News media, asset prices and capital flows: Evidence from a small open economy

Galen Sher

Supervised by Dave Strugnell

Dissertation submitted in partial fulfilment of the requirements for the degree of Master of Business Science

11 March 2017

Department of Actuarial Science

Commerce Faculty

University of Cape Town
The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.
Abstract

Objectives

This work investigates the role for the content of print news media in determining asset prices and capital flows in a small open economy (South Africa). Specifically, it examines how much of the daily variation in stock prices, bond prices, trading volume and capital flows can be explained by phrases in the print news media. Furthermore, this work links such evidence to the existing theoretical and empirical literature.

Methods

This work employs natural language processing techniques for counting words and phrases within articles published in national newspapers. Variance decompositions of the resulting word and phrase counts summarise the information extracted from national newspapers in this way. Following previous studies of the United States, least squares regression relates stock returns to single positive or negative ‘sentiment’ factors. New in this study, support vector regression relates South African stock returns, bond returns and capital flows to the high-dimensional word and phrase counts from national newspapers.

Results

I find that domestic asset prices and capital flows between residents and non-residents reflect the content of domestic print news media. In particular, I find that the contents of national newspapers can predict 9 percent of the variation in daily stock returns one day ahead and 7 percent of the variation in the daily excess return of long-term bonds over short-term bonds three days ahead. This predictability in stocks and bonds coincides with predictability of the content of domestic print news media for net equity and debt portfolio capital inflows, suggesting that the domestic print news media affects foreign residents’ demand for domestic assets. Moreover, predictability of domestic print
news media for near future stock returns is driven by emotive language, suggesting a role for ‘sentiment’, while such predictability for stock returns further ahead and the premium on long-term bonds is driven by non-emotive language, suggesting a role for other media factors in determining asset prices. These results do not seem to reflect a purely historical phenomenon, finite-sample biases, reverse causality, serial correlation, volatility or day-of-the-week effects. The results support models where foreign agents' short-run beliefs or preferences respond to the content of domestic print news media heterogeneously from those of domestic agents, while becoming more homogeneous in the medium term.

Keywords: Asset pricing, capital flows, newspapers, machine learning, sentiment

JEL Classification: G12, G15, G17, C58
Chapter 1

Introduction

For many people, the mass media is their only source of news information. However, the news media have also been found to affect what people think is important [McCombs and Shaw, 1972]. It is therefore not surprising that the media has been found to influence voting behaviour [DellaVigna and Kaplan, 2007]. This ability to influence people's beliefs and preferences should also have implications for asset prices and cross-border capital flows.

Despite this potential importance of print news media, few studies have explored the ways in which it could influence asset prices, and none that I am aware of have linked it to capital flows. Those explorations of the link between the content of print news media and asset prices have focused on the ‘sentiment’ or emotional content of print news media, the closed economy context and the stock market. In this work, I broaden the literature on the role for print news media in affecting stock and bond prices and I fill the gap in the literature on the role for print news media in affecting capital flows, using data from South Africa.

Reproducing the analyses of seminal papers, Tetlock [2007] and Garcia [2013], on an archive of some fifteen thousand news articles published in South African national newspapers between 2008 and 2014, I cannot find evidence to support the role for
their single ‘sentiment’ print news media factor in explaining stock returns in-sample. However, when I generalise the single factor representation of newspaper articles to a multi-factor representation, I find strong out-of-sample evidence supporting a role for the print news media in explaining stock prices. Using such a multi-factor representation of the content of newspaper articles, I find that 9 percent of the daily out-of-sample variation in stock prices can be accounted for by the content of newspaper articles, and these results do not reflect purely historical phenomena, finite-sample biases, omitted variable biases, reverse causality or proxying for other sources of predictability like lagged returns, volatility or day-of-the-week effects. This finding also eliminates some uncertainties with findings in the existing literature.

I then explore the explanatory power of different types of language through different multi-factor representations of the content of print news media. In particular, I contrast emotive language with non-emotive language, and simple vocabularies based on individual words with complex vocabularies based on collections of words. This process allows me to investigate whether the print news media affects asset prices through sentiment or other fundamental channels. I also explore the role for multi-factor representations of print news media in explaining bond prices and net portfolio capital flows up to five days ahead.

The multi-factor representation of domestic print news media can explain out-of-sample variation in aggregate stock returns, aggregate stock trading volumes and net portfolio equity capital inflows one and two days ahead. This finding suggests that the domestic print news media affects aggregate stock prices through foreign demand with a lag of one or two days. This predictability also appears to be driven by emotive language, which suggests that the domestic print news media influences foreign demand through sentiment, specifically. The predictability holds out of sample, limiting concerns about finite sample biases, and after controlling for lagged returns, volatility and day-of-the-week effects, which limits concerns about proxying for these other potential
sources of predictability.

Three and four days ahead, the multi-factor representation of domestic print news media similarly shows excess predictive power for aggregate stock returns, but not for trading volume or net equity portfolio capital inflows. This finding suggests that foreign and domestic agents’ beliefs and preferences are affected by the content of domestic print news media three and four days prior, but that these two groups of agents respond similarly to such content. The predictability three and four days ahead is driven by more complex non-emotive vocabularies, suggesting that the response of beliefs and preferences to domestic print news media does not operate through sentiment.

In addition to its role in determining stock prices, I find that the multi-factor representation of print news media can explain 7 percent of the daily variation in the excess return on long-term bonds over short-term bonds three days ahead. This finding suggests that the domestic print news media affects the price of long-term domestic bonds relative to short-term domestic bonds with a lag of three days. The three-day lag coincides with predictability of the domestic print news media for net portfolio debt capital inflows three days ahead, suggesting that foreign demand for domestic long-term bonds relative to short-term bonds depends on the domestic print news media three days prior. This predictability also appears to be driven by non-emotive language, which further suggests that if the domestic print news media influences foreign demand for domestic long-term bonds, then it does so without affecting foreign sentiment.

This work is structured as follows. Chapter 2 motivates the subsequent empirical analysis by investigating the role for print news media in theoretical asset pricing models. This chapter also reviews the empirical evidence available in the existing literature. Chapter 3 describes the sources of print news media and financial data. It also describes transformations applied to the financial variables and their summary statistics in anticipation of the subsequent empirical work. Chapter 4 explains the method of matching news publications to financial variables, the specifications used in the analysis
and the method of estimating these specifications. Chapter 5 describes the main results of applying the methodology from Chapter 4. In addition, Chapter 5 also compares against results from an application of another method in the literature to these data, and provides further supplementary analyses to assist in interpreting the main results. Chapter 6 concludes with implications of the results for theoretical models of asset pricing and offers ideas for future research.
Chapter 2

Theory and existing evidence

2.1 Closed economy asset pricing model

In the classic closed economy consumption asset pricing model, risky asset prices are determined by preferences over rates of intertemporal substitution and risk aversion, and beliefs about expected future returns and volatility. Investor psychology therefore plays an integral role in asset pricing. In the context of this model, news takes the form of changes in the state of the economy, or in changes to uncertain future preferences or beliefs. Therefore, news should by definition affect equilibrium asset prices. However, it is also known that expected returns in this model vary through time. By informing agents about the current state of the economy, by affecting agents’ expectations about the future or by affecting their risk aversion, the print news media could determine these time-varying expected returns.

To formalise this intuition, consider the agent at date $t$ that seeks to maximise current utility

$$E_t \sum_{i=0}^{\infty} \beta^i u(c_{t+i})$$

(2.1)
subject to the budget constraints

\[ c_t = e_t - p_t \xi_t \] (2.2)

\[ c_{t+i} = e_{t+i} + d_{t+i} \xi_t \text{ for } i > 0 \] (2.3)

by choosing the quantity \( \xi_t \) of an infinitely divisible asset at this date \( t \), where \( E_t \) denotes the expectation operator conditional on information up to and including date \( t \).  \(^1\) \( \beta < 1 \) is the constant rate of time preference between any dates \( s \) and \( s + 1 \), \( u \) is an increasing and concave function that describes utility over consumption, \( c_s \) is random consumption at any date \( s \), \( e_s \) is the agent’s deterministic endowment or non-asset income at any date \( s \), \( p_t \) is the price of the asset at this date \( t \), and \( d_s \) is the random cashflow from the asset at any date \( s \). The utility (2.1) captures the agent’s beliefs through the weight that the expectation \( E_t \) places on the different states of the world. It also captures the agent’s preference for earlier consumption through parameter \( \beta \) and risk aversion through the concavity of \( u \). The potential choice \( \xi_t = 0 \) corresponds to the agent’s option to consume the endowment \( e_s \) at each date \( s \). If \( p_t > 0 \) and \( d_s > 0 \) almost surely for all dates \( s \), then choosing a larger quantity \( \xi_t \) increases consumption at future dates \( t + 1, t + 2, t + 3, \ldots \) at the cost of lower consumption at the current date \( t \). The asset is therefore a technology for deferring consumption from the current date \( t \) to all future dates \( t + 1, t + 2, t + 3, \ldots \) or for advancing consumption from all future dates \( t + 1, t + 2, t + 3, \ldots \) to the current date \( t \).

These concepts define a standard convex optimisation problem in one choice variable \( \xi_t \) subject to two linear equality constraints. By differentiating the Lagrangian function

---

\(^1\)There is a single implicit probability space on which all random variables are defined and there is a single filtration to which all stochastic processes are adapted. The operator \( E_t \) is a conditional expectation with respect to this filtration.
for this problem, we obtain the first order condition

$$p_t u'(c_t) = E_t \sum_{i=1}^{\infty} u'(c_{t+i}) d_{t+i}$$  \hspace{1cm} (2.4)$$

where $u'$ is the derivative of the function $u$ with respect to its argument. The agent chooses the quantity $\xi_t$ of the asset held to satisfy the condition (2.4). The left side of equation (2.4) defines the marginal cost of holding an additional unit of the asset, in terms of the foregone utility from lost consumption at the current date $t$. The right side of equation (2.4) defines the marginal benefit of holding an additional unit of the asset, in terms of the increased utility from gained consumption at dates $t + 1, t + 2, t + 3, \ldots$ due to the asset proceeds $d_{t+1}, d_{t+2}, d_{t+3}, \ldots$ at those dates. The agent uses the financial asset technology to trade off current and future consumption until this marginal cost and benefit are equal, because if one were greater than the other at a given choice of $\xi_t$ then utility could be improved by adjusting this choice.

In the representative agent context where all agents have the same beliefs, preferences and endowments, the condition (2.4) can be seen as determining the current price $p_t$ of the asset jointly with consumption $c_t, c_{t+1}, c_{t+2}, \ldots$ at dates $t, t + 1, t + 2, \ldots$. Cochrane [2009] extends this argument to the heterogeneous agent context where each agent has specific beliefs, preferences and endowments. In this heterogeneous agent context, the equation (2.4) holds for each agent because each agent chooses a specific $\xi_t$ to trade current consumption off against future consumption, and the same market price $p_t$ appears in each such equation.\(^2\) Therefore, even in the heterogeneous agent context, the condition (2.4) for any agent can be seen as determining the current price $p_t$ of the asset jointly with consumption $c_t, c_{t+1}, c_{t+2}, \ldots$ at dates $t, t + 1, t + 2, \ldots$.

\(^2\)If some other numerical value $v_t > p_t$ were to satisfy an individual agent’s condition (2.4), then the agent would be able to improve her utility by choosing a higher quantity $\xi_t$, which would result in a lower current consumption $c_t$ and hence a higher $u'(c_t)$ and, through the condition (2.4), a lower $v_t$. Similarly $v_t < p_t$ would require a higher $v_t$ to solve (2.4). These stable conditions ensure that the market price $p_t$ solves (2.4) for any individual agent.
By iterating condition (2.4) one date ahead to obtain an expression for $p_{t+1}$, we can see that asset returns defined as $r_{t+1} := (p_{t+1} + d_{t+1})/p_t - 1$ satisfy

$$1 = E_t \left[ \beta u'(c_{t+1})\left(1 + r_{t+1}\right) \right].$$

(2.5)

By expanding equation (2.5) using the identity $\text{cov}_t(X, Y) = E_t XY - (E_t X)(E_t Y)$ and cancelling terms we obtain

$$E_t r_{t+1} = -1 + \frac{u'(c_t)}{\beta E_t u'(c_{t+1})} - \frac{\text{cov}_t[u'(c_{t+1}), r_{t+1}]}{E_t u'(c_{t+1})}$$

(2.6)

where $\text{cov}_t$ denotes the covariance operator conditional on information up to and including date $t$.\(^3\) Equation (2.6) shows how the equilibrium returns from the closed economy consumption asset pricing model vary through time in a way that depends on each agent’s preferences and beliefs. It is therefore no surprise that the print news media could determine asset prices in a closed economy by informing agents about the current state of the economy $c_t$, influencing agents’ expectations over future states of the economy $c_{t+1}$ or influencing agents’ preferences through the shape of the marginal utility function $u'$.

Campbell et al. [1993] formalise the intuition that preferences influence asset prices in the context of a heterogeneous agent model, where some agents have uncertain future levels of risk aversion and other agents cannot fully offset these agents’ buying or selling pressure because they themselves are risk averse. News in the form of changes in the first group’s risk aversion can therefore affect risky asset prices, even when agents are rational in the sense that their beliefs are consistent with the model of the economy in which they operate. Similarly, De Long et al. [1990] formalise the intuition that beliefs influence asset prices through a heterogeneous agent model where some agents have irrational and uncertain future beliefs about forthcoming returns or volatility, in the

---

\(^3\)Specifically, $\text{cov}_t(X, Y) := E_t(X - E_t X)(Y - E_t Y)$ for any well-defined random variables $X, Y$. 
sense that their beliefs are not consistent with the model of the economy in which they operate, and when other agents have constraints like finite investment horizons or risk aversion that prevent them from offsetting the buying or selling pressure of irrational agents. The authors emphasise that it is the combination of both the anticipation of uncertainty in irrational agents' future beliefs and the above constraints on rational investors that leads to lower equilibrium risky asset prices.

2.2 Evidence on the role for media in a closed economy

Several papers study the role for news in determining closed economy stock prices by evaluating the extent to which observed measures of news can explain, in a statistical sense, variation in observed stock returns. In an early work, Cutler et al. [1988] find that at most one third of the daily variation in stock prices can be accounted for by the unexpected component of contemporaneous variation in several macroeconomic indicators. The authors conclude that stock prices must be determined by factors other than future cashflows or discount rates, which suggests an important role for preferences and beliefs.

Two closely related papers, Tetlock [2007] and Garcia [2013], study the role for ‘sentiment’ in determining stock prices. Using data for the period 1984–1999, Tetlock [2007] finds that the words appearing in the Wall Street Journal’s “Abreast of the Market” column on a given morning provide a leading indicator of the change in the Dow Jones Industrial Average Index on that day. Specifically, Tetlock [2007] defines a single media factor for each day as the difference between the fractions of all words appearing on that day that are emotionally ‘negative’ and ‘positive’, according to the Harvard IV-4 dictionary classification. The author bases his findings on in-sample hypothesis tests. Since negative word counts are negatively associated with that day’s
stock return, whether measured between consecutive days’ closing prices or between 10 a.m. and the time of market close on the same day, and since the association between negative word counts and future days’ stock returns reverses sign at longer forecast horizons, the author concludes that there is some evidence that stock prices reflect media sentiment rather than information about future cashflows or discount rates. One uncertainty in this conclusion is the potential role for finite sample bias, given the large number of parameters being estimated relative to the number of observations. Garcia [2013] applies the method of Tetlock [2007] to find a similar association between word counts from articles published in the New York Times on the morning of a given day and changes in the Dow Jones Industrial Average on that day, over the 1905–2005 period. The author also documents a procyclicality in the magnitude of this relationship. The large number of observations in Garcia [2013] limits concerns about finite sample biases present in Tetlock [2007]. However, two explanations acting together could account for the results in Garcia [2013]. The predictability of news for changes in stock prices could reflect a pre-WWII phenomenon, and it could reflect reverse causality from financial market developments after the NASDAQ closed at 4 p.m. to the content of news articles published the next morning. The author investigates each potential cause in isolation, and the results weaken appreciably after allowing for the second potential explanation, but the author does not allow for both explanations together. More detail on Tetlock
[2007] and Garcia [2013] appears in the replication exercise of Section 5.1.

Other authors have investigated the informativeness of other news sources for stock returns. Antweiler and Frank [2004] find that the bullishness of messages on stock message boards provides only limited information about future stock returns of individual companies. Bullishness on a given day is an aggregate of the bullishness of each individual message that was posted on that day. In turn, the bullishness of each individual message is measured using a statistical algorithm that extrapolates from subjective bullishness ratings for a sample of messages. The extrapolation procedure is based on frequencies of occurrence of individual words. The authors also find that disagreement between the messages is informative about trading volume, and that both the number and bullishness of messages are informative about future volatility. Luss and d’Aspremont [2015] find that individual company press releases are informative about future intraday company-specific stock volatility, but not about the direction of future company-specific stock returns. The authors’ word list is ad hoc, but has the advantage of containing some pairs of words like didn’t increase. Da et al. [2015] find that the volume of internet search queries related to words like ‘recession’, ‘bankruptcy’ and ‘unemployment’ is informative about future stock return reversals, stock return volatility and equity mutual fund outflows. The list of individual words that the authors use to compute this volume on a given day is based on the Harvard IV-4 dictionary of emotively negative words, adjusted to the terms used in search queries and to those that covary with stock returns.

These empirical studies specify word lists based on emotive words, which are suitable for studies of the role of sentiment in asset pricing, but leave open the role of other print news media factors in affecting asset prices. In particular, the emotive words commonly used ignore macroeconomic terms, which may be important for affecting agents’ beliefs and preferences. For example, a discussion involving macroeconomic terms in the print

McDonald [2011], but claims to produce results that are qualitatively similar to those of Tetlock [2007].
news media could alter agents’ beliefs about the current or future states of the economy. The print news media can be informative for stock returns and can affect asset pricing without necessarily affecting sentiment. Furthermore, if it is really the case that other print media factors turn out to be important for asset pricing, then the measured effect of the single, sentiment-related print media factor in the existing literature would be biased by such omitted variables.

2.3 Theory and evidence on the role of print media in an open economy

In an open economy asset pricing model, foreigners’ preferences and beliefs determine their demand for domestic assets, which affects domestic asset prices. If foreigners are at an informational disadvantage relative to domestic residents about domestic assets, foreigners may rely more on such publicly available information as print news media to inform their beliefs about future returns and volatility of domestic assets, while domestic residents have access to additional sources. Hence, foreigners could behave like the irrational agents in the model of De Long et al. [1990] in having their beliefs determined by such publicly available information as print news media that does not necessarily provide information not already incorporated into asset prices. In the context of a small open economy like South Africa, where domestic asset values are small relative to those of the rest of the world and controls are limited, foreigners have the potential to influence domestic asset prices substantially.

Solnik [1974] is the seminal contribution on international asset pricing. The author studies the situation where investor $\kappa$ in country $k$ maximises lifetime utility to achieve the value

$$J^\kappa(W^\kappa, t) = \max E_t \left[ \int_t^{T_\kappa} U^\kappa(C^\kappa, s) ds + \tilde{B}^\kappa(W^\kappa(T^\kappa), T^\kappa) \right]$$

(2.7)
at time $t$, subject to the budget constraint

$$dW^\kappa(t) = W^\kappa(t)\sum_{i=1}^{n} w_i^\kappa(t) \frac{dI_i^k(t)}{I_i^k(t)} + W^\kappa(t)\sum_{i=1}^{n} v_i^\kappa(t) \frac{dB_i^k(t)}{B_i^k(t)} + (Y^\kappa(t) - C^\kappa(t)) dt \quad (2.8)$$

where

- $T^\kappa$ is the random date defining the terminal investment horizon of investor $\kappa$;
- $U^\kappa$ is a date-specific utility function over consumption for investor $\kappa$;
- $\tilde{B}^\kappa$ is a date-specific bequest function defining utility over terminal wealth for investor $\kappa$;
- $W^\kappa(t)$ is the wealth of investor $\kappa$ at date $t$ in the domestic, country $k$ currency;
- $n$ is the number of countries with assets available for purchase and also the number of currencies;
- $w_i^\kappa(t)$ and $v_i^\kappa(t)$ are the fractions of wealth of investor $\kappa$ at date $t$ that are allocated to stocks and bonds respectively;
- $I_i^k(t)$ and $B_i^k(t)$ are the prices of stocks and bonds respectively in country $i$ at date $t$ in the currency of country $k$;
- $Y^\kappa(t)$ is the labour or non-asset income of investor $\kappa$ at date $t$ in the domestic, country $k$ currency; and
- $C^\kappa(t)$ is the consumption of investor $\kappa$ at date $t$ in the domestic, country $k$ currency.\(^7\)

In this setting, investors each have their own specific preferences described by $J^\kappa(W^\kappa, t)$. However, they are assumed to have common beliefs about normally distributed local

\(^7\)Clearly the notation is general enough to incorporate $W, I, B, Y, C$ in nominal currency units or in real consumption units.
currency stock returns and about deterministic local currency rates of return on bonds. Solnik [1974] shows that investor \( \kappa \) demands a quantity

\[
d^\kappa_i(t) = \frac{W^\kappa(t)}{A^\kappa(t)} \sum_{j=1}^{n} s_{ij}(\alpha_j - R_j) \tag{2.9}
\]

of stocks of country \( i \) at date \( t \), where \( \alpha_j \) is the instantaneous expected rate of return on stocks of country \( j \) in the currency of country \( j \), \( R_j \) is the instantaneous rate of return on bonds of country \( j \) in the currency of country \( j \),

\[
A^\kappa(t) = -W^\kappa(t) \left( \frac{\partial^2 J^\kappa}{\partial (W^\kappa)^2} \right) \left( \frac{\partial J^\kappa}{\partial W^\kappa} \right) \tag{2.10}
\]

is the coefficient of relative risk aversion of investor \( \kappa \) at that date and \( s_{ij} \) is element \((i, j)\) of the inverse of the \( n \times n \) matrix of instantaneous covariances associated with the \( n \times 1 \) vector \((I^1(t) \ I^2(t) \ \cdots \ I^n(t))\) of stock returns in domestic currencies.\(^8\) If investor \( \kappa \) of country \( k \) is representative of all investors of country \( k \), then equation (2.9) can be aggregated to obtain the quantity

\[
D_i := \sum_{k=1}^{n} d^k_i(t) = \left( \sum_{k=1}^{n} \frac{W^k(t)}{A^k(t)} \right) \sum_{j=1}^{n} s_{ij}(\alpha_j - R_j) \tag{2.11}
\]

demanded by the world for the stock of country \( i \).\(^9\) Equation (2.11) is a standard downward-sloping function describing the relationship between the quantity demanded and the price of the stock of country \( i \), because the parameter \( \alpha_i \) on the right hand side of equation (2.11) is decreasing in the price of the stock of country \( i \). We can invert

\(^8\)It is notable that foreign exchange risk between the currencies of countries \( i \) and \( k \) does not affect the quantity (2.9) that investor \( \kappa \) demands of the stock of country \( i \). This phenomenon occurs because investor \( \kappa \) is assumed to be able to borrow at the rate of return on bonds in country \( i \). Hence the value of every unit of currency of country \( i \) that investor \( \kappa \) holds in stocks of country \( i \) can be costlessly hedged against such currency fluctuations by borrowing an equivalent number of units of currency of country \( i \) at the rate of return on the bonds of country \( i \).

\(^9\)\( W^k \) and \( A^k \) are aggregates, across all investors \( \kappa \) in country \( k \), of \( W^\kappa \) and \( A^\kappa \) respectively.
(2.11) to obtain
\[ \alpha_i - R_i = \left( \sum_{k=1}^{n} \frac{W^k(t)}{A^k(t)} \right)^{-1} \sum_{j=1}^{n} \sigma_{ij} D_j \] \tag{2.12}

where \( \sigma_{ij} \) is the element \((i, j)\) of the \(n \times n\) matrix of instantaneous covariances associated with the \(n \times 1\) vector \((I_1^1(t) \ I_2^2(t) \ \cdots \ I_n^n(t))\) of stock returns in domestic currencies.\(^{10}\)

Equation (2.12) shows how stocks are priced in this model. Equilibrium in this model occurs when the market clears by equating demand for each country’s stocks with supply of that country’s stocks. In Solnik [1974], the supply of each country’s stocks is fixed at its market capitalisation, so that the equilibrium quantities demanded on the right hand side of (2.12) are fixed at these levels, and the price of a country’s stocks adjusts so that expected returns \(\alpha_1, \ldots, \alpha_n\) satisfy (2.12).

In this model, we can analyse counterfactual scenarios for aggregate demand for the stock of country \(i\) caused by counterfactual levels of wealth \(W^k\) or risk aversion \(A^k\) for any country \(k\). For any two countries \(i\) and \(k\), increasing \(A^k\) would shift inward the demand function (2.11) for stocks of country \(i\). However, since the equilibrium aggregate demands \(D_1, \ldots, D_n\) are fixed at the levels of supply, the expected return \(\alpha_i\) must increase to satisfy (2.12). To effect the increase in \(\alpha_i\), the price of the stocks of country \(i\) must fall. The effect of increasing \(A^k\) will be larger if \(W^k\) is larger. Therefore, if domestic print news media in some way affects the risk aversion of foreign investors, it could affect foreign demand for domestic stocks. This would in turn lead to a relationship between the domestic print news media, equity capital inflows and returns on domestic stocks. While it seems unlikely that the South African print news media could significantly

\(^{10}\)Stack equation (2.11) across countries \(i\) to obtain
\[
\begin{pmatrix}
D_1 \\
\vdots \\
D_n
\end{pmatrix} = \left( \sum_{k=1}^{n} \frac{W^k(t)}{A^k(t)} \right) \begin{pmatrix}
\sigma_{11} & \cdots & \sigma_{1n} \\
\vdots & \ddots & \vdots \\
\sigma_{n1} & \cdots & \sigma_{nn}
\end{pmatrix}^{-1} \begin{pmatrix}
\alpha_1 - R_1 \\
\vdots \\
\alpha_n - R_n
\end{pmatrix}
\]
and then invert the linear system to solve for the \(n\)-vector \((\alpha_1 - R_1 \ \cdots \ \alpha_n - R_n)\). Row \(i\) of the resulting inverted system gives equation (2.12).
affect the risk aversion of investors in the rest of the world, even very small such effects could have substantive consequences for South African asset prices, given the large size of foreign wealth relative to South African wealth.

A more interesting counterfactual would be the effect of changing foreigners’ beliefs about the expected return or volatility of domestic stocks. Such shocks would lead to shocks to aggregate demand from foreigners for domestic stocks. However, we cannot analyse such shocks in the model of Solnik [1974], because beliefs in this model are homogeneous across countries.

We next turn to the empirical evidence on the role for print media in open economy asset pricing. To the best of my knowledge, this is the first work to present such evidence. Empirical literature on the role for news in determining capital flows is limited, even though the literature on capital flow surges and sudden stops makes qualitative references to the important role for “market sentiment”. Fratzscher [2012] finds a role for news, in the sense of deviations in macroeconomic variables from median expectations expressed in Bloomberg surveys, in explaining capital flows. However, the macroeconomic variables considered, including the trade balance, gross domestic product, industrial production and unemployment, do not provide information about the role for the print media in determining asset prices.

2.4 Contrasting to additional papers in the literature

In this subsection, I compare and contrast the research question in this work to the questions studied in other papers. This discussion should provide the reader with some comfort that I have not omitted other, in some cases prominent, papers from a review of the literature.
Tetlock et al. [2008] study the relationship between the content of news articles that mention specific S&P 500 firms and subsequent changes in the stock prices of those firms. On each calendar date, the authors calculate the fraction of words, appearing in news articles in the Wall Street Journal and Dow Jones News Service on that date, which are classified as ‘negative’ by the Harvard IV-4 dictionary. They find that this fraction of negative words is negatively (partially) correlated with future individual stock returns even after controlling for lagged individual stock returns and earnings forecasts. Moreover, this fraction of negative words is negatively (partially) correlated with future abnormal returns from a Fama and French [1993] three-factor model, which suggests that the fraction of negative words is a previously undiscovered source of priced risk.

Tetlock et al. [2008] is concerned with the relative pricing of individual company stocks. An asset pricing factor that affects the cross-section of expected returns is not necessarily relevant for the time variation in expected returns on the aggregate stock market. We should primarily be interested in aggregate economic implications, and therefore cross-sectional phenomena are of secondary importance, except where there is reason to believe they have such aggregate implications. In this work, I focus instead on the pricing implications for the aggregate stock market. I consider the role for the print news media more generally, rather than some notion of sentiment, and I explore the open-economy angle by examining the relationship between the print news media and capital flows.

Baker and Wurgler [2007] provide an introduction to the role for ‘sentiment’ in the pricing of individual company stocks. Again, the cross-section of expected stock returns is a fundamentally different question from the behaviour of aggregate stock returns that I pursue here. The authors offer a number of definitions of sentiment, including responses from telephone surveys and market based measures like mutual fund flows, trading volume, the premium of dividend-paying stocks over other stocks,
the discount to net asset value on closed-end equity funds, first-day returns from initial public offerings and option-implied volatility. The authors do not describe measures of sentiment derived from computational linguistics methods applied to print media.

Mo et al. [2016] analyse the relationship between stock returns on trading date $s$ and sentiment on trading date $t$ where $t = s + 1, s + 2, s + 3, s + 4$ and $s + 5$. The authors study an archive of 12 million articles, from newspapers including the New York Times and Wall Street Journal, published between 2012 and 2015. They compute sentiment on a given day by adding up the weights, provided by Baccianella et al. [2010], associated with each word published in their newspaper archive on that day. They find a positive association between the returns on several stock market indices and future days’ sentiment. This finding is qualitatively the same as the finding in Table 3 of Tetlock [2007] and is only a sign that the sentiment factor they calculate exhibits similar properties to the pessimism factor in the latter paper. The potential role for stock returns in influencing the content of newspaper articles is recognised as early as Tetlock [2007], which is why claims about the potential effect of sentiment on stock returns are only made after controlling for lagged stock returns.
Chapter 3

Data

The data for this project are text from published news articles, historical market index levels from the Johannesburg Stock Exchange (JSE) and historical capital flows. I construct an archive of 15,584 South African news articles published in The Times, Business Day and Financial Mail between 11 December 2008 and 6 February 2014. I obtain these articles from Factiva, a news provider owned by Dow Jones & Company, with the search term ‘economy’.

The articles are time stamped with their day of publication, so text can be analysed at a daily or lower frequency.

These three newspapers are distributed in print nationally and are available free of charge online. All three are owned by Times Media Group, headquartered in Johannesburg. Between August and October 2016, the websites of The Times, Business Day and Financial Mail received 19.7, 1.2 and 1.1 million visits respectively. Data on numbers of website visits are provided by SimilarWeb Ltd, a digital market intelligence company.

1It may be that other news articles, which are excluded from this search, are informative about asset prices or capital flows. This possibility would only strengthen the case for the role of the print news media that I document in this work.

2Print editions of The Times and Business Day are published every weekday and those of the Financial Mail are published weekly on Fridays. According to figures released by the Audit Bureau of Circulations, between July and September 2016 the circulations of these three
print publications were 59, 23 and 13 thousand copies per publication date respectively, down from 109, 26 and 15 thousand copies per publication date a year earlier.

I obtain minute-by-minute Top 40 futures contract prices from the data provider Portara CQG.\textsuperscript{3,4} Open market trading of equity derivatives on the JSE takes place between 8:30 a.m. and 5:30 p.m. local time. The 8:30 a.m. opening price is determined by an auction conducted between 8:25 a.m. and 8:30 a.m. This 8:30 a.m. price is therefore a post-open price in the sense that it reflects information released since the close of trading on the previous trading day. I calculate post-open stock returns using changes in the natural logarithm of the Top 40 index futures contract price between start times of 8:30 a.m., 10:30 a.m. and 12:00 p.m. and an end time of 5:30 p.m.\textsuperscript{5} Summary statistics on these post-open returns appear in panel 3.1a of Table 3.1.

I obtain daily equity total return indices and trading volume of the JSE All Share Index and Top 40 Index from Bloomberg.\textsuperscript{6} From the same source, I obtain Bloomberg South Africa bond price indices for 10 and 1-3 year maturities.\textsuperscript{7} I obtain the GOVI daily historical total return index of South African government bond prices from Thomson Reuters Datastream. This index is calculated and published by the JSE and consists of the ten largest and most liquid South African government bonds by market capitalisation and clean consideration turnover respectively. I calculate daily stock and bond returns as the change from one day to the next in the natural logarithm of the respective total return index. I compute daily changes in trading volume as the change

\footnotetext[3]{http://www.portaracqg.com/}
\footnotetext[4]{The minute-by-minute futures contract prices refer to Top 40 Index futures contracts traded on the JSE. These prices are back-adjusted as at 31 October 2016 and the contracts are rolled over one day before expiry to create a continuous time series of futures prices.}
\footnotetext[5]{Tetlock [2007] uses changes in the Dow Jones Industrial Average between 10 a.m. and 4 p.m., Garcia [2013] uses changes in this index between 11 a.m. and 4 p.m. Trading on the NYSE and NASDAQ, which contain the stocks that make up the Dow Jones Industrial Average, takes place between 9:30 a.m. and 4 p.m. Trading on the JSE takes place between 8:30 a.m. and 5:30 p.m. The JSE has an administration period for allocations and reporting between 5:30 and 6:15 p.m.}
\footnotetext[6]{https://www.bloomberg.com/}
\footnotetext[7]{The Bloomberg tickers for these securities are JALSH Index, TOP40 Index, BSAFR10 Index and BSAFR13 Index respectively.}
from one day to the next in the natural logarithm of the sum of one and trading volume. I calculate the daily equity premium as the difference between the daily stock and bond returns, and I calculate the long bond premium as the daily change in the natural logarithm of the 10 year maturity bond price index minus the daily change in the natural logarithm of the 1-3 year maturity bond price index.

I obtain daily net portfolio equity and debt capital flows into South Africa in millions of US dollars from the Institute for International Finance.\textsuperscript{8} These data on capital flows cover business days only, not weekends. The daily asset prices and capital flows data cover the same period as the newspaper archive, while the intraday equity futures prices cover the period between 14 October 2010 and the last date of the newspaper archive. Summary statistics of the daily stock and bond returns, changes in trading volume and capital flows appear in panel 3.1b of Table 3.1.

\textsuperscript{8}https://www.iif.com/
Table 3.1: Summary statistics of the data used in this work. The calculation and source of all variables follows the description in Chapter 3. Trading volume refers to changes in trading volume, as described in that section. Total capital flows refers to the sum of equity and debt capital flows. The last three columns of each subtable show the p-values from Ljung and Box [1978] tests for serial correlation, where the null hypothesis is the absence of serial correlation between lags 1 and $m$ inclusive.

(a) Post-open returns

<table>
<thead>
<tr>
<th>opening time</th>
<th>obs.</th>
<th>mean</th>
<th>std dev.</th>
<th>min</th>
<th>max</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30</td>
<td>721</td>
<td>0.000</td>
<td>0.008</td>
<td>-0.030</td>
<td>0.039</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10:30</td>
<td>800</td>
<td>0.000</td>
<td>0.006</td>
<td>-0.024</td>
<td>0.038</td>
<td>0.81</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>12:00</td>
<td>792</td>
<td>0.000</td>
<td>0.006</td>
<td>-0.019</td>
<td>0.034</td>
<td>0.04</td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

(b) Daily returns, trading volume and capital flows

<table>
<thead>
<tr>
<th>variable</th>
<th>obs.</th>
<th>mean</th>
<th>std dev.</th>
<th>min</th>
<th>max</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>JALSH Index return</td>
<td>1223</td>
<td>0.000</td>
<td>0.011</td>
<td>-0.037</td>
<td>0.056</td>
<td>0.86</td>
<td>0.15</td>
<td>0.50</td>
</tr>
<tr>
<td>TOP40 Index return</td>
<td>1223</td>
<td>0.000</td>
<td>0.012</td>
<td>-0.039</td>
<td>0.062</td>
<td>0.89</td>
<td>0.11</td>
<td>0.40</td>
</tr>
<tr>
<td>GOVI return</td>
<td>1283</td>
<td>0.000</td>
<td>0.004</td>
<td>-0.021</td>
<td>0.015</td>
<td>0.01</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Long bonds premium</td>
<td>1023</td>
<td>0.000</td>
<td>0.006</td>
<td>-0.024</td>
<td>0.061</td>
<td>0.09</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>JALSH premium</td>
<td>1223</td>
<td>0.000</td>
<td>0.011</td>
<td>-0.040</td>
<td>0.061</td>
<td>0.61</td>
<td>0.47</td>
<td>0.86</td>
</tr>
<tr>
<td>Top 40 premium</td>
<td>1223</td>
<td>0.000</td>
<td>0.012</td>
<td>-0.043</td>
<td>0.068</td>
<td>0.85</td>
<td>0.39</td>
<td>0.80</td>
</tr>
<tr>
<td>JALSH trading volume</td>
<td>1230</td>
<td>-0.025</td>
<td>0.634</td>
<td>-8.928</td>
<td>9.082</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Top 40 trading volume</td>
<td>1230</td>
<td>-0.023</td>
<td>0.611</td>
<td>-8.422</td>
<td>8.643</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Equity net capital inflow</td>
<td>1232</td>
<td>7.6</td>
<td>96</td>
<td>-594</td>
<td>1324</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Debt net capital inflow</td>
<td>1232</td>
<td>22.7</td>
<td>172</td>
<td>-1022</td>
<td>999</td>
<td>0.10</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Total net capital inflow</td>
<td>1229</td>
<td>29.4</td>
<td>196</td>
<td>-1091</td>
<td>1037</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

(c) Word frequencies

<table>
<thead>
<tr>
<th>factor</th>
<th>obs.</th>
<th>mean</th>
<th>std dev.</th>
<th>min</th>
<th>max</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative</td>
<td>815</td>
<td>0.023</td>
<td>0.005</td>
<td>0.010</td>
<td>0.049</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>positive</td>
<td>815</td>
<td>0.012</td>
<td>0.003</td>
<td>0.003</td>
<td>0.038</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>pessimism</td>
<td>815</td>
<td>0.011</td>
<td>0.006</td>
<td>-0.023</td>
<td>0.042</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Chapter 4

Method

4.1 Matching news to returns

As explained in Chapter 3, I consider two sets of dependent variables. The first are intraday stock (futures) returns and the second are daily stock returns, bond returns, percentage changes in trading volume, equity premia, long bond premia and capital flows. I match each set of dependent variables slightly differently to articles published in the national newspapers. By tailoring the matching procedure to each set of dependent variables, I am able to ensure that I minimise any overlap between publication times of newspaper articles and market trading times.

Consider first the matching of newspaper articles to intraday stock returns. I match all articles published on the previous trading day and any intervening calendar days to the post-open stock return on any given day. In particular, I match newspaper articles published on Friday, Saturday and Sunday to the intraday stock return on Monday. I match newspaper articles published on Monday to the intraday stock return on Tuesday, and so on. This matching scheme avoids any potential for the content of the news articles to reflect the changes in stock prices that they are to be used to predict. I am able to match 815 days of intraday returns to newspaper articles in this
way.

Next consider the matching of newspaper articles to the dependent variables observed at the daily frequency. For ease of exposition I explain with the example of daily close-to-close stock returns. If we enumerate the trading days on which we observe closing stock prices as \( \ldots, t - 1, t, t + 1, t + 2, \ldots \), then I match the stock return \( y_{t+1} \) between trading days numbered \( t \) and \( t + 1 \) to those newspaper articles published on the date of the trading day numbered \( t \) and, if applicable, any non-trading calendar days earlier than the date of the day numbered \( t \) but (strictly) later than the date of the trading day numbered \( t - 1 \). This means that I match newspaper articles published on Friday to the close-to-close stock return computed between Friday close and Monday close. It also means that I match newspaper articles published on Saturday, Sunday and Monday to the close-to-close stock return computed between Monday close and Tuesday close. I am able to match 1283 daily stock returns to newspaper publication days in this way. In what follows, I also consider the ability of print news media to explain stock returns up to 5 days ahead. The matching procedure is analogous to the one-step ahead case, replacing \( y_{t+1} \) by \( y_{t+n} \) for \( n = 1, 2, 3, 4, 5 \).

Note that I match newspaper articles published on Sunday to Monday’s post-open stock return and the daily return between Monday close and Tuesday close. This matching procedure ensures that in each case, newspaper articles published on Sunday are matched to the nearest available future return that does not coincide with Sunday. A procedure of matching Sunday’s newspaper articles to daily stock returns between Friday close and Monday close, for example, would allow for the possibility that any financial market developments between the close of the domestic stock market on Friday and the publication of a newspaper article on Sunday could influence the content of that newspaper article.\(^1\) I would like to rule out such reverse causality interpretations of any

\(^1\)The JSE closes for stock trading at 5 p.m. according to the Coordinated Universal Time + 2 hours time zone, which is 6 hours before the New York Stock Exchange closes, at 4 p.m. according to the Coordinated Universal Time - 5 hours time zone.
role for print news media in the pricing of assets. Although this matching procedure minimises the potential for reverse causality in the case of one-step ahead daily returns, it does not eliminate this interpretation in this case because a newspaper article time-stamped with a given day may have been published after the close of trading on that day. This potential for reverse causality is the motivation for beginning with an analysis of post-open stock returns, where such an interpretation is precluded. Nevertheless, the results for intraday and one-step ahead daily returns are similar, which shows that the reverse causality explanation is not important in the case of one-step ahead daily returns. Such a reverse causality interpretation is also precluded for the 2−, 3−, 4− and 5−day ahead daily stock return predictions that I analyse.

4.2 A multi-factor representation of print news media

I convert the text of these matched articles into numeric features. If multiple articles are matched to a given stock return, the bodies of those articles are concatenated into one document, so that I end up with a time series of documents for the purposes of extracting features. I convert each document to numerical features based on word lists that I refer to as vocabularies. A vocabulary could be a list of individual words, a list of individual words and pairs of consecutive words, or a list of individual words, pairs of consecutive words and triples of consecutive words. For consistency with the existing literature, three vocabularies to consider are the positive, negative and combined positive and negative individual words as defined in Loughran and McDonald [2011] and used in Garcia [2013]. Another natural vocabulary is the union of all individual words occurring in any of the articles. To allow for terms with qualifying words like not good, I also consider an extension of the preceding vocabulary that includes all individual words.

2This vocabulary is called the bag-of-words model in computational linguistics.
and all pairs of consecutive words occurring in any of the articles. Finally, to allow for terms with qualifying words that may be separated from the word they qualify, like *not very good*, I consider an additional vocabulary based on all individual words, all pairs of consecutive words and all triples of consecutive words appearing in any of the articles. I therefore end up with six vocabularies, each one of which will produce a set of explanatory variables that can be used in a predictive regression model.

Following Luss and d’Aspremont [2015], I calculate term frequency–inverse document frequency (TF-IDF) features from these vocabularies. The TF-IDF feature for each vocabulary item (i.e. an individual word, word pair or word triple) in a given document is $TF \times \log(m/DF)$ where $TF$ is the number of occurrences of this item in this document, $m$ is the total number of documents and $DF$ is the number of documents in which this item appears.\(^3\) For a given document, this procedure produces one feature number for each item in a vocabulary so that each document is represented by a vector. For vocabularies with many items, these vectors can be long. The Loughran and McDonald [2011] dictionaries contain 354 positive words and 2355 negative words, and the collection of all articles contains 67 thousand individual words, 1.6 million individual words and pairs of consecutive words, and 5.8 million individual words, pairs of consecutive words and triples of consecutive words.

Any such vector of numeric TF-IDF features is a multi-factor representation of the print news media. I use this terminology throughout this work. To summarise, each observed stock return is associated with exactly one document, which is a collection of all the newspaper articles matched to that stock return, and each document is transformed into a vector of up to 5.8 million numeric features that describe its content. These stock returns and numeric features can then be used to train a prediction model.\(^4\) I explain

\(^3\)The TF-IDF features therefore downweight vocabulary items that occur many times in across all newspaper articles, because these items are less likely to be useful for discriminating between documents associated with high and low stock returns.

\(^4\)Note that we have more explanatory variables than observations, which makes the estimation of classical regression models like ordinary least squares infeasible, but other regression models that
the model specifications in Section 4.3 and estimation procedure in Section 4.4.

4.3 Model specification

I consider the role for print media factors in explaining stock returns, bond returns, the equity premium, the long bond premium, changes in stock trading volume and capital flows. Let the value of any of these dependent variables on the $i^{th}$ ordered trading date in the sample be denoted $y_i$, where the exact dependent variable that is meant should be clear from the context. I consider intraday (post-open) stock return and daily stock return dependent variables, which should also be identifiable from the context. For each of the dependent variables measured at the daily frequency, I consider models for the dependent variable $h$ days ahead for $h = 1, 2, 3, 4, 5$.

For each dependent variable, I consider models with print media explanatory variables only, with a set of non-print media explanatory variables only and with a combined set of print media and non-print media explanatory variables jointly. Let us denote the print news media explanatory variables that are matched to the dependent variable on trading date numbered $i + 1$ by $f_i$. Then the single ‘sentiment’ factor representation of the print news media relates to the case where $f_i$ is a scalar, and the multi-factor representation of print news media relates to the case where $f_i$ is a vector.\footnote{I explain the multi-factor representation of print news media in Section 4.2.}

For the stock return, bond return, equity premium and long bond premium dependent variables, I use five lags of the dependent variable, five lags of the squared dependent variable and day-of-the-week indicator variables as non-print media explanatory variables. For the stock return dependent variable, these explanatory variables match those used in Tetlock [2007] and Garcia [2013].\footnote{Following Garcia [2013], the stock returns used to construct lagged explanatory variables follow the same definition as the stock returns used for the dependent variable. Therefore, for models where the dependent variable is the change in the natural logarithm of the futures price between 10:30 a.m. and 5:30 p.m., the lagged return explanatory variables would be calculated from these returns.} They capture the potential role for employ regularisation techniques remain feasible.
serial correlation, volatility and day-of-the-week effects, which have been documented in earlier work, in determining stock returns. By including these explanatory variables in the models for the bond return, equity premium and long bond premium, I allow for the possibility of serial correlation, volatility and day-of-the-week effects in these dependent variables, too. Given the trading date numbered \( i + 1 \), these five lags of the dependent variable are \( y_i, y_{i-1}, y_{i-2}, y_{i-3} \) and \( y_{i-4} \). Let us denote the day-of-the-week effect for the \((i + 1)\)th trading date by \( \theta_{i+1} \). Therefore, the models for the stock return, bond return, equity premium and long bond premium \( h \) days ahead can be written as

\[
y_{i+h} = \alpha + f_i^T \beta + \sum_{k=1}^{5} \phi_k y_{i-k} + \sum_{k=1}^{5} \gamma_k y_{i-k}^2 + \theta_{i+1} + e_{i+1} \quad \text{for all } i, \tag{4.1}
\]

for any \( h = 1, 2, 3, 4, 5 \), which defines the parameters \( \alpha, \beta, \phi, \gamma \) and the error terms \( e_i \) for all \( i \).

When stock trading volume, equity capital flow or total equity and debt capital flow are the dependent variable, I use five lags of stock returns, five lags of squared stock returns and day-of-the-week indicator variables as the non-print media explanatory variables. When debt capital flow is the dependent variable, I use five lags of bond returns, five lags of squared bond returns and day-of-the-week indicator variables as the non-print media explanatory variables. These choices allow for potential returns-chasing behaviour in capital flows. The models for stock trading volume, equity capital flow, debt capital flow or total equity and debt capital flow can be written as

\[
y_{i+h} = \alpha + f_i^T \beta + \sum_{k=1}^{5} \phi_k r_{i-k} + \sum_{k=1}^{5} \gamma_k r_{i-k}^2 + \theta_{i+1} + e_{i+1} \quad \text{for all } i, \tag{4.2}
\]

for any \( h = 1, 2, 3, 4, 5 \), where \( r_i \) is the appropriate asset return on any trading date numbered \( i \).
4.4 Model estimation

Note that specifications (4.1) and (4.2) can be written as

\[ y_i = b + w \cdot x_i + e_i \quad \text{for all } i \]  

(4.3)

where \( \cdot \) is the dot product, \( b := \alpha \) and \( w := (\phi_1 \cdots \phi_5 \gamma_1 \cdots \gamma_5 \theta^T \beta^T) \). Depending on the context, we could have \( x_i := (y_{i-h} \cdots y_{i-h-4} y_{i-h}^2 \cdots y_{i-h-4}^2 d_i^T f_i^T) \) or we could have \( x_i := (r_{i-h} \cdots r_{i-h-4} r_{i-h}^2 \cdots r_{i-h-4}^2 d_i^T f_i^T) \), where \( d_i \) is a binary vector indicating the day of the week of trading date numbered \( i \).

To estimate the vector of parameters \( w \) and the scalar parameter \( b \), I use a machine learning technique called support vector regression (SVR) due to Vapnik [1995].\(^7\) This technique has been found to have good performance in the context of predicting future volatility of individual stock returns based on individual word and pair of consecutive words TF-IDF features extracted from companies’ own 10-K filings to the US Securities and Exchange Commission [Kogan et al., 2009]. Luss and d’Aspremont [2015] use an earlier version of this technique, which is designed for predicting binary dependent variables, to predict the sign of stock returns from news features.

The parameters \( w, b \) must be estimated from a sample of data

\[ (y_1, x_1), (y_2, x_2), \ldots, (y_l, x_l) \]  

(4.4)

and an SVR provides one method for doing so with good theoretical properties and good results in applications. At a high level, we can think of an SVR as a generalisation of simple median regression combined with a ridge regression penalty term. Intuitively, the median regression component makes the parameter estimates robust to outliers, and the penalty component both improves the generalisability of the fitted linear regression.

\(^7\)In this work, I consider the simple case of linear SVR.
function to unseen (i.e. out-of-sample) cases and ensures that a unique estimator exists even with more explanatory variables than observations.

An SVR introduces two free parameters, also called tuning parameters, denoted $\epsilon$ and $C$. Using the parameter $\epsilon$ we denote the $\epsilon$-insensitive error function

$$
|\xi|_\epsilon = \begin{cases} 
0 & \text{if } |\xi| < \epsilon \\
|\xi| - \epsilon & \text{otherwise.}
\end{cases} 
$$

(4.5)

for any scalar $\xi$. This $\epsilon$-insensitive error function is a generalisation of the absolute value function, because the former converges to the latter at each $\xi$ as $\epsilon$ approaches zero. For any positive $\epsilon$, the value of the $\epsilon$-insensitive error function at a point $\xi$ is zero when $\xi$ is within $\epsilon$ of zero, and when $\xi$ departs from zero by an amount more than $\epsilon$, the error function equals this amount. For a given choice of $(w, b)$, all sample points $(y_i, x_i)$ that produce errors $y_i - w \cdot x_i - b$ satisfying $|y_i - w \cdot x_i - b| < \epsilon$ contribute zero to the $\epsilon$-insensitive error function. It is therefore not surprising that such points $(y_i, x_i)$ turn out not to contribute to the linear SVR estimator for $w$.

This error function can be compared to the standard squared error loss function $\xi^2$. The $\epsilon$-insensitive error function does not penalise errors within a tolerance of $\epsilon$, while the squared error loss function imposes a small penalty on such errors that decays as $|\xi| \to 0$. As the error $|\xi|$ grows, the squared error loss function imposes a penalty that grows quadratically fast, which makes its parameter estimates sensitive to outliers, while the $\epsilon$-insensitive error function imposes a penalty that only grows linearly, which produces parameter estimates that are more robust to outliers. In this work, I choose $\epsilon := 0$, which is related to median regression of Huber [1964].

It is possible that better predictive performance could be obtained with alternative choice of $\epsilon$. This would strengthen the case for the role of the print news media that I find with $\epsilon = 0$.

It is notable that Garcia [2013] uses median regression as a robustness check.

With these definitions, a linear SVR then defines estimators for $w, b$ as those that
minimise
\[ C \sum_{i=1}^{l} |y_i - w \cdot x_i - b|_\epsilon + \frac{1}{2} w \cdot w, \] (4.6)

which is a linear combination of \( \epsilon \)-insensitive errors implied by the chosen parameters \( w, b \) and a penalty term measuring the squared Euclidean distance of the parameters \( w \) from zero. The penalty term is the same as the penalty term from ridge regression. When \( x_i \) is long relative to \( l \), or in other words when we have many explanatory variables relative to the sample size, there may be infinitely many choices of \( w \) that each evaluate the sum of fitted \( \epsilon \)-insensitive errors to zero. Since the penalty term is strictly convex, it has a unique minimum at \( w = 0 \), and hence including the penalty term ensures that the estimator for \( w \) is unique. In repeated samples, the penalty term biases the parameter estimates of \( w \) toward zero, which biases the linear regression model toward a model with a constant term only. Zero is a useful parameter vector to bias toward if the explanatory variables \( x_i \) provide only a small amount of information about the dependent variable \( y_i \) relative to their noise.

The parameter \( C \) defines the weight in the linear combination placed on the fitted \( \epsilon \)-insensitive errors relative to the weight of \( \frac{1}{2} \) in the linear combination placed on the penalty term. Larger values of \( C \) increase the relative importance of choosing \( w, b \) to minimise the \( \epsilon \)-insensitive errors on the sample data, while smaller values of \( C \) increase the relative importance of making predictions that are similar to the constant-only model. The parameter \( C \) is therefore inversely proportional to the amount of bias introduced into the estimator of \( w \) by the penalty term. In repeated samples, the introduction of even a small amount of bias in a parameter estimator can provide a large reduction in mean square error of that estimator through a large reduction in its variance. After some experimentation, in this work I consider \( C = \frac{1}{2^9}, \frac{1}{2}, 1 \) and 2.\(^{10}\)

\(^{10}\)I choose a range values of \( C \) to allow for both high and low levels of regularisation. However, I do not claim these values of \( C \) to be optimal. They achieve some predictability, as shown in Chapter 5, but it is possible that even more predictability could be achieved by improving these choices of \( C \).
Having defined the linear SVR estimator in (4.6), I turn here to the practicality of computing this estimator. If we set $\epsilon$ to be very large or if there are many explanatory variables relative to the sample size, the fitted $\epsilon$-insensitive errors in the first term in (4.6) could each be zero at the optimal choice of $w$. However, Vapnik [1995] shows that we can allow for the general case of strictly positive $\epsilon$-insensitive error terms at the cost of additional notation by introducing $2l$ slack variables denoted $\xi_1, \ldots, \xi_l$ and $\xi_1^*, \ldots, \xi_l^*$, which are non-negative variables that loosely correspond to the sizes of the positive and negative $\epsilon$-insensitive error associated with each observation $i = 1, \ldots, l$. Using this extra notation, Vapnik [1995] states that the minimisation problem (4.6) can be written as

$$\min_{w, b, \xi_1, \ldots, \xi_l, \xi_1^*, \ldots, \xi_l^*} \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

subject to $y_i - w \cdot x_i - b \leq \epsilon + \xi_i$,

$$w \cdot x_i + b - y_i \leq \epsilon + \xi_i^*$$

and $\xi_i, \xi_i^* \geq 0$ for all $i = 1, \ldots, l$, \hspace{1cm} (4.7)

which is a standard quadratic optimisation problem subject to inequality constraints.

To aid in the comparison of the slack variables in (4.7) with the $\epsilon$-insensitive errors in (4.6), I provide the following result.

**Lemma 1.** At the solution to (4.7), $\xi_i + \xi_i^* = |y_i - w \cdot x_i - b|_\epsilon$ for $i = 1, \ldots, l$.

**Proof.** Let $w, b, \xi_1, \ldots, \xi_l$ and $\xi_1^*, \ldots, \xi_l^*$ solve (4.7) with $C > 0$ and $\epsilon >= 0$. Fix an arbitrary choice of $i$. First we prove that $\xi_i > 0 \Rightarrow \xi_i^* = 0$. Suppose $\xi_i > 0$. Then $y_i - w \cdot x_i - b = \epsilon + \xi_i$ because by definition of $w, b$ it is not possible to reduce $\xi_i$ further and maintain $y_i - w \cdot x_i - b \leq \epsilon + \xi_i$. Combining these facts we have $y_i - w \cdot x_i - b > \epsilon >= 0$. Therefore $w \cdot x_i + b - y_i - \epsilon < -\epsilon <= 0 \leq \xi_i^*$. However, $\xi_i^*$ should be as small as possible to solve (4.7), so we must have $\xi_i^* = 0$. \hspace{1cm} 32
Second, we note by symmetry that \( \xi_i^* > 0 \Rightarrow \xi_i = 0 \). Third, note that if \(|y_i - w \cdot x_i - b| < \epsilon\), which could occur for example if \( \epsilon \) were large, then we must have \( \xi_i = \xi_i^* = 0 \) to minimise the objective. Putting these three statements together, we have

\[
\xi_i + \xi_i^* = \begin{cases} 
0 & \text{if } |y_i - w \cdot x_i - b| < \epsilon \\
\max\{y_i - w \cdot x_i - b - \epsilon, w \cdot x_i + b - y_i - \epsilon\} & \text{otherwise.}
\end{cases}
\]

and \( \max\{y_i - w \cdot x_i - b - \epsilon, w \cdot x_i + b - y_i - \epsilon\} = |y_i - w \cdot x_i - b| - \epsilon \).

The Lagrangian function for the problem (4.7) is

\[
L := \frac{1}{2} w \cdot w + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) - \sum_{i=1}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*) - \sum_{i=1}^{l} \alpha_i (\epsilon + \xi_i - y_i + w \cdot x_i + b) - \sum_{i=1}^{l} \alpha_i^* (\epsilon + \xi_i^* + y_i - w \cdot x_i - b) .
\]

where \( \eta_i, \eta_i^*, \alpha_i, \alpha_i^* \) for all \( i \) are the Lagrange multipliers. The standard Karush–Kuhn–Tucker method for solving the optimisation problem with inequality constraints (4.7) is to minimise \( L \) with respect to \( w, b, \xi_1, \ldots, \xi_l, \xi_1^*, \ldots, \xi_l^* \) and maximise \( L \) with respect to the Lagrange multipliers, subject to the constraint that all Lagrange multipliers be non-negative. The first order conditions for a minimum with respect to the parameters in (4.7) are:

\[
0 = \frac{\partial L}{\partial b} = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \quad (4.9)
\]

\[
0 = \frac{\partial L}{\partial w} = w - \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) x_i \quad (4.10)
\]

\[
0 = \frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i \quad (4.11)
\]

\[
0 = \frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* . \quad (4.12)
\]
By using (4.10), (4.11) and (4.12) to eliminate $\eta_i$, $\eta_i^*$ and $w$ for all $i$ in (4.8), other terms cancel and we obtain

$$L = -\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_ix_j - \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*).$$

(4.13)

Equation (4.13) is the Lagrangian function that depends only on $\alpha_i, \alpha_i^*$ for $i = 1, \ldots, l$, with other parameters having been concentrated out. The non-negativity of the Lagrange multipliers from (4.8) combined with equations (4.11) and (4.12) give

$$0 \leq \alpha_i, \alpha_i^* \leq C$$

for $i = 1, \ldots, l$.

(4.14)

Maximising (4.13) with respect to $\alpha_i, \alpha_i^*$ for all $i$, and subject to the constraints (4.9) and (4.14) defines the dual problem to the primal problem (4.7). This is a simple quadratic programming problem over a rectangular region.\(^{11}\)

4.5 Performance assessment

I assess the performance of this model by computing out-of-sample prediction errors. In particular given a full sample (4.4) with size $l = m$, I consider fitting the model (4.7) on 100 expanding window subsamples of size $l < m$ for $l \in \{\lfloor \frac{m-1}{100} \rfloor : l' = 1, 2, \ldots, 100 \}$ and obtaining each model forecast $\hat{y}_{l+1}$ given the input $x_{l+1}$. I then compute the out-of-sample cumulative root mean square prediction errors

$$\sqrt{\frac{1}{l'} \sum_{l''=1}^{l'} \left( y_{\lfloor \frac{m-1}{100} \rfloor l'' + 1} - \hat{y}_{\lfloor \frac{m-1}{100} \rfloor l'' + 1} \right)^2}$$

(4.15)

for each $l' = 1, 2, \ldots, 100$.

\(^{11}\)Further details on the SVR method can be found in the introduction by Smola and Schölkopf [2004].
Against this model I also consider a naive benchmark returns forecasting model that predicts that the next day’s return will be the historical mean return observed up to the date at which the prediction is made. Replacing the term in (4.15) involving $\hat{y}$ by $\sum_{l=1}^{\tilde{l}} y_l / \tilde{l}$ where $\tilde{l} = \lfloor \frac{m-1}{100} \rfloor$, I obtain the cumulative root mean square prediction error for a benchmark model that forecasts the historical mean return. This benchmark model is agnostic about the future direction of returns above or below the historical mean, so outperforming this model indicates in particular that we tend to predict the deviations of future returns from the historical mean better than could be expected by pure chance. This historical mean benchmark model also has the virtue that its mean square is simply the (out-of-sample) variance of the dependent variable, so if a candidate model outperforms this benchmark model in terms of mean square prediction error, then the candidate model produces smaller (out-of-sample) errors on average than the (out-of-sample) variation of the returns being modelled. In the results presented below, I express the cumulative root mean square prediction error (4.15) as a fraction of the cumulative root mean square prediction error of the historical mean model. A ratio less than unity indicates that the candidate forecasting model outperforms the benchmark historical mean model in terms of out-of-sample mean square prediction error.\textsuperscript{12}

\textsuperscript{12}I use the errors from the benchmark forecasting model as a device for rescaling the errors from the main candidate forecasting models. Some comparisons, like those between model specifications with the same dependent variable, are invariant to the choice of benchmark model used to rescale the root mean square errors.
Chapter 5

Results

5.1 A single media factor

I estimate the parameters in the specification

\[ y_{i+1} = \alpha + \beta f_i + \sum_{k=1}^{5} \phi_k y_{i-k} + \sum_{k=1}^{5} \gamma_k y_{i-k}^2 + \theta_{i+1} + \epsilon_{i+1} \text{ for all } i \quad (5.1) \]

where \( y_i \) denotes the post-open stock return on trading date numbered \( i \), \( f_i \) is the single media factor and \( \theta_i \) is a set of day-of-the-week indicator variables, by ordinary least squares. Following Tetlock [2007] and Garcia [2013], the media factor \( f_i \) is the fraction of occurrences of negative or positive words, or the difference (called ‘pessimism’) between the fractions of occurrences of negative and positive words, in newspaper articles matched to stock returns on trading day \( i + 1 \).\(^1\) Summary statistics on these three individual media factors appear in panel 3.1c of Table 3.1. This panel shows that about 2 percent of words are classified as negative, and 1 percent of words are classified as positive, on a given day. These factors show some persistence as well.

\(^1\)The matching procedure is explained in Chapter 4 and ensures that articles matched to returns of a given trading day are published on the preceding calendar day or earlier. I use positive and negative words as defined in Loughran and McDonald [2011].
Figure 5.1: White [1980] $t$-statistics associated with ordinary least squares estimates of the coefficient $\beta$ in regressions of type shown in equation (5.1) and estimated in Garcia [2013]. The horizontal axis indicates the starting time that I use to compute post-open returns. I compute post-open returns until 5:30 p.m. A dotted horizontal line appears at 1.96, the upper 97.5$^{th}$ percentile of the standard normal distribution.
I compute $t$-statistics using White [1980] standard errors and present the $t$-statistics associated with the coefficient $\beta$ in Figure 5.1. The $t$-statistic with the largest absolute value is 1.74 and occurs when counts of positive words are used to explain the next day’s change in stock index futures price between 12:00 p.m. and 5:30 p.m. As Figure 5.1 makes clear, the relationship between the media factor and stock returns found in Tetlock [2007] and Garcia [2013] is not replicated on these data, for two reasons. First, the signs of the estimated $\beta$ coefficient from equation (5.1) do not always agree with their hypothesised sign. The number of negative words is positively associated with the next day’s 8:30 and 10:30 a.m. post-open stock return, and the ‘pessimism’ factor, calculated as the difference between the numbers of negative and positive words, is positively associated with the next day’s 8:30 a.m. post-open stock returns. The number of positive words is negatively associated with the next day’s 10:30 a.m. post-open stock return. Second, not one of the $\beta$ coefficients from equation (5.1) are estimated to be statistically different from zero at the 5% significance level.\footnote{Starting times other than the three presented in Figure 5.1 and ending times other than 5:30 p.m. produce $t$-statistics that are closer to zero in absolute value and similarly incorrectly signed.} If a lack of significance were due to fewer observations in my sample,\footnote{I have a maximum of 759 observations in these regressions, while Tetlock [2007] and Garcia [2013] have 3709 and 19184 observations respectively.} then I would expect to see similar estimates of the coefficient $\beta$ in my sample to those obtained in the earlier studies, while obtaining larger standard errors. However, I obtain estimates of $\beta$ of between $-2.1$ and 4.1 basis points per unit of standard deviation of $f_i$, which is about half the size of the estimates obtained in these earlier studies.

The fact that the results of Tetlock [2007] and Garcia [2013] are not replicated on my data could reflect problems with their estimation like finite sample bias or reverse causality, or the fact that the single print media factor model is a purely historical phenomenon, as explained in Chapter 1. Alternatively, their results could be specific to the newspapers that they study, which is suggested by the lack of predictability found
for internet stock message boards and company press releases [Antweiler and Frank, 2004, Luss and d’Aspremont, 2015], or specific to the large economy that is the United States.

A final explanation for the lack of explanatory power of the single media factor on this sample is that specification (5.1) omits other important media factors. The effect that earlier studies attribute to the single media factor could then reflect to some extent the role for such omitted factors. To explore such an explanation, I present results from estimating (5.1) with a generalised representation of the print media factor \( f_i \) that nests the single print media factor as a special case.

### 5.2 The information in national newspapers

In this section, I explore the extent to which the single media factor could be missing important media factors through a variance decomposition exercise. Let \( X \) denote a matrix with 1283 rows and 67 thousand columns. Each row and column of \( X \) could contain a TF feature for that particular date and individual word in the archive of newspaper articles. Alternatively, the matrix \( X \) could contain TF-IDF features in place of TF features. Below, I explore both options for \( X \). Given a choice for \( X \), let

\[
X = U\Sigma V^T
\]  

(5.2)

denote the singular value decomposition of \( X \) into the 1283 \( \times \) \( k \) orthonormal matrix \( U \), the \( k \times k \) diagonal matrix \( \Sigma \) and the orthonormal matrix \( V \) with 67 thousand rows and \( k \) columns, where \( k \) is the rank of \( X \) and \( T \) is the transpose.\(^4\)\(^5\) Equation (5.2) and the orthonormality property imply that \( X^T XV = V\Sigma \). In turn, this system of equations

\(^4\)Orthonormality of \( U \) and \( V \) means \( U^TU \) and \( V^TV \) are identity matrices.

\(^5\)This application of singular value decomposition is called latent semantic analysis in the field of computational linguistics. Introductory treatments can be found in Landauer et al. [1998] and Chapter 18 of Manning et al. [2008].
can be written out more explicitly as

\[ X^T X v_i = \sigma_i^2 v_i \text{ for } i = 1, \ldots, k, \]  

(5.3)

where \( \sigma_i \) is the \( i^{th} \) entry along the diagonal of \( \Sigma \) and \( v_i \) is the \( i^{th} \) column of \( V \).\(^6\) Note that the left-hand side of (5.3) can be grouped as \( X^T(Xv_i) \), which is the uncentered sample covariance between the columns of \( X \) and the vector \( Xv_i \). Then equation (5.3) says that this covariance is proportional to the unit length vector \( v_i \), with constant of proportionality \( \sigma_i^2 \).

Suppose we order the columns of \( U, \Sigma \) and \( V \) so that the entries along the diagonal of \( \Sigma \) are in descending order. Then the first principal component of \( X \) is defined to be the vector \( Xv_1 \). This component is the linear combination of columns of \( X \) that explains the largest share of variation in \( X \), in the sense that for \( i = 1 \), the uncentered sample covariance (5.3) between the columns of \( X \) and the component \( Xv_i \) is maximally proportional to the unit length vector \( v_i \). In this sense, no other single factor constructed as a linear combination of the columns of \( X \) could explain a greater share of the uncentered sample covariance in \( X \). When \( X \) contains TF features, each of the three options for the single media factor from Section 5.1 can be written as linear combinations of the columns of \( X \) and thus would explain, in this sense, less of the variation in \( X \) than the first principal component.\(^7\)

Figure 5.2 plots the fraction \( (\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2)/\sum_i \sigma_i^2 \) of the variation in \( X \) that is explained by the first \( k \) principal components against \( k = 1, 2, \ldots, 300 \).\(^8\) Panel 5.2a

---

\(^6\)Representation (5.3) shows that \( \sigma_i \) and \( v_i \) are also the \( i^{th} \) eigenvalue and eigenvector of the uncentered sample covariance matrix \( X^T X \), respectively.

\(^7\)Denote by \( v \) a vector of 67 thousand elements, with entries equal to 1 where the corresponding column in \( X \) refers to a Loughran and McDonald [2011] negative word, with entries equal to \(-1\) where the corresponding column in \( X \) refers to a Loughran and McDonald [2011] positive word, and with all other entries equal to 0. If \( X \) contains TF features then \( Xv \) is the linear combination of columns of \( X \) that produces the ‘pessimism’ single media factor of Garcia [2013].

\(^8\)I choose to stop at the first 300 principal components because Landauer et al. [1998] present evidence, from other contexts, that word associations based on approximations of \( X \) by its first 300 components perform best on multiple-choice tests of synonyms. In addition, this number represents
Figure 5.2: Fractions of variation in the features $X$ extracted from the print news media that are explained by the first $k$ principal components, for $k = 1, 2, \ldots, 300$.

shows these fractions for the matrix $X$ of TF features, while Panel 5.2b shows these fractions for the matrix $X$ of TF-IDF features. The first principal component explains 6.6 percent of the variation in TF features, whence any of the individual single print media factors cannot explain at least $93 < 100 - 6.6$ percent of this variation. Since the literature has focused on single print news media factors, much of the interesting variation in the content of print news media has not been explored.

In Figure 5.2b, the first principal component of the matrix of TF-IDF features can explain 2.1 percent of its variation, suggesting that TF-IDF features are richer than TF features. Figure 5.2 also shows that the first 300 principal components can explain 60 percent of the variation in the matrix of TF features and 26 percent of the variation in the matrix of TF-IDF features, which further confirms the interesting extra variation in TF-IDF features.\(^9\)

We can view the vector $v_i$ defining the $i^{th}$ principal component of $X$ as a list of weights to be applied to the words associated with the columns of $X$. The words about a quarter of the maximum possible number of principal components available in the matrix $X$, which is limited to 1283 rows. Finally, computational considerations prevent the analysis of more components.

\(^9\)This richness may be related to why TF-IDF features are usually preferred in other applications.
that attract the largest such weights summarise the information conveyed by the $i^{th}$ principal component of $X$. It is interesting to see whether the resulting words are grouped together into coherent ‘concepts’ in this way.

The words that attract the largest weights in vector $v_1$ associated with the first principal component of the TF-IDF matrix are $SA, will, said, year, economy$ and $growth$. These words contain the original search query ‘economy’ used to find the newspaper articles. The three newspapers considered are all South African, but the frequent occurrence of the abbreviation $SA$ for South Africa suggests that the reported news is primarily domestic. In addition, the word $growth$ suggests a reference to economic growth, and the word $will$ suggests a discussion about the future. The word $said$ could indicate that someone is being quoted. The word $year$, like $will$ again references time, although in this case it is not clear whether past ($last\ year$), present ($this\ year$) or future ($next\ year$) is meant. The first principal component therefore seems to have some kind of interpretation. For example, it could refer to the concept of quoted opinions about future economic growth.

The words that attract the largest weights in the second principal component of the TF-IDF matrix are $year, last, said, bank, quarter, inflation, recession, sales$ and $month$. This group of words alludes to quotes about recent economic indicators. It is not obvious that the second principal component is completely conceptually distinct from the first, even though they are uncorrelated in sample by definition. Nevertheless, the second principal component places negative weights on $will$ and $SA$ and therefore appears to focus more on recent specific economic data than on future headline growth for the country.

In the third principal component of the TF-IDF features, the words attracting the largest weights are $says, SA, will, market, banks, property, prices, countries, years, investment, global$ and $capital$. This group of words alludes to a concept of quotes about financial issues or global intermediated investment. Interestingly, the words $said$
and *yesterday* attract large negative weights in this component, which suggests that the third principal component focuses on quotes that are current rather than even a day old.

Similarly, the fourth principal component concentrates on words like *engine*, *says*, *rear*, *car*, *vehicle* and places negative weight on *said*, *ANC*, *last* and *Zuma*. This component seems to describe recent quotes about motor vehicles, which could include transportation or manufacturing. The fifth component concentrates on words like *will*, *ANC*, *inflation*, *government*, *Zuma*, *policy*, *budget*, *Gordhan* and *tax*, while placing negative weights on *said*, *Africa*, *China*, *mining*, *companies*, *yesterday* and *business*. This fifth component includes references to the governing political party, the president and the finance minister. It therefore alludes to discussion or opinion about future domestic government policy, but contrasts against quotes, or discussion of historical events and corporate or foreign affairs.

The multi-factor representation of print news media described in Section 4.2 therefore offers a much richer characterisation than the single factor representation described in Section 5.1. The variance decomposition exercise presented in this section quantifies the minimum extent of the variation unexplained by the single print news media factor, and offers some interpretable groupings of individual words. However, the conceptual interpretations for these groupings are also messy, not least because every word attracts a weight in every grouping.

### 5.3 Multiple media factors

The poor performance of the single media factor specification (5.1) on these data motivates the consideration of media factors that could be omitted from that specification. For example, rather than counting the number of occurrences of all positive words on a given day, we could count the number of occurrences on that day of each positive word
Table 5.1: Minimum relative cumulative root mean square (out-of-sample) prediction errors for intraday returns on Top 40 index futures contracts. The rows of the table index different models, in the sense of different sets of explanatory variables used for prediction. All models with lagged returns also include lagged squared returns and day-of-the-week explanatory variables, as discussed in the text. The columns of the table indicate whether intraday returns being predicted are computed from 8:30 a.m., 10:30 a.m. or 12:00 p.m., while all intraday return windows end at 5:30 p.m. Each entry in the table gives the minimum, across the various newspaper vocabularies described in the text and tuning parameters \( C = 2^{-8}, 2^{-1}, 1, 2 \), of the root mean square prediction error, expressed relative to the root mean square prediction error of a mean-only model.

<table>
<thead>
<tr>
<th>model</th>
<th>8:30</th>
<th>10:30</th>
<th>12:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>best model using news only</td>
<td>0.91</td>
<td>0.98</td>
<td>1.04</td>
</tr>
<tr>
<td>best model using returns only</td>
<td>1.01</td>
<td>1.03</td>
<td>0.98</td>
</tr>
<tr>
<td>best model using returns and news</td>
<td>1.01</td>
<td>1.02</td>
<td>0.96</td>
</tr>
</tbody>
</table>

in a predefined list of positive words. By considering a vocabulary made up of positive and negative words, I obtain a richer set of explanatory variables that nest the single factor time series explanatory variable in specification (5.1).\(^{10}\) Rather than restricting attention to whether a single factor constructed from news can price past levels of the stock market, I consider which groups of words can price future levels of the stock market before having observed them.\(^{11}\) I present the results on the informativeness of print news media for post-open returns in Section 5.3.1 and for daily returns, capital flows and trading volumes in Section 5.3.2.

5.3.1 Informativeness for post-open returns

Table 5.1 presents the minimum, across the six newspaper vocabularies and tuning parameters \( C = 2^{-8}, 2^{-1}, 1, 2 \), of the full-sample cumulative root mean square prediction error, given in display (4.15) with \( l' = 100 \). The results in Table 5.1 are expressed as a ratio to the full-sample cumulative root mean square prediction error of a benchmark.

\(^{10}\)I explain the multi-factor representation of print news media in Section 4.2, and the full model specifications in Section 4.3.

\(^{11}\)I explain the out-of-sample performance assessment that I use here in Section 4.5.
model that forecasts the historical mean return, so that values less than 1 indicate outperformance. The variables being predicted are changes in the natural logarithm of Top 40 futures prices over an interval starting at 8:30 a.m., 10:30 a.m. or 12:00 p.m. and ending at 5:30 p.m. The best performing model in the table predicts the intraday return between 8:30 a.m. and 5:30 p.m. and achieves a cumulative root mean square prediction error that is 91% of that of the model that forecasts using the historical mean return. This relative root mean square prediction error can be interpreted as 9% lower than the variation of the post-open stock returns being predicted, or equivalently as 9% lower than the root mean square prediction error of a naive model that forecasts using the historical mean return. The vocabulary for this best-performing model contains all words appearing in the newspaper archive and the model excludes lagged return features (i.e. it excludes lagged returns, lagged squared returns and day-of-the-week indicator variables). Note that this vocabulary outperforms vocabularies constructed from pairs and triples of words and those vocabularies constructed from the Loughran and McDonald [2011] dictionary. This suggests that the predictive content of news for future returns can be distilled into individual words, rather than pairs or triples of words, but cannot be reduced simply to those individual words with positive or negative connotations.\footnote{Note that there is no concern of reverse causality in the relationship between news and returns described here because only news from the previous calendar day or earlier is matched to the intraday return on a given day. As described in Chapter 3, the opening price at 8:30 a.m. is determined by an opening auction and differs from the previous day’s close price. Hence it is unlikely that any news published after the previous trading day’s 5:30 p.m. market close would not already be incorporated into the 8:30 a.m. post-auction opening price.}

By reading across the first row of Table 5.1, we see that predictability of the previous day’s news for intraday returns declines as the trading day progresses, so that by noon the previous day’s news is incorporated into the Top 40 futures price. The results in the first row of Table 5.1 demonstrate the predictive content of news for 8:30 and 10:30 a.m. post-open stock returns. In addition, we may be interested to know whether this predictability can be explained by the predictive content of lagged
returns, volatility or day-of-the-week effects. In particular, the content of news articles could reflect past stock returns and past volatility, which could be related to future stock returns. If we were to find that news features provide predictive content by proxying for other features of stock returns, this would not change the predictive content of news features, but could offer an explanation of the mechanism by which news is informative.\textsuperscript{13}

The second row of Table 5.1 shows the minimum, across the tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2,$ of the full-sample cumulative root mean square prediction error of a model based on lagged returns, volatility and day-of-the-week effects only relative to that of a model that forecasts the historical mean return, for each post-open return window. The last row of Table 5.1 also presents such a relative prediction error, but takes the minimum over the six newspaper vocabularies and four choices for the tuning parameter $C$ and applies to a model containing all of the above explanatory variables. By comparing the entries in the last two rows in any given column, we see that the inclusion of news does not worsen the fraction of the out-of-sample standard deviation of post-open returns that can be explained by a model containing lagged returns, volatility and day-of-the-week explanatory variables. This confirms that the above predictability of news for stock returns does not appear to proxy for known sources of predictability like lagged returns, volatility or day-of-the-week effects.\textsuperscript{14}

A further observation from Table 5.1 is that the best performing model for each type of post-open return always involves news features. This suggests that no matter the definition of post-open returns, it is always advantageous to have news for prediction. More specifically, the relative prediction error for 8:30 a.m. post-open stock returns using returns only is larger than the that of the best performing model involving news features.

\textsuperscript{13}Further analysis would be required, if this were to be the case.

\textsuperscript{14}Note that the out-of-sample predictive performance of a model does not necessarily improve with the addition of more explanatory variables. This behaviour is similar to that of the well known adjusted $R^2$, which can decline with the addition of more explanatory variables.
Figure 5.3: Relative cumulative root mean square prediction error based the best performing model specifications in Table 5.1. The figure plots the expression in display (4.15) as a ratio of the cumulative root mean square prediction error of a model that forecasts the historical mean return, against \( l' \). The figure shows these prediction errors for post-open returns between 8:30 a.m. and 5:30 p.m., between 10:30 a.m. and 5:30 p.m. and between 12:00 p.m. and 5:30 p.m.

features by an amount equal to 10 percent of the variation of such stock returns.\(^{15}\) For 10:30 a.m. post-open returns, this figure falls to 5 percent, and for 12:00 p.m. post-open returns it falls again to 2 percent. In this sense, the informativeness of news features for intraday stock returns declines with time.\(^{16}\)

The preceding discussion concerned full-sample root mean square prediction errors, i.e. those in display (4.15) with \( l' = 100 \), but this does not give an indication of the stability or instability of such prediction errors. Figure 5.3 plots the relative cumulative root mean square prediction errors for post-open returns against the time index \( l' = 1, 2, \ldots, 100 \). At each date on the horizontal axis, the height of a line in Figure 5.3 shows the (relative) cumulative out-of-sample root mean square prediction error up to that

\(^{15}\) That is, 10% = 1.01 − 0.91 in the first column of Table 5.1.

\(^{16}\) The calculations are 5% = 1.01 − 0.98 and 2% = 0.98 − 0.96 respectively.
date, based on the errors from fitting models over expanding windows and predicting out of sample one day ahead. The right endpoint of the lines in Figure 5.3 correspond to the best full-sample relative cumulative root mean square prediction errors presented in Table 5.1, and the trajectory of the lines shows more detail about the stability of these full-sample estimates. The estimates of relative cumulative root mean square prediction error for post-open stock returns stabilise from about February 2012. Furthermore, the relative cumulative root mean square prediction errors for 10:30 a.m. and 12 p.m. post-open returns decline with the inclusion of extra data between February 2012 and February 2014. The relative cumulative root mean square prediction error of 8:30 a.m. post-open returns increases somewhat over this period, but remains well below unity. This stability suggests that these errors are relatively precisely estimated and would not increase dramatically with the addition of more data. This stability provides some comfort that these results are robust, and that not even the out-of-sample predictive power evidenced by the full-sample relative root mean square prediction error is subject to data snooping [White, 2000]. Therefore, the predictive content of news articles for post-open stock returns does not seem to be explained by overfitting in-sample or out-of-sample.

Among the collection of news-only models in the first row of Table 5.1, the vocabulary that produces the best relative cumulative root mean square prediction error is the list of all individual words. For the combined model of news and lagged returns, the vocabulary based on all individual words and pairs of consecutive words performs equally as well as the vocabulary based on all individual words, pairs of consecutive words and triples of consecutive words. These vocabularies perform best for of the choices of post-open return in Table 5.1. In no case do the three vocabularies of Loughran and McDonald [2011] emotive words outperform the three vocabularies of all individual, pairs and triples of words.\textsuperscript{17} Therefore, the informativeness of news for post-open

\textsuperscript{17}In the case of 8:30 a.m. post-open returns and news-only models, models based on all six vocabu-
Table 5.2: Informativeness of the multiple media factors for daily stock returns, stock trading volume, bond returns, the equity premium, the long bond premium and capital flows $n$ steps ahead for $n = 1, 2, 3, 4, 5$. Each entry in the table represents the difference between one and the minimum, over the six vocabularies and four tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2$, of the ratio of the root mean square (out-of-sample) prediction error of the news-only prediction model to the root mean square (out-of-sample) prediction error of a historical mean forecast model. The entries are multiplied by 100 to be expressed in percentage point units. Entries that would be negative are not shown with the understanding that the news-only model is not more informative than the historical mean model in these cases. The calculation of each variable is described in Chapter 3.

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 40 return</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Top 40 change in log volume</td>
<td>56</td>
<td>54</td>
<td>54</td>
<td>55</td>
<td>44</td>
</tr>
<tr>
<td>GOVI return</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>Equity premium</td>
<td>13</td>
<td>22</td>
<td>17</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td>Long bond premium</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Debt net inflow</td>
<td>-</td>
<td>1</td>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Equity net inflow</td>
<td>6</td>
<td>24</td>
<td>8</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Debt and equity net inflow</td>
<td>-</td>
<td>2</td>
<td>12</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

stock returns seems to come from the non-emotive words.\textsuperscript{18} This finding suggests that the single media factor model in specification (5.1) performs poorly due to an omitted factor.

### 5.3.2 Informativeness for daily returns, trading volume and capital flows

The preceding discussion concerned the informativeness of print news media for post-open stock returns. I turn next to the informativeness of print news media for daily stock and bond returns, change in trading volume, the equity premium, the long bond

\textsuperscript{18}I provide some examples of the words that receive the largest weight in the estimation in Section 5.3.2.
premium and capital flows. I investigate the informativeness of print news media for these variables up to 5 trading days ahead.

Table 5.2 shows the reduction in variance of each variable at each forecast horizon that is achievable using only the multi-factor representation of print news media. The entry in this table for a given variable and forecast horizon \( n = 1, 2, 3, 4, 5 \), shows the maximum, over the six vocabularies and four tuning parameters \( C = 2^{-8}, 2^{-1}, 1, 2 \), of the percentage reduction in daily variance of the variable that is achievable using only the multi-factor representation of the print news media \( n \) days earlier. The overall impression from the table is that the print news media is informative about a number of these variables. I discuss this table row-by-row below, including how the results change with the addition of the extra explanatory variables specified in Section 4.3.

The first row of Table 5.2 shows that the multi-factor representation of print news media is informative for daily stock returns up to four days ahead, but not five days ahead. The multiple media factors reduce the one-day ahead out-of-sample variation of daily stock returns by 12\% of their original level, which is slightly larger than the 9\% achieved above for post-open returns, suggesting that the difference is due to reverse causality between out-of-hours financial market developments and print media content. Not shown in the table, a large portion of this reduction in daily variation does not seem to be explained by serial correlation, volatility or day-of-the-week effects in the sense that the best model with only serial correlation, volatility and day-of-the-week explanatory variables only achieves a 3\% reduction in daily variation.\(^{19}\) For two-, three- and four-day ahead returns, none of the 15\%, 16\% and 7\% reduction in daily variation shown in Table 5.2 seems to be attributable to serial correlation, volatility or day-of-the-week effects in the sense that the best models in each case with only serial correlation, volatility and day-of-the-week effects cannot explain any of the daily

\(^{19}\)This number can be calculated as \( 1 - 0.996 \) where 0.996 appears in the first row and second column of the subtable for one day ahead returns in Table A.1 in the appendix.
variation in returns.\textsuperscript{20,21}

The vocabularies that achieve the best performance shown in first row of Table 5.2 also vary with the forecast horizon $n$. The most informative vocabulary for one-day ahead returns, which achieves the 12\% reduction in daily variance shown in the first row and column of Table 5.2, is the list of negative words. On this list of negative words, the words that receive the largest coefficients by absolute value are \textit{slowed}, \textit{failed} and \textit{defensive}. For two-day ahead returns, there is not much difference between the six vocabularies, with four achieving the same 15\% reduction in daily variation shown in Table 5.2. However, for three- and four-day ahead returns, the best performing vocabularies are the more complex non-emotive vocabularies. The vocabulary of all individual words and pairs of consecutive words and the vocabulary of all individual words, pairs of consecutive words and triples of consecutive words both achieve the 16\% reduction in daily variance for three-day ahead returns in Table 5.2. For the latter vocabulary, the vocabulary items that receive the largest coefficients by absolute value in a full-sample estimation are \textit{pic}, \textit{gold}, \textit{cwp}, \textit{busa} and \textit{gepf}. These items include acronyms for the Public Investment Corporation, Community Work Programme, Business Unity South Africa and Government Employees’ Pension Fund.\textsuperscript{22} The most informative vocabulary for four-day ahead returns is the vocabulary of all individual words, pairs of consecutive words and triples of consecutive words. Some of the vocabulary items receiving the

\textsuperscript{20}The root mean square prediction error of the best models with serial correlation, volatility and day-of-the-week effects, relative to the root mean square prediction error of the historical mean model, appear in the first row and second column of each subtable in Table A.1 in the appendix. For two, three and four day ahead returns, these numbers are greater than unity, showing an absence of evidence for predictive content.

\textsuperscript{21}I obtain very similar results using daily returns on the All-Share Index.

\textsuperscript{22}The Public Investment Corporation is an asset management company owned by the government with assets under management of some 44\% of GDP (see \url{www.pic.gov.za}). The Community Work Programme is a government programme that provides stipends to unemployed or under-employed people, in exchange for work in their communities, to assist in their job search (see \url{http://www.gov.za/CommunityWorkProgramme}). Business Unity South Africa is an association of South African private sector firms with coordination and advocacy objectives (see \url{https://www.busa.org.za/}). The Government Employees’ Pension Fund is pension scheme with 1.2 million active members and over 400 thousand pensioners and beneficiaries (see \url{http://www.gepf.gov.za/}). Its assets are managed by the PIC.
largest coefficients by absolute value in a full-sample estimation are *business*, *kumba*, *water*, *climate change* and *brokers*. I provide further discussion of these vocabulary items in Section 5.3.3.

It therefore seems that the print news media plays an important role in the pricing of stocks. It takes about four days for the effects of the print news media to be incorporated into stock prices, and the effects are clearest three days ahead. Emotive words are the most important characteristics of print news media in affecting stock prices one day ahead, while more complex non-emotive features play a bigger role three and four days ahead.

Closed economy models of heterogeneous agents predict that trading volume is determined by the extent of disagreement between these agents. The second row of Table 5.2 shows that the multi-factor representation of print news media can explain 56% of the out-of-sample variation in trading volume changes one day ahead, and this fraction declines to 44% of the variation five days ahead. These fractions are large, but a large proportion could be attributed to the variation explicable by lagged stock returns, volatility and day-of-the-week effects. Not shown in the table, the best models with lagged stock returns, volatility and day-of-the-week effects explain 47% and 50% of the variation in trading volume changes one and two days ahead respectively, leaving only the remaining 9% and 4% of the variation in trading volume changes explicable by print news media only. A similar calculation reveals no substantive extra explanatory power three, four or five days ahead. Therefore, I find a role for print news media in affecting trading volume up to two days ahead, which is consistent with a role for print news media in heterogeneous agent models, but I cannot find a role for print news

---

23 *Kumba* refers to Kumba Iron Ore, a large iron-ore mining company in South Africa.
24 For example, see the discussion in Section 2.1
25 The exact fractions of (out-of-sample) variance explicable by the best models with only lagged stock returns, volatility and day-of-the-week explanatory variables are 56%, 53% and 44% for three-, four- and five-day ahead returns respectively. These fractions can be obtained from the difference between unity and the numbers 0.437, 0.471 and 0.557 appearing in the second row and second column of each subtable of Table A.1 in the appendix.
media in affecting trading volume three days ahead where the effect of print news media on stock prices is strongest.

The multi-factor representation of print news media is even more informative about daily variation in trading volume changes for stocks in the All-Share Index. However, the extra explanatory power of print news media over the lagged return, volatility and day-of-the-week explanatory variables is smaller in the case of the All-Share Index. This finding suggests that the role for print news media in driving trading in heterogeneous agent models is primarily through large stocks. This finding favours an interpretation where the print news media affects some agents’ beliefs about large stocks, rather than affecting some agents’ risk aversion toward all risky assets.

The third row of Table 5.2 shows limited evidence of the informativeness of print news media for government bond prices. By contrast, the fourth row of this table shows that the multi-factor representation of print news media is appreciably informative for the equity premium, with the greatest reduction in (out-of-sample) variance being achieved two days ahead. These results suggest that the print news media is primarily informative for the prices of domestic risky assets rather than domestic risk-free assets. This supports the interpretation of the concepts in domestic print news media as a source of time-varying risk, against which risky assets are priced.

If we regard long-term bonds as the risky asset and short-term bonds as the risk-free asset, we may expect the print news media to affect the relative pricing of such bonds for the same reasons as we would expect the print news media to affect the pricing of stocks relative to bonds above. The fifth row of Table 5.2 shows that the multi-factor representation of print news media is informative about (out-of-sample) variation in the excess return of long-term bonds over short-term bonds one, three, four and five days ahead. After allowing for serial correlation, volatility and day-of-the-week effects,

\[26\text{None of the (out-of-sample) variation in the equity premium is explained by serial correlation, volatility or day-of-the-week effects. This is evidenced by the fact that the numbers in the fourth row and second column of each subtable in Table A.1 are greater than unity.}\]
7%, 3% and 2% of the three-, four- and five-day ahead predictability remain, and none of the one-day ahead predictability remains.\textsuperscript{27} Hence, there appears to be a role for print news media in affecting the price of long-term bonds relative to that of short-term bonds, and this effect is strongest three days ahead.

The last three rows of Table 5.2 show that the multi-factor representation of print news media is informative about net portfolio capital flows into South Africa. This informativeness is strongest for equity flows two days ahead and debt flows three days ahead. The informativeness of print news media for total debt and equity net portfolio inflows mirrors such informativeness for debt flows because debt flows tend to be several multiples larger than equity flows.

The fraction of the (out-of-sample) variance of two-day ahead equity portfolio flows that can be explained by the multi-factor representation of print news media but not by lagged stock returns, volatility or day-of-the-week variables is 4%.\textsuperscript{28} The vocabularies of positive words and combined positive and negative words are the most informative for equity portfolio flows two days ahead. The words attracting the largest coefficients by absolute value are greater, improved and strong. These findings should be compared with the informativeness of print news media for two-day ahead stock returns and two-day ahead stock trading volume, which are also driven by emotive vocabularies. These findings are consistent with foreign residents affecting domestic stock prices by adjusting their demand for domestic stocks on any given day based on the emotive content of domestic print media two days earlier.

The best model for three-day ahead debt portfolio flows using only lagged bond

\textsuperscript{27}These three numbers come from the fact that the root mean square prediction errors of the best models including print news media explanatory variables fall below those of the best models with only lagged returns, lagged squared returns and day-of-the-week explanatory variables by the fractions 0.982 − 0.914 ≈ 7%, 0.946 − 0.912 ≈ 3% and 0.875 − 0.895 ≈ 2% of the root mean square prediction error of the historical mean model. These fractions appear in the fourth row of each subtable in Table A.1 in the appendix.

\textsuperscript{28}This number can be calculated as 0.801 − 0.761 based on the fifth row of the two day ahead subtable of Table A.1 in the appendix.
returns, volatility or day-of-the-week variables cannot explain any of the variation in this dependent variable. Therefore, the 11% of the variance of three-day ahead bond portfolio flows explicable by the multi-factor representation of print news media does not seem to reflect lagged bond returns, volatility or day-of-the-week effects. This explanatory power also coincides with the informativeness of print news media for the three-day ahead long bond premium, which suggests that foreign residents’ demand for long-term relative to short-term domestic bonds on any given day depends on the content of domestic print media three days earlier.

5.3.3 Most informative words

Section 5.3.2 identifies some of the words that are most informative for stock returns, trading volume and capital flows in the sense of attracting the largest coefficients by absolute value in a full-sample estimation. This section delves deeper into the informativeness of these individual words by presenting the full-sample coefficients on such words and some of the other words that often co-occur with them.

Table 5.3 shows the most informative vocabulary items for stock returns three and four days ahead, as listed in Section 5.3.2. The most informative vocabulary items presented for three day ahead stock returns attract large coefficients of between 1 and 1.6 percent. If these coefficients are to be interpreted causally, in the sense of achieving substantive out-of-sample prediction accuracy, then they suggest that the occurrence of the item in the print news media leads to an increase in the stock return three days ahead by an amount of between 1 and 1.6 percent per unit of inverse document frequency log(m/DF).

However, this interpretation ignores the collinearity between

---

29 This idea can be seen in the three days ahead subtable of Table A.1 in the appendix, where the best relative root mean square prediction error of the models with lagged return, volatility and day-of-the-week explanatory variables is 1.086 > 1.

30 No particular vocabulary seems to drive the predictive content of the multi-factor representation of domestic print media for three-day ahead portfolio debt flows.

31 The caption to Table 5.3 describes its rows and columns.

32 This notation for inverse document frequency follows the usage and explanation in Section 4.2.
Table 5.3: The first column of this table presents the vocabulary items listed in Section 5.3.2 that attract some of the largest coefficients by absolute value in a full-sample estimation of the best-performing models for 3- and 4-day ahead stock returns. These vocabulary items correspond to the vocabulary of all individual words, pairs of consecutive words and triples of consecutive words. The coefficients correspond to a model without lagged stock return explanatory variables and with the best choice of C tuning parameter. For each vocabulary item, columns 3-8 show the other items in the vocabulary that correlate highest, in TF-IDF features over time, with the given item. These correlations appear underneath each other vocabulary item. Columns 3-8 exclude near duplicates, one-word versions of multi-word phrases, and parts of the acronyms in the cases of *pic*, *cwp*, *busa* and *gepf*. All words are displayed after converting to lower case.

(a) three day ahead returns

<table>
<thead>
<tr>
<th>item</th>
<th>coefficient</th>
<th>most associated items by TF-IDF features</th>
</tr>
</thead>
<tbody>
<tr>
<td>pic</td>
<td>0.016</td>
<td>masilela, maqha, dlamini, says pic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.76, 0.72, never, objective, pic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>chairman, pic still, gold, gold fields</td>
</tr>
<tr>
<td>gold</td>
<td>0.013</td>
<td>gold price, oz, miners, gold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>anglogold, ashanti, physical, gold</td>
</tr>
<tr>
<td></td>
<td></td>
<td>use, legal, use, responsibilities, legal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>use, legal, legal, legal, legal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>use, legal, legal, legal, legal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>use, legal, legal, legal, legal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>use, legal, legal, legal, legal</td>
</tr>
<tr>
<td>cwp</td>
<td>0.011</td>
<td>operative, banks, baml, grocery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>spend, use, banking, responsibilities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>multiple, reasons, multiple, reasons</td>
</tr>
<tr>
<td></td>
<td></td>
<td>multiple, reasons, multiple, reasons</td>
</tr>
<tr>
<td></td>
<td></td>
<td>multiple, reasons, multiple, reasons</td>
</tr>
<tr>
<td></td>
<td></td>
<td>multiple, reasons, multiple, reasons</td>
</tr>
<tr>
<td>busa</td>
<td>0.010</td>
<td>busa will, business, politics sa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>says busa, within, busa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>jerry, vilakazi, busa, busa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.52, 0.49, 0.48, 0.47, 0.46, 0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.52, 0.49, 0.48, 0.47, 0.46, 0.46</td>
</tr>
<tr>
<td>gepf</td>
<td>0.010</td>
<td>developmental, investment, riverside</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gepf pic, says pic, sound investments</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cancun, decisions, cancun, decisions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.89, 0.86, 0.84, 0.83, 0.83, 0.81</td>
</tr>
</tbody>
</table>

(b) four day ahead returns

<table>
<thead>
<tr>
<th>item</th>
<th>coefficient</th>
<th>most associated items</th>
</tr>
</thead>
<tbody>
<tr>
<td>business</td>
<td>0.002</td>
<td>small, says, sa</td>
</tr>
<tr>
<td></td>
<td></td>
<td>business, school</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.36, 0.34, 0.31, 0.31, 0.30, 0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>kumba, said</td>
</tr>
<tr>
<td></td>
<td></td>
<td>arbitration, process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.76, 0.75, 0.72, 0.68, 0.66, 0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>water, use, supply</td>
</tr>
<tr>
<td></td>
<td>water</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sa, water, utilities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>water, use, supply</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.63, 0.63, 0.60, 0.60, 0.56, 0.55</td>
</tr>
<tr>
<td></td>
<td>climate change</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>emissions</td>
<td>kyoto, protocol</td>
</tr>
<tr>
<td></td>
<td></td>
<td>climate change, response</td>
</tr>
<tr>
<td></td>
<td></td>
<td>greenhouse gas, cmp7, carbon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.59, 0.56, 0.55, 0.54, 0.53, 0.52</td>
</tr>
<tr>
<td></td>
<td>brokers</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>labour</td>
<td>areas, play</td>
</tr>
<tr>
<td></td>
<td></td>
<td>largest, insurance, brokers, cib, retain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.69, 0.62, 0.62, 0.61, 0.58, 0.55</td>
</tr>
</tbody>
</table>
vocabulary items. Individual words, for example, do not occur in isolation. When
they occur, they occur as parts of sentences with some grammatical structure. This
grammatical structure forced certain combinations of words to co-occur.\footnote{Section 5.2 shows evidence on the co-occurrence of words in the newspaper archive analysed here.} Therefore, when interpreting the coefficient on a particularly informative vocabulary item, it is important to remember that this coefficient reflects the informativeness for three day ahead stock returns of other co-occurring vocabulary items.

Table 5.3 shows the six other vocabulary items whose TF-IDF features are most correlated over time with the given most informative vocabulary item. The table shows that *pic* occurs most frequently with *masilela* and *maghawe dlamini*, referring to the CEO and a senior manager of the Public Investment Corporation (PIC) respectively. The item *pic chairman* also refers to a leadership position within the organisation and occurs often with *pic*. The item *says pic* refers to a quote from someone holding a position at the PIC and *never objective* refers to the investment objectives of the organisation. These co-occurring vocabulary items seem likely to be informative about future actions that the PIC would take. Moreover, given that the PIC manages a large quantity of assets relative to the size of the stock market, such actions could amount to sizable changes in demand for stocks. The positive coefficient on the PIC suggests that the organisation tends to be mentioned in the print news media when it intends to exercise its controlling rights in a positive way or intends to increase its demand for stocks. This interpretation of its usage in print news media is confirmed by the sample of occurrences presented in Figure 5.4.

The item *gold* is positively associated with stock returns three days ahead. It tends to appear together with the items *gold price* and *oz*, which allude to discussions of the gold market, and the items *gold miners*, *anglogold ashanti* and *gold fields*, which allude to discussions of gold mining activity. Together, these suggest that gold tends to appear in the print news media in reference to increased mining activity and favourable
rcise its control, be it through the PIC or as a buyer of services, has incre-

t the Public Investment Corporation (PIC) holder of a 12.2% stake in Stand-

t on, and that he would advise the PIC to oppose it at next months annual g-

ility provided by Holcim and gave the PIC a 20% equity stake in AfriSam. Brian

chairman of Barloworld in which the PIC was invested resigned days after Mr

ations was the means through which the PIC approached responsible investing in r-

mple. In the internal portfolio, the PIC owns 152 shares, the smallest of whi-

998 it had a staff of just eight. The PIC has never had an objective of profit.

THE Public Investment Corporation (PIC) announced on Friday that it had rea-
e than 80% of the note holders, the PIC said. PIC CEO Elias Masilela said: &

the form of a basic income grant. The CWP is cleverly designed and departs sign-

t estimated at more than R20bn. In the CWP, participants are paid R60 for a six-

ious cycle of unemployment, that the CWP has been designed. Initiated by gov-

ty organiser, Mario Wanza. Among the CWP projects in the area is a block watch

the broader community. Small bands of CWP workers in branded uniforms can be se-

For retailers and Capitec bank (all CWP participants must have bank accounts

olved, though, the arrival of the CWP has done more than bring an inflow of

ant community spirit. Philip says the CWP has provided an institutional mechan-

was closely involved in piloting the CWP when it was introduced in the second

e percentage) of 65%. There are 74 CWP sites around the country, with more

ould be huge, Business Unity SA (Busa) warned yesterday. It said the econ-

sons told the committees that while Busa supports deficit financing as an ess-

loan guarantees. Business Unity SA (Busa) and the National Union of Minework

arguing that this makes it unlikely Busa will stand up to the government in f-

month or an SX4 from R2957 a month. Busa warns that fuelling inflation will a-

ning point will occur is uncertain. Busa has raised concerns that Eskoms 35

den access to post-school education. Busa is to discuss education and training

st economic growth and job creation. Busa also said yesterday it was concer-

an against the wind in this regard. Busa shared the Banks concerns that elec-

yesterday. President Jacob Zuma told Busa that job losses early this year were

ble way to use national savings. The GEPF already provides support to govern-

eration of ethical investing at the GEPF, the largest pension fund in Africa

developmental and responsible. The GEPF board of trustees agrees that invest-

Thanks to these strong returns the GEPF is now fully funded. With R1trillion

stones of the relationship with the GEPF is the development investment policy

d world and Sasol in particular, the GEPF, the PICs biggest client with 90%

r Nhlanhla Nene said yesterday. The GEPF announced yesterday that R13bn would

vernance practices. According to the GEPF, the environmental sustainability f

Government Employees Pension Fund (GEPF), John Oliphant, by its board thr

is manner, it is essential that the GEPF board, and the finance ministry, e

Figure 5.4: Quotes from the archive of print news media containing representative
occurrences of the acronym vocabulary items pic, cwp, busa and gepf.
prospects for these firms.

References in the print news media to the Community Work Programme *cwp* tends to occur with references to demand for banking services and retail goods. The item *baml* references Bank of America Merrill Lynch, and tends to be used in relation to economic analysis or forecasts produced by the bank. The positive coefficient on *cwp* therefore suggests that the occurrence of this item in the print news media increases perceptions about future earnings in banking and retail sectors. The quotes from the print news media archive presented in Figure 5.4 support this interpretation.

Business Unity South Africa (*busa*) tends to be referenced in the context of actions that the organisation will take *busa will*, or in the context of statements made by the organisation *says busa* or its CEO *jerry vilakazi*. These actions and positions are evident in the sample of occurrences presented in Figure 5.4. Its positive coefficient suggests that these actions or positions are perceived as beneficial for future earnings of firms. The Government Employees’ Pension Fund is mentioned often in conjunction with the PIC, and also in relation to strategic investment decisions. These strategic investment decisions appear to relate to economic development and environmental obligations, with a reference to the Cancun climate change conference of 2010. These strategic investment topics are evidenced in the quotes shown in Figure 5.4.

The vocabulary items listed in Section 5.3.2 that attract large coefficients by absolute value in a large sample estimation for four day ahead returns are *business, kumba, water, climate change* and *brokers*. However, Table 5.3 shows that the coefficients on these vocabulary items are about one tenth of the size of those on the items that predict three day ahead returns. This reflects the lower choice of tuning parameter, $C = 2^{-8}$, for four day returns than the $C = 1$ chosen for three day returns. The item *climate change* attracts a negative coefficient, suggesting that the news on this topic

---

34 These two particular choices of tuning parameter are the two that perform best, out of the four options for $C$ considered in this work, for three and four day ahead returns respectively.
tends to be negative for firms. The item brokers refers most often to labour brokers, which are intermediaries between workers and firms. They have a negative reputation domestically for exploiting workers, which may explain the item’s negative coefficient.
Chapter 6

Conclusion

6.1 Contribution and implications

The print news media could play a role in asset pricing by affecting agents’ beliefs or preferences. These affected beliefs and preferences could be of all agents or of only some agents, and these agents in turn could be foreign or domestic.\(^1\) Previous work has explained the role for print news media in the pricing of stocks in terms of emotive language affecting either the beliefs of some irrational domestic agents, or the preferences of some susceptible domestic agents, in a large closed economy. I document some uncertainties with this conclusion owing to potential finite sample biases, reverse causality or the discovery of a purely historical phenomenon.

In this work, I use data on a small open economy to show evidence that confirms the importance of print news media for the pricing of stocks and bonds. Specifically, I find that a multi-factor representation of print news media can predict about 9 percent of the variation in daily stock returns (one day ahead) and 7 percent of the variation in the daily excess return of long-term bonds over short-term bonds (three days ahead), but only limited quantities of the variation in daily aggregate bond returns. This role

---

\(^1\)A role in asset pricing for heterogeneous agents should be understood to be accompanied by constraints on the agents whose beliefs or preferences are not subject to outside influences.
for the print news media does not seem to reflect a purely historical phenomenon, finite-sample biases, reverse causality, serial correlation, volatility or day-of-the-week effects, which therefore rules out the sources of uncertainty associated with previous work. To the best of my knowledge, this is the first out-of-sample evidence on the role for the print news media in determining the prices of assets and the first evidence in an open economy context.

I also present some evidence of three mechanisms by which these overall effects could operate. First, the excess predictability of print news media for stock returns up to two days ahead is driven by emotive language and large stocks, and it coincides with the excess predictability of print news media for equity capital inflows and trading volume up to two days ahead. To the best of my knowledge, this is the first evidence that the print news media could determine capital flows. This finding supports an interpretation of heterogeneous agents that trade with each other on the basis of recent (i.e. two days' prior) emotive language in the print news media. The finding also suggests that some of the agents are located abroad, which is consistent with an information asymmetry between foreign and domestic residents about domestic stocks. The clearer effect for larger stocks suggests either that emotive language in the domestic print news media influences foreign residents' beliefs about large stocks in particular, or that foreign residents act on their beliefs about all domestic stocks through those that have lower transaction costs, but is more difficult to reconcile with interpretations of the print news media affecting foreign residents’ risk aversion because risk aversion would affect both large and small stocks.

Second, the excess predictability of print news media for stock returns three and four days ahead is driven by non-emotive language and large stocks, and does not coincide with any excess predictability of print news media for equity capital inflows or trading volume at those forecast horizons. This finding supports an interpretation where agents’ beliefs and preferences respond homogeneously to non-emotive language
in the print news media three to four days prior. While the preceding finding suggests a short-term role for heterogeneous ‘sentiment’ between domestic and foreign residents, this finding suggests that other ‘non-sentiment’ factors associated with the print news media drive its medium-term effects on stock returns.

Third, the excess predictability of print news media for the excess return of long-term bonds over short-term bonds three days ahead is driven by non-emotive language and coincides with excess predictability of the print news media for portfolio debt net inflows from abroad. This finding supports an interpretation where foreign agents’ demand for domestic long-term bonds depends on non-emotive language three days prior, where foreign agents finance their purchases of domestic long-term bonds through selling domestic short-term bonds, and where foreign agents invest their proceeds from selling domestic long-term bonds in domestic short-term bonds.

### 6.2 Future research

This work explores the relationship between the print news media, asset prices and capital flows in a small open economy setting. We obtain several new findings, but many interesting questions remain. In closing, I set out ideas for future research and sketch how they could be approached.

The empirical results in this work demonstrate predictability for stock returns, long-term bond returns, trading volume and capital flows based on past information extracted from the print news media. The predictability is demonstrated through out-of-sample experiments, rather than through in-sample tests of statistical significance. Such tests of significance are almost always based on approximate asymptotic distributions, which can lead to problems of finite sample bias when there are many parameters to be estimated relative to the sample size.²

²I explain this problem, in the context of Tetlock [2007], in Footnote 4.
Nevertheless, even the out-of-sample estimates of mean square prediction error provided here are subject to sampling uncertainty. I provide some reassuring evidence in Section 5.3.1 that this out-of-sample prediction performance also holds in earlier subsamples. It would be interesting to compare the 9 and 7 percent reductions in variation that I obtain for stock returns and excess returns of long-term bonds over short-term bonds, given print news media information, to reductions in variation that could typically be expected to occur merely through sampling uncertainty alone. That is, it would be interesting to perform a formal hypothesis test on the out-of-sample relative root mean square prediction error that I obtain from the SVR estimation procedure.

Formal hypothesis testing in this context is not straightforward. While there have been recent advances in high-dimensional central limit theorems, I am not aware of out-of-sample applications or applications to the SVR context. In the absence of further theoretical work, one way to approach this question would be through a Monte Carlo experiment. A resampling procedure, for example the block bootstrap of Kunsch [1989], could be used to simulate a stochastic process for asset returns with weak dependence. By combining this procedure with the traditional so-called ‘i.i.d.’ bootstrap applied to the multi-factor representation of print news media, one could randomly resample new datasets. The collection of SVR estimators and out-of-sample root mean square prediction errors over all datasets would produce distributions against which the above statistics of 9 and 7 percent could be compared in a formal hypothesis test.

It would be interesting to extend the analysis performed here to the cross-sections of stock and bond returns. The equity premium and excess return of long-term bonds over short-term bonds correspond to two of the five factors that Fama and French [1993] find to be highly informative for the cross-section of stock and bond returns. Given that I find predictability here for these two factors, I would expect to see some predictability for such cross-sections. The other three common factors identified in Fama and French [1993] are the excess return of the smallest stocks in the stock market over the largest...
stocks in the stock market, the excess return of the stocks with the highest ratios of book value equity to market value equity over the stocks with the lowest ratios of book value equity to market value equity, and the excess return of corporate bonds over long-term government bonds. By constructing such factors for the South African market, the same methods used above could be applied to study the informativeness of the South African print news media for these factors. Evidence of predictability for these extra three factors would strengthen the case for the print news media in explaining the cross-sections of asset returns.

Finally, the most important and challenging extension to this work would be further investigation of the mechanism by which the print news media could be determining risky asset prices. I begin exploring this question in Section 5.3.3, which reveals that the frequency of occurrence of specific phrases in the print news media could give information about demand and supply for assets. This section also makes it seem more likely that three and four day ahead returns react to fundamental information than emotive content in domestic print news media. One as yet unexplored avenue for further insight into the mechanism is to identify phrases that have common predictive power for the returns on both stocks and bonds. If there are risk factors in the multi-factor representation of the print news media that are important for general asset prices, then this set of risk factors should be common to the returns on both stocks and bonds.
Appendix A

Informativeness of news for daily variables

This appendix provides more detail on the results that are presented in Section 5.3.2. Table A.1 shows the best out-of-sample predictive performance of classes of models with different groups of explanatory variables. In that table, there is one subtable for each forecast horizon. The first three columns of each subtable show the minimum, over the six vocabularies and four tuning parameters $C = 2^{-8}, 2^{-1}, 1, 2$, of the ratio of the root mean square (out-of-sample) prediction error of the news-only prediction model to the root mean square (out-of-sample) prediction error of a historical mean forecast model. The columns labelled ‘news’ apply to the class of models with print news media explanatory variables only; those labelled ‘returns’ apply to the class of models with lagged returns, lagged squared returns and day-of-the-week explanatory variables only; and columns labelled ‘both’ apply to the class of models with both these sets of explanatory variables. The column labelled $\rho$ in each subtable is calculated as the difference between one and entry in the ‘news’ column, if this difference is positive. Further descriptive details appear in the caption to the table.
Table A.1: Informativeness of the multiple media factors for the dependent variables of daily stock returns, stock trading volume, bond returns, the equity premium, the long bond premium and capital flows $n$ days ahead for $n = 1, 2, 3, 4, 5$. The subtables and their entries are explained in the text of Appendix A. The collection of all columns labelled $\rho$ are shown in Table 5.2 in the main text. Entries for $\rho$ that would otherwise be negative are not shown with the understanding that the news-only model is not more informative than the historical mean model in these cases. The calculation of each dependent variable is described in Chapter 3.

<table>
<thead>
<tr>
<th></th>
<th>one day ahead</th>
<th>two days ahead</th>
<th>three days ahead</th>
<th>four days ahead</th>
<th>five days ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>news returns</td>
<td>both $\rho$</td>
<td>news returns</td>
<td>both $\rho$</td>
<td>news returns</td>
</tr>
<tr>
<td>Top 40 return</td>
<td>0.881</td>
<td>0.966</td>
<td>0.960 0.12</td>
<td>0.849 1.025</td>
<td>0.996 0.15</td>
</tr>
<tr>
<td>Top 40 change in log volume</td>
<td>0.440</td>
<td>0.532</td>
<td>0.515 0.56</td>
<td>0.456 0.493</td>
<td>0.493 0.54</td>
</tr>
<tr>
<td>GOVI return</td>
<td>0.953</td>
<td>0.918</td>
<td>0.922 0.05</td>
<td>1.068 0.881</td>
<td>0.869 -</td>
</tr>
<tr>
<td>Equity premium</td>
<td>0.875</td>
<td>1.004</td>
<td>1.000 0.13</td>
<td>0.781 1.106</td>
<td>1.056 0.22</td>
</tr>
<tr>
<td>Long bond premium</td>
<td>0.938</td>
<td>0.931</td>
<td>0.931 0.06</td>
<td>1.410 0.957</td>
<td>0.947 -</td>
</tr>
<tr>
<td>Debt net inflow</td>
<td>1.012</td>
<td>1.157</td>
<td>1.156 -</td>
<td>0.994 0.908</td>
<td>0.908 0.01</td>
</tr>
<tr>
<td>Equity net inflow</td>
<td>0.937</td>
<td>0.881</td>
<td>0.881 0.06</td>
<td>0.761 0.801</td>
<td>0.797 0.24</td>
</tr>
<tr>
<td>Debt and equity net inflow</td>
<td>1.092</td>
<td>0.797</td>
<td>0.795 -</td>
<td>0.975 0.816</td>
<td>0.815 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.839 1.011</td>
<td>1.008 0.16</td>
</tr>
<tr>
<td>Top 40 change in log volume</td>
<td>0.456</td>
<td>0.437</td>
<td>0.437 0.54</td>
<td>0.452 0.471</td>
<td>0.472 0.55</td>
</tr>
<tr>
<td>GOVI return</td>
<td>0.969</td>
<td>0.919</td>
<td>0.925 0.03</td>
<td>1.012 0.961</td>
<td>0.960 -</td>
</tr>
<tr>
<td>Equity premium</td>
<td>0.834</td>
<td>1.040</td>
<td>1.035 0.17</td>
<td>0.933 1.084</td>
<td>1.084 0.07</td>
</tr>
<tr>
<td>Long bond premium</td>
<td>0.914</td>
<td>0.982</td>
<td>0.979 0.09</td>
<td>0.912 0.946</td>
<td>0.922 0.09</td>
</tr>
<tr>
<td>Debt net inflow</td>
<td>0.885</td>
<td>1.086</td>
<td>1.084 0.11</td>
<td>1.149 0.990</td>
<td>0.987 -</td>
</tr>
<tr>
<td>Equity net inflow</td>
<td>0.915</td>
<td>0.876</td>
<td>0.870 0.08</td>
<td>0.839 0.801</td>
<td>0.798 0.16</td>
</tr>
<tr>
<td>Debt and equity net inflow</td>
<td>0.880</td>
<td>0.834</td>
<td>0.832 0.12</td>
<td>1.028 0.953</td>
<td>0.953 -</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.062 1.127</td>
<td>1.127 -</td>
</tr>
<tr>
<td>Top 40 change in log volume</td>
<td>0.557</td>
<td>0.557</td>
<td>0.554 0.44</td>
<td>0.557 1.027</td>
<td>1.000 0.06</td>
</tr>
<tr>
<td>GOVI return</td>
<td>0.935</td>
<td>1.027</td>
<td>1.000 0.06</td>
<td>1.058 1.066</td>
<td>1.066 -</td>
</tr>
<tr>
<td>Equity premium</td>
<td>1.058</td>
<td>1.066</td>
<td>1.066 -</td>
<td>0.916 0.895</td>
<td>0.875 0.08</td>
</tr>
<tr>
<td>Long bond premium</td>
<td>1.307</td>
<td>1.006</td>
<td>1.006 0.08</td>
<td>0.905 1.611</td>
<td>1.607 0.10</td>
</tr>
<tr>
<td>Debt net inflow</td>
<td>0.974</td>
<td>1.020</td>
<td>1.019 0.03</td>
<td>1.020 1.019</td>
<td></td>
</tr>
</tbody>
</table>
Bibliography


