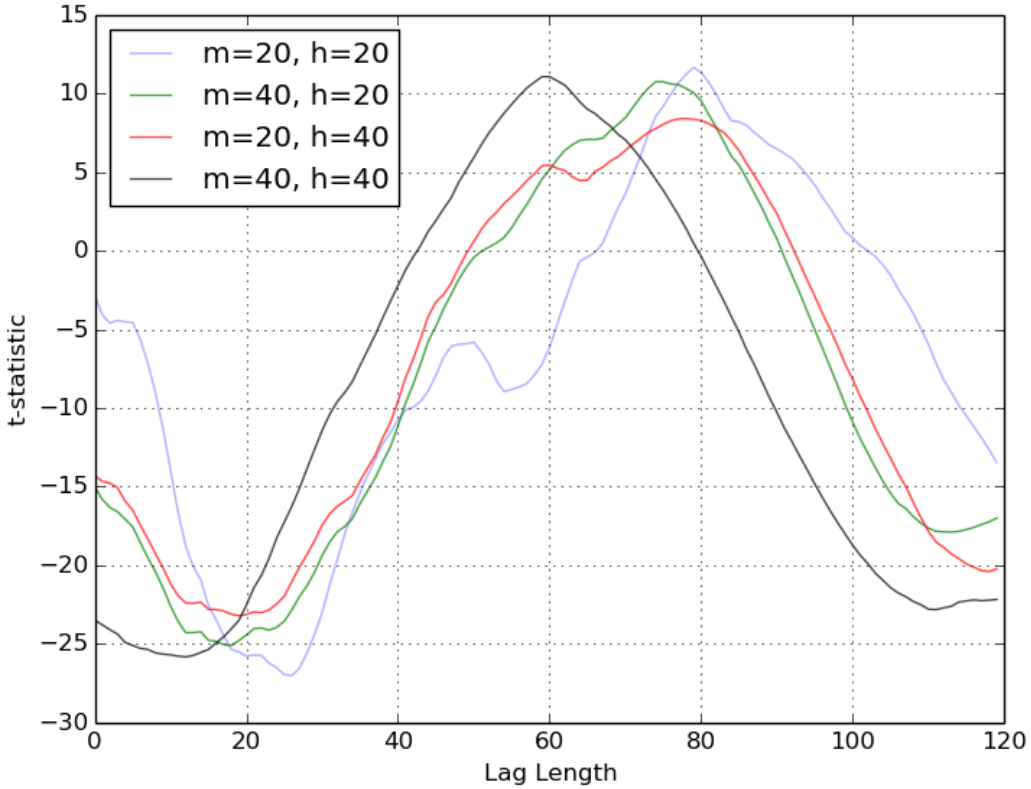

Figure 3: Cross correlations of raw returns with different holding periods: Similarly to figures 1 and 2, the x-axis shows the lag length τ, and the y-axis shows the value of the t-statistic for the fitted parameter of equation (8). The key in the upper left-hand corner shows the different values of the prediction period m and holding period h displayed on the plot. The t-statistics are calculated using the generalised least squares algorithm (Aitken 1934) to control for heteroskedasticity. We use the version provided in the statsmodels library of the python programming language. We see that the inclusion of a holding period significantly smooths the shape of the oscillations we observed in figures 1 and 2. The size of the t-statistics also indicates that the result has been significantly strengthened.

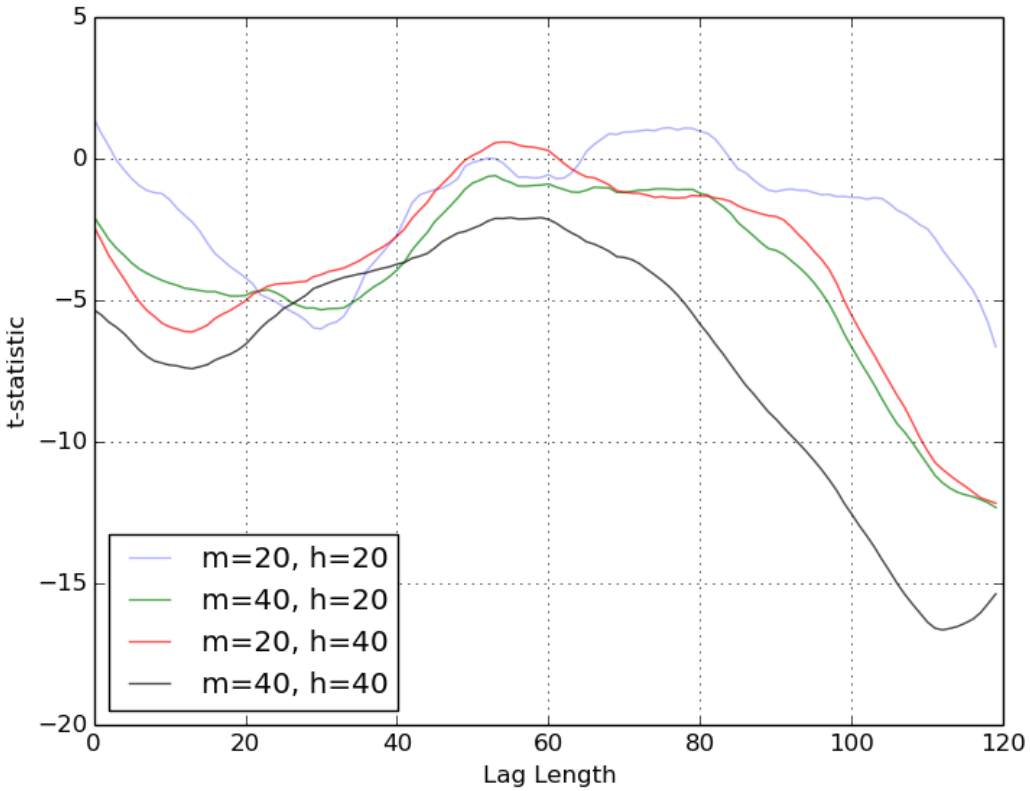

Figure 4: Cross correlations of abnormal returns with different holding periods: Similarly to figures 1, 2 and 3, the x-axis shows the lag length τ, and the y-axis shows the value of the t-statistic for the fitted parameter of equation (8). The key in the upper left-hand corner shows the different values of the prediction period m and holding period h displayed on the plot. The t-statistics are calculated using the generalised least squares algorithm (Aitken 1934) to control for heteroskedasticity. We use the version provided in the statsmodels library of the python programming language. As with figure 3, we see from the size of the t-statistics that the sentiment-price relationship is significantly stronger with the inclusion of a holding period. We also see, unique to this plot, that almost all of the t-statistics are negative in sign, and they decrease as the lag length increases. This suggests that as the lag length increases as the sentiment-price relationship becomes predominantly negative.

3.4.3. Prediction and holding periods

A feature of figures 1-4 is that increasing the prediction window and holding periods appears to have a larger effect on the strength of the relationship than varying the lag length. To analyse this in more detail we run regressions for different prediction windows and holding periods setting the lag length $\tau = 1$ so that the two windows do not overlap. We choose prediction windows of $m = 20$ to $m = 100$ trading days at 20 trading day intervals, so approximately 1- to 5-months in trading days. We use holding periods of $h = 10$, $h = 20$ and $h = 40$ trading days. Again, we have run the same analysis with various other values of m and h but find the results we present representative of other model specifications.

Table 1 summarises the results of these regressions for both raw and abnormal returns. Coefficient values are reported in standard deviations. For raw returns, we see that the initial relationship is positive in sign, with the strongest relationship being the $m = 20$, $h = 20$ model, where $R^2 = 2.27\%$. This changes for longer prediction windows, where the signs of the relationship are all negative. In these cases, the longest holding period always has the strongest relationship, with the highest being the $m = 60$, $h = 40$ model where $R^2 = 4.11\%$.

For abnormal returns, the relationships are all weaker, with the positive short-term relationship having disappeared completely. At longer prediction windows, we see the relationship returns has the same sign as the raw returns, although the relationship is not as strong. Again, we see that the relationship is strongest for longer holding periods, where the highest R^2 value is now the model, where $m = 80$ and $h = 40$, $R^2 = 1.37\%$.

The evidence from table 1 suggests that the long-term aggregate of information has a predominantly negative correlation with returns. This is in contrast to the finding of Chen et al. (2014), where the authors reported the long-run relationship between sentiment and price had a positive sign. Table 1 illustrates how, although there is sign-strength-timing variability present in the relationship, there are trends in the relationship which lead to long term correlations. This is somewhat similar to the situation described by the momentum literature, where theories, such as Hong and Stein (1999) model momentum as an oscillatory structure in the information-price relationship, but empirical work by authors like Moskowitz et al. (2012) evidence long run relationships that occur as a consequence of trends in the oscillatory structure.

	Returns		Abnormal Returns	
Model	β_q	$\% Adj\ R^2$	β_q	$\% Adj\ R^2$
m=20,h=10	0.1	0.99	-0.001	-0.01
t-statistic	(10.354)		(-0.083)	
m=20,h=20	0.151	2.27	-0.002	-0.01
t-statistic	(15.737)		(-0.176)	
m=20,h=40	-0.046	0.2	-0.037	0.13
t-statistic	(-4.718)		(-3.813)	
m=40,h=10	-0.023	0.04	-0.02	0.03
t-statistic	(-2.338)		(-2.07)	
m=40,h=20	-0.051	0.25	-0.045	0.19
t-statistic	(-5.265)		(-4.599)	
m=40,h=40	-0.174	3.01	-0.084	0.69
t-statistic	(-18.165)		(-8.655)	
m=60,h=10	-0.057	0.32	-0.047	0.22
t-statistic	(-5.918)		(-4.895)	
m=60,h=20	-0.093	0.86	-0.076	0.57
t-statistic	(-9.669)		(-7.868)	
m=60,h=40	-0.203	4.11	-0.112	1.26
t-statistic	(-21.333)		(-11.654)	
m=80,h=10	-0.064	0.39	-0.044	0.18
t-statistic	(-6.552)		(-4.522)	
m=80,h=20	-0.1	0.99	-0.077	0.59
t-statistic	(-10.359)		(-7.993)	
m=80,h=40	-0.163	2.65	-0.117	1.37
t-statistic	(-17.024)		(-12.174)	
m=100,h=10	-0.024	0.05	-0.04	0.15
t-statistic	(-2.521)		(-4.103)	
m=100,h=20	-0.032	0.09	-0.069	0.47
t-statistic	(-3.3)		(-7.148)	
m=100,h=40	-0.098	0.95	-0.114	1.29
t-statistic	(-10.121)		(-11.82)	

Table 1: Prediction and holding periods: The left-hand column gives the model specification, where *m* is the prediction window, and *h* is the holding period. Both numbers indicate the length of the window in trading days. The betas are given in standard deviations of the return, and the regressions are fitted with generalised least squares. In all cases, the sample size *n* = 10,695 stock trading days. Both the mean return and abnormal return are 0 to 5 decimal places, and the standard deviation of the raw return is 0.017 and 0.014 for abnormal returns. So for example, the expected log percent return for a standard deviation increases in sentiment due to the *m* = 100, *h* = 40 raw return, the model would be calculated as 100*(0.017*-0.098) = -0.167%. We see from the results that the strength of the abnormal returns relationship strengthens as both the prediction window and holding period increase. At the m=20 prediction window, we observe no relationship. For raw returns, we see that there is a positive relationship at *m* = 20 and *h* = 10 or *h* = 20. The relationship is then negative at other prediction windows and holding periods. The middle prediction windows *m* = 40 to *m* = 80 show stronger relationships, and the longest holding period *h* = 40 always shows the strongest relationship.

3.4.4. Comparison to time series momentum

Since we have made heavy use of existing momentum evidence and under reaction theories in the explanation of our results, it is natural to consider the possibility that textual sentiment is simply a proxy for existing forms of price momentum. To test this, we repeat the analysis from table 1, including lagged returns as an explanatory variable over the same prediction window, that is

$$H(r_{i,t}, h) = \alpha + \beta_q MA(q_{i,t-\tau}, m)' + \beta_{tsmom} MA(r_{i,t-\tau}, m)' + \varepsilon_t \qquad (27)$$

where the subscript *q* indicates the sentiment parameter, and *tsmom* indicates the time-series momentum parameter of the model. Table 2 reports the results from fitting equation (27) to our data for the same specifications of *h* and *m* as before and where τ is still 1 to ensure that the prediction window and the holding period do not overlap.

The results from table 2 evidence the fact that momentum and sentiment are different and are complementary predictors of price. Looking at raw returns, we see that all sentiment parameters are still significant. We also see that many of the lagged returns parameters are also significant and that the adjusted R^2 values for the regression in most cases are several times larger than using sentiment alone. The highest of these results being the $m = 100$, $h = 40$ model, where $R^2 = 14.37\%$.

We see largely the same story for abnormal returns. We see no relationship between the short prediction and short holding period models for sentiment, but there is now a significant relationship between sentiment at a 20-day prediction window and a holding period of 40 days. This relationship is heavily influenced by a positive relationship between the momentum factor and price at this model specification. We see that, when considered together, both sentiment and momentum are mostly negative predictors of prices. The higher R^2 values for these models again all come where the holding period is 40 trading days. The highest of these results is for the $m = 100$, $h = 40$ model, where the adjusted $R^2 = 11.58\%$.

A feature of the results reported in table 2 is that sentiment plays a greater role in the results of prediction windows of 40 to 80 trading days, and momentum becomes much more of a factor in the 100–trading day models. This is consistent with theory as the momentum literature typically concerns a longer-run relationship than the one we

document here. The result also suggests that one of the other sources of variability, not linked to investor psychology, may be the cause of some of the variability in the sentiment-price relationship.

	Returns			Abnormal Returns		
Model	β_q	β_{tsmom}	% $Adj\ R^2$	β_q	β_{tsmom}	% $Adj\ R^2$
m=20,h=10	0.101	-0.161	3.59	0	-0.032	0.08
t-statistic	(10.62)	(-16.935)		(0.017)	(-3.304)	
m=20,h=20	0.153	-0.214	6.86	-0.002	-0.007	-0.01
t-statistic	(16.292)	(-22.879)		(-0.155)	(-0.693)	
m=20,h=40	-0.045	-0.099	1.18	-0.038	0.029	0.2
t-statistic	(-4.664)	(-10.302)		(-3.902)	(2.942)	
m=40,h=10	0.006	-0.191	3.59	-0.019	-0.033	0.13
t-statistic	(0.623)	(-19.768)		(-1.906)	(-3.352)	
m=40,h=20	-0.027	-0.16	2.75	-0.044	-0.015	0.2
t-statistic	(-2.784)	(-16.518)		(-4.52)	(-1.525)	
m=40,h=40	-0.16	-0.09	3.79	-0.083	-0.025	0.75
t-statistic	(-16.624)	(-9.344)		(-8.522)	(-2.583)	
m=60,h=10	-0.04	-0.122	1.78	-0.044	-0.052	0.48
t-statistic	(-4.103)	(-12.576)		(-4.504)	(-5.38)	
m=60,h=20	-0.068	-0.179	4.01	-0.071	-0.071	1.06
t-statistic	(-7.057)	(-18.664)		(-7.346)	(-7.294)	
m=60,h=40	-0.175	-0.192	7.72	-0.103	-0.126	2.83
t-statistic	(-18.608)	(-20.402)		(-10.78)	(-13.126)	
m=80,h=10	-0.043	-0.176	3.45	-0.029	-0.112	1.41
t-statistic	(-4.425)	(-18.356)		(-3.019)	(-11.524)	
m=80,h=20	-0.069	-0.266	7.94	-0.056	-0.168	3.34
t-statistic	(-7.3)	(-28.303)		(-5.784)	(-17.395)	
m=80,h=40	-0.127	-0.306	11.85	-0.087	-0.235	6.78
t-statistic	(-13.806)	(-33.266)		(-9.202)	(-24.818)	
m=100,h=10	0.01	-0.247	6.03	-0.003	-0.189	3.57
t-statistic	(1.054)	(-25.982)		(-0.321)	(-19.431)	
m=100,h=20	0.016	-0.342	11.55	-0.021	-0.25	6.46
t-statistic	(1.705)	(-37.072)		(-2.167)	(-26.074)	
m=100,h=40	-0.046	-0.37	14.37	-0.051	-0.327	11.58
t-statistic	(-5.084)	(-40.769)		(-5.426)	(-35.128)	

Table 2: Comparison of sentiment and momentum: Similar to table 1, the left-hand column gives the model specification, where m is the prediction window, and h is the holding period. Both numbers indicate the length of the window in trading days. Both betas are given in standard deviations of the return, and the regressions are fitted with generalised least squares. In all cases, the sample size n = 10,695 stock trading days. We see from the t-statistics that sentiment relationships are complementary to momentum effect. We also see across the board that the adjusted R^2 values are strengthened by the inclusion of both variables. Particularly for abnormal returns, the sentiment-price relationship is significantly stronger than the momentum-price relationship. In the case of raw returns, the relationship, judged by the size of the respective t-statistics, is approximately equal apart from m = 100, where the relationship is dominated by momentum. Interestingly, this is not the case for the m = 100 model for abnormal returns, suggesting an asymmetry in the way stock-specific rather than more general sector or market information is processed by investors.

3.5. Returns to buying sentiment winners

A simple test of how economically meaningful this information is, is to see whether a simple trading rule can produce excess returns. Our results suggest the stronger sentiment-price relationships, particularly for abnormal returns, to be the ones where we have long prediction and holding periods and the lag length parameter $\tau = 1$. As a result, we will form portfolios of stocks based on models where $\tau = 1$. Given that we know that there is a stronger relationship at longer holding periods, we utilise a trading strategy developed by Jegadeesh and Titman (1993) in their early work on price momentum but adopted for use with a sentiment input signal rather than a lagged price input signal.

The strategy is as follows: at each time period, we sort the stock based on their j trading day average sentiment. The following trading day, we purchase an even-weighted portfolio of the top 20 or 30 stocks with the intention of holding them for g trading days. The following day, we purchase the next set of stocks according to the same criteria. At $t + g$ trading days, we then sell the first portfolio we purchased as the holding period expires. We continue along the same lines until the end of the series so that after we purchase the first g days' worth of portfolios, we are holding $20g$ or $30g$ open positions on any single trading day, although the same stock may appear in several of these open positions. We report returns as the average return between t and $t - g$ for all open positions we close on a given trading day so that we have a single number indicative of the rolling return we would realise from the strategy.

Table 3 shows the results of applying this strategy for sentiment averaging periods of 60, 80, and 100 trading days and holding periods of 10, 20, and 40 trading days. The returns are reported as log percent. Firstly, we see that the results clearly strengthen as we lengthen the holding period, echoing the results of our regression analysis. For any averaging period, we see that the best returns and highest t-statistics are for holding periods where $g = 40$. We also see that the results strengthen as we lengthen the averaging period, with the best results coming where $j = 100$. In cases where $j = 80$ and $j = 100$, most of the portfolios have t-statistics high enough to still be significant after applying the Bonferroni correction, adjusting for the fact we have run many tests. For example, the probability of achieving a t-statistic of > 3 after adjusting for the fact we show the results of 18 portfolios is 0.0243.

The highest result we see is for the 20 stock portfolios with a 100-day averaging

period and a 40-day holding period. For this portfolio, we see an excess return, calculated as the return of the portfolio minus the return from holding the market portfolio over the same period of 0.042%. To put this into perspective, an excess return of that size over a full year would equate to excess returns of 10.577% APR.

3.5.1 Transaction costs

Assessing these results in terms of the reasonable cost of pursuing these strategies, Clarkson et al. (2006) report the average transaction costs for trading stocks with an online broker is between 0.15% and 0.2% of the trade value. Taking the $j = 100$, $g = 40$ strategy where we buy the top 20 stocks as an example, at the close of each holding period, we sell 20 stocks each time period. The 40-day average return of a single stock over this time period is 1.692%, so subtracting the 0.2% transaction fees; this leaves a gross log profit from the trade of 1.492% per completed trade. Dividing by 40 leaves a gross log profit per day of 0.037% for pursuing this strategy, which is equivalent to an APR of 9.776% (calculated as $100(e^{1+(0.00037*250)} - 1)$) in excess of the market portfolio and a gross APR of 15.606% overall after reasonable costs have been deducted.

	20 Stock Portfolio		30 Stock Portfolio	
	Return	**Excess Return**	**Return**	**Excess Return**
	t-stat(market)	**t-stat(strategy)**	**t-stat(market)**	**t-stat(strategy)**
$j = 60, g = 10$	N=4060		N=6090	
Return	0.037	-0.019	0.041	-0.015
t-statistic	(-2.897)	(-1.34)	(-2.802)	(-0.904)
$j = 60, g = 20$	N=3860		N=5790	
Return	0.042	-0.004	0.043	-0.003
t-statistic	(-0.907)	(-1.12)	(-0.844)	(-1.102)
$j = 60, g = 40$	N=3460		N=5190	
Return	0.06	0.019	0.06	0.019
t-statistic	(6.287)	(2.633)	(7.498)	(3.123)
$j = 80, g = 10$	N=3660		N=5490	
Return	0.054	0.014	0.048	0.008
t-statistic	(1.978)	(1.7)	(1.354)	(1.015)
$j = 80, g = 20$	N=3460		N=5190	
Return	0.057	0.016	0.056	0.015
t-statistic	(3.5)	(2.824)	(3.927)	(3.114)
$j = 80, g = 40$	N=3060		N=4590	
Return	0.06	0.03	0.063	0.033
t-statistic	(8.701)	(4.551)	(11.684)	(6.669)
$j = 100, g = 10$	N=3260		N=4890	
Return	0.049	0.016	0.054	0.021
t-statistic	(2.239)	(2.633)	(3.464)	(3.94)
$j = 100, g = 20$	N=3060		N=4590	
Return	0.052	0.022	0.05	0.02
t-statistic	(4.32)	(4.208)	(4.793)	(4.656)
$j = 100, g = 40$	N=2660		N=3990	
Return	0.063	0.042	0.058	0.038
t-statistic	(11.338)	(7.963)	(12.073)	(8.05)

Table 3: Returns to buying sentiment winners: The left-hand column denotes the portfolio. We report results for 20 and 30 stock portfolios. Returns are given as daily log percent average returns for the whole portfolio. The excess return is defined as the return minus the return of holding an even weighted portfolio of the 93 underlying stocks. We calculate the t-statistic(market) as a comparison of the portfolio return per stock per day versus the market return per stock per day. The t-statistic(strategy) is calculated as the return per stock per day versus the return of holding the market portfolio using the same rolling investment strategy, thus buying 93 stocks each day and holding them for G days. We note that several of the t-statistics would be robust after applying the Bonferroni correction for the fact we have analysed 18 portfolios. For example, the probability of achieving a t-statistics of greater than three after applying the Bonferroni correction is $p = 0.0243$.

3.6. Conclusions, limitations, and future work

In the introduction we argued that there were both empirical and theoretical reasons for expecting variability in the sign, strength and timing of the sentiment-price relationship. We use the largest dataset gathered for a study of this type to document how this type of variability is indeed a key feature of the relationship. Importantly, we also show the strongest evidence available in the literature to date that online textual sentiment can be used to forecast future prices and that this relationship is strong enough to have genuine economic meaning for investors. We build on the evidence documented by Chen et al. (2014), by showing that there is a long run relationship between sentiment and price. We extend this finding by showing that the relationship takes the form of an oscillating wave, as the price moves away from a given point in time on the sentiment series.

Because of this feature, we draw a link to the behavioural finance literature and empirical work on price momentum. We use multivariate regression analysis to show that, although the price momentum is similar to the sentiment-price relationship, sentiment contains incremental information over that which is already recorded in the serial correlation of prices. This suggests that some of the sign-strength-timing variability in the sentiment price relationship may be due to factors, such as the trust relationship between content authors and investors, which are difficult or impossible to measure effectively.

In terms of future work, a methodological consideration we raise is that we have made extensive use of linear regression models. Given the different sources of variability in the relationship we outline in section 3.1, and our empirical observations of this variability we present in our results, this leads to quite a crude analysis as we have to make decisions about the different lag lengths as well as prediction and holding periods to use for each of these models. The problem is that there does not seem to be any theoretical reason for the choice of prediction window or holding periods. Ideally, we would like to be able to identify the best model so that we could consider all the possible lag lengths and prediction and holding periods at once. Also, our analysis considers the parameters for these models to be fixed. There appears to be no theoretical reason why the lag lengths and prediction and holding periods in these models would be static over time. We suggest a major area of further investigation for both the textual sentiment and momentum literature to develop techniques for fitting models more suitable for the types of relationship we observe.

Appendix

Ticker	Query String	Messages	Return	σ Return	Sentiment	σ Sentiment
AAPL	(AAPL AND NASDAQ) OR "Apple Inc."	694157	42.72	1.45	-23.26	0.78
MO	(MO AND NYSE) OR "Altria Group"	63307	40.56	0.9	145.5	0.87
GD	(GD AND NYSE) OR "General Dynamics"	34101	38.9	1.15	20.17	0.86
UNH	(UNH AND NYSE) OR "United Health Group Inc."	37977	38.5	1.25	61.99	0.94
UNP	(UNP AND NYSE) OR "Union Pacific"	20005	37.21	1.27	0.85	0.69
CVS	(CVS AND NYSE) OR "CVS Caremark"	45246	37.06	0.92	102.29	0.91
LLY	(LLY AND NYSE) OR "Lilly (Eli) & Co."	34880	36.71	1.09	4.12	0.86
SPG	(SPG AND NYSE) OR "Simon Property Group"	28223	36.57	0.84	356.93	1.39
LOW	Lowe's	25083	36.27	1.24	121.21	1.04
GILD	(GILD AND NASDAQ) OR "Gilead Sciences"	59548	35.42	2.1	-11.74	0.74
BIIB	(BIIB AND NASDAQ) OR "BIOGEN IDEC Inc."	48708	34.01	2.31	54.95	0.81
HPQ	(HPQ AND NYSE) OR "Hewlett-Packard"	67812	32.94	1.54	2.25	0.67
EXC	(EXC AND NYSE) OR "Exelon Corp."	28150	32.02	1.27	91.37	0.95
ALL	Allstate Corp	24885	31.26	0.9	68.28	1.13
LMT	(LMT AND NYSE) OR "Lockheed Martin"	58504	30.55	1.09	44.9	0.77
INTC	(INTC AND NASDAQ) OR "Intel Corp."	183908	29.87	1.41	26.7	0.79
WBA	(WBA AND NASDAQ) OR "Walgreens Boots Alliance"	19680	29.8	1.79	64.32	0.96
AMGN	(AMGN AND NASDAQ) OR "Amgen"	61224	29.27	1.56	-7.33	0.82
HD	(HD AND NYSE) OR "Home Depot"	53998	28.94	1.13	-53.72	0.96
SO	(SO AND NYSE) OR "Southern Company" OR " Southern Co"	19071	28.64	0.85	100.63	1.09
MDT	(MDT AND NYSE) OR "Medtronic"	48944	28.59	1.21	-3.41	0.78
COST	Costco	28083	27.76	0.97	-28.74	0.49
TXN	(TXN AND NASDAQ) OR "Texas Instruments"	39445	26.66	1.27	157.79	0.91
CSCO	(CSCO AND NASDAQ) OR "Cisco Systems"	69137	23.97	1.1	77.09	0.73
MRK	(MRK AND NYSE) OR "Merck & Company" OR " Merck & Co"	73582	23.96	1.22	-52.91	0.64
DIS	(DIS AND NYSE) OR "The Walt Disney Company"	70870	22.27	1.11	69.07	0.51
BRK.B	(BRK.B AND NYSE) OR "Berkshire Hathaway"	49175	22.22	0.88	-4.52	0.67
FDX	(FDX AND NYSE) OR "FedEx Corporation"	39200	21.81	1.19	-5.3	0.8
TGT	(TGT AND NYSE) OR "Target Corp."	24193	21.47	1.3	-39.48	0.88
ABBV	(ABBV AND NYSE) OR "AbbVie"	40923	20.99	1.59	-44.59	0.74
MMM	(MMM AND NYSE) OR "3M Company"	12879	20.73	0.97	-9.27	0.73
RTN	(RTN AND NYSE) OR "Raytheon Co."	33505	20.49	1.22	49.18	0.89
WFC	(WFC AND NYSE) OR "Wells Fargo"	193937	19.92	0.95	-83.17	0.77
ABT	(ABT AND NYSE) OR "Abbott Laboratories"	26751	19.55	1	-14.63	0.68
TWX	(TWX AND NYSE) OR "Time Warner Inc."	105619	19.37	1.79	-21.55	0.61
NKE	(NKE AND NYSE) OR "Nike"	59276	18.86	1.32	13.11	0.63
PEP	(PEP AND NYSE) OR "PepsiCo Inc."	51080	18.67	0.85	97.63	0.72
BMY	(BMY AND NYSE) OR "Bristol-Myers Squibb"	31634	16.95	1.44	14.35	0.76
V	Visa Inc	98614	16.92	1.33	-27.75	0.73
DD	(DD AND NYSE) OR "Du Pont (E.I.)"	57664	16.82	1.06	32.18	0.78
SBUX	(SBUX AND NASDAQ) OR "Starbucks Corp."	63171	15.55	1.21	10.85	0.63
NSC	(NSC AND NYSE) OR "Norfolk Southern Corp."	19156	15.49	1.35	71.55	0.93
JNJ	(JNJ AND NYSE) OR "Johnson & Johnson" OR " Johnson and Johnson"	65997	14.56	0.95	22.17	0.78
ORCL	(ORCL AND NYSE) OR "Oracle Corp."	98175	14.37	1.26	95.58	0.79
MSFT	(MSFT AND NASDAQ) OR "Microsoft"	410677	14.01	1.38	52.72	0.83
HON	(HON AND NYSE) OR "Honeywell Int'l Inc."	39515	11.93	1.03	112.5	0.87
WMT	(WMT AND NYSE) OR "Wal-Mart Stores"	98972	11.3	0.9	-42.81	0.72
ACN	(ACN AND NYSE) OR "Accenture"	41030	10.58	1.07	229.09	0.99
PG	(PG AND NYSE) OR "Procter & Gamble"	62884	10.32	0.78	54.05	0.7

BK	(BK AND NYSE) OR "The Bank of New York Mellon Corp."	62018	10.24	1.19	-96.34	0.86
CL	(CL AND NYSE) OR "Colgate-Palmolive"	18299	9.99	0.91	-5.25	0.75
BA	(BA AND NYSE) OR "Boeing Company"	193094	9.82	1.28	51.48	0.72
APC	(APC AND NYSE) OR "Anadarko "	29781	9.64	2.1	-19.18	0.91
CMCSA	(CMCSA AND NASDAQ) OR "Comcast Corp."	117113	9.1	1.2	34.15	0.82
EMC	(EMC AND NYSE) OR "EMC Corp."	43541	8.76	1.18	54.76	0.91
PFE	(PFE AND NYSE) OR "Pfizer Inc."	94617	8.44	1.02	-50.81	0.68
USB	(USB AND NYSE) OR "U.S. Bancorp"	23469	8.34	0.95	-30.19	0.88
UTX	(UTX AND NYSE) OR "United Technologies"	50835	7.66	0.98	53.43	0.9
DOW	Dow Chemical	28486	6.09	1.57	56.26	0.75
DVN	(DVN AND NYSE) OR "Devon Energy Corp."	12436	5.63	1.86	41.25	0.75
F	Ford Motor" OR " Ford	203008	5.52	1.37	-3.93	0.75
MON	(MON AND NYSE) OR "Monsanto Co."	46712	5.39	1.05	-120.46	0.79
KO	(KO AND NYSE) OR "The Coca Cola Company"	95534	5.3	0.95	66.04	0.58
T	AT&T" OR " AT and T	227148	4.86	0.9	150.31	0.86
MDLZ	(MDLZ AND NASDAQ) OR "Mondelez International"	22409	3.34	1.1	66.95	0.93
GS	(GS AND NYSE) OR "Goldman Sachs"	332367	3.18	1.16	-143.84	0.74
VZ	(VZ AND NYSE) OR "Verizon Communications"	167099	3.12	0.95	61.47	0.83
AIG	(AIG AND NYSE) OR "American International Group"	40286	2.15	1.13	-164.42	0.71
MA	(MA AND NYSE) OR "Mastercard"	61868	1.25	1.51	90.37	0.94
COP	(COP AND NYSE) OR "ConocoPhillips"	36417	0.81	1.44	-36.7	0.7
MCD	(MCD AND NYSE) OR "McDonald's Corp."	71680	0.78	0.86	-38.11	0.75
EBAY	(EBAY AND NASDAQ) OR "eBay Inc."	113884	0.57	1.43	27.31	0.76
JPM	(JPM AND NYSE) OR "JPMorgan Chase & Co."	228061	0.05	1.22	-111.12	0.33
SLB	(SLB AND NYSE) OR "Schlumberger Ltd."	28547	-0.32	1.55	-9.03	0.67
BAC	(BAC AND NYSE) OR "Bank of America"	238181	-0.56	1.46	-115.4	0.74
UPS	(UPS AND NYSE) OR "United Parcel Service"	15660	-0.59	1.08	-23.55	0.78
MET	(MET AND NYSE) OR "MetLife Inc."	23542	-1.65	1.39	8.06	0.91
COF	(COF AND NYSE) OR "Capital One Financial"	24620	-2.16	1.11	-29.23	0.9
FOXA	(FOXA AND NASDAQ) OR "Twenty-First Century Fox" OR " Twenty Firs	21654	-2.53	1.44	-18.02	0.83
CAT	(CAT AND NYSE) OR "Caterpillar Inc"	69169	-3.41	1.37	-46.05	0.75
XOM	(XOM AND NYSE) OR "Exxon Mobil"	73226	-5.02	1.1	-60.86	0.74
AXP	(AXP AND NYSE) OR "American Express Co"	57402	-5.23	1.17	45.74	0.83
OXY	(OXY AND NYSE) OR "Occidental Petroleum"	15320	-6.14	1.44	-112.99	0.91
C	Citigroup	227549	-6.61	1.36	-186.72	0.72
QCOM	(QCOM AND NASDAQ) OR "QUALCOMM Inc."	74712	-6.95	1.45	83.15	0.92
GE	(GE AND NYSE) OR "General Electric"	148786	-8.19	0.96	92.15	0.7
CVX	(CVX AND NYSE) OR "Chevron Corp."	77541	-9.02	1.24	-77.73	0.6
EMR	(EMR AND NYSE) OR "Emerson Electric Company"	15285	-12.63	1.16	133.65	1.01
HAL	(HAL AND NYSE) OR "Halliburton Co."	53379	-13.19	2.04	-61	0.96
IBM	(IBM AND NYSE) OR "International Business Machines"	170544	-13.36	1.13	105.02	0.83
GM	(GM AND NYSE) OR "General Motors"	132773	-15.18	1.51	-54.86	0.9
APA	(APA AND NYSE) OR "Apache Corporation"	32175	-21.26	1.87	11.59	0.82
FCX	(FCX AND NYSE) OR "Freeport-McMoran" OR " Freeport McMoran"	17170	-71.83	2.02	-127.54	0.63

Appendix: Query Strings and Descriptive Statistics. The table shows the ticker symbol, query string and some descriptive statistics for the data in our sample. The statistics are for the cumulative return and cumulative sentiment over the sample period of 273 trading days. We also show message numbers, and standard deviations for either series. The correlation coefficient between cumulative sentiment and returns is 0.33, indicating that sentiment does provide some useful information in differentiating over the cross-section of stocks.

Chapter 4: Signal Diffusion Mapping: Optimal Forecasting with Time Varying Lags

4.1. Introduction

Conventional time series methodologies for financial forecasting are limited in their ability to handle lags. They can only accommodate lags fixed integer length, e.g. a one-period lag, a two-period lag. We argue that the relationship between real time series can often have a richer, more dynamic time structure than has been implicitly assumed in the traditional methodological approaches. We introduce a new forecasting methodology, signal diffusion mapping (SDM), which extracts the maximum possible information from the modelled relationship and produces the best possible forecast at each point in time, in circumstances where the lag-length is time-varying.

The notion of time-varying lag-lengths may strike some researchers as odd. Indeed, we suspect that the fact that lags have been almost invariably modelled as fixed-integer-lags in time-series financial modelling has probably conditioned most financial forecasters to believe that financial data actually behaves in the manner that their models seek to measure. However, we contend that such a conclusion would be a classic case of ‘if all you have is a hammer, everything looks like a nail’.

Real-world temporal relationships are less well ordered than we might like to think. The history of financial markets is replete with colourful examples of convicted insider traders who exploited access to privileged information for pecuniary gain. This can include information about companies' earnings announcements, future takeover targets, price-sensitive macroeconomic news, etc. An insider trader trading prior to a news event will reverse the time-order that financial theory conventionally assumes, i.e. information first, price- reaction second. On the other hand, nobody would accept the assertion that insider-trading occurs prior to each and every information event. In other words, a reasonable conclusion would be that occasional cheating happens and this distorts the information-price time lag. If a fixed-integer lag model were imposed on such a scenario, it would result in the measured correspondence between information and price being lower than if

the modeller had a means of adjusting for the time distortion.

Insider trading is not the only source of lag-length distortion. Cheung et al. (2004) survey the opinions of foreign exchange dealers and conclude that ‘news’, ‘speculative forces’ and ‘bandwagon effects’ are the main drivers of price within a day and that fundamentals, which mainstream finance theory assumes to be the key drivers of price, are perceived as only relevant in determining returns over the long term (i.e. over 6 months). These speculative forces and bandwagon effects are perhaps best captured in Soros' reflexivity theory which posits a bi-directional symbiotic feedback relationship between information and price (Soros 2009). Importantly, Soros argues that some reflexivity relationships are sustained for extended periods, while others can fizzle out after a short time. Trying to extract the relationship between such variables with a fixed-length-lag model is like trying to eat soup with a fork – quite simply the wrong tool for the job.

The key contribution of this paper is the introduction of a new model (SDM) to the forecasters' arsenal which can handle time-varying lag-lengths of the type we describe above, retrieving the maximum possible information about the relationship between two time series. Our method builds upon well-established, published methods from the data analysis and econophysics literature, which were developed to measure optimal relationships in related series of fixed lengths, e.g. Keogh and Pazzani (1999), Sornette and Zhou (2005) and Zhou and Sornette (2007). However, in their original form none of those methods were appropriate for forecasting. We adapt those methods to the problem of forecasting in a manner analogous to a Kalman filter or recursive Bayes estimation with time-varying-lags, which we name signal diffusion mapping. We test SDM with synthetic data and we demonstrate that it is able to recover the true optimal relationship between series where we have deliberately distorted the temporal relationship by shortening/lengthening the lag at different points in time.

Note that SDM does not rule out the possibility that the true relationship between two time series could actually be a fixed length lag. An important feature of our proposed methodology is that it will identify the optimal underlying relationship between the two series whether the lag structure is fixed of varying in time. In other words, if the best possible relationship is obtained by lagging one of the series by two time-periods, then that is what the model will find. On the other hand, if an even better link between the two series could be shown by allowing series 1 to lead series 2 some of the time and to lag series 2 at other times, our method will identify this as the best model.

The remainder of our paper is structured as follows: section 4.2 introduces some concepts and notations from the relevant literatures we will use in defining our approach. Section 4.3 will describe a Bayesian interpretation of the time-sequencing problem. In section 4.4 we present the SDM algorithm in full. In section 4.5, we present results from testing the SDM algorithm on simulated data. Finally, section 4.6 presents our conclusions.

4.2. Background

In the introduction we highlight the work of Soros (2009) and Cheung et al. (2004) as examples of work which highlights the temporal complexity of the information-price relationship. In the case of Soros, this complexity is captured as a constantly shifting feedback mechanism, capable of rapid changes; for example, periods of constant growth lasting years can shift to huge market crashes in the space of less than a week. In the case of Cheung *et al.* the emphasis is on scale, where short term information effects (intra-day) are augmented with long term effects of 6 months or more.

As described, this implies the information-price relationship has a complex causality structure. Statistical causality in the financial literature is mostly considered in terms of Granger causality (Granger 1969), so that where the time-sequencing of a relationship is unclear, Granger causality tests are employed as the main tool in determining the nature of the statistical causality.

In section 4.2.1, we introduce the Granger causality model and cases where it has been applied in the study of information-price relationship. We also discuss some other approaches to determining statistical causality. We then to discuss the limitations of these models in terms of the dynamism and scaling of the relationship we wish to measure. In section 4.2.2, we discuss how the Bayesian framework is a good fit for problems of this type. We briefly highlight other work where Bayes Estimators have been used for model fitting in a financial context and draw the link to other state space techniques used in time-series analysis.

4.2.1. Time-sequencing in financial forecasting

To give the issue of variable lag-lengths a mathematical framing, the standard relationship

that is considered in the Granger causality approach is that some autoregressive series of the form

$$x_t = \alpha x_{t-1} + N(0,1)$$

where α is the autoregressive parameter, $|\alpha| < 1$ so the process is stationary, and $N(0,1)$ indicates a normally distributed random variable with a mean of 0 and variance of 1. We then consider another variable related to **x** as;

$$y_t = \beta_1 x_{t-1} + \beta_2 x_{t-2} \ldots \beta_\tau x_{t-\tau} + u_t \tag{28}$$

where

$$u_t = N(0, \sigma_u^2)$$

is the noise term obfuscating the relationship.

Granger causality in the bi-variate case is simply a regression model specified on the lagged values of **x**. Throughout this paper bold-face type will denote a vector in the usual manner, where **β** are the coefficients values or parameters of the model; in some cases we might also consider lags of **y** but here we omit them for brevity. The parameters of the model are then estimated with the ordinary least squares (OLS) algorithm. If any of the **β** coefficients are statistically significant, researchers can then state that there is evidence to support the hypothesis that **x** Granger causes **y**.

This is by far the most common form of analysis of the lagged dependence between time-series, example studies of information price relationships using this methodology include; (Bollen et al. (2011) and Sprenger et al. (2014), both studying the effect of public sentiment on stock prices; Hong et al. (2009), modelling risk spillover between international financial markets; and Hiemstra and Jones (1994), studying the causality between stock prices and volume using the Granger causality test specified in equation (28) and a non-linear variant.

Even in cases where more complex methodologies are applied, usually the treatment of the temporal relationship between the variables reduces to some variant of Granger causality. An example of this is Sharkasi et al. (2005), who use wavelet decomposition to study spillover effects from international stock indices, then use linear regression to fit models on the decomposed wavelets.

One of the attractions of the Granger causal approach is that it gives a clear pathway to constructing a forecasting model. Based on the results of equation (28) it is

possible to identify lags with significant *p*-values then construct a second regression using these lags. This second model can then be used to determine estimates of future values of **y** based on the coefficient values in the second regression model.

The situation Soros (2009) and Cheung et al. (2004) describe, however, is one of significant variation in the temporal relationship between variables, both in terms of the speed by which the relationship can change but also the fact that the relationship can exist at varied scales. In Soros's description of reflexivity he states how financial markets can switch from bull to bear periods very quickly; these switching moments are characterised by changes in the information-price relationship that happen over the course of days - suggesting any model would have to adjust almost immediately to the change in circumstance. Cheung *et al.* report how professional traders consider information effects at different scales, from intra-day effects to long run effects, > 6 months.

Now consider trying to capture this with the Granger causal approach, firstly tackling scale: Cheung et al.'s conclusions indicate we would need to fit a model with a large number of parameters to capture each different scale considered. Considering daily intervals, for example, this would lead to a model with ~180 parameters, leading to very low statistical power for the test due to the 'curse of dimensionality'. Secondly, assuming the parameters of equation (28) are time-varying, we could employ a version of equation (28) on some rolling window. In order to achieve reasonable results, however, requires specifying a reasonable sized window – yet theory suggests that change in the relationship will occur almost instantly.

4.2.2. Parameter estimation using Bayesian inference

To be specific, the issue we have is that given the Granger causal model we believe the model specification would have too many parameters - and that these parameters would change too quickly for us to be able to fit the model we need to test the available financial theory with the OLS algorithm. We are not arguing that the basic structure of the Granger causal model is deficient in its ability to characterise the information-price relationship per se - just that we cannot fit the model in the way we would like.

Recently, there have been a number of papers in the finance and econometrics literatures using various types of recursive Bayes estimator (RBE) to estimate model

parameters. Examples of this work include, Carvalho and Lopes (2007), who present an RBE for dynamically parameterised stochastic volatility models, and Carvalho et al. (2009), who fit the parameters of a dynamic, conditionally linear model (see Arulampalam et al. (2002) and Lopes and Tsay (2011) for reviews of the filed). As yet, however, no work exists attempting to fit causality models in a similar fashion.

From a Bayesian perspective the parameter values for the lags in equation (28) can be considered as a state space, where the parameters are represented as probabilities that a given lag could be causally influencing the value of y_t. Simplistically, this means that we can update the parameters based on their relative probabilities as they transition from one time-period to another, rather than on their goodness-of-fit over a large number of previous observations. We will show how this can allow the lag-structure to vary much more dynamically than it could do if fitted with OLS and circumvents issues of dimensionality.

This state space representation also links to other areas of the time-series analysis literature where variable lag paths have been considered. The dynamic time warping (DTW; see, for example, Keogh and Pazzani 1999; Senin 2008; Warren Liao 2005; Sakurai et al. 2005) and optimal thermal causal path (OTCP; Sornette and Zhou 2005; Zhou and Sornette 2007) literature both study historical lagged relationships using similar state space techniques. Much of section 4.4.2 concerns integrating these ideas into the RBE framework, so that we can forecast with variable lag-structures rather than view them historically.

The key contribution we make in this paper, is to show it is possible to dynamically fit the model in equation (28) using a RBE; we will begin by fitting the simple linear model and expand to more exotic cases in later sections. The specific RBE algorithm we present to complete this task is the signal diffusion mapping (SDM) algorithm. The rationale for the name is because we will map the diffusion gradient of information flowing between the two series over the state space of possible lags; this results in us being able to plot the time evolution of this relationship on a heatmap, i.e. the map of the flow of the signal between the series.

In section 4.3 we will further define the time-sequencing problem in Bayesian terms and introduce the notational conventions we will adhere to throughout this paper.

4.3. A Bayesian view of the time-sequencing problem

In this section we are going to interpret the Granger causal model in terms of the common general dynamical model (GDM), this a more general form of the normal linear dynamical model which forms the basis of the Kalman filter (Lopes and Tsay 2011). We then outline the general method of solving the GDM equations recursively using Bayes' theorem. Finally, we describe the criteria by which an estimator can be seen as optimal, which will leave a clear pathway to introducing SDM as the optimal solution to these equations.

The Bayesian approach to statistical decision making is based around the definition of relative beliefs about a set of different events. These beliefs are represented as a state space of probabilities associated with each possible event; given equation (28), our beliefs are about a series of discrete causal relationships between **x** and **y**. We are going to hold this set of beliefs in a probability vector containing an entry for each of the lagged values under consideration; we denote these options as an N_s length probability vector $\mathbf{w}_t$ where N_w is the number of considered lags (read 'number of states') at *t*, that is

$$\mathbf{w}_t : [w_t^1 \dots w_t^{N_s}]$$

where

$$\sum_{i=1}^{i=N_s} w_t^i = 1$$

The convention we adhere to throughout this paper is that subscripts will denote a vectors position in time and superscripts denote the relative position of a value in the vector, so w_t^i, would be read the *i*th position on vector **w** at time *t*. In terms of equation (28), these probability weights are conceptually similar to the parameter values **β** – we differentiate them in the notation to save confusion as they are probability weights rather than regression coefficients and will further define this difference in section 4.4.2.

The second aspect of the Bayesian approach is to then define a measure of how well these beliefs map on to some measurements of the processes under study. Let $\mathbf{d}_t$ be an N_s length measurement vector $\mathbf{d}_t : [d_t^1 \dots d_t^{N_s}]$ containing some measure of the **x**, **y** relationship corresponding to each of the values of $\mathbf{w}_t$. For now the reader simply has to understand this as a comparative measure of the relationship, we will make this definition concrete in the following section.

We note that it is typical in the RBE literature to use **x** for the state vector and **y** for the measurement vector. We have differed from this notation because in the econometrics literature **x** and **y** are typically the time-series under study. As we believe SDM is mainly aimed at the financial forecasting community, we have sided with the econometric notational convention.

The GDM is then given by two equations: the first, usually referred to as the *system model*, governs the way in which our beliefs about the system propagate forward through time; this is defined in probabilistic terms as

$$\mathbf{w}_t \propto Pr(\mathbf{w}_t \mid \mathbf{w}_{t-1}) \tag{29}$$

where the operator should be read as 'varies in proportion to'; so the equation states that the probability densities of $\mathbf{w}_t$ vary proportionally to a probability mass function applied to the densities at *t* - 1. The second, usually called the *measurement model*, governs the way we are going to interpret the measurement vector in terms of the state probabilities:

$$\mathbf{d}_t \propto Pr(\mathbf{d}_t \mid \mathbf{w}_t) \tag{30}$$

which should be read as 'the measurements vary proportionally to the likelihood of the measurement given the state vector probabilities'.

If the system model can be characterised as a Markov chain, i.e. the values of **w** at *t* + 1 depend only on the values at *t*, then we can define a Bayesian prediction model for the updated values of $\mathbf{w}_t$ as

$$w_{t|t-1}^{i} = \sum_{j=1}^{j=N_s} (w_t^i \mid w_{t-1}^j)\, w_{t-1}^j \tag{31}$$

where $w_{t|t-1}^{i}$ denotes the estimate of the probability density of the *i*th position on the probability vector **w** given the prior densities and the system model. All that equation (31) really states is that, if we have a system model holding the transition probability of the *i*th lag holding useful information given the preceding vector of lags, then we can sum over these probabilities to estimate the next *t*'s value for the lag. It is common in the literature to see equation (31) written as an integral; the reason for the summation in our case is that the state space over the lags is discrete rather than continuous.

Given an estimation of the weightings of **w** based on the known densities at *t* - 1 and the system model, we then receive a set of measurements $\mathbf{d}_t$ that we use to update this

forward projection of the densities based on the observed evidence. Given the measurement model (30) we can write this due to a combination of Bayes' theorem and the law of total probability as

$$w_t^i = \frac{Pr(d_t^i \mid \mathbf{w}_t) w_{t|t-1}^i}{\sum_{j=1}^{j=N_s} Pr(d_t^j \mid \mathbf{w}_t) w_{t|t-1}^j} \tag{32}$$

which gives the likelihood of the measurement given the data, multiplied by the prior probability if the *i*th lag normalised over all N_s possible lags.

This yields the basic prediction and update structure of a *grid-based filter*, which is a type of RBE defined on a discrete set of possible states – in this case lags - and forms the basis of a number of RBE algorithms, two of the best known being the bootstrap filter (Gordon et al. 1993) and the Auxiliary Particle Filter (Pitt and Shephard 1999). The attraction of this formulation is that, providing we can valid functional forms for equations (29) and (30), the filter will evolve to the optimal calculation of the probability densities of **w** over repeated iterations of equations (31) and (32) (Arulampalam et al. 2002; Lopes and Tsay 2011).

Optimality in this sense means the probability densities which minimise the measurement error. Thus we need a measurement model which is consistent with reasonable assumptions about the **x**, **y** relationship. We also need to specify a probability density function for the values of these measurements, since equation (32) requires that we calculate the likelihoods for these observations given our beliefs about lag probabilities. Finally, we require a probability mass function for equation (30) which captures the time evolution of the lag structure in a theoretically justifiable way.

If these criteria are met, we can claim the estimator is optimal under the specified assumptions. The trade-off in defining the estimator is then how to posit relaxed enough assumptions about the forms of equations (29) and (30) to make the estimator generally applicable to a range of forecasting tasks.

What we are going to show in introducing SDM is that we can specify the form of equation (29) using very weak assumptions. These assumptions are well supported by other areas of the literature. This effectively removes time-varying lags as an issue and allows us to plug in any relevant measurement model available in the econometrics literature. This makes the SDM algorithm useful in a large number of practical

applications, since the researcher can still utilise all of the current modelling approaches for bivariate series and plug in the SDM estimator as the time-sequencing test supplementing the specified model.

4.4. Signal diffusion mapping

In the following subsections we will being by describing the system model we are going to use for equation (29). This is the key operation in the SDM algorithm and remains invariant despite the specification of the measurement model. We will then define a measurement model to fit the simple Granger causal model (28). After illustrating this simple case we will then go on to describe more exotic variants of the SDM algorithm, i.e. bi-directional causality structures and positive-negative switching causality structures. Finally, we will give an algorithmic implementation of the SDM algorithm and discuss implementation issues and computational complexity.

4.4.1. System model

To frame the description that follows consider again the Granger causal model; each lag in the model is considered independent of the others, and models are fitted to ascertain if there is something peculiar to a particular lag that characterises the relationship between variables. In the real world people are not likely to consider time-periods this way. As an example, a common investment strategy 'discovered' by Jegadeesh and Titman (1993) is to buy portfolios of winners: stocks that show high positive returns over a 3-12 month prior period. These portfolios on average produce higher returns than the market.

An investor following such a strategy must decide on how long to wait before deciding a stock has momentum. In the literature this is assumed to be between 3 and 12 months, leaving quite a lot of room for manoeuvre. Over time, our investor will likely review their strategy and take decisions as to whether they should wait longer of shorter amounts of time during the next investment period. It is likely at different points that some adjustments will be made to the waiting period. In the context of the Granger causal model, the information contained in these adjustments is lost since any relationship between time-periods has already been assumed away.

We take the position here that a better way to model this situation is to assume the lag-lengths are interdependent, but the temporal variation in the lag structure as a whole is finite. Returning to our example, if we knew the time horizon our momentum investor was using to place trades at *t*, then we might infer that at $t + 1$ if they altered their investment strategy then it wouldn't be by much, i.e. 27-28 days, for example, with the size of an expected change being proportional to the amount of time that has passed since *t*. It is with this assumptions in mind that we specify the system model, in an effort to allow the modelling of the temporal relationship between variables, which is (more) consistent with the way we expect people to alter decisions about the timing of their actions.

Purely from a forecasting perspective, clearly we would like to be able to define a model where the densities over the lags do not change over time. In this case we would be able to estimate when the relationship was likely to occur perfectly. Another way of stating this would be that the error in the forward projection of our beliefs in the systems state is 0 and $\mathbf{w}_t = \mathbf{w}_{t-1}$.

If there is variation in our beliefs over time then we need to posit an equation describing this variation. Here, there are two important quantities: *firstly*, the structure this variation takes – i.e. how this variation deforms the densities of $\mathbf{w}$. *Secondly*, the magnitude of the variation – i.e. how much deformation in the density vector we expect. After defining this function we then need to find the parameterisation of the function which minimises the total error for the system model over time.

In terms of the structure of the variation, there is already a large body of literature dealing with variations in historical lag structures. We are going to follow these literatures in defining the structure of the temporal variation similarly – just as a probabilistic forward projection of the state vector probabilities, rather than a historical representation.

Both the DTW (see, for example, Keogh and Pazzani 1999; Senin 2008; Warren Liao 2005; Sakurai et al. 2005) and OTCP (Sornette and Zhou 2005; Zhou and Sornette 2007) literatures use state space methods to study historical lead-lag relationship between time-series. The basic premise of either algorithm is to define a matrix of all of the pairwise relationships between the variables on a matrix containing a measure of the relationship in each square. Then traverse the matrix from a fixed start to fixed end point in such a way as to reveal the optimal or lowest cost relationship between the variables based on the measure.

In order to do this, both algorithms make the assumption that the structure of the lagged relationship can vary only slowly in time. Slow varying in this sense means that a lag path can vary by only 1 unit time period for every unit increase in *t*. Taking DTW as an example (although both algorithms are derived from the same premise), given two time-series **c**: $[c_1 \dots c_k]$ and **q**: $[q_1 \dots q_k]$, we can construct a *kk* matrix of the pairwise distances, i.e. $|c_i - q_j|$, between the two series for each lag length. The DTW algorithm then seeks the lowest distance, continuous 1 to 1 mapping between the two series from a fixed start point to a fixed end point.

This task is completed by understanding that the lowest distance path to a point on the matrix must be the summation over the lowest distance pathways up to this point – which justifies the recursive relation

$$\varepsilon_{i,j} = |c_i - q_j| + MIN(\varepsilon_{i-1,j}, \varepsilon_{i-1,j-1}, \varepsilon_{i,j-1}) \tag{33}$$

where $\varepsilon_{i,j}$ is interpreted as the relative cost of the pathway, so that the minimum distance pathway is the lowest cost pathway up to a given point on the matrix.

Defined in this way the lag path has some desirable properties we would expect in a realistic lagged relationship:

(i) A relationship is defined for each *t* – there are no large jumps where no relationship is defined for a set of time-periods.

(ii) The relationship is a continuous 1:1 mapping – if this was not the case there could be overhangs or cliffs in the lag path implying the causal relationship ran from the future into the past. (Sornette and Zhou, 2005)

Figure 5 demonstrates the permissible variations in the lagged relationship due to equation (33). Consider these properties in the context of the momentum trader example, because the paths over the matrix are constrained to move only 1 unit *t* per jump; the variation in the average position of all of the paths between time *t* and *t* + *n* is finite and increasing with *t*. We might think of the paths as individual investment strategies being played out by different investors in the same asset; a lag length where many paths have congregated is akin to a mutual decision or thought process shared by a group of investors that a certain time horizon is optimal for momentum investing.

The way we are going to implement these ideas in the SDM estimator is to make the same assumption – that the time evolution of the lag structure is relatively slow. Given

that the ideal forecasting scenario is that our beliefs about the lag probabilities are accurate this means characterising any deviation, or error, in these beliefs as a relatively slow-varying function of the state vector.

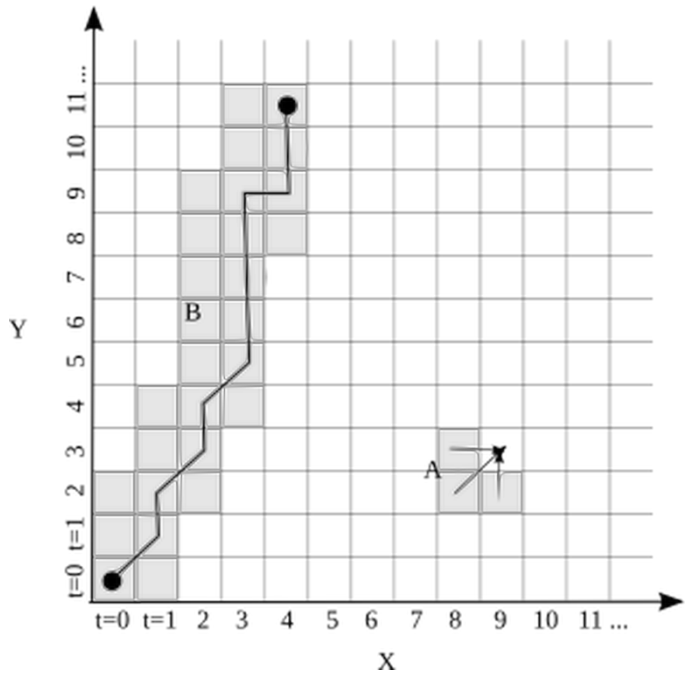

Figure 5: Slow varying lag paths. (A) describes the three squares in the matrix via which we can reach square (9, 3) due to the recursive relation used in the DTW and OTCP algorithms. (B) then shows an example of a path defined as a continuous mapping between points (0,0) and (4,11).

In our notation, equation (33) implies allowing a given $\mathbf{w}_t$ to be influenced by one of the following 3 values:

$$(w_{t-1}^{i-1}, w_{t-1}^{i}, w_{t}^{i+1}) \tag{34}$$

Assume that we knew the amount of variation in the lag structure with certainty, so we could fix a parameter θ, which captured the total magnitude of the variation. The parameter controls the rate of change per unit t for each path. Given that we are expecting $\mathbf{w}_t = \mathbf{w}_{t-1}$ in the 0 error case, equation (34) tells us the structure of the variation, so that we could substitute into (29) and write

$$w_t \propto w_{t-1} + error$$

so

$$w_t^i \propto w_{t-1}^i + \frac{\theta}{3}(w_{t-1}^{i-1} + w_{t-1}^i + w_t^{i+1}) \tag{35}$$

Dividing the magnitude of the error by 3 is just a normalisation to take care of the times where we reach the sides of the matrix; i.e. when there is no information for the w_t^{i+1} lag, in these cases we divide by 2. Equation (35) shouldn't be confused with the weighting schema discussed by Sornette and Zhou (2005) in introducing OTCP, the weighting function in the SDM/recursive Bayes case is given in the measurement model specification later in this paper.

Clearly, in most interesting cases we are not going to be able to assume a constant amount of variation, so a better characterisation would be as a random sequence of perturbations in the lag structure **v**: $[v_1 \dots v_t]$. We could define a mass or density function for **v** but as we will show this is not necessary and **v** can be any arbitrary sequence. We would then like, at each *t*, to choose a parameter value for θ which minimises the global error over repeated iterations, i.e. minimises the amount of variation in our beliefs about the lag structure.

To do this we can define some properties of **v** due to the properties of the probability vectors. Firstly, we note that $0 \leq |\mathbf{w}_t - \mathbf{w}_{t-1}| \leq 2$, so assuming no error we can also state $|\mathbf{w}_t| - |\mathbf{w}_{t-1}| = 0$. It follows then that the magnitude of the error process at *t* is the distance between the vectors $\mathbf{v}_t = |\mathbf{w}_t - \mathbf{w}_{t-1}|$. Further, we can also infer that $\mathbf{v}_t$ always has an expected value, since the expectation of a positive random variable is always defined, and that this expectation must lie on the bounded interval [0, 2].

Under these circumstances, the optimal choice of value for θ is the median of the distribution of the errors up to this point. This is due to the fact that the median is the minimiser of the distance function where the expected value of the function is defined; so we can define the optimal choice of value for θ as

$$\theta_t = \widetilde{\mathbf{v}_{1:t-1}} \tag{36}$$

where the tilde denotes the median of the vector of observed errors.

Substituting into equation (31) then yields the system model for the forward projection of the state vector as

$$w_{t|t-1}^i = w_{t-1}^i + \frac{\theta_t}{3}(w_{t-1}^{i-1} + w_{t-1}^i + w_t^{i+1}) \tag{37}$$

An important note is that at the first and last positions, i.e. the boundaries, on **w** not

all of the 3 positions will be defined since we have no data for either w_{t-1}^{i-1} or $w_t^{N_s+1}$, here we simply set the probability of these positions to 0. A further point of note is that equation (35) implies $\sum \mathbf{w}_{t|t-1}$takes values > 1 where $\theta > 0$, so is no longer a probability vector. At this stage a better interpretation of these estimates is as a set of weights defined by their relative likelihoods. We could normalise this vector to 1 by dividing over the sum of the weights but this is unnecessary as we will compute the actual probabilities using the update step (32).

In terms of the optimality conditions we state at the end of section 4.2, equation (37) is always defined for any pair of series irrespective of their spatial relationship. This follows because the information about the spatial relationship between the series is held as probabilities; so for the reasons already specified we can always calculate the median of the distribution and this will always be the optimal estimate for θ_t. The assumption we make is that the time evolution of the causality structure is relatively slow, this assumption is justified by a the large literature on DTW algorithms and the nascent OTCP literature.

The attractiveness of this approach is it allows us to concentrate on modelling the complexity of the distance relationship between the series independently of the temporal relationship; to show this we will describe how to fit equation (28) using this system model in the next section, then expand to a number of other cases.

4.4.2. Measurement models

Returning to our hypothetical momentum trader, the system model accounts how decisions about the timing of a trade at *t* are likely to affect decisions at *t* + 1, the measurement model is where we model how likely it is that an observed distance in the matrix is due to a shift in trading strategy. The analogy here is a group of investors all using similar momentum strategies, at each *t* they assess their past behaviours and make small adjustments to their strategies based modelled by the system model. The measurement model is where we look for evidence of these changes in strategy; this evidence comes in the form of the empirical distances we observe between the time series.

Expanding on the original example, consider a case a population of investors a variable proportion of which are pursuing a momentum strategy at any point in time. If 100% of the population is purely momentum investing then we would expect 100% of the

observed changes in price to be caused by momentum strategies, if the market is 50% momentum 50% value investors then only 50% of the price movement would be expected to be caused by momentum on average. If we had some empirical measure of momentum, a weekly survey or an online forum for example, the empirical distribution of the distances between this measure and asset prices, converted to a probability distribution, models the analytical probability that the observed distance between the price series and the momentum measure is due to a greater proportion of the population momentum investing.

Given that the state vector $\mathbf{w}_t$ holds our current beliefs about the timings the population of investors are using to enact their momentum strategies, it makes sense to only want to alter these beliefs after assessing the strength of the evidence. If there is compelling evidence that the changes in price at $t + 1$ are due to changes in momentum at $t + 1$ then we could like to incorporate all of this evidence into our state vector beliefs. If there is not, we may wish to disregard most of this information.

There are many ways we could calculate the empirical distribution of **d**; in the examples that follow we assume **d** to follow an exponential distribution and take the maximum likelihood estimators to ascertain the parameters of the model on a rolling basis. There are other options: we could use any distribution function for the distances as long as the parameters are estimable; or if we believed the relationship had a non-standard distribution we could estimate the empirical probabilities using a particle filter methodology (see Arulampalam et al. (2002), for a review of potential options).

To complete the SDM estimator we need to define a measure of the **x**, **y** relationship and posit a functional form for mapping these measures into the same probability space as the system model. This means making some assumptions about the form of the distances between the two series then using these assumptions to calculate the likelihood of the measurement. We than substitute these likelihoods into equation (32) and the estimator is complete. We will begin in section 4.4.2.1 by describing how to fit equation (28) where **x** and **y** are assumed to be positively correlated. We will then show, in subsequent sections, how we can relax these assumptions so that we can fit bi-directionally causality structures and models where there are positive-negative regime shifts in the causality structure.

4.4.2.1. Simple linear causal models

If we assume that **x** and **y** are measured in the same units, for example they may have been rescaled using their respective means and standard deviations, then if there was one lagged value of **x** causally related to **y** we could write this as;

$$y_t = x_{t-\tau} + u_t \tag{38}$$

If there is more than one lagged value of **x** causally related to **y** the OLS approach to model fitting is to interpret further lags as independent random variables. Thus the model expands by adding more random variables into the regression. The difference in approach using SDM is that we are going to maintain the assumption that there is only one causal relationship between the variables but that this relationship is distributed over a number of lags. In this way we can think of the probabilities as a weighted average over a partitioned interpretation of **x**, that is

$$y_t = u_t + \sum_{i=1}^{i=N_s} w_t^i x_{t-i} \tag{39}$$

The importance of equation (39) is that it contains only one global error distribution. The estimator can then focus on recursively estimating this single distribution rather than attempting to calculate the parameters of multiple distributions leading to dimensionality issues.

We will use the squared distance between **x** and **y** as the measure of the relationship, so that in effect the SDM estimator becomes the least squares estimator for a relationship with time-varying lags. Rearranging the terms in (39) yields

$$u_t = y_t - \sum_{i=1}^{i=N_s} w_t^i x_{t-i} \tag{40}$$

squaring the terms gives

$$u_t^2 = \left(y_t - \sum_{i=1}^{i=N_s} w_t^i x_{t-i} \right)^2 \tag{41}$$

then, setting the squared distance as the measure of the relationship so that $d_t^i = (y_t - x_{t-i})^2$, we can write equation (41) in a more compact form:

$$u_t^2 = \mathbf{w}_t \cdot \mathbf{d}_t \tag{42}$$

where the dot indicates the scalar product of the two vectors. If both **x** and **y** have been rescaled to mean 0 and u_t is $N(0,1)$, then we can assume $\mathbf{w}_t \cdot \mathbf{d}_t$ should equal zero given the correct probability weights. This justifies the measurement model being given as the departure of the empirical distribution of **d**, from zero, which is equivalent in this case to the distribution of u^2.

The assumption of the model (28) is that u_t is a mean 0 random variable with unknown variance. As a result we would expect the squared values of u_t^2 to be drawn from the exponential distribution. The density function for this distribution is given as.

$$Pr(u_t^2 \mid \lambda) = \lambda e^{-\lambda u_t^2} \tag{43}$$

Where λ is the rate parameter with the maximum likelihood estimator given as the mean of the observed values of u^2, that is

$$\lambda_t = \frac{1}{\frac{1}{tN_s}\sum_{s=1}^{s=t}\sum_{i=1}^{i=N_s} d_s^i} \tag{44}$$

We have included the time-varying subscript for λ_t as we are going to calculate the parameter of the distribution through repeated samples over time. Given an estimate for λ_t we can then calculate the likelihood of a given distance causally influencing the values of **y** using the likelihood function conditional on the value of λ_t.

$$Pr(d_t^i \mid \lambda_t) = \lambda_t e^{-\lambda_t d_t^i} \tag{45}$$

Substituting these likelihoods into the system model equation (32) then yields

$$w_t^i = \frac{p(d_t^i \mid \lambda_t) w_{t|t-1}^i}{p(\mathbf{d}_t \mid \lambda_t) \cdot \mathbf{w}_{t|t-1}} \tag{46}$$

Note that, unlike equation (32), we have used the scalar product notation for the denominator of (46). Combining equations (37) and (46) will then give the optimal recursive estimator of equation (28), under the specified assumptions.

4.4.2.2. Bi-directional causality structures

In the introduction we discussed certain occasions where the time ordering of information-price relationships may be reversed, i.e. when insider trading occurs in the run-up to an

information event. The estimator we have presented so far only considers causality running from information to price. In this section we will show how it is easy to generalise the SDM estimator to cases of bi-directional causality. A simple bi-directional analogue to equation (28) is the system of equations

$$y_t = \beta_1 x_{t-1} + \beta_2 x_{t-2} \ldots \beta_\tau x_{t-\tau} + u_t \tag{47}$$

$$x_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} \ldots \alpha_\tau y_{t-\tau} + \mu_t$$

where $\mu = N(0, \sigma_\mu^2)$ and $u = N(0, \sigma_u^2)$.

As there are now probability weightings associated with the lagged values of either variable, to simplify the notation we introduce the probability matrix **W** indexed at *t*, but containing a column of values for either series. The first column will contain the lagged values of **x** and the second the lagged values of **y** so that the entry $w_t^{i,1}$ would indicate the *i*th lag of **x** at time *t* and $w_t^{i,2}$ indicates the *i*th lag of **y** at time *t* so that

$$\mathbf{W}_t = \begin{bmatrix} w_t^{1,1} & w_t^{1,2} \\ \ldots & \ldots \\ w_t^{N_s,1} & w_t^{N_s,2} \end{bmatrix} \tag{48}$$

where since $\mathbf{W}_t$ is a probability matrix the sum of all elements in the matrix at *t*, equals 1 by definition.

We will define the distance measures associated with each of the lagged values in a matrix with corresponding entries to $\mathbf{W}_t$, that is

$$\mathbf{D}_t = \begin{bmatrix} d_t^{1,1} & d_t^{1,2} \\ \ldots & \ldots \\ d_t^{N_s,1} & d_t^{N_s,2} \end{bmatrix}$$

where $d_t^{i,1} = (y_t - x_{t-i})^2$ and $d_t^{i,2} = (x_t - y_{t-i})^2$.

Equation (63) enables us to write the squared distances for system of equations described in equation (47) as $u_t^2 = \sum_{i=1}^{i=N_s} w_t^{i,1} d_t^{i,1}$ and $\mu_t^2 = \sum_{i=1}^{i=N_s} w_t^{i,2} d_t^{i,2}$, which implies

$$u_t^2 + \mu_t^2 = \left(\sum_{i=1}^{i=N_s} w_t^{i,1} d_t^{i,1} \right) + \left(\sum_{i=1}^{i=N_s} w_t^{i,2} d_t^{i,2} \right) = \mathbf{W}_t \cdot \mathbf{D}_t \tag{49}$$

Since the summation of two exponential distributions is another exponential distribution, the maximum likelihood estimator for the sum of the squared error terms u and μ is the mean of the globally observed errors - due to equation (65) we can then write this as

$$\lambda_t = \frac{1}{\frac{1}{2tN_s}\sum_{s=1}^{s=t}\sum_{i=1}^{i=N_s} d_s^{i,1} + d_s^{i,2}} \tag{50}$$

The system model is then applied independently to either column of the matrix and the measurement model is applied to the **D** and **W** matrices using

$$w_t^{i,j} = \frac{Pr(d_t^{i,j} \mid \lambda_t) w_{t|t-1}^{i,j}}{Pr(\mathbf{D}_t \mid \lambda_t) \cdot \mathbf{W}_{t|t-1}} \tag{51}$$

The estimator is again optimal under the same assumptions as the unidirectional case.

4.4.2.3. Positive-negative regime shifts

The formulation of the estimator we introduce in section 4.4.2.2 shows how we can allow different functional representations of the **x**, **y** relationship to compete against each other for a share of the global probability density. Another case we might consider is where there is unidirectional causality running from **x** to **y**, but there are regime shifts from positive to negative correlation in the nature of the relationship. To capture this structure we can define a distance matrix containing the positive and negative squared distances so that

$$\mathbf{D}_t = \begin{bmatrix} d_t^{1,1} & d_t^{1,2} \\ \dots & \dots \\ d_t^{N_s,1} & d_t^{N_s,2} \end{bmatrix}$$

where $d_t^{i,1} = (y_t - x_{t-i})^2$ and $d_t^{i,2} = (y_t + x_{t-i})^2$.

The rest of the estimator is constructed in the same way as the bi-directional case described in section 4.4.2.2. and remains optimal under the same assumptions.

4.4.2.4. Forecasting

Once the probabilities of the either the vector or matrix of values for **w** are defined, we can then generate forecasts for the values of the leading series based on the probability mass over the lagging values. In the simple linear causal case we just multiply the values of the

lagging series through by the probability mass vector at *t* - 1, that is

$$\widehat{y_t} = \sum_{i=1}^{i=N_s} x_{t-i} w_{t-1}^{i} \tag{52}$$

It might at first appear odd that we don't first apply the Bayesian prediction step before making the forecast. The reason for this is that we are assuming any deviation from $\mathbf{w}_t = \mathbf{w}_{t-1}$ is due to noise obfuscating the relationship. We clearly don't wish to include this noise in our estimate; omitting it then just leaves the previous mass vector at *t* - 1.

In the positive-negative regime-switching case we can construct the estimate in a similar fashion by including both the positive and negative values of **x**, that is

$$\widehat{y_t} = \sum_{i=1}^{i=N_s} \left(x_{t-i} w_{t-1}^{i,1}\right) + \left(-x_{t-i} w_{t-1}^{i,2}\right) \tag{53}$$

In the bi-directional causality case an estimate for either series can be constructed using (52). The issue with this is that the probability mass on either column of the matrix may not sum to 1, so that either series may be causally influencing the other at the same time. In this case the estimate will be lower due to the reduced probability mass not the actual structure of the relationship.

To counter this we first calculate the mass of either column to give the relative probability either series is lagging the other

$$Pr(x \Rightarrow y|t) = \sum_{i=1}^{i=N_s} w_{t-1}^{i,1} \tag{54}$$

or

$$Pr(y \Rightarrow x|t) = \sum_{i=1}^{i=N_s} w_{t-1}^{i,2}$$

where the arrow should be read 'probability *x* lags *y*'. Then re-normalise the weights of either series to a unit mass (i.e. their conditional probabilities) before making the forecast

$$\widehat{y_t} = \sum_{i=1}^{i=N_s} x_{t-i} \frac{w_{t-1}^{i,1}}{Pr(x \Rightarrow y|t)} \tag{55}$$

or

$$\widehat{x_t} = \sum_{i=1}^{i=N_s} y_{t-i} \frac{w_{t-1}^{i,2}}{Pr(y \Rightarrow x|t)}$$

We can use the results of equation (54) as confidence weightings for either forecast, based on the probability one series is causally influencing the other.

4.4.3. Algorithmic implementation

Up to this point we have assumed that we would include every lag of **y** in the forecast of **x**. In most forecasting scenarios this would lead to a large tail of extremely low probability lags as more time-periods are considered. It also means that the computational complexity of the SDM algorithm would scale exponentially with *t*. Clearly in most forecasting tasks we only wish to consider a finite number of lagged values of the either series and so we can bound N_s to some reasonably small number; in this case the algorithm scales linearly with *t*. Consider the following pseudo-code implementation of the algorithm for the simple unidirectional linear causality model described in section 4.4.2.1.

The implementation of the algorithm as described is then no more complex than running three N_s length loops. Note the necessity for the first for loop in the code to be reversed running from N_s to 1; this is due to the boundary condition that where, w^{Ns+1} is not defined, so without reversing the order of the loop the probability of the w^{i+1} would always be undefined.

A second implementation consideration is that over some lags will return infinitesimal probabilities; in most programming languages this will either, result in an error as the floating point numbers overflow the memory limit of the language, or in some cases this will results in the probability being set to 0. A simple fix for this issue is to set a very small lower limit for the probability of each lag. Not doing this, results in suboptimal forecasts as the lag structure becomes increasingly path dependent as more lags are set to 0 over time.

Pseudo-code

Inputs: **x**: $[x_1 \dots x_t]$, **y**: $[y_1 \dots y_t]$, $\mathbf{w}_{t=0}$: $\left[w_{t=1}^1 \dots w_{t=1}^{N_s}\right]$ where $w_0^i = 1/N_s$

While $\leq t$:

Calculate parameter values

$\theta_s = \widetilde{v_{1:s-1}}, \lambda_s = \overline{u_{1:s-1}^2}, g = 0$

Apply system model

For $\in [N_s : 1]$: $w_{s|s-1}^i = w_{s-1}^i + \frac{\theta_t}{3}\left(w_{s-1}^{i-1} + w_{s-1}^i + w_s^{i+1}\right)$

Update probabilities based on new evidence

For $\in [1 : N_s]$: $d_s^i = (y_s - x_{s-i})^2, \widehat{w_s^i} = p\left(\, d_s^i \mid \lambda_s \right) w_{s|s-1}^i, g = G + \widehat{w_s^i}$

Normalise probability weights and update parameter vectors

$v_s = 0, u_s = 0$

For $\in [1 : N_s]$: $w_s^i = \widehat{w_s^i}/G\,, v_s = v_s + \left|w_s^i - w_{s-1}^i\right|, u_s = u_s + w_{s-1}^i d_s^i$

$s = s + 1$

4.5. Experiments on simulated data

In this section we are going to show the results of testing the SDM algorithm on a number of series constructed based on the assumptions of the Granger Causality model, as specified in equation (49), with the type of causality structures described by Cheung et al. (2004) and Soros (2009). The purpose of these simulations is not to define exact mathematical models representative of these theoretical insights, but rather to provide a number of cases that could arise in empirical work aimed at evidencing this area of financial theory. In section 4.5.1 we discuss the construction of these examples and present the equations used to generate the test series. We will then present results in section 4.5.2 of the root mean squared error forecasts we achieve using the SDM procedure to predict the values of the simulated price series and show forecasts for a range of different noise levels.

4.5.1. Construction of simulated series

In the work of Cheung et al. (2004) and Soros (2009) the notion of complex causality structures is only discussed in qualitative terms; this leaves no clear guidance as to what the best mathematical model to describe such series would be. To make sure we are covering a large range of possibilities we are going to test the SDM algorithm on five different series constructions. For each, we make the same assumptions as the Granger causality model described in equation (28) that the lagging series, in each of our example cases this will be the information series, is an autoregressive series constructed using

$$x_t = 0.9x_{t-1} + N(0,1) \tag{56}$$

where the autoregressive term is 0.9, so that the series exhibits significant but not infinite memory. This follows the testing framework used by Sornette and Zhou (2005). Using this basic series construction we hypothesise a number of different types of lag structure variation, which we describe in detail below.

We will focus on unidirectional forecasts. This greatly simplifies our results, as otherwise we have to show two sets of statistics for each test: one for either series, depending on which is leading in time. We note, however, that we would expect the same quality forecasts in bi-directional or positive-negative regime shifting cases as in the simple unidirectional case.

4.5.1.1. Step function models

There are a number of reasons why we might expect a jump in the lag length between time periods. Sticking to the example of momentum trading, consider a situation in which a population of investors are investing in price momentum on average *s* days from a news event. There are some news events that are so vast in their implications that they would cause a compete re-evaluation from the market over the strategies that are being pursued. The Japanese earthquake of 2011 would be one example, investors in Japanese stocks on the day of the earthquake would not adjust their strategies slowly, they are likely to jump from *s* days lag to 0 days lag on receipt of this type of information. We would also expect the vast majority of investors to act in a similar manner, creating a discontinuity in an otherwise continuous lag structure.

We construct two step function models. The first is a simple step function where

there is a single lag at each *t* causally influencing the value of the price series. The second is a model where there are multiple lags influencing the price series. For the first model the price series $f_1(x)$ is constructed as

$$f_1(x,t,\tau) = y_t = x_{t-\tau} + u_t \tag{57}$$

where the lag length τ varies as

$$\tau_t = \begin{cases} 5 & if \quad 0 < t < 200 \\ 20 & if \quad 201 < t < 400 \\ 10 & if \quad 401 < t < 600 \end{cases} \tag{58}$$

The error term u_t is a mean 0 Gaussian noise, so that σ_u is equal to the root mean square error (RMSE) we would expect from fitting a model of the **x**, **y** relationship if we knew the complete lag structure in advance. For the second model, $f_2(x)$, we will also include the lagged values local to τ in the series construction that is

$$f_2(x,t,\tau) = y_t = \frac{1}{7}\sum_{i=-3}^{i=3} x_{(t-\tau)+i} + u_t \tag{59}$$

Thus the price series is dependent on the three values of **x** immediately before and after τ in time.

4.5.1.2. Random walk models

We are also going to consider two models where the lag structure varies due to a random walk model. This type of model intuitively corresponds with the momentum trader example we gave earlier. Each investor is trading on a lag-length that varies slowly as a function of *t*. To model this, we construct the series similarly to the step function models but replace the lag length function (58) with a trinomial random walk of the form

$$\tau_t = \tau_{t-1} + z_t \tag{60}$$

where

$$f_z(z) = \begin{cases} 1/2 & z \in [-1,0] \quad if \quad \tau \geq 25 \\ 1/3 & z \in [-1,0,1] \quad if \quad 5 < \tau < 25 \\ 1/2 & z \in [0,1] \quad if \quad \tau \leq 5 \end{cases}$$

indicating that the random walk is bounded so that the lag-length can only vary between 5

and 25 time periods. The series $f_3(x)$ is then constructed similarly to equation (57), i.e. a random walk model where only one lag is causally influencing the price series. A second series $f_4(x)$ is constructed where the averaging function (59) is also applied to the series.

4.5.1.3. Fixed-integer-lag model

A final model we consider is a fixed-integer-lag model where

$$f_5(x,t) = y_t = x_{t-5} + u_t$$

We include this case as a baseline to show how well the algorithm forecasts in cases where there is no obfuscating lag variation.

4.5.1.4. Descriptive statistics and model fitting

For each of the models we use the SDM algorithm exactly as specified in the pseudo-code example given in section 4.3 where we set N_s to 30 time periods. To generate the forecast we take the probability weightings given to each of the lagging values, that is, each of the values of $\mathbf{w}_{t-1}$ and multiplying them by the corresponding values of **x** using

$$\widehat{y_t} = \mathbf{w}_{t-1} \cdot \mathbf{x}_{t-N_s:t-1} \tag{61}$$

which is identical to equation (52) given in scalar product notation.

The benefit of using simulated data is that we know what the optimal forecast value would be for a series of this type. In our case the error in the model is equal to σ_u since in all cases y_t-x_t = u_t, and u is a standard normal variate with mean 0. This can be verified since calculating the RMSE for this forecast against the observed values of **y** we get

$$RMSE(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = \sqrt{\frac{1}{t}\sum_{i=1}^{i=t}(\widehat{y_i} - y_i)^2} \tag{62}$$

which, in the optimal case where $\widehat{\boldsymbol{y}} = \boldsymbol{x}$, can be stated as

$$\sqrt{\frac{1}{t}\sum_{i=1}^{i=t}\{(x_i + u_i) - x_i\}^2} = \sqrt{\frac{1}{t}\sum_{i=1}^{i=t}u_i^2} \tag{63}$$

where the right-hand side of the equation is the same as σ_u. Since we know that the expected error between the series is equal to σ_u we also calculate a statistic for the deviation from expected forecast error as

$$FE(\hat{\boldsymbol{y}}, \boldsymbol{y}) = RMSE(\hat{\boldsymbol{y}}, \boldsymbol{y}) - \sigma_u \tag{64}$$

4.5.2. Experimental results

Figure 6 presents the results, each point on either plot represents the average over 500 trials for each model and noise level, σ_μ. The left hand plot shows the RMSE in its raw form calculated using equation (62). The right hand plot shows the FE statistic calculated using equation (64).

We can see from the left-hand plot that the RMSE statistic tracks the value of σ_u closely in all cases. For series where there is significant temporal variation, the random walk model $f_3(x)$ for example, we can see a period where the noise level is low $\sigma_u < 1$, there is a significant difference between the SDM model and the perfect fit. This is explicable by the fact that for these models the expected fit does not take into account the extra noise from the temporal variation in the model structure.

The right hand plot also supports this theory; we see that for the fixed-integer lag model the SDM generated model fits very close to the maximum achievable < 0.1 standard deviations difference for all noise levels. As the temporal variance of the models increases, we see the initial (low-noise) fits generated by the SDM algorithm getting increasingly poor - although the model fits are clearly still very good.

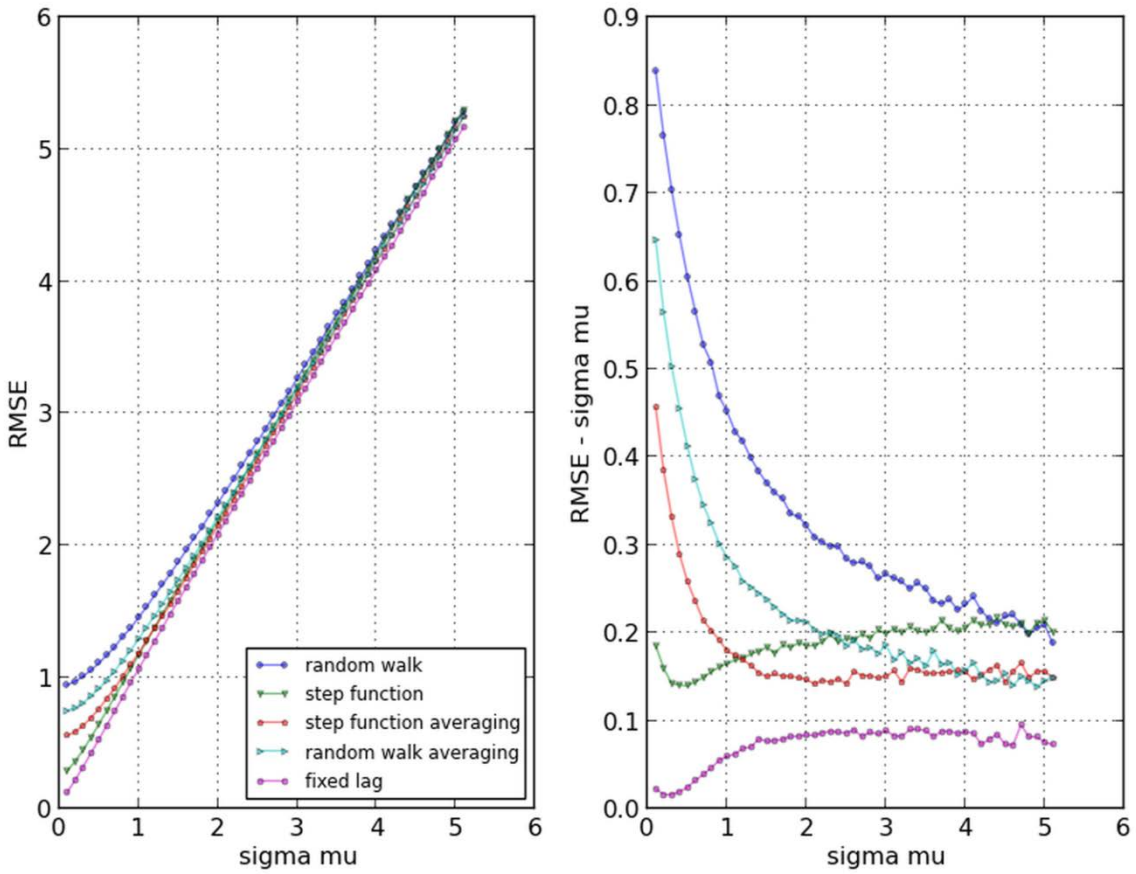

Figure 6: Experimental results. The left hand plot shows the RMSE for the SDM model fit. The key indicates each of the different model types so $f_1(x)$ = step function, $f_2(x)$ = step function averaging, $f_3(x)$ = random walk, $f_4(x)$ = random walk averaging and $f_5(x)$ = fixed lag. The right hand plot shows the FE statistic for each model.

4.6. Conclusions

The work of Cheung et al. (2004) and Soros (2009) highlights a key issue with the way most financial forecasting research deals with time-sequencing. Participants in financial markets do not see the relationship between information and price as temporally fixed, but describe a situation where fluctuations in the timing of the information-price relationship are key drivers of the variation seen in asset prices. In this situation, the authors describe how understanding when information is likely to affect price is key to successful forecasting. In this paper we have introduced a methodology that can capture the type of fluctuation the authors describe.

We have shown that, under the same assumptions as the standard Granger causality approach to time sequencing, the SDM algorithm produces the optimal Bayesian estimate of the forecasting distribution over the values of the lagging series. Importantly, the SDM algorithm we present will provide optimal Bayesian forecasts of the leading series in situations where the lag structure is not time varying, but also in situations where it is. As we have noted, given the construction of the system equations we provide in this paper it would be possible to substitute the Gaussian measurement model we use for a range of other more complex models of bivariate relationships present in the literature. The result is an estimation framework for time-varying lags, which is optimal under the conditions we describe, but also highly extensible to other cases.

There are a variety of ways SDM could be used to advance financial theory. For example, a classic approach in empirical finance is to take the aggregated values of an indicator series as a portfolio optimisation function. Taking the momentum theory work of Jegadeesh and Titman (1993) as an example, here the authors take the lagged 3-12 month return of a series as an indicator of its future 3-12 month return. If we consider this model mathematically we are assuming a uniform distribution of the lagged returns is a predictor of future returns. Why would a uniform distribution be optimal in this case? Surely there are other distributions that could be considered. With the introduction of SDM we provide forecasters with the ability to find the optimal distribution, even if this distribution is time varying. We propose that any work that uses this type of aggregated lagged indicator would benefit from being repeated with the optimal estimating distribution for comparison.

Another currently popular avenue of research is the temporal synchronicity of high frequency trades, Alsayed and McGroarty (2014), for example, use the Hayashi-Yoshida

(HY) cross-correlation estimator to study sub-second arbitrage opportunities in high-frequency index futures data. This work could be supplemented by using the SDM estimator to ascertain the changing distribution over time, to analyse if certain market conditions or events lead to greater arbitrage opportunities.

The second example suggests a key way this research could be extended in future. Currently the state vector **w** is assumed in the model to be a discrete set of binned probabilities. There is, however, no particular reason why we couldn't model the lag length probabilities as a continuous distribution, or use a particle filter type approach to estimate the empirical probability of an event occurring at a given distance from the last event. This could be a particularly useful addition to the estimator for working with asynchronous data or models that assume continuous return curves.

Chapter 5: Variability in Textual Sentiment-Price Relationships

5.1. Introduction

In this chapter we address a deficiency in the time series model we applied to model the sentiment-price relationship in chapter three. Based on the evidence from the cross-correlation plots in figures 1-4 of chapter three, there is a lagged relationship between sentiment and price that takes the form of an oscillator, of waveform. We used a simple moving average process to capture this relationship, which in light of the oscillatory structure of the relationship misses much of the structure shown in figures 1-4. In this chapter, we address this issue by applying the SDM estimator to model the relationship.

In chapter three of this thesis we describe how communities of self-declared investors are now going online to exchange information about stocks. This is happening in large enough numbers for Goldman Sachs (2015) to estimate there is a $4 trillion addressable market for online social investment services. A number of research articles have attempted to assess whether the content of these services can be used to model and forecast prices (Antweiler and Frank 2004; Das and Chen 2007; Sprenger et al. 2014; Chen et al. 2014). In these studies, researchers define a ‘textual sentiment metric’, a quantitative measure of the tone of a document based on features of the language used in the text. This is done by identifying some sets of words that are expected to indicate a generally positive or negative outlook for a stock. For example, the presence of the word ‘bullish’ may be considered an indicator of positive sentiment and ‘bearish’ may be an indicator of negative sentiment based on the lists of sentiment carrying words provided by Loughran and McDonald (2011b). A feature of these methodologies is the assumption that the semantic meaning conveyed by a piece of content is fully characterised by the presence of certain words in the text. The definition of the sentiment measure is some function of the frequency of a group of words.

Using this type of approach, several studies have documented a statistical relationship between online textual sentiment and prices, although the nature of the

reported relationship varies significantly over studies. For example, Das and Chen (2007) and Antweiler and Frank (2004) report a contemporaneous relationship between sentiment and returns but not lagged relationship. In contrast, Sprenger et al. (2014) reports a negative correlation between sentiment and returns at lag lengths of up to eight trading days, whereas, Chen et al. (2014) documents a positive correlation between sentiment and returns of up to three years.

In chapter three of this thesis we also document a long run relationship between sentiment and price. We go on to show, using cross-correlation plots, that prices oscillate around sentiment in a predictable manner: first there is a short-run (20- to 40-trading days) negative correlation between sentiment and price, followed by a midterm (80- to 100-trading days) positive correlation, ending in longer-term (100- to 120-trading days) negative correlation. We show that using a very simple trading strategy, of buying stocks based on their 100-trading average sentiment and holding them for 40-trading days, returns profits well in excess of the market benchmark once reasonable transaction costs have been accounted for.

Based on our results from chapter three and the inconsistencies in the findings of previous studies, we suggest that variability in sign, strength and timing may be a feature of the relationship between sentiment and price. In chapter three we discussed three potential sources of variability:

Variability in interpretation, where investor's cognitive biases lead them to interpret the word in the text differently based on their current state of mind. Evidence for this is drawn from the behavioural asset pricing literature, where authors such as Daniel et al. (1998), Daniel et al. (2002), Barberis et al. (1998), Hong and Stein (1999), Grinblatt and Han (2005) and Frazzini (2006), describe how variability in the sign, strength and timing of the relationship between information and price can vary as a function of the cognitive biases of investors.

Variability in authorship, where there may be variation in how much investors trust particular sources of information, or how easily investors can process the semantic meaning of an author's writing style. Evidence in support of this type of variability comes from the literature on managers use of language in corporate filings, where authors, such as Davis and Tama-Sweet (2011), Huang et al. (2012), Loughran and Mcdonald (2011), Larcker and Zakolyukina (2012) and Rogers et al. (2011), document how investors may be

misled by managers use of deceptive language. This creates a feedback loop between investor beliefs about a manager's honesty and observed relationships between prices and managers previous statements.

Variability in context, where there may be contextual factors that affect the semantic meaning of the content that are not explicitly referenced in the text. Several authors, such as Henry and Leone (2009) and Loughran and McDonald (2011), have described how finance specific words work better as a classifier of financial content than standard lists of sentiment carrying words. Extending this line of reasoning, timing is an important aspect of whether a statement is good or bad. The statement 'Dow Jones to top 17,000 points by Wednesday', takes on a different semantic meaning if viewed on Tuesday or Thursday. Sentiment measurement techniques which focus on the polarity of a message alone will miss this type of contextual information.

Of these three sources of variability, only one has a natural measure that has been reported in the literature. In the momentum and reversal literatures, empirical work by authors such as Jegadeesh (1990), Jegadeesh and Titman (1993), Thaler and De Bondt (1985) and Moskowitz et al. (2012), has contributed to the evidence on similar oscillating patterns in serial correlations in prices. This empirical evidence has driven the development of a range of behavioural asset pricing models that produce this type of oscillatory structure as a consequence of investors innate cognitive biases. Examples of this work include, Daniel et al. (1998), Daniel et al. (2002), Barberis et al. (1998), Hong and Stein (1999), Grinblatt and Han (2005) and Frazzini (2006). We use multivariate regression analysis in chapter three to show that, whilst some of the relationship between sentiment and price can be explained by serial correlations in prices, sentiment contains significant incremental information about future prices that cannot be explained by price momentum.

We are then left with the fact that there are significant sources of variability in the sentiment-price relationship which are not measurable in any meaningful way. For example, it is difficult to see how the level of trust investors place in a given author could be measured effectively in most cases. This leaves a gap in the literature concerning how this relationship can be modelled effectively in light of the latent variability that appears to be one of its characteristic features. More concretely, the empirical results we present in chapter three document an oscillatory relationship between sentiment and price, yet the strategy we employ in chapter three to trade online sentiment is significantly less complex

than the sentiment-price relationship we observe. The relationship we document takes the form of an oscillating waveform, yet the trading strategy we employ uses a simple average of the lagged sentiment series as the signal to trade.

To fill this gap, we return to modelling the same relationship tackles in chapter three, this time armed with the new time series analysis methodology we presented in chapter four. Signal diffusion mapping (SDM), is a recursive Bayes estimator which we show in chapter four is the optimal Bayesian estimator of complex, time varying, lagged relationships we documented in chapter three. The estimator attempts to find the best average over the lagged values of a time series to forecast the values of the leading series. Consequently, we can directly compare the results we get from using the simple moving average strategy we introduced in chapter three against the SDM averaging distribution. Our hypothesis is that, if the sentiment-price relationship is time varying or more complex than a simple average, SDM should identify a better distribution and subsequently provide a better forecast.

We show that SDM does indeed outperform a simple average by a substantial margin; we evaluate this performance using a similar method to the Diebold-Mariano test and by comparing the directional forecasting accuracy of both SDM and the simple average. We go on to present a portfolio of ‘SDM winners’, which returns an excess APR of 19.424% over the market benchmark after reasonable transaction or trading costs have been accounted for. This is more than double the excess return we generated using a simple average in chapter three, and suggests SDM can successfully identify significant, economically meaningful information about future prices.

In terms of interpreting these results, SDM is an upgrade on a simple averaging model because it allows the sentiment-price relationship to vary, across stocks, and over time. We are not alone in reporting that allowing for extra variability on either of these dimensions in the data leads to greatly improved forecasts. For example, Jegadeesh and Wu (2013) document how assigning weights to dictionaries of sentiment carrying words greatly improves the performance of their classifier. The authors fit linear regression models to find parameter estimates for the size of the return associated with the frequency of a given word. This forms a stock specific set of parameter estimates for a list of words. The authors document that some apparently positive sentiment carrying words have negative weights and vice versa, leading them to conclude that identifying the correct term

weights is as important as identifying the correct words for the sentiment dictionary.

In another example, Daniel and Moskowitz (2011) document how returns to momentum portfolios experience short periods of significantly negative returns. These momentum crashes occur during short market downturns and significantly reduce portfolio performance. The authors show that introducing a time varying aspect to their stock selection strategy doubles the return of portfolios by avoiding these crashes. Although the predictor variable in this case is price momentum rather than textual sentiment, the example serves to illustrate how a similar type of price predictability is affected by change over time.

This raises the question whether we can ascertain which type of variability is most responsible for the improvement in performance: variability in the averaging distribution across stocks, or over time. To answer this question, we develop a novel hypothesis testing framework based on the output of the SDM estimator. We document that stock specific distributions over the lagged values of the sentiment series are responsible for most of the improvement in estimator performance. We also find strong evidence that the distributions vary significantly over time. We conclude that latent variables do significantly alter the way content is interpreted, that they are variable over time, and that there are contextual factors that are specific to different stocks.

In conclusion; we present strong evidence that the sentiment-price relationship exhibits significant variability. We show that using SDM we can capture much more information about the relationship than traditional methods allow. Using this information, we document excess returns from trading sentiment that are large enough to have significant economic meaning for investors.

The rest of this paper is presented as follows: section 5.2 describes our data and variable construction; section 5.3 introduces the SDM estimator and how we intend to apply it to our data; section 5.4 describes empirical findings from comparing the forecasting accuracy of SDM over a simple average, and shows results from portfolios optimised with the SDM algorithm. In section 5.5 we explore the structure of returns predictability using SDM and section 5.6 concludes.

5.2. Data and variables

So that our results are comparable, we use exactly the same data and variable construction as we used for chapter three. Here we provide descriptions of key variables again for reference, but omit some of the corresponding justifications for brevity since these remain the same. As a recap, for prices we are using the component stocks of the Standard and Poor's 100 (SP100) stock index. Of the 100 stocks we find that 93 have messages covering every day in the sample period so we omit the remaining 7 stocks from the rest of this study to prevent low sampling rates from biasing our results. We downloaded the daily adjusted closing price for each stock from the Yahoo Finance website.

For text data we have collected over 10.2 million messages, each of which specifically references one of the companies in our sample, from the period running January 2014 to March 2015. These messages are drawn from 45,516 different content sources across the web adding up to over 6 billion separate words. As a result we believe this to be at least one order of magnitude more content than is typically considered for a study of this type. For example, Antweiler and Frank (2004) use a corpus of 1.5 million messages from the Yahoo Finance and Raging Bull message boards, this is the second largest corpus that has been considered in the online textual sentiment literature. Full details of the query strings we to collect this sample and descriptive statistics can be found in the Appendix of chapter three.

Following standard practice, we convert closing prices to log returns using

$$r_{i,t} = log\left(\frac{p_{i,t}}{p_{i,t-1}}\right) \tag{65}$$

where $p_{i,t}$ indicates the adjusted closing price of the *i*th stock at time *t* and $r_{i,t}$ is the corresponding log return.

To measure sentiment, we use the Loughran and McDonald (2011) lists of sentiment carrying words. We count the number of positive and negative sentiment carrying words that occur on each day in the study period, we then define sentiment as

$$q_{i,t} = log\left\{\frac{\left(1 + lmp_{i,t}\right)}{\left(1 + lmn_{i,t}\right)}\right\} \tag{66}$$

where $lmp_{i,t}$ indicates the count of the positive sentiment carrying words from the

Loughran and McDonald dictionary which occurred in messages referencing the *i*th stock at time *t*, $lmn_{i,t}$ indicates the corresponding negative sentiment carrying words. In this paper we consider raw returns only, rather than abnormal returns, as we are primarily interested in testing the performance of SDM against the simple moving average we used to model the sentiment-price relationship in chapter three, particularly the returns to sentiment trading strategies based on either modelling approach.

5.3. Methodology

In this section, we introduce the model we employed as part of the sentiment trading strategy we describe in chapter three, then discuss problems with this model in light of the argument we make about time variability being characteristic of the sentiment-price relationship. We then introduce the SDM algorithm and describe how it is appropriate for solving these problems. The focus of this description is on the application of SDM to the specific task at hand, rather than a discussion of the methodology per se, as full technical details of the estimator, its derivation and necessary assumptions for optimality are covered in chapter four. Throughout we adhere to standard vector notation, so lowercase italic *x* denotes a variable, lowercase bold a vector **x** and upper case bold a matrix **X**.

5.3.1. Moving average models

The strategy we employ in chapter three was based on simple averages, we document, using linear regression analysis, that the average 60-, 80- or 100-trading-days sentiment is negatively correlated to the average 20- to 40-trading-day return. We subsequently base our trading strategy around the model

$$E\left[\frac{1}{h}\sum_{s=1}^{s=h} r_{i,t+s}\right] = \frac{1}{(m+1)}\sum_{k=0}^{k=m} -q_{i,t-k} \tag{67}$$

where *h* and *m* are the holding period and moving average period respectively, so in our analysis we find *m* = 60, *m* = 80, *m* = 100 and *h* = 40, *h* = 20 to be the better performing values. This type of model is now common in the financial literature under the heading of

momentum trading strategies, for example Moskowitz et al. (2012). The difference in our case is that the momentum literature considers the stocks relationship to its own prior price, here, we substitute prior price for the sentiment measure. Also, the momentum literature considers a positive relationship between prior returns and future returns, whilst here we consider a negative relationship between sentiment and future returns.

The issue with equation (67) is that in the first paper of this thesis we report that the lagged relationship between sentiment and price appears to be more like an oscillator or waveform. Mathematically, the flat average over the sentiment series does not capture this at all. We also note that this issue is typical of the momentum literature, where empirical evidence from authors, such as Moskowitz et al. (2012), and theories such as Hong and Stein (1999), suggest price momentum forms an oscillatory effect on prices, yet researchers still tend to use simple averages as the basis of momentum portfolios, in spite of the fact this would be expected to miss important dynamic properties of the relationship.

To account for this type of variation, we can expand equation (67) by considering a more general functional form, replacing the average with a series of probability weights, assigned to the lagged values of each stock conditional on the specific stock and time, that is

$$E\left[\frac{1}{h}\sum_{s=1}^{s=h} r_{i,t+s}\right] = \sum_{k=0}^{k=m} w_{i,t}^{k}\left(-q_{i,t-k}\right) \tag{68}$$

where

$$\sum_{k=0}^{k=m} w_{i,t}^{k} = 1 \tag{69}$$

so that w^k indicates the kth position on vector $\mathbf{w}_{i,t}$. Since $\mathbf{w}_{i,t}$ is a probability vector, equation (69) states that the values of $\mathbf{w}_{i,t}$ must some to 1. Under the Bayesian interpretation of probability we use in this paper, these weights are measuring the strength of our beliefs that a given lag $q_{i,t}$ is influencing the sentiment-price relationship at a given point in time.

Via the same logic, we can then extend the model to include both positive and negative values of **q** using

$$E\left[\frac{1}{h}\sum_{s=1}^{s=h} r_{i,t+s}\right] = \sum_{k=0}^{k=m} w_{i,t}^{k,0}\left(-q_{i,t-k}\right) + w_{i,t}^{k,1} q_{i,t-k} \tag{70}$$

where

$$\sum_{k=0}^{k=m} w_{i,t}^{k,0} + w_{i,t}^{k,1} = 1 \tag{71}$$

So that $w^{k,j}$ is a probability weight attached to the *k*th lag and *j* is an indicator variable taking 0 if the probability weight is associated with negative sentiment and 1 where the weight is associated with negative sentiment. Equation (71) states that, since $\mathbf{W}_{i,t}$ is a probability matrix, the values must sum to 1.

The form of (69) is then flexible enough to generate values ranging over the full scale from $MIN(-q_{i,k})$ to $MAX(q_{i,k})$, and which is allowed to vary in time. This means that equation (71) can, in principle, capture the variability over the lagged values of the sentiment series we report in chapter three. In the next section we are going to show how we can define the optimal values for these weights using the SDM algorithm.

5.3.2. Preliminaries

Before we begin we are going to introduce some definitions and notational conventions that will greatly simplify the following description. To make the definition of the probability matrix explicit; we define $\mathbf{W}_{i,t}$ as a pair of column vectors, the first contains probability weights attached to negative values of the sentiment series and the second contains weights attached to the positive values. A different matrix of weights is defined for each stock and time period, so that

$$\mathbf{W}_{i,t} = \begin{bmatrix} w_{i,t}^{0,0} & w_{i,t}^{0,1} \\ \dots & \dots \\ w_{i,t}^{m,0} & w_{i,t}^{m,1} \end{bmatrix} \tag{72}$$

represents a full matrix of probability weights for the *i*th stock at time *t*. One interpretation of this matrix is as the conditional distribution over the lagged values of the sentiment series for a given stock and time period. As we are considering both positive and negative

values of the sentiment series, to simplify notation we are also going to group these by stock and time in the corresponding matrix to **W**. that is

$$\mathbf{\Psi}_{i,t} = \begin{bmatrix} -q_{i,t} & q_{i,t} \\ \cdots & \cdots \\ -q_{i,t-m} & -q_{i,t-m} \end{bmatrix} \tag{73}$$

Where the dot product of the matrices, denoted $\mathbf{W}_{t,i} \cdot \mathbf{\Psi}_{t,i}$, is indicative of the summation over the product of each pair of elements in either matrix, that is

$$\mathbf{W}_{i,t} \cdot \mathbf{\Psi}_{i,t} = \sum_{k=0}^{k=m} \sum_{j=0}^{j=1} w_{i,t}^{k,j} \psi_{i,t}^{k,j} \tag{74}$$

We are then going to define a further matrix $\mathbf{D}_{i,t}$, containing the a measure of the distance between the sentiment series and the holding period return. First defining this distance as

$$d_{i,t}^{k,j} = \psi_{i,t}^{k,j} - \left(\frac{1}{h} \sum_{s=1}^{s=h} r_{i,t+s} \right) \tag{75}$$

this leads to the definition of the distance matrix as

$$\mathbf{D}_{i,t} = \mathbf{\Psi}_{i,t} - \left(\frac{1}{h} \sum_{s=1}^{s=h} r_{i,t+s} \right) = \begin{bmatrix} d_{i,t}^{0,0} & d_{i,t}^{0,1} \\ \cdots & \cdots \\ d_{i,t}^{m,0} & d_{i,t}^{m,1} \end{bmatrix} \tag{76}$$

The usefulness of these definitions is that we can now state equation (70) in the simpler form

$$E\left[\frac{1}{h} \sum_{s=1}^{s=h} r_{i,t+s} \right] = \mathbf{W}_{i,t} \cdot \mathbf{\Psi}_{i,t} \tag{77}$$

or alternatively using the distance matrix

$$\varepsilon_t = \mathbf{W}_{i,t} \cdot \mathbf{D}_{i,t} \tag{78}$$

where ε_t is the noise obfuscating the relationship.

5.3.3. Grid based filters

SDM is a type of recursive Bayes estimator where the beliefs about the state of the system form a grid or matrix, in our case these are the probability weights contained in the probability matrices, estimators of this form make up a family of estimators commonly referred to in the literature as grid based filters. The optimal solution to the grid based filtering equations is well defined under the assumption that the evolution of the system's state follows a Markov chain. The Markov chain assumption is typically framed in terms of the general dynamical model (GDM), which in our notation is the system of equations

$$\mathbf{W}_{i,t} \propto Pr\left(\mathbf{W}_{i,t} \middle| \mathbf{W}_{i,t-1}\right) \tag{79}$$

and

$$\mathbf{D}_{i,t} \propto Pr\left(\mathbf{D}_{i,t} \middle| \mathbf{W}_{i,t}\right) \tag{80}$$

Equation (79) models the time evolution of our beliefs about the state of the system, held as probability weights in the matrix. Since this is a Markov chain this time evolution is only conditionally dependent on the previous time-period. Equation (80) models some probabilistic measure of the system's state. Conceptually, $\mathbf{W}_{i,t}$ models the combined beliefs of a population of investors as to which lags of the sentiment series are most relevant for forecasting future prices. Equation (79) models the form and the amount of change we expect in these beliefs between time $t-1$ and time t. $\mathbf{D}_{i,t}$ is a measure of the new evidence we receive about the relationship at t. As we assume the sentiment-price relationship may be noisy these measurements will contain some error, equation (80) states the amount of error we are expecting. The solution to these equations is found recursively, by first projecting the system's state forward in time using the probability model given in equation (79), then Bayes rule to update the probability weights based on the new measurements at t.

Firstly, we begin by initiating the system's state as a set of even probability weights. Since we have no prior beliefs about the sentiment-price relationship at $t = 0$, the weighting of each value in $\mathbf{W}_{i,0}$ is an even split over the number of possible values, so that $w_{i,0}^{k,j} = 1/2\,(m+1)$. We then define a *system model*, to project this state forward in time based on some assumptions about how investor beliefs alter over time. The forward projection of the system's state in our notation is defined as

$$w_{i,t|t-1}^{a,b} = \sum_{k=0}^{k=m} \sum_{j=0}^{j=1} Pr\left(w_{i,t}^{a,b} \middle| w_{i,t-1}^{k,j}\right) w_{i,t-1}^{a,b} \tag{81}$$

so that for a given probability weight $w^{a,b}$, we calculate the updated weight as the summation over the weights at t - 1 multiplied by the conditional probability that each of these weights could be influencing the relationship at t.

To update these beliefs we include a *measurement model* which governs how we update the probability weightings in light of new measurements. We do this via a combination of Bayes rule and the law of total probability, that is

$$w_{i,t}^{a,b} = \frac{Pr\left(d_{i,t}^{a,b} \middle| \mathbf{W}_{i,t}\right) w_{i,t|t-1}^{a,b}}{Pr\left(\mathbf{D}_{i,t} \middle| \mathbf{W}_{i,t}\right) \cdot \mathbf{W}_{i,t|t-1}} \tag{82}$$

where the denominator is just a normalising term which contains the sum of the probabilities over all possible states. This ensures the weights sum to 1. In our case, the probability weights are indicative of the current relationship between the sentiment series at a particular lag and returns. What equation (81) models, is how much we expect this lagged relationship to change in light of new information about the relationship. Conceptually, the relationship between sentiment and price is a function of how investors are currently interpreting information in the sentiment variable. In these terms, what equation (81) captures is how much this interpretation would be expected to change over time. Equation (82) captures the strength of the new evidence in favour of a change in the relationship.

There are then two conditional distributions in equations (81) and (82) for which we have not posed a functional form. If the functional forms for these distributions are an accurate reflection of the way the system functions, then these equations represent the optimal Bayesian solution to the estimation of the GDM equations. The signal diffusion mapping algorithm poses a general form to these equations based on quite relaxed assumptions about the underlying data generating process. The following section describes this in terms of the sentiment-price relationship we wish to measure.

5.3.4. Signal diffusion mapping

The SDM approach to solving these equations is conceptually quite straightforward. We are simply going to assume that given some distribution over the lags of the sentiment series, this distribution is only going to change relatively slowly over time. The model we present for the conditional distribution in equation (81) is one that enforces a relatively slow rate of change on the system's state. This is appropriate in this context because we know there are oscillations in the lagged relationship between sentiment and price, as documented in chapter three. We also suspect that there may be changes to the lagged relationship over time, due to latent factors affecting the relationship that are not recorded explicitly in the text. For either of these types of variability we would expect there to be some finite rate of change. For example, we would not expect investors to completely change the heuristics they are using to make decision over a single day. Rather, we would expect this to be a gradual process over time. Similarly, if there are latent factors concerning the wider context content was published in that are not reflected explicitly in the text, we would not expect that these would change completely over the course of a day, rather, they are likely to trend over time.

To do this, we first consider a situation where we know exactly what values for the probability weights represent the optimal relationship between sentiment and price at time *t* - 1. Further, we also know that this is not expected to change between now and *t*, in this case we would simply expect that the system's state at *t* was a copy of the system's state at *t* - 1 so that

$$w_{i,t}^{a,b} = w_{i,t-1}^{a,b} \tag{83}$$

the issue with this is that we are expecting there to be some variation in the lag distribution over time and we wish to model how this change is likely to occur. In introducing SDM in chapter four we argue that a reasonable model of this change is to assume the lag structure will change by one time period for every unit increase in *t*. The way we apply this using SDM is to assume that the lag structure can change, but this change is locally bound to the immediately adjacent lags on the probability matrix. We do this using

$$w_{i,t|t-1}^{a,b} = w_{i,t-1}^{a,b} + \frac{1}{3}\left(w_{i,t-1}^{a-1,b} + w_{i,t-1}^{a,b} + w_{i,t-1}^{a+1,b}\right) \tag{84}$$

so that we are only assuming dependence between lags that are immediately adjacent in

time and have the same sign. The sense in equation (84) in the case of the online textual sentiment-price relationship is that, if investors are currently informing their trading decisions based on textual sentiment at $t - 5$, it is unlikely all traders investing in the stock will change their minds completely and start investing based on information from another lag. More likely, there will be a gradual change over time. SDM models this by assuming that the probability mass moves slowly away from its current location.

Given this system model, to complete the estimator we need to propose some functional form for the conditional distribution in equation (82). Here we are going to assume the distances are normally distributed and so the probability density function of the distances is given as

$$Pr(d_{i,t}^{k,j} \mid \mathbf{W}_{t,i}) = Pr(d_{i,t}^{k,j} \mid \mu_i, \sigma_i^2) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{\frac{-(d_{i,t}^{k,j} - \mu_i)^2}{2\sigma_i^2}} \quad (85)$$

Substituting the conditional distributions in equations (81) and (82) for (84) and (85) then provides the optimal Bayesian estimator of the lagged relationship providing the assumptions are correct.

5.3.5. Forecasting

The result of applying SDM to the data is a conditional probability matrix $\mathbf{W}_{i,t}$, where each entry in the matrix contains the probability of a value in $\mathbf{\Psi}_{i,t}$ influencing the sentiment-price relationship for the *i*th stock at time *t*. From this matrix we can derive a number of different probability distributions for forecasting.

Firstly, we might assume that there is a single distribution, common to all stocks, characterising the way in which information is processed by investors. This is similar to the assumption of the simple averaging model in that we are assuming there is a fixed distribution that characterises the relationship, and that this distribution does not vary over time. The difference in this case is that the form of the distribution is more flexible than the flat average.

As the estimator is a Markov chain, the optimal estimate of this distribution given at end of the sample period. Taking the average of these distributions across all stocks

yields

$$\mathbf{W}_{i\in\mathbf{i},N_t} = \frac{1}{N_i}\sum_{i=1}^{i=N_i} \mathbf{W}_{i,N_t} \tag{86}$$

where N_t is the length of the sample period and N_i is the total number of stocks in the sample. All equation (86) states is that to derive the distribution for all stocks in the set of stocks **i**, we simply take the average of all of the 93 distributions. As each probability matrix contains the conditional distribution over the lags for a given time period this calculation corresponds to taking the marginal distribution for the final time-period over all stocks in the sample.

For forecasting purposes this distribution replaces the moving average in (67) so the forecast for the expected value of the return over the holding period for the *i*th stock is

$$E\left[\frac{1}{h}\sum_{s=1}^{s=h} r_{i,t+s} \mid \mathbf{W}_{i\in\mathbf{i},N_t}\right] = \mathbf{W}_{i\in\mathbf{i},N_t} \cdot \left(\mu_i + \mathbf{\Psi}_{i,t}\right) \tag{87}$$

where the notational convention $E[r \mid w]$ means the expected return based on some set of probabilistic beliefs.

Note that μ_i is included in the forecast, in chapter three we showed how to estimate the conditional mean of the distance matrix as part of the estimation procedure. In this paper we are going to assume we know the mean of the distribution already, as estimating multiple parameters takes data and we have only a finite time period of data for this study. Our reasoning is similar to that of Diebold (2015) who argues that studies should clearly state their modelling assumptions and maximise the use of their data, rather than attempt to make realistic pseudo out of sample tests. So here we are assuming that the distance matrix has a definable mean, and that this is estimable from prior data. Similarly, in chapter three we include a parameter modelling the expected rate of change of the probability distribution. Again we omit this parameter as there is limited data to estimate it.

Defining the forecast based on the state vector distribution at t_{N_t} assumes there is one distribution which characterises the relationship. Part of our argument is that this distribution may vary significantly over time. Also, the forecast given in equation (87) suffers from look-ahead bias as we are using information from the full sample period. A

more realistic forecasting scenario is to use only the values of $\mathbf{W}_{i,t}$ we could have calculated based on prior data. A complicating factor in this case is that we are fitting SDM to the front running return period, so effectively the estimates for $\mathbf{W}_{i,t}$ are inclusive of information about future values of the price series over the holding period running $t + 1{:}t + h$. To prevent this type of look-ahead bias, we need to lag the estimating distribution by h time periods, that is

$$\mathbf{W}_{i\in\mathbf{i},t-h} = \frac{1}{N_i}\sum_{i=1}^{i=N_i}\mathbf{W}_{i,t-h} \tag{88}$$

leading to the forecast

$$E\left[\frac{1}{h}\sum_{s=1}^{s=h} r_{i,t+s} \mid \mathbf{W}_{i\in\mathbf{i},t-h}\right] = \mathbf{W}_{i\in\mathbf{i},t-h}\cdot\left(\mu_i + \mathbf{\Psi}_{i,t}\right) \tag{89}$$

In section 1 we argued that there are at least three major sources of variability in the sentiment price relationship that could be latent factors affecting the relationship between our measure of sentiment and price. There are aspects of this variability that could reasonably be expected to be stock specific, for example, if there are contextual factors which modify the semantic meaning of certain words then we might expect these to be different for different industry sectors. As a result, we are also going to consider forecasts for local distributions for each stock. These are very similar so we just do not need to average the forecasting distribution over all stocks. The corollary of (86) for the *i*th stock is then

$$E\left[\frac{1}{h}\sum_{s=1}^{s=h} r_{i,t+s} \mid \mathbf{W}_{i,N_t}\right] = \mathbf{W}_{i,N_t}\cdot\left(\mu_i + \mathbf{\Psi}_{i,t}\right) \tag{90}$$

similarly the forecast

$$E\left[\frac{1}{h}\sum_{s=1}^{s=h} r_{i,t+s} \mid \mathbf{W}_{i,t-h}\right] = \mathbf{W}_{i,t-h}\cdot\left(\mu_i + \mathbf{\Psi}_{i,t}\right) \tag{91}$$

indicates a forecast with the local distribution for the *i*th stock where any look-ahead bias has been removed.

5.3.6. Comparing forecast accuracy

In order to compare the predictive accuracy of the SDM estimates against the moving average model we use in chapter three we apply two hypothesis tests. As the hypothesis we are testing is that the SDM estimator produces significantly better forecasts than the moving average model, the null hypothesis is that there is no statistically significant different between the forecasts across all stocks using either model. We test this in two ways: the first is though a t-test on the difference between the forecast errors produced by either model similar to the Diebold and Mariano statistic (Diebold and Mariano 1995). This is one of the most widely used statistical tests for comparing competing forecasting methodologies. The second test we employ is based on the directional forecast accuracy of either model.

For the first test we define a quantity $\delta_{i,t}$ denoting the forecast error between the estimate of the returns series and the forecasting model. That is

$$\delta_{i,t} = \left|E\left[r_{i,t} \mid SDM\right] - r_{i,t}\right| - \left|E\left[r_{i,t} \mid MA\right] - r_{i,t}\right| \tag{92}$$

where $E\left[r_{i,t} \mid SDM\right]$ indicated the expected return forecast by the SDM algorithm and $E\left[r_{i,t} \mid MA\right]$ indicates the forecast from the moving average model (equation (67)), the bar brackets indicate the absolute value function. Conceptually $\delta_{i,t}$ is a comparison of the distance between either estimator and the observed values of the returns series, this type of test has been widely popularised in the forecasting literature by Diebold and Mariano (1995). If there is no difference between the accuracy of either forecast we would expect the mean of the forecast errors to be approximately 0. This gives a natural hypothesis test that the mean of these errors is significantly different from 0. In the Diebold and Mariano test significant effort is made to model auto-correlation in the forecast errors, as we are going to compare forecast accuracy over 93 different stocks, we expect that this type of auto-correlation is not an issue. Our test is then a simple z-test of the difference in forecast errors.

The second test we employ is based on the directional forecasting accuracy of either model. For this we count every time the forecast has the same sign as the observed return. We then calculate the percentage accuracy of the model and compare the difference in forecast accuracy to the Bernoulli distribution.

5.4. Empirical results

In this section we present results from applying the SDM and moving average models to our data.

5.4.1. Comparing forecast accuracy

For the moving average model we use equation (67), for the SDM model we use equations (86) to (91). We test the models on for maximum lagged windows of $m = 60$, $m = 80$ and holding periods of $h = 20$ and $h = 40$, as these correspond to the best performing moving average models, as shown in chapter three.

Table 4 summarises these results. Each of the 4 models described in section 3.5 are listed in the first column of the table, we use the same *m* and *h* parameters are for both the SDM and moving average models. $Dist(model)$ is the average distance between the estimate and the actual values of the return series, that is the mean of $\left|E\left[r_{i,t} \mid model\right] - r_{i,t}\right|$.

We can see that in all cases SDM models produce smaller mean distances over the sample set. The t-statistics are all negative, indicating SDM has a smaller average error than the moving average and are all large enough to be significant at infinitesimal *p*-values. Similarly we see that the directional forecast accuracy is in most cases is more than 2% better using SDM than using a moving average. Again, the significance of these results is extremely high, a 2% difference between estimates over 10,000 samples has a *p*-value approaching 0.

Comparing the average error distances for the different SDM models, it appears the estimator works better with a holding period of 40-trading-days, although there are fairly similar results across SDM model specifications. In terms of our initial hypothesis, table 4 provides strong evidence in support of the claim that SDM is a better estimator of the relationship than a simple average. This suggests that variability in the sentiment-price relationship is a feature of the data which has not been captured in previous studies.

Model	*m*	*h*	Dist(MA)	Dist(SDM)	t-stat	N	Dir(MA)	Dir(SDM)
$\mathbf{W}_{i,N_t}$	80	40	0.161	0.112	(-49.597)	14322	55.09%	66.51%
$\mathbf{W}_{i,t-h}$			0.155	0.124	(-26.182)	10509	55.65%	61.08%
$\mathbf{W}_{i\in I,N_t}$			0.161	0.121	(-40.39)	14322	55.09%	60.28%
$\mathbf{W}_{i\in I,t-h}$			0.155	0.122	(-28.837)	10509	55.65%	62.25%
$\mathbf{W}_{i,N_t}$	60	40	0.172	0.117	(-53.461)	16182	58.02%	64.00%
$\mathbf{W}_{i,t-h}$			0.17	0.131	(-32.744)	12369	56.54%	58.11%
$\mathbf{W}_{i\in I,N_t}$			0.172	0.121	(-48.924)	16182	58.02%	59.30%
$\mathbf{W}_{i\in I,t-h}$			0.17	0.124	(-38.489)	12369	56.54%	59.85%
$\mathbf{W}_{i,N_t}$	80	20	0.221	0.188	(-31.514)	16182	54.24%	61.83%
$\mathbf{W}_{i,t-h}$			0.226	0.211	(-13.851)	14229	52.92%	55.01%
$\mathbf{W}_{i\in I,N_t}$			0.221	0.199	(-23.056)	16182	54.24%	56.53%
$\mathbf{W}_{i\in I,t-h}$			0.226	0.204	(-21.154)	14229	52.92%	56.53%
$\mathbf{W}_{i,N_t}$	60	20	0.236	0.188	(-41.404)	18042	54.51%	60.21%
$\mathbf{W}_{i,t-h}$			0.236	0.21	(-22.357)	16089	54.68%	51.84%
$\mathbf{W}_{i\in I,N_t}$			0.236	0.196	(-35.882)	18042	54.51%	54.72%
$\mathbf{W}_{i\in I,t-h}$			0.236	0.2	(-30.087)	16089	54.68%	54.97%

Table 4: Comparing forecast accuracy: The table summarises the results of comparing the SDM based forecasts with their moving average counterparts. Each of the SDM models are listed on the left hand side, the summary statistics document the results of applying different tests of forecast accuracy to the results of the SDM model vs. the moving average using the same values of *m* and *h*. We calculate the summary statistics by taking distances for each of the forecasts over all stocks and all time-periods and pooling these into one test. The t-statistics are then calculated as the mean of the pooled distance values for the moving average forecast minus the SDM forecast, so that the large negative results indicates that SDM forecasts had much smaller distance errors than moving average forecasts. The direction statistics are calculated by simply counting the number of times the pooled forecast results and the same sign as the observed values of *r*, then dividing by the total. Again we see that SDM forecasts the direction of the relationship much better than the moving average model.

5.4.2. Trading strategy

Given the results in table 4 we would expect the SDM models to produce higher returns based on the same type of trading strategy we use in chapter three. As the strategy we employ here is the same as in chapter three, we include only a brief description here as a recap. At the start of each trading day, we sort the stocks based on the results of the SDM forecast for the next *h* trading days. We then form a portfolio of the top 20 or 30 sentiment winners, that is the 20 or 30 stocks with the highest SDM forecast, and purchase these stocks. We then hold these stocks for the full *h* trading day period before selling them irrespective of whether we lose or gain by the sale. On the following trading day we repeat the process, purchasing a further 20 or 30 stocks for our portfolio, as a result, at any point in time we have 20*h* open positions, although any individual stock may be included in up to *h* of these positions at one time. We report the return of the portfolio as the average daily

return of stocks we sell at time *t* over the period we held them for, so effectively the average realised daily return of all positions we opened as time *t* - *h*.

Table 5 summarises the results from this strategy for *m* = 60, *m* = 80 and *h* = 20, *h* = 40, excess return figures and t-statistics are calculated based on the return from holding an un-weighted portfolio of the 93 stocks in our reduced SP100 index. First, we see that the simple moving average portfolio is outperformed by the SDM portfolios in all cases; this is as expected given the increased predictability we observed from table 4.

In terms of the expected return we might access from forming an actual trading strategy around the SDM results, the more realistic estimate of strategy performance is given by the rolling estimates, since these do not suffer from the same look-ahead bias as the final distribution strategies. Of these, the best preforming example is the 20 stock portfolio based on the average distribution over all stocks. Here, we see a total return of 0.137 log percent per day, based on a holding period of *h* = 40 trading-days and an averaging period of *m* = 80 trading-days. Viewing this result in light of reasonable costs, Clarkson et al. (2006) report that the average cost of realising a trade via an online broker 0.2% of the value of the trade. Taking this into account, the average log return per stock, per trade from the SDM strategy equates to $0.00137 * 40 = 0.0548$. subtracting transaction costs of $log(1.002)$ gives a log return per stock, per trade minus costs of 0.0528, or $0.0528/40 = 0.00132$ per stock, per day, resulting in a projected APR of $100\left(e^{(0.00157*250)} - 1\right) = 39.099\%$. Performing the same calculation on the excess return we see a log return per day minus costs of 0.076% equating to a 19.424% APR for the excess return.

While these results at first appear very high, there are sound reasons why this may be the case: *firstly*, it is logical that, since a simple average can be used to form profitable trading strategies as we showed in chapter three, an optimal average provided via the SDM algorithm should generate superior results. *Secondly*, in the momentum literature, Daniel and Moskowitz (2011) have shown that by augmenting a momentum trading model with a parameter that allows for positive-negative regime switches returns to momentum strategies double over long periods. This serves as an example of how accounting for extra variability, as we have done here via the use of SDM, can lead to dramatic improvements in portfolio performance.

			20 Stock Portfolio			30 Stock Portfolio		
	m	*h*	**Return**	**Excess**	**t-stat**	**Retrurn**	**Excess**	**t-stat**
MA	80	40	0.06	0.026	(7.711)	0.063	0.029	(10.432)
$\mathbf{W}_{i,N_t}$			0.149	0.115	(39.418)	0.128	0.095	(38.986)
$\mathbf{W}_{i,t-h}$			0.126	0.066	(17.65)	0.119	0.058	(20.254)
$\mathbf{W}_{i\in I,N_t}$			0.127	0.093	(30.9)	0.11	0.077	(31.27)
$\mathbf{W}_{i\in I,t-h}$			0.137	0.076	(20.708)	0.116	0.055	(18.423)
MA	60	40	0.06	0.019	(6.419)	0.059	0.019	(7.608)
$\mathbf{W}_{i,N_t}$			0.134	0.094	(34.595)	0.119	0.079	(34.241)
$\mathbf{W}_{i,t-h}$			0.09	0.024	(7.242)	0.085	0.019	(7.072)
$\mathbf{W}_{i\in I,N_t}$			0.115	0.074	(24.787)	0.105	0.065	(28.451)
$\mathbf{W}_{i\in I,t-h}$			0.116	0.05	(13.942)	0.101	0.035	(12.878)
MA	80	20	0.057	0.017	(3.661)	0.056	0.016	(4.074)
$\mathbf{W}_{i,N_t}$			0.152	0.111	(25.609)	0.132	0.092	(26.007)
$\mathbf{W}_{i,t-h}$			0.093	0.04	(8.526)	0.079	0.026	(6.772)
$\mathbf{W}_{i\in I,N_t}$			0.116	0.075	(17.191)	0.1	0.059	(16.678)
$\mathbf{W}_{i\in I,t-h}$			0.11	0.057	(11.875)	0.099	0.046	(11.693)
MA	60	20	0.042	-0.007	(-1.499)	0.043	-0.006	(-1.554)
$\mathbf{W}_{i,N_t}$			0.142	0.093	(22.376)	0.124	0.075	(22.384)
$\mathbf{W}_{i,t-h}$			0.077	0.016	(3.526)	0.07	0.01	(2.55)
$\mathbf{W}_{i\in I,N_t}$			0.111	0.063	(14.461)	0.101	0.053	(15.354)
$\mathbf{W}_{i\in I,t-h}$			0.115	0.054	(11.736)	0.102	0.041	(11.291)

Table 5: Returns to buying SDM winners: The table shows the results of the SDM and moving average based trading strategies. The different models are specified on the left hand side, MA denotes the moving average model the others are the different specifications of SDM model. We see that in all cases SDM outperforms a simple average, most of the time the performance increases by several times. We also see that each of the SDM portfolios is statistically significant with very high t-statistics. Generally, as would be expected the 20 stock portfolios outperform the 30 stock portfolios. We also see that in general the SDM portfolios generated using the final time-periods distribution tend to outperform the rolling distributions for individual stocks, but also that the rolling distribution taken over all stocks outperforms the stock specific rolling distribution. We think this may be because there is either not enough data, or too much noise in the sample to estimate stock specific distributions accurately in most cases. t-statistics are calculated for each portfolio against the benchmark of holding the market portfolio.

5.5. Examining the structure of returns predictability

The results from section 5.4 strongly suggest that SDM forecasts better than a simple average and can be used to generate profitable portfolios. What we do not see from the statistics of forecast accuracy is exactly where this extra predictability comes from. SDM is a methodological upgrade on a simple moving average because it considers stock specific distributions over the lagged values of the sentiment series, and it allows these distributions to change over time. In light of the arguments we make about variability in the introduction and in chapter three, it could be that there are latent factors affecting how words are interpreted in the context of a specific stock. Alternatively, latent factors could be affecting how words are interpreted at different times during the sample period.

In this section we introduce a hypothesis testing framework for assessing which of these factors contributes most to the increased predictability we document in section 5.4 For this section we consider the results from the estimator where the maximum lag length is $m = 80$ and where the holding period $h = 40$, as these correspond to some of the better results we observed in section 5.4.

5.5.1. Analysis preliminaries

In defining our hypothesis testing framework, we take advantage of the fact that, since each $\mathbf{W}_{i,t}$ is a conditional probability distribution, we can define the four dimensional joint distributions of both dimensions of the matrix, different stocks and time periods as

$$Pr(k, \varphi, i, t) = \mathbf{W}\, Pr(t)\, Pr(i) = \frac{\mathbf{W}}{(N_t+1)N_i} \tag{93}$$

where k is the lag length dimension, φ is the positive or negative sign, t it the time dimension and i is the stock dimension of the distribution. This characterisation allows us to define a number of marginal distributions in terms of their log odds ratio. This is important because the properties of the ratio allow for the construction of t-statistics for hypothesis testing purposes.

To do this we first define the log odds ratio function *LOR* as the natural logarithm of the positive half of the joint distribution given in equation (93) divided by the negative

half, that is

$$LOR(k,i,t|\varphi=1) = log\left\{\frac{Pr(k,i,t|\varphi=1)}{Pr(k,i,t|\varphi=0)}\right\} \tag{94}$$

The importance of this definition is that *LOR* has a well-defined mean and variance and so should tend to the normal distribution asymptotically (Agresti 2003). Further, if the two series are uncorrelated we would expect the mean of this distribution to be 0. With these properties defined, we can treat the log odds ratio of different marginal distributions as we would a normally distributed sample in a standard z-test.

5.5.2. Evidence of fixed lag distributions

The first aspect of the relationship we examine is the SDM equivalent of the cross-correlation plots shown in figures 1 to 4 from chapter three. As this distribution is taken across all stocks it is conceptually the average lagged oscillation in the sentiment-price relationship across the 93 stocks in our reduced S&P100 index. We examine this distribution again to see if we can confirm the results we reported in chapter three using SDM. The log odds ratio in this case is then

$$t\_statistic(k|\, i \in \mathbf{i}, t \in \mathbf{t}, \varphi = 1) = \frac{\frac{1}{N_i(N_t+1)}\sum_{s=0}^{s=N_t}\sum_{j=1}^{j=N_i} LOR(k,j,s|\varphi=1)}{\sigma_{LOR(k|t\in\mathbf{t},i\in\mathbf{i},\varphi=1)}/\sqrt{N_i(N_t+1)}} \tag{95}$$

where the numerator takes the average log odds ratio for each lag, over all stocks and all time periods. The denominator is simply the standard error of the numerator where σ represents the standard deviation. All equation (95) really states is that if you take the average over all stocks and time periods for each lag length *k*, then divide this by the standard error of the mean to retrieve the t-statistic for a given lag.

Figure 7 plots the t-statistics for each lag length *k* calculated using equation (95). The y-axis of figure 7 holds the t-statistic, and the x-axis holds the lag-length *k*. What we see is that there does appear to be a statistically significant fixed distribution over all stocks. The shape of this distribution shows a negative initial relationship between sentiment and prices, followed by a reversal at a three-month or greater lag length, which

tallies approximately with the distribution of t-statistics we documented in chapter three figure 9. In terms of the available literature, this is similar to the short term reversal effect documented by Jegadeesh (1990), although the length of the reversal period is slightly longer than the one month period documented in these studies.

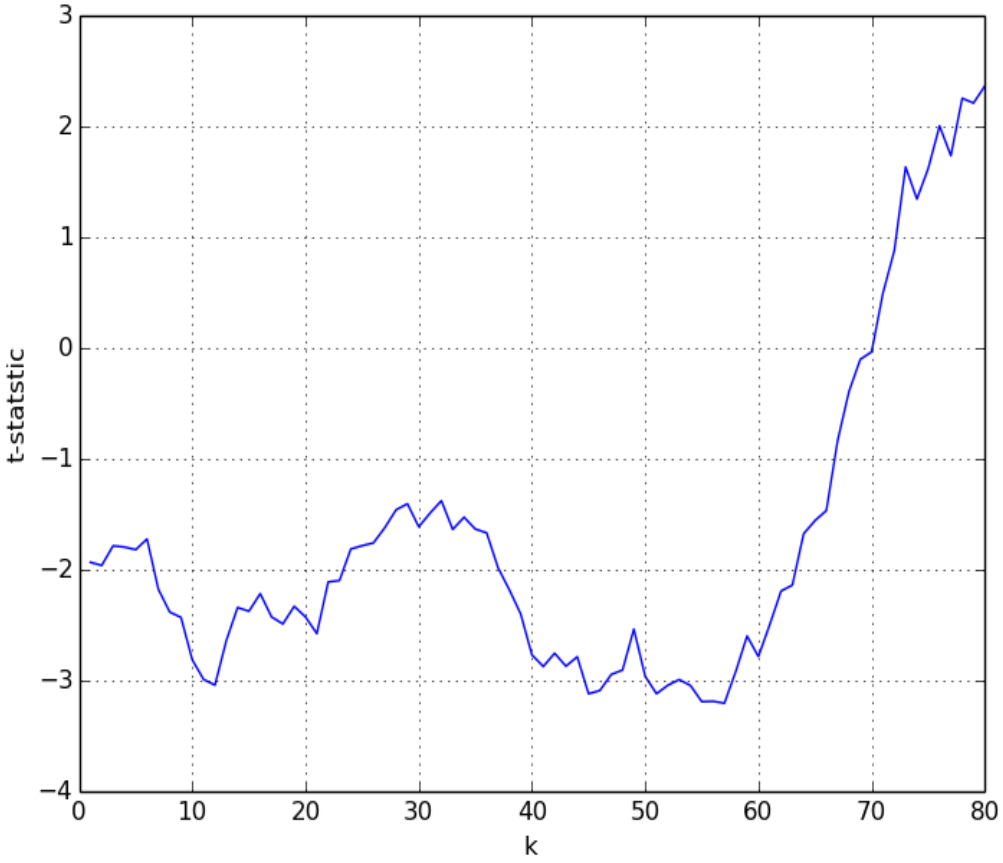

Figure 7: Cross correlation plot: the plot shows the average lagged relationship between all stocks and their corresponding sentiment series. The plot is very similar to the cross-correlation plot shown in chapter three figure 3. The x-axis holds the lagged values of the sentiment series and the y-axis holds the value of the t-statistic calculated using equation (95). We see that there are similar oscillations in the relationship to the ones we observed in chapter three.

5.5.3. Lag distributions across stocks

The second aspect of the relationship we consider is whether there is significant variability in the form of the lagged distribution across stocks. To analyse this property in detail we consider the t-statistics of a given lag and stock over time, that is

$$t\_statistic(k,i|t \in \mathbf{t}, \varphi = 1) = \frac{\frac{1}{(N_t+1)}\sum_{s=0}^{s=N_t} LOR(k,i,s|\varphi = 1)}{\sigma_{LOR(k,i|t\in\mathbf{t},\varphi=1)}/\sqrt{(N_t+1)}} \tag{96}$$

which is identical to the calculation in (95), except we take the mean and the standard error of *k* and *i* over all values of *t*.

Figure 8 shows a heat-map of the values of (96). On the y-axis are listed all of the stocks we consider in this study, the x-axis show the different lag-lengths *k*. Each horizontal row contains the distribution over the lags, the corollary of the distribution shown in figure 7. The colours show the magnitude of the t-statistic with the scale given in the bar on the right hand side of the plot.

What we see immediately from figure 8 is the t-statistics are much higher than those reported in figure 7, indicating that the stock specific distribution is a much stronger indicator of prices than the aggregate distribution over all stocks. We also see that there is a large range of different shaped distributions; these range from stocks like Ebay (EBAY) and Boeing (BA), which have similar shapes to figure 7, to stocks such as Twentieth Century Fox (FOXA) and Texas Instruments (TXN) which show the reverse, a positive relationship followed by a negative one. Other stocks, such as McDonalds (MCD) and MetLife (MET), show distributions that have several swings from positive to negative.

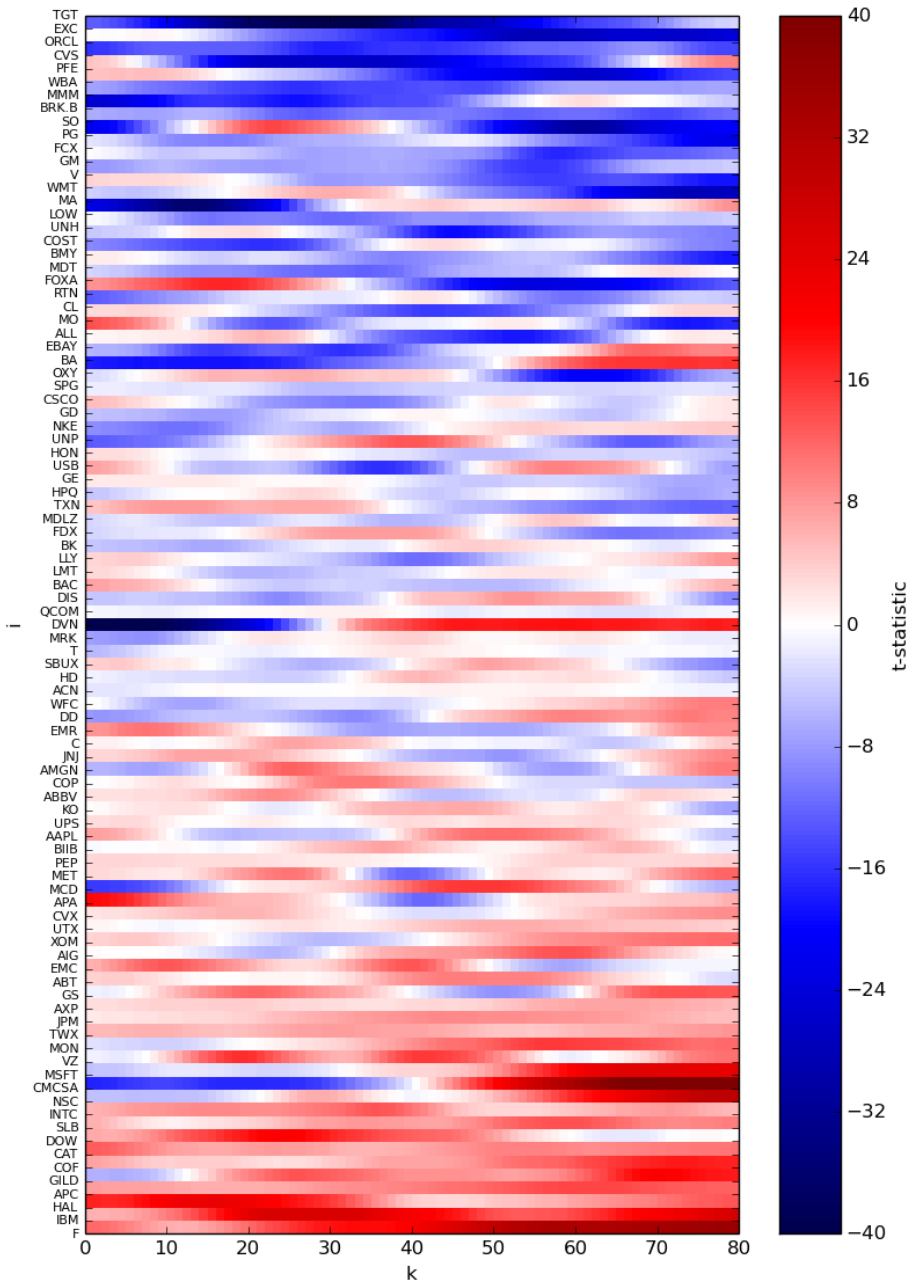

Figure 8: Stock specific lag distributions: the plot shows a heat-map of the distribution shown in figure 1 for each of the 93 stocks in the sample. The y-axis records the ticker of each of the stocks and the x-axis records the lag length. The colour-bar on the right hand side shows the colour associated with a given t-statistic value, dark blue being a strong negative relationship, and dark red being a strong positive relationship. T-statistics are calculated using equation (96). We see that there is significant variability in the form of the distributions across stocks, and that the t-statistics are very high in many cases.

5.5.4. Lag distributions over time

Figure 8 shows that there is significant variability in the lag distributions over stocks. Another potential source of variability is over time. To test for this we consider whether there are statistically significant changes in the sign of the distribution over time, considering the t-statistic

$$t\_statistic(t,i|k \in \mathbf{k}, \varphi = 1) = \frac{\frac{1}{(m+1)}\sum_{v=0}^{v=m} LOR(v,i,t|\varphi = 1)}{\sigma_{LOR(i,t|k\in\mathbf{k},\varphi=1)}/\sqrt{(m+1)}} \tag{97}$$

which is again identical to (95) and (96), just now the mean and standard errors are taken over the values of *k*, where *m* refers to the maximum lag considered in the analysis.

Figure 9 shows the t-statistics from (96) displayed on a similar heat-map to figure 8, just with the x-axis now containing the value of *t*, rather than the lag-lengths *k*. We see from the magnitude of the t-statistics in figure 9 that the larger values are about half the magnitude of the t-statistics in figure 8, but still much higher than those in figure 7. We see, in general, that stock prices have either a positive or negative relationship to sentiment, Ford (F) for example, shows a strong, consistent, positive relationship between sentiment and price. On the other hand Target Group (TGT) shows a strong, consistent, negative correlation between sentiment and prices over time. There are, however, some more complex cases, Apache Corp. (APA) or Emerson Electric (EMR) both show relationships that oscillate from positive to negative over time.

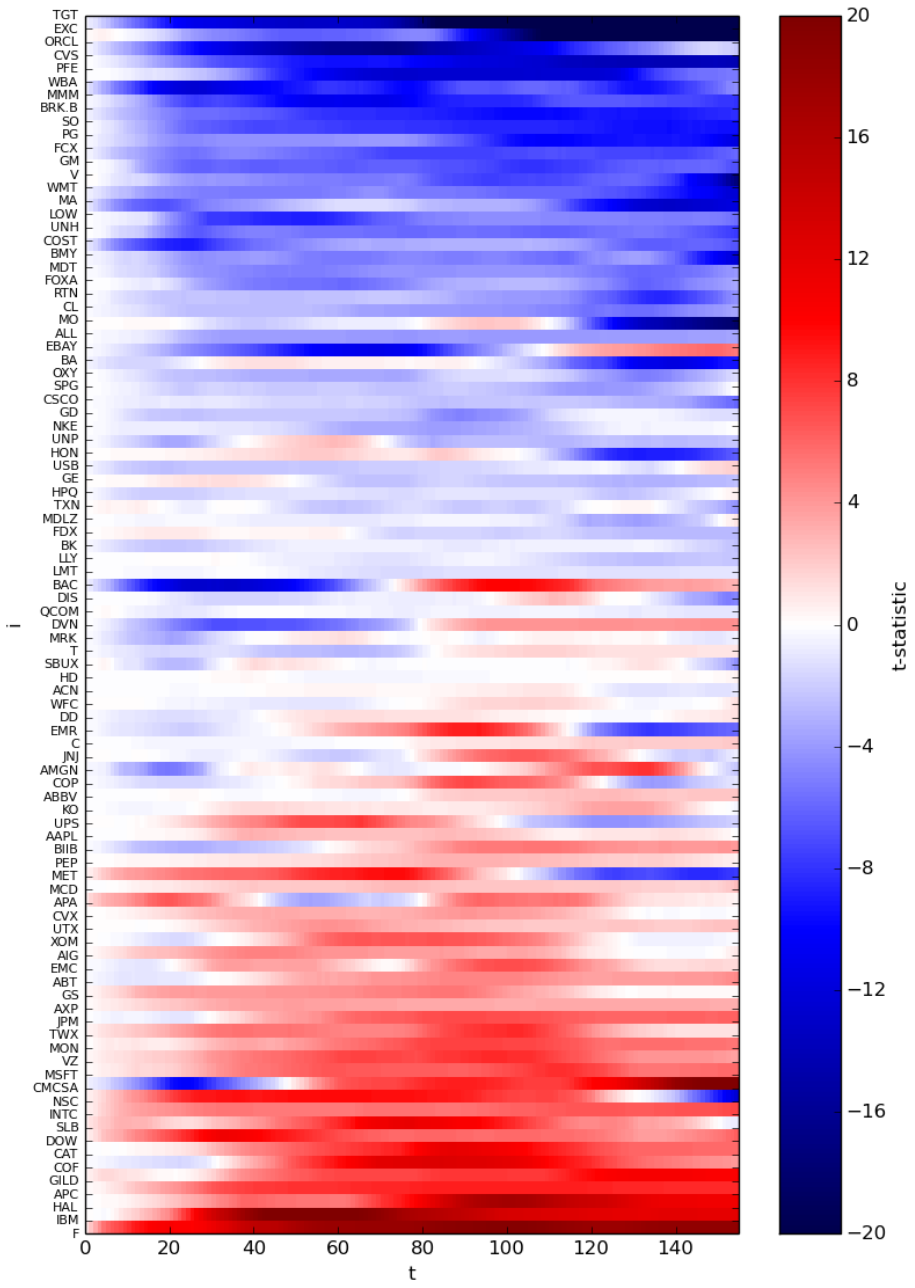

Figure 9: Changes in distributions over time: the plot shows whether the relationship between a stock and the lagged values of the sentiment series is predominantly positive or negative, calculated using equation (97). The y-axis records the ticker of each of the stocks and the x-axis records the lag length. The colour-bar on the right hand side shows the colour associated with a given t-statistic value, dark blue being a strong negative relationship, and dark red being a strong positive relationship. We see from the plat that the t-statistics are in general lower than in figure 2 although still highly significant. There are also a number of cases where the sign of the correlation changes over time from positive to negative.

5.6. Conclusions, limitations and future work

We argued in the introduction that there are latent factors that are likely to influence the way the information measured in sentiment metrics are incorporated into prices. The crux of this issue is that there are a large number of factors which affect a person's reading of a text, for example, how they feel at the time, what they think of the author, or their understanding of other factors not explicitly referenced in the text that affect its intended meaning. The potential problem this poses when attempting to model the sentiment-price relationship is that these factors are not explicit in the actual words in the text, and so subsequently latent from any measure of textual sentiment. The result is that the time series relationship between sentiment and price is likely to be affected by significant sources of latent variability. The methodologies that have been applied in this area to date by authors, such as Antweiler and Frank (2004), Das and Chen (2007), Sprenger et al. (2014), Chen et al. (2014) and in our own work in chapter three, are not capable of modelling latent variability of this type.

To bridge this gap, in this paper we show that introducing a new time series analysis methodology, SDM, to the study of the sentiment-price relationship which allows us to model some of this variability. We show that SDM significantly outperforms the methodology we used in chapter three and allows for the construction of sentiment trading strategies that generate significant excess returns. We document that the lagged relationship between sentiment and price is specific to each stock, and also changes over time. We consider this evidence that there are contextual factors which cause changes in the way which words are interpreted in with respect to a particular stock at a particular point in time. Subsequently, we consider this to be evidence in support of our central hypothesis, which is that there are significant sources of latent variability which characterise the sentiment-price relationship.

Chapter 6: Conclusions

This thesis has two aims: *firstly*, to examine whether sign-strength-timing variability is a significant factor influencing the online textual sentiment-price relationship. *Secondly*, to introduce methodological innovations that can inform future work in this area by allowing this variability to be accurately modelled. I am going to open this section by discussing the major findings of this thesis in this regard.

The evidence I present in chapter three and chapter five gives strong indication that there is sign-strength-timing variability present in the online textual sentiment-price relationship. In chapter three I document how the relationship between a lagged sentiment indicator and prices forms an oscillating pattern on a cross-correlation plot. Some of these oscillations are strong enough to be statistically significant. This result suggests that for all stocks, sign-strength-timing variability does exist. The large performance increases from using the SDM estimator documented in chapter five support this conclusion. In chapter five, however, there is evidence from the odds ratios plots that not only are the sign, strength and timing of the relationship variable with respect to each other, but these factors are also variable with respect to time. The success of the SDM estimator in forecasting the sentiment-price relationship shows that this variability over time must be slow enough to have been captured by the estimator.

Considering this in terms of the available literature; Chen et al. (2014) provide the only other evidence of long run relationships between sentiment and price available in the literature to-date. The authors consider the relationship between sentiment and cumulative abnormal returns at time horizons of up to three years. They classify two forms of sentiment, sentiment expressed in articles, and sentiment expressed in comments on articles, for the financial content site Seeking Alpha. They find that there is a monotonically increasing positive correlation between sentiment and abnormal returns for both article and comments content. Interestingly, for comments content they show some oscillatory structure in the lagged relationship between sentiment and price[10], similar to the

[10] This oscillatory structure is shown in figure 3 panel B of Chen et al.'s paper. Please note, this is my inference, the authors do not discuss oscillatory structure in the paper. From the graph there is an initial

oscillatory structure I report in chapter three. The key difference between the findings of Chen et al. and my findings in chapter three is that the sign of the relationship is different in either case. This is explicable due to my findings in chapter five that sign-strength-timing variability can also itself change over time.

The picture this paints of the sentiment-price relationship is one where investor interpretations of the semantic meaning in text are significantly complex. I have argued throughout this thesis that this is to be expected, because much of the contextual information an investor will be using to judge the semantic content of a piece of text is latent from existing sentiment measures. I am not making any claim here that I know, or have shown, what causes this variability specifically. Instead, I have given three broad examples to show potential driving factors: *variability in interpretation*, *variability in authorship* and *variability in context*, but there may be more. What I aim to demonstrate with these examples is that some of these factors are likely to be difficult or impossible to find suitable proxies for. As a result, the variability I document in chapter three and chapter five should be considered a feature of the relationship that needs modelling in its own right, rather than an indication that more variables need to be found to measure different features of the relationship.

This being true, the second aim of this thesis is to provide methodological innovations for modelling relationships of this type. To this end I have introduced SDM in chapter four. SDM is a good fit for problems of this type because it is capable of modelling sign-strength-timing variability, providing the inter-relationships between these factors change relatively slowly over time. This assumption is particularly intuitive in the case of the sentiment-price relationship because the variability in the relationship is a function of how investors interact with information. Imagine if this relationship did not change slowly over time but was stochastic, this would indicate that a group of investors could interpret that same words completely differently each time they read a piece of content. In effect, we would be saying that the semantic meaning in the content is stochastic, that words do not carry any fixed meaning. Clearly this does not make sense; words do carry meaning that most people interpret in a similar way. The result is that semantic meaning in text cannot be stochastic, but the evidence suggests it is variable. The conclusion I draw is that this variability must be relatively slow. As a result, SDM provides a key methodological

---

downwards curve between 1 month and 3 months, followed by an upwards curve between 3 months and 6 months, followed by a downwards curve again.

innovation for further work in this area. I have extended this work in chapter five, by showing how the output of the estimator can be used to develop a powerful hypothesis testing framework for different types of variability.

On a more practical level, a key contribution this thesis makes to the literature is to show that the online textual sentiment-price relationship is potentially much stronger than previous studies suggest. I have documented trading strategies, both using SDM and a simpler model, that return well in excess of the market benchmark once reasonable costs have been accounted for. How accessible are these returns in practice? There are some overheads that may prevent these strategies from becoming commonplace in the near future. For example, there are costs associated with purchasing large datasets of the type I used in this thesis. I would estimate that yearly running costs for acquiring datasets of this size would be between £50,000 and £100,000. This is clearly not enough to put off a reasonable size fund or bank, but may be too much for many smaller investors.

The fact that there is material information about a stocks future price contained in online messages also has implications for both policy makers and financial theorists. Since the advent of the web there has been an exponentially increasing quantity of information available online. Services, such as Google search, have been developed to address the fact that a person cannot be expected to sift the vast quantities of available information by themselves. From the sample I have gathered for this study it is clear that there is also too much financial content available online to reasonably expect any person to read it. This means that there is now a gap in the knowledge discovery process surrounding price of a brokerage service between raw content and investors. This service is in some ways nothing new, analysts already produce reports for investors and traders to help them make allocation decisions. What is new is that online services are explicitly aimed at investors themselves rather than a third party information provider.

The key issue with this from a policy perspective is that insider trading law explicitly forbids investors from making money from material non-public information. As the people creating the information about future stock prices are also profiting from the interpretation of this information this may be illegal under U.S. law. It also illustrates the clear conflict of interest between users of online services as authors of material information and investors. This is problematic from a theoretical perspective because there is no theoretical explanation as to why individuals would reveal valuable information to other

investors. By nature, revealing information to the market in this way should reduce its value and so reduce the original holder of the information's profits.

The result is two equally unpalatable options, either: investors could be going online to share material information at the expense of their profits. This would mean that theories assuming the presence of rational, profit maximising investors are incorrect. The second option is that investors are making profits from sharing information they create. This would mean that there is large scale criminality occurring online in these communities.

This issue mimics issues the web has created in a number of other industries, particularly the publishing industry. As the web enables massive scale sharing of information at low-no cost laws that rely on there being structural barriers to information sharing are difficult to implement effectively. In the case of sharing financial information online, tracking down all of the users of online services to see if they have profited from information sharing would be incredibly difficult and so it is unlikely there will be mass arrests for online insider trading in the near future.

I would consider a more likely way forward is for the both researchers and policy makers to relax their assumptions about the information-price relationship. Both U.S. insider trading law and notions of rational investors have come from an assumption that stock price moves based on material 'facts' about a company. Increasingly, with each new financial crisis we become more aware of the fact that this is a tenuous argument at best. Once you do away with this idea conceptually the movement of prices and the behaviour of investors online become much clearer. If Keynes is correct and prices move based on a recursive 'beauty contest' where each investor attempts to guess what other investors think, then the real source of price relevant information is contained in the thoughts, writings and conversations of investors. In this case we would expect a good investor to also be a master of building social relationships, and of guessing or manipulating other investors' intentions through communication. If this is that case then it may be better to develop a legal framework for the markets that is forgiving of this aspect of behaviour, rather than attempting to force an alien structure on investor behaviour.

In summation, I consider this thesis to show evidence that there is good predictability between online textual sentiment and price. A key issue with modelling this relationship is that current sentiment analysis methodologies are a long way from being

able to classify content with the contextual understanding a person is capable of. To counter this issue, it is possible to use SDM to model the variability caused by a sentiment measures lack of contextual awareness.

As a final remark, there are several other fields of study which consider the relationship between textual data and some other variable. Examples include, predicting elections with social media sentiment measure (Tumasjan et al. 2010) or predicting box office revenues for films using the sentiment of online content (Asur and Huberman 2010). Clearly these literatures will face many of the same issues as the financial literature, so the methodological innovations I present in this thesis are applicable to a broad range of literatures.

References

Agresti, A., 2003. Asymptotic Theory for Parametric Models. In *Categorical Data Analysis*. pp. 576–599.

Alsayed, H. and McGroarty, F., 2014. Ultra High Frequency Statistical Arbitrage Across International Index Futures. *Journal of Forecasting*, 33(6), pp.391–408. Available at: http://papers.ssrn.com/abstract=2225753.

Antweiler, W. and Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59(3), pp.1259–1294. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2004.00662.x/full [Accessed November 15, 2011].

Arulampalam, M.S. et al., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), pp.174–188. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=978374.

Asness, C.S. et al., 2014. Fact , Fiction and Momentum Investing. *Journal of Portfolio Management*, 2014, pp.1–26.

Asness, C.S., Moskowitz, T.J. and Pedersen, L.H., 2013. Value and Momentum Everywhere. *The Journal of Finance*, 68(3), pp.929–985. Available at: http://doi.wiley.com/10.1111/jofi.12021 [Accessed May 23, 2014].

Asur, S. and Huberman, B. a., 2010. Predicting the Future with Social Media. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 1.

Baker, M. and Wurgler, J., 2007. Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2), pp.129–151. Available at: http://www.nber.org/papers/w13189.

Barberis, N., Shleifer, A. and Vishny, R., 1998. A model of investor sentiment. *Journal of Financial Economics*, 49(3), pp.307–343. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0304405X98000270\nhttp://www.sciencedirect.com/science/article/pii/S0304405X98000270.

Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp.1–8. Available at: http://www.sciencedirect.com/science/article/pii/S187775031100007X.

Brock, W., Lakonishok, J. and LeBaron, B., 1992. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *Journal of Finance*, 47(5), pp.1731–1764.

Available at: http://proxy2.hec.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4653571&site=bsi-live.

Buehlmaier, M.M.M., 2013. The role of the media in takeovers: Theory and evidence. *SSRN Electronic Journal*.

Cambria, E. et al., 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2), pp.15–21.

Carhart, M.M., 1997. On persistence in mutual fund performance. *The Journal of Finance*, 52(1), pp.57–82.

Carretta, A., Farina, V. and Graziano, E., 2011. Does investor attention influence stock market activity? The case of spin-off deals. *SSRN Electronic Journal*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1930861.

Carvalho, C., Lopes, H. and Polson, N., 2009. Particle learning for generalized dynamic conditionally linear models. , pp.1–19. Available at: http://apps.olin.wustl.edu/faculty/conferences/sbies2009/uploads/Lopes_Hedibert.pdf [Accessed August 29, 2014].

Carvalho, C.M. and Lopes, H.F., 2007. Simulation-based sequential analysis of Markov switching stochastic volatility models. *Computational Statistics and Data Analysis*, 51(9), pp.4526–4542. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167947306002349 [Accessed August 29, 2014].

Chen, H. et al., 2014. Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Rev. Financ. Stud.*, 27(5), p.hhu001–. Available at: http://web.ics.purdue.edu/~bhwang/wisdom of crowds.pdf [Accessed February 10, 2014].

Cheung, Y.W., Chinn, M.D. and Marsh, I.W., 2004. How do UK-based foreign exchange dealers think their market operates? *International Journal of Finance and Economics*, 9(4), pp.289–306.

Clarkson, P.M., Joyce, D. and Tutticci, I., 2006. Market reaction to takeover rumour in internet discussion sites. *Accounting and Finance*, 46, pp.31–52.

Cogent Research, 2008. *Social Media's Impact on Personal Finance and Investing*,

Daniel, K., Hirshleifer, D. and Subrahmanyam, A., 1998. Investor Psychology and Security Market Under- and Overreactions. *The Journal of Finance*, LIII(6).

Daniel, K., Hirshleifer, D. and Teoh, S.H., 2002. Investor psychology in capital markets: evidence and policy implications. *Journal of Monetary Economics*, 49(1), pp.139–209. Available at: <Go to ISI>://000178277200010.

Daniel, K. and Moskowitz, T., 2011. Momentum Crashes. *CBS Working Paper*, pp.1–38. Available at:

http://www.columbia.edu/~kd2371/papers/unpublished/mom4.pdf\npapers2://publication/uuid/8CC56886-0412-41E7-B03A-BB0FE59568D9.

Das, S.R. and Chen, M.Y., 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9), pp.1375–1388. Available at: http://algo.scu.edu/~sanjivdas/chat_FINAL.pdf [Accessed November 15, 2011].

Datamonitor, 2010. *Social media in financial services: The customer as the advisor*

Davis, A.K., Piger, J.M. and Sedor, L.M., 2012. Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language. *Contemporary Accounting Research*, 29(3), pp.845–868.

Davis, A.K. and Tama-Sweet, I., 2011. Managers' Use of Pessimistic Tone Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A Angela K. Davis. *SSRN Electronic Journal*.

Deloitte, 2007. *Most customers read and rely on online reviews; companies must adjust*,

Demers, E. and Vega, C., 2011. Linguistic Tone in Earnings Press Releases: News or Noise? *Working Paper*, pp.1–65.

Diebold, F., 2015. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. *Journal of Business and Economic Statistics*, 33(1), p.16. Available at: http://www.nber.org/papers/w18391.

Diebold, F.X. and Mariano, R.S., 1995. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 20(1), pp.134–144.

Doran, J.S., Peterson, D.R. and Price, S.M., 2012. Earnings Conference Call Content and Stock Price: The Case of REITs. *Journal of Real Estate Finance and Economics*, 45(2), pp.402–434.

Engelberg, J. et al., 2008. Costly Information Processing : Evidence from Earnings Announcements. In *AFA 2009 San Francisco Meetings Paper*

Engelberg, J.E., Reed, A. V. and Ringgenberg, M.C., 2012. How are shorts informed?. Short sellers, news, and information processing. *Journal of Financial Economics*, 105(2), pp.260–278.

Fama, E. and French, K., 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*. Available at: http://www.sciencedirect.com/science/article/pii/0304405X93900235 [Accessed August 28, 2012].

Fama, E.F., 1970. Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 25, pp.383–417.

Fama, E.F., 1965a. Random Walks in Stock Market Prices. *Financial Analysts Journal*, 51(1), pp.75–80.

Fama, E.F., 1965b. The Behaviour of Stock Market Prices. *The Journal of Business*, 38(1), pp.34–105.

Fama, E.F. and French, K.R., 2004. The Capital Asset Pricing Model: Theory and Evidence. *The Journal of Economic Perspectives*, 18(3), pp.25– 46.

Feldman, R. et al., 2008. The Incremental Information Content of Tone Change in Management Discussion and Analysis.

Ferguson, N., Philip, D. and Guo, J.M., 2014. Media Content and Stock Returns: The Predictive Power of Press. *In Midwest Finance Association 2013 Annual Meeting Paper*, 86(February), pp.1–31. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2111352.

Ferris, S.P., Hao, Q. and Liao, M.Y., 2013. The effect of issuer conservatism on ipo pricing and performance. *Review of Finance*, 17(3), pp.993–1027.

Frazzini, A., 2006. The disposition effect and underreaction to news. *The Journal of Finance*, 61(4), pp.2017–2046.

García, D., 2013. Sentiment during Recessions. *Journal of Finance*, 68(3), pp.1267–1300.

Goldman Sachs, 2015. *The Future of Finance Part 3: The Socialization of Finance*,

Gordon, N., Salmond, D. and Smith, A., 1993. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal …*, 140, pp.107–113. Available at: http://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0015 [Accessed September 1, 2014].

Granger, C.W.J., 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3), pp.pp. 424–438. Available at: http://www.jstor.org/stable/1912791.

Grassi, M. et al., 2011. Sentic Web: A New Paradigm for Managing Social Media Affective Information. *Cognitive Computation*, 3(3), pp.480–489.

Grinblatt, M. and Han, B., 2005. Prospect theory, mental accounting, and momentum. *Journal of Financial Economics*, 78(2), pp.311–339.

Gutierrez, R.C. and Kelley, E.K., 2008. The long-lasting momentum in weekly returns. *Journal of Finance*, 63(1), pp.415–447.

Henry, E., 2008. Are Investors Influenced By How Earnings Press Releases Are Written? *Journal of Business Communication*, 45(4), pp.363–407.

Henry, E., 2006. Market Reaction to Verbal Components of Earnings Press Releases: Event Study Using a Predictive Algorithm. *Journal of Emerging Technologies in Accounting*, 3(1), pp.1–19.

Henry, E. and Leone, A.J., 2009. Measuring Qualitative Information in Capital Markets Research. *Social Science Research Network*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1470807.

Hiemstra, C. and Jones, J., 1994. Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation. *The Journal of Finance*, 49(5), pp.1639–1664. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1994.tb04776.x/abstract [Accessed August 28, 2014].

Hong, H., Kubik, J. and Stein, J.C., 2004. Social interaction and stock market participation. *The Journal of Finance*, 59(1), pp.137–163. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2004.00629.x/full.

Hong, H. and Stein, J.C., 1999. A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *The Journal of Finance*, 54(6), pp.2143–2184. Available at: http://www.blackwell-synergy.com/doi/abs/10.1111/0022-1082.00184.

Hong, Y., Liu, Y. and Wang, S., 2009. Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics*, 150(2), pp.271–287. Available at: http://www.sciencedirect.com/science/article/pii/S0304407608002248 [Accessed August 29, 2014].

Huang, A., Zang, A. and Zheng, R., 2012. Large Sample Evidence on the Informativeness of Text in Analyst Reports. *SSRN Electronic Journal*, pp.1–57. Available at: http://www.researchgate.net/publication/228259513_Large_Sample_Evidence_on_the_Informativeness_of_Text_in_Analyst_Reports/file/32bfe5100734e6e9e4.pdf.

Ivković, Z. and Weisbenner, S., 2007. Information diffusion effects in individual investors' common stock purchases: Covet thy neighbors' investment choices. *Review of Financial Studies*, 20(4), pp.1327–1357.

Jagannathan, R. and Wang, Z., 1996. The Conditional CAPM and the Cross-Section of Expected Returns. *Journal of Finance*, 51(1), pp.3–53.

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. *The Journal of Finance*, 45(3), pp.881–898. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1990.tb05110.x/abstract.

Jegadeesh, N. and Titman, S., 1993. Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *The Journal of Finance*, 48(1), p.65. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1993.tb04702.x/full [Accessed March 10, 2015].

Jegadeesh, N. and Wu, D., 2013. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3), pp.712–729. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1787273 [Accessed February 10, 2014].

Kahneman, D. and Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), pp.263–292. Available at: http://www.princeton.edu/~kahneman/docs/Publications/prospect_theory.pdf.

Kearney, C. and Liu, S., 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33(Cc), pp.171–185. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1057521914000295 [Accessed October 31, 2014].

Keenan, D., 2012. My thwarted attempt to tell of Libor shenanigans. *The Financial Times*.

Keogh, E. and Pazzani, M., 1999. Scaling up dynamic time warping to massive datasets. *Principles of Data Mining and Knowledge …*, (Derriere). Available at: http://link.springer.com/chapter/10.1007/978-3-540-48247-5_1 [Accessed January 31, 2014].

Kothari, S.P., Li, X. and Short, J.E., 2009. The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. *The Accounting Review*, 84(5), pp.1639–1670. Available at: http://aaajournals.org/doi/abs/10.2308/accr.2009.84.5.1639.

Larcker, D.F. and Zakolyukina, A. a., 2012. Detecting Deceptive Discussions in Conference Calls. *Journal of Accounting Research*, 50(2), pp.495–540.

Liu, B. and McConnell, J.J., 2013. The role of the media in corporate governance: Do the media influence managers' capital allocation decisions? *Journal of Financial Economics*, 110(1), pp.1–17.

Lo, A.W., 2007. Efficient Markets Hypothesis. *SSRN Electronic Journal*, pp.1–28.

Lo, A.W., 2008. *Finance Theory I (Massachusetts Institute of Technology: MIT OpenCourseWare)*,

Lo, A.W., 2005. Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis. *Journal of Investment Consulting*, 7(2), pp.21–44. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1702447.

Lo, A.W., 2004. The Adaptive Markets Hypothesis. *The Journal of Portfolio Management*, 30(5), pp.15–29.

Lopes, H. and Tsay, R., 2011. Particle filters and Bayesian inference in financial econometrics. *Journal of Forecasting*, 30(1), pp.168–209. Available at: http://onlinelibrary.wiley.com/doi/10.1002/for.1195/full [Accessed August 28, 2014].

Loughran, T. and McDonald, B., 2011a. Barron's Red Flags: Do They Actually Work? *Journal of Behavioural Finance*, 12(4), pp.90–97.

Loughran, T. and McDonald, B., 2013. IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, 109(2), pp.307–326.

Loughran, T. and McDonald, B., 2011b. When is a Liability not a Liability ? Textual Analysis , Dictionaries , and 10-Ks. *The Journal of Finance*, 66(1), pp.35–65.

Mandelbrot, B., 1966. Information theory and psycholinguistics: a theory of words frequencies. In P. Lazafeld and N. Henry, eds. *Readings in Mathematical Social Science*. Cambridge, Massachusetts: MIT Press.

Mangen, C. and Durnev, A., 2010. The real effects of disclosure tone: Evidence from restatements. *SSRN eLibrary*, 1(514), pp.1–55. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1650003.

Markowitz, H., 1952. Portfolio Selection. *The Journal of Finance*, 7(1), pp.77–91.

Markowitz, H.M., 1999. The Early History of Portfolio Theory: 1600-1960. *Financial Analysts Journal*, 55(4), pp.5–16.

McIntyre, D., 2009. Turning Wall Street on its head. *TIME magazine*.

Menkhoff, L. and Schmidt, U., 2005. The use of trading strategies by fund managers: some first survey evidence. *Applied Economics*, 37(21), pp.1719–1730. Available at: http://www.tandfonline.com/doi/abs/10.1080/00036840500217606 [Accessed March 4, 2015].

Merton, R.C., 1973. An Intertemporal Capital Asset Pricing Model. *Econometrica*, 41(5), pp.867–887.

Mollenkamp, C. and Whitehouse, M., 2008. Study Casts Doubt on Key Rate. *The Wall Street Journal*.

Moskowitz, T.J., Ooi, Y.H. and Pedersen, L.H., 2012. Time series momentum. *Journal of Financial Economics*, 104(2), pp.228–250.

Mullen, T. and Collier, N., 2004. Sentiment analysis using support vector machines with diverse information sources. *Proceedings of EMNLP-04, 9th Conference on Empirical Methods in Natural Language Processing*. Available at: http://edu.tsuda.ac.jp/~mullen/Papers/emnlp_corrected.pdf.

Pitt, M.K. and Shephard, N., 1999. Filtering via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association*, 94(446), pp.590–599. Available at: http://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474153.

Price, S.M. et al., 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking and Finance*, 36(4), pp.992–1011.

Rogers, J.L., Van Buskirk, A. and Zechman, S.L.C., 2011. Disclosure tone and shareholder litigation. *Accounting Review*, 86(6), pp.2155–2183.

Rosenthal, S. et al., 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*.

Sakurai, Y., Yoshikawa, M. and Faloutsos, C., 2005. FTW: fast similarity search under the time warping distance. *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 1, pp.326–337. Available at: http://dl.acm.org/citation.cfm?id=1065210 [Accessed June 3, 2013].

Senin, P., 2008. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu*, pp.1–23. Available at: http://spoken-number-recognition.googlecode.com/svn/trunk/docs/Dynamic time warping/08-04.pdf [Accessed July 29, 2014].

Sharkasi, A., Ruskin, H.J. and Crane, M., 2005. Interrelationships among international stock market indices: Europe, Asia and the Americas. *International Journal of Theoretical and Applied Finance*, 8(5), pp.603–622. Available at: http://www.worldscientific.com/doi/abs/10.1142/S0219024905003190 [Accessed August 29, 2014].

Sinha, N., 2010. Underreaction to news in the US stock market. *Available at SSRN*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1572614.

Sornette, D. and Zhou, W., 2005. Non-parametric determination of real-time lag structure between two time series: the "optimal thermal causal path"method. *Quantitative Finance*, (February 2008). Available at: http://www.tandfonline.com/doi/abs/10.1080/14697680500383763 [Accessed April 24, 2013].

Soros, G., 2009. *The new paradigm for financial markets. The credit crisis of 2008 and what is means*,

Sparck Jones, K., 1972. a Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28(1), pp.11–21.

Sprenger, T. et al., 2014. Tweets and Trades: The information content of stock microblogs. *European Financial Management*, 20(5), pp.926–957. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-036X.2013.12007.x/full [Accessed March 13, 2014].

Tetlock, P.C., 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3), pp.1139–1168. Available at: http://doi.wiley.com/10.1111/j.1540-6261.2007.01232.x.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S., 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, 63(3), pp.1437–1467. Available at: http://doi.wiley.com/10.1111/j.1540-6261.2008.01362.x.

Thaler, R. and De Bondt, W.F.M., 1985. Does the Stock Market Overreact? *Journal of Finance*, 40(3), pp.793–805. Available at: http://www.jstor.org/stable/2327804\nhttp://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1985.tb05004.x/full.

Tumasjan, A. et al., 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Munich. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852.

Twedt, B. and Rees, L., 2012. Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy*, 31(1), pp.1–21.

Urquhart, a. and Hudson, R., 2013. Efficient or adaptive markets? Evidence from major stock markets using very long run historic data. *International Review of Financial Analysis*, 28, pp.130–142.

Warren Liao, T., 2005. Clustering of time series data—a survey. *Pattern Recognition*, 38(11), pp.1857–1874. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0031320305001305 [Accessed October 31, 2012].

Zhou, W. and Sornette, D., 2007. Lead-lag cross-sectional structure and detection of correlated–anticorrelated regime shifts: Application to the volatilities of inflation and economic growth rates. *Physica A: Statistical Mechanics and its Applications*, (February 2008), pp.1–16. Available at: http://www.sciencedirect.com/science/article/pii/S0378437107001781 [Accessed May 6, 2013].