A tale of two sentiment scales: Disentangling short-run and long-run components in multivariate sentiment dynamics

Danilo Vassallo* Giacomo Bormetti† Fabrizio Lillo§‡

Abstract

The digitalization of news and social media provides an unprecedented source to investigate the role of information on market dynamics. However, the observed sentiment time-series represent a noisy proxy of the true investor sentiment. Moreover, modeling the joint dynamics of different sentiment series can be beneficial for the assessment of their economic relevance. The main methodological contribution of this paper is twofold: i) we filter the latent sentiment signals in a genuinely multivariate model; ii) we propose a decomposition into a long-term random walk component, named long-term sentiment, and a short-term component driven by a stationary Vector Autoregressive process of order one, named short-term sentiment. The proposed framework is a dynamic factor model describing the joint evolution of the observed sentiments of a portfolio of assets. Empirically, we find that the long-term sentiment co-integrates with the market price factor, while the short-term sentiment captures transient and firm-specific swings. By means of quantile regressions, we assess the significance of the explanatory power of filtered present sentiment on future returns. Then, we demonstrate how the lagged relation can be successfully exploited in a portfolio allocation exercise.

Keywords: Investment analysis; Sentiment analysis; Kalman filter; expectation maximization; quantile regressions

JEL codes: C30; C58; G11

*Scuola Normale Superiore, Piazza dei Cavalieri 7, Italy. Corresponding author. E-mail: danilo.vassallo@sns.it

†University of Bologna, Italy E-mail: giacomo.bormetti@unibo.it

‡University of Bologna and Scuola Normale Superiore, Italy E-mail: fabrizio.lillo@unibo.it

§The authors thank Thomson Reuters for kindly providing Thomson Reuters MarketPsych Indices time series. We benefited from discussion with Giuseppe Buccheri, Fulvio Corsi, Luca Trapin, as well as with conference participants to the Quantitative Finance Workshop 2019 at ETH in Zurich and the AMASES XLIII Conference in Perugia. Fabrizio Lillo acknowledges support by the SOBIGDATA++ project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871042.

1 Introduction

Nowadays, as Ignacio Ramonet wrote in The Tyranny of Communication, "a single copy of the Sunday edition of the New York Times contains more information than an educated person in the eighteenth century would consume in a lifetime". This huge amount of information cannot be read by a single person. Recent developments in machine learning algorithms for sentiment analysis help us to categorise and extract signals from text data and pave the way for a new area of research. The use of these new sources of textual data has become popular to analyse the relationship between sentiment and other economic variables using econometric techniques. (Algaba et al., 2020) refer to this new strand of literature as *Sentometrics*. A vast literature focuses on the forecast of future market returns or volatility using sentiment. For instance, (Groß-Klußman and Hautsch, 2011) study the impact of unexpected news on the displayed quotes in a limit order book, (Sun et al., 2016; Schnaubelt et al., 2020) show that intraday S&P 500 index and constituent returns are predictable using lagged investor sentiment, (Peterson, 2016) investigates the trading strategies based on sentiment, (Tetlock, 2007; Garcia, 2013) consider the Dow Jones Industrial Average (DJIA) index predictability using sentiment, (Calomiris and Mamaysky, 2019) show how the predictability can be exploited in different markets around the world, while Kim and Kim (2014) finds no evidence of predictability coming from messages on Yahoo! Finance. In addition to market returns forecast, sentiment indicators have been used to forecast other market and economic variables. (Antweiler and Frank, 2004; Borovkova and Mahakena, 2015; Allen et al., 2015; Smales, 2015; Jiao et al., 2020) study the impact of sentiment on volatility, (Ranco et al., 2015) analyse the impact of social media attention on market dynamics, (Borovkova, 2015) develops risk measures based on sentiment index, (Nyman et al., 2021; Mai et al., 2019) find predictive power of sentiment indices on financial crisis and firms' bankruptcy respectively, (Feuerriegel and Gordon, 2019) use financial news to predict macroeconomic indicators and (Lillo et al., 2015) show that different types of investors react differently to news sentiment. Huynh et al. (2021) shows that media coverage and investor sentiment was able to predict stock returns and volatility during the recent COVID-19 pandemic.

As observed by Zygmunt Bauman in Consuming Life, as the amount of information increases also the amount of useless information increases, and the noise becomes predominant. Two different non-exclusive methods have been explored in the literature to remove or, at least, mitigate the impact of useless information. In the first case, a general-to-specific approach is used directly on the textual data. The amount of information can be reduced selecting only verified news (i.e. eliminating fake news), considering only the words which are closely related to the topic of interest, considering the importance of any news (e.g. Da et al. 2011), selecting only news which appear for the first time (e.g. Thomson Reuters News Analytics engine uses the novelty variable, see Borovkova et al. 2017), or weighting a news by means of a measure of attention (e.g. with the number of clicks it receives when published in a news portal Ranco et al. 2016). Obviously, the selection of the relevant data is application-specific. For instance, fake news may be irrelevant to forecast the GDP of a country but may be crucial to forecast the results of an election (e.g. Allcott

and Gentzkow 2017).
In the second case, sentiment time series are directly considered, rather than the text source they are built from. The observed sentiment is noisy and various approaches have been proposed to filter it and to recover the latent signal. (Thorsrud, 2018) applies a 60-day moving average, (Peterson, 2016) uses the Moving Average Convergence-Divergence methodology proposed in (Appel, 2003) and (Borovkova and Mahakena, 2015; Audrino and Tetereva, 2019; Borovkova et al., 2017) introduce the Local News Sentiment Level model (LNSL), a univariate method which takes inspiration from the Local Level model of (Durbin and Koopman, 2012). In spite of its convenience from a practical perspective, the moving average approach is not statistically sound and the window length is usually chosen following rules of thumb, which have been tested empirically but lack a clear theoretical motivation. The methods based on the Kalman-Filter techniques present a natural and computationally simple choice to extract informative signal. Unfortunately, when multiple assets are considered in the analysis, the LNSL model does not exploit the multivariate nature of the data. One goal of this paper is to show that the covariance structure is very informative in sentiment time series analysis.

The first contribution of this paper is to extend the existing time series methods in the latter stream of literature. We propose to model noisy sentiment disentangling two different sentiment signals. In our approach, the observed sentiment follows a linear Gaussian state-space model with three relevant components[1]. The first component, named *long-term sentiment* is modeled as a random walk, the second component is termed *short-term sentiment* and follows a VAR(1) process, and the last component is an i.i.d Gaussian *observation noise* process. We name the novel sentiment state-space model Multivariate Long Short Sentiment (MLSS). In the empirical section of the paper, we separately model the sentiment coming from newspapers (news sentiment) and the sentiment extracted from social media (social sentiment). We show that the decomposition provides a better insight on the nature of sentiment time series, linking the long-term sentiment to the long-term evolution of the market – proxied by the market factor – while the short-term sentiments reflect transient swing of the market mood and is more related to the market idiosyncratic components. Specifically, we find that i) the second news long-term sentiment cointegrates with the first market factor extracted via PCA; ii) the correlation structure of the short-term sentiment explains a significant and sizable fraction of correlation of return residuals of a CAPM model. Finally, we show that the multivariate local level model provides the best description of the data with respect to alternative models, such as the LNSL.

The second contribution of the paper is to unravel the relation between news and market returns conditionally on quantile levels. We perform various quantile regressions showing that sentiment has good explanatory power of returns. When contemporaneous effects are considered, the result is expected and holds for all models at intermediate quantile levels. However, when the analysis is focused on abnormal days – i.e. days for which returns belong to the 1% and 99% quantiles – neither the noisy sentiment nor the filtered sentiment from an LNSL model explain the observed

[1]In the empirical application we find that the Gaussian assumption is appropriate for the investigated data

market returns. The only model achieving statistical significance is the MLSS. This result shows that it is essential to filter the noisy sentiment according to the MLSS, which exploits both the multivariate structure of the data and disentangles the long- and short-term components. Moreover, a test performed on the single components confirms the intuition that the short-term sentiment is the one responsible for the contemporaneous explanatory power. The empirical evidence in favor of the MLSS becomes even more compelling when lagged relations are tested. When a single day lag is considered, i.e. one tests whether yesterday sentiment explains today returns, the significance of all models, but MLSS, drops to zero. This result holds across all quantile levels. Instead, for quantiles smaller than 10% and larger than 90%, the returns predictability for the MLSS model is highly significant. As before, the decomposition in two time scales is essential and the short-term component is the one responsible of the effect. The analysis extended including lagged sentiment – up to five days – confirms previous findings by (Garcia, 2013) that past sentiment contributes in predicting present returns. Interestingly, this is true for quantiles between 5% and 10%, both negative and positive, but neither in the median region nor for extreme days. In light of this findings, we finally investigated whether media and social news immediately digest market returns and whether this relation depends on the sign of returns. Our results provide a clear picture showing that i) the impact of market returns on sentiment is significant up to five days in the future when negative extreme returns – i.e. belonging to quantiles from 1% to 10% – are considered, ii) when positive returns are considered the impact rapidly fades out and is significant only for quantiles smaller than 5%, iii) previous findings become not significant if the MLSS sentiment is replaced by the observed noisy sentiment. Consistently with the intuition provided by these results, we test whether the returns predictability of the MLSS model can be exploited in a portfolio allocation exercise. We show that the portfolios generated with the MLSS sentiment series have higher Sharpe ratio and lower risk than similar portfolios constructed with raw sentiment or sentiment filtered with the univariate LNSL model. Our model outperforms also the benchmark constituted by the buy-and-hold equally weighted portfolio. This result remains true when transaction costs are included.

The rest of the paper is organized as follows. In section 2, we develop the multivariate model for the sentiment and discuss the estimation technique. In section 3, we introduce the TRMI sentiment index and describe the data used in the analysis. In section 4, we report the empirical findings and discuss the advantages of the multivariate approach. In section 5, we compare the various techniques and report the performances of the long-short sentiment decomposition in explaining daily returns. Section 6 describes the portfolio allocation strategies using different filtering techniques and assesses the superiority of the MLSS filter among the others. Section 7 draws the relevant conclusions and sketch possible future research directions.

2 The Model

Consider K assets and the corresponding K observed daily sentiment series S_t^i where $i = 1, \ldots, K$. The observed daily sentiment S_t^i quantifies the opinions of investors and consumers about company

i. In most cases, the observed sentiment is a continuous number in a compact set.

The Local News Sentiment Level model (LNSL), presented in (Borovkova and Mahakena, 2015) and subsequently used in (Audrino and Tetereva, 2019), reads as follows

$$\begin{aligned} S_t^i &= F_t^i + \epsilon_t, \qquad \epsilon_t \overset{d}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^i\right), \\ F_t^i &= F_{t-1}^i + v_t, \quad v_t \overset{d}{\sim} \mathcal{N}\left(0, \sigma_v^i\right). \end{aligned} \tag{2.1}$$

for every $i = 1, \ldots, K$. This model is a univariate specification of the Local Level model of (Durbin and Koopman, 2012). The latent sentiment series F_t^i are considered as slowly changing components, modeled as independent random walks and the parameters σ_ϵ^i and σ_v^i are estimated via maximum likelihood (MLE).

Since the LNSL model does not consider the correlations of the innovations among the K assets, we can easily derive its multivariate version as

$$\begin{aligned} S_t &= F_t + \epsilon_t, \qquad \epsilon_t \overset{d}{\sim} \mathcal{N}\left(0, R\right), \\ F_t &= F_{t-1} + v_t, \quad v_t \overset{d}{\sim} \mathcal{N}\left(0, Q\right). \end{aligned} \tag{2.2}$$

where $S_t = \left[S_t^1, \ldots, S_t^K\right]'$ and $F_t = \left[F_t^1, \ldots, F_t^K\right]'$ are K dimensional vectors, Q is a $K \times K$ symmetric matrix and R is a $K \times K$ diagonal matrix. We refer to the multidimensional LNSL model as MLNSL. The synchronous correlation among the innovations of the latent sentiment are described by the covariance matrix Q, while the correlations among the observation noises are assumed to be 0. Clearly, the LNSL model is a special case of the MLNSL model when the matrix Q is diagonal. Since the number of parameters for this model scales as K^2, the MLE of the MLNSL model is computationally demanding. For this reason, we use the Kalman-EM approach described in (Corsi et al., 2015).

The idea of the LNSL and MLNSL models is that the latent sentiment is a slowly changing component with a Gaussian disturbance. In their empirical studies, (Audrino and Tetereva, 2019) observe that the signal to noise ratio $\frac{\sigma_v^2}{\sigma_\epsilon^2}$, obtained using the LNSL filter, is very small. This finding indicates that the majority of the daily changes in the sentiment series can be considered as noise. One possible explanation of this result is that the Local Level specification of these models is not sufficiently rich to capture all the signals from the observed sentiment. Indeed, in newspapers and social media there is a consistent amount of articles and opinions which represent fast trends or rapidly changing consumer preferences. Following the recent strand of literature on persuasion (Gerber et al., 2011; Hill et al., 2013), these fast trends have strong but short-lived effects on consumer preferences. Since the (M)LNSL model, introduced in Borovkova and Mahakena (2015) and Audrino and Tetereva (2019), considers the sentiment as an integrated series, these mean-reverting signals are treated as noise.

The main contribution of this paper is to define a new model which disentangles the slowly changing sentiment from a rapidly changing sentiment, that we name short-term sentiment, and the observation noise. We build our model on the existing approach of Borovkova (2015) and

Audrino and Tetereva (2019) and we specify the slowly changing components as a unit root process. In addition, it is reasonable to think that the slowly changing components of a set of firms with common characteristics, for instance belonging to the same sector, market, or country, should be affected by the same trends and shocks. For this reason, in our model we consider a number $q \leqslant K$ of factors driving the slow component of the sentiment dynamics. We name these factors as long-term sentiment. We do not fix the number q *a priori*, but we select it by means of information criteria.

To provide a more quantitative intuition behind our modeling specification, let us consider the true, but unobserved, daily investor's mood M_t^i of asset i. We hypothesize today's daily mood can be written as

$$\text{Mood}_t^i = \text{Long-term Mood}_t^i + \text{Short-term Mood}_t^i. \tag{2.3}$$

Based on Audrino and Tetereva (2019) and Borovkova and Mahakena (2015), the Long-term Mood is composed by yesterday's Long-term Mood plus a shock $s_t^{i,\,\text{long}}$, which is usually small but permanent, i.e.

$$\text{Long-term Mood}_t^i = \text{Long-term Mood}_{t-1}^i + s_t^{i,\,\text{long}}.$$

On the contrary, the Short-term Mood is short-lived, but with a strong and highly influential impact. In particular, the Short-term Mood is composed by a residual part of the yesterday Short-term Mood plus a shock $s^{i,\,\text{short}}$, i.e.

$$\text{Short-term Mood}_t^i = \phi^i \text{Short-term Mood}_{t-1}^i + s_t^{i,\,\text{short}}.$$

In this framework, the long-term shocks permanently change the investor's mood while the short-term shocks has an exponentially decaying persistence in the investor's mood. Equation (2.3) can be rewritten as

$$\text{Mood}_t^i = \text{Long-term Mood}_{t-1}^i + s_t^{i,\,\text{long}} + \phi^i \text{Short-term Mood}_{t-1}^i + s_t^{i,\,\text{short}}. \tag{2.4}$$

Considering the whole story and the dynamic of the two sentiments shocks, we can rewrite equation (2.4) as

$$\text{Mood}_t^i = \underbrace{\sum_{k=-\infty}^{t} (\phi_i)^{t-k} s_k^{i,\,\text{short}}}_{\text{Short-term Mood}_t^i} + \underbrace{\sum_{k=-\infty}^{t+1} s_k^{i,\,\text{long}}}_{\text{Long-term Mood}_t^i},$$

where we assumed $\text{Mood}_{-\infty}^i$ to be negligible and equal to zero. In full generality, the multivariate version of model (2.3) can be formulated as follows

$$\text{Mood}_t = A \times \text{Long-term Mood}_t + B \times \text{Short-term Mood}_t\,,$$

with A and B being $K \times K$ matrices. However, in light of the considerations in the previous paragraph, we restrict the matrix B to be the identity matrix. In this way, the Short-term Mood

is purely company-specific. We replace $A\,\text{Long-term Mood}_t$ with the product between a factor loading matrix and a limited number of factors, that is we rewrite the previous equation as

$$\text{Mood}_t = \Lambda \times \text{Long-term Factor Mood}_t + \text{Short-term Mood}_t, \tag{2.5}$$

where Λ belongs to $\mathbb{R}^{K\times q}$ with $q \leq K$. It is important to notice that the significance of Λ can be statistically tested and the selection of the number q of factors can be performed by means of AIC and BIC criteria. Following Audrino and Tetereva (2019), we assume that the observed sentiment S_t is a noisy observation of the investors Mood_t, and we formulate a state-space model for S_t consistent with the intuition provided by model (2.5). The Multivariate Long Short Sentiment model (MLSS) for the observed sentiment model, assuming a Gaussian specification for the short-term sentiment shock, long-term sentiment shock and the observation noise, reads

$$\begin{aligned} S_t &= \Lambda F_t + \Psi_t + \epsilon_t, \quad && \epsilon_t \overset{d}{\sim} \mathcal{N}(0, R), \\ \Psi_t &= \Phi\Psi_{t-1} + u_t, && u_t \overset{d}{\sim} \mathcal{N}(0, Q_{short}), \\ F_t &= F_{t-1} + v_t, && v_t \overset{d}{\sim} \mathcal{N}(0, Q_{long}), \end{aligned} \tag{2.6}$$

where $R \in \mathbb{R}^{K\times K}$ is the diagonal covariance matrix of the observation noise ϵ_t, $\Phi \in \mathbb{R}^{K\times K}$ is the matrix of autoregressive coefficients, $Q_{short} \in \mathbb{R}^{K\times K}$ is the covariance matrix of the short-term sentiment innovations, and $Q_{long} \in \mathbb{R}^{q\times q}$ is the covariance matrix of the random walk innovations. Section B in the online material shows an alternative specification for the long-term component, but the implementation of this model is left for future work. In equation (2.6), F_t and Ψ_t are the latent processes which proxy the Long-term Factor Mood and Short-term Mood in (2.5), respectively. Please notice that the essential difference between equation (2.5) and equation (2.6) is that the observed sentiment, and its components, are noisy versions of the investors' mood and its long and short components. Finally, in this paper, we force a diagonal structure on the matrix Φ, thus neglecting the possible lead-lag effects among sentiments. This restriction is introduced to limit the curse of dimensionality of the model. As an alternative approach allowing for possible lead-lag effects, one should consider to implement well-known regularization techniques. Chen et al. (2017) propose an Expectation Regularization Maximization algorithm to estimate high-dimensional state space model. However, this methodology can not be used in the case of augmented state-space models. A possible extension of regularization techniques for augmented state-space models is an interesting research problem, whose investigation is beyond the scope of the current work. After model (2.6) is estimated on data (see Section 4), we perform the standard tests on the residuals finding that they are approximately normally distributed, serially uncorrelated, as well as their absolute values [2]. All these results indicate that a Gaussian state space model is an appropriate specification for the sentiment dataset.

It is worth noticing that, if the observed sentiment S_t lies in a compact set, the LNSL, MLNSL

[2] Detailed results are available upon request.

and MLSS models, in their current specification, do not consider the upper and lower bounds. This issue can be accounted for with a non linear transformation of the data, e.g. by Fisher transforming the data, which maps them on the whole real line as commonly done for correlation time series. In this paper, we do not apply any non linear transformation since most of the daily sentiment observations are far from the bounds and, as we verified, the Fisher transform mildly affects our analysis. In addition, the definition of unit root processes in presence of bounds is non standard but studied in the literature and it can be tested using ad-hoc unit root tests (Cavaliere and Xu, 2014).

The estimation of the unknown parameters is based on a combination of the Kalman filter with Expectation Maximization (Kalman, 1960; Shumway and Stoffer, 1982; Wu et al., 1996; Harvey, 1990; Banbura and Modugno, 2014; Jungbacker and Koopman, 2008). Given that model (2.2) is a special case of model (2.6), in the next session we only consider the estimation procedure of model (2.6).

2.1 Estimation procedure

The estimation of model (2.6) is performed using the Kalman filter (Kalman, 1960) and the Expectation Maximization (EM) method in (Dempster et al., 1977) and (Shumway and Stoffer, 1982) which was proposed to deal with incomplete or latent data and intractable likelihood. The EM algorithm is a two-step estimator. In the first step, we write the expectation of the likelihood considering the latent process as observed. In the second step, we re-estimate the static parameters maximizing the expectation obtained in the first step. This routine is repeated until some convergence criterion is satisfied. To cast (2.6) in a standard state-space representation, we use the same procedure of (Banbura and Modugno, 2014) and define the augmented states $\tilde{\Lambda}$, $\tilde{F}$, $\tilde{\Phi}$ and $\tilde{Q}$ s.t.

$$\begin{aligned} S_t &= \tilde{\Lambda}\tilde{F}_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, R), \\ \tilde{F}_t &= \tilde{\Phi}\tilde{F}_{t-1} + v_t, \quad v_t \sim \mathcal{N}\left(0, \tilde{Q}\right), \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} &\tilde{\Lambda} = \begin{bmatrix} \Lambda & I_K \end{bmatrix} \in \mathbb{R}^{K\times(q+K)}, \quad \tilde{F}_t = \begin{bmatrix} F_t \\ \Psi_t \end{bmatrix} \in \mathbb{R}^{(q+K)\times 1}, \quad \tilde{\Phi} = \begin{bmatrix} I_q & 0 \\ 0 & \Phi \end{bmatrix} \in \mathbb{R}^{(q+K)\times(q+K)}, \\ &\tilde{Q} = \begin{bmatrix} Q_{long} & 0 \\ 0 & Q_{short} \end{bmatrix} \in \mathbb{R}^{(q+K)\times(q+K)}. \end{aligned} \tag{2.8}$$

The EM renders the approach feasible in high dimension. Indeed, while a direct numerical maximization of the likelihood is computationally demanding, the EM algorithm, thanks to the Kalman filtering and smoothing recursions in Appendix A, can be formulated using the closed-form equations reported in Appendix B. In particular, it allows to disentangle the long-term sentiment F_t and the short-term sentiment Ψ_t. To derive the EM steps we consider the log-likelihood $l\left(S_t, \tilde{F}_t, \theta\right)$

where θ denotes the set of static parameters $\tilde{\Lambda}$, $\tilde{\Phi}$, $\tilde{Q}$ and R. The EM proceeds in a sequence of steps:

1. E-step: it evaluates the expectation of the log-likelihood using the estimated parameters from the previous iteration $\theta(j)$:

$$G\left(\tilde{\Lambda}(j), \tilde{\Phi}(j), \tilde{Q}(j), R(j)\right) = E\left[l\left(S_t, \tilde{F}_t, \theta(j)\right) | S_1, \ldots, S_T\right]. \tag{2.9}$$

   The E-step strongly relies on equations (A.1). The details are explained in Appendix A.

2. M-step: the parameters are estimated again maximizing the expected log-likelihood with respect to θ:

$$\theta(j+1) = \arg\max_{\theta} G\left(\tilde{\Lambda}(j), \tilde{\Phi}(j), \tilde{Q}(j), R(j)\right). \tag{2.10}$$

   The M-step is performed updating the static parameters following equations (B.2)-(B.5). Details are provided in Appendix B.

We initialize the parameters $\theta(0)$ and repeat steps 1 and 2 until we reach the convergence criterion

$$\frac{|l\left(S_t, \tilde{F}_t, \theta(j)\right) - l\left(S_t, \tilde{F}_t, \theta(j-1)\right)|}{|l\left(S_t, \tilde{F}_t, \theta(j)\right) + l\left(S_t, \tilde{F}_t, \theta(j-1)\right)|} < \frac{\epsilon}{2}. \tag{2.11}$$

We set $\epsilon = 10^{-3}$. As observed in (Harvey, 1990), the dynamic factor model (2.7) is not identifiable. Indeed, if we consider a non singular invertible matrix M, then the parameters $\theta_1 = \{\Lambda, R, Q\}$ and $\theta_2 = \{\Lambda M^{-1}, R, MQM'\}$ are observationally equivalent, then starting from S_t we cannot distinguish θ_1 from θ_2. We solve this identification problem using the approach proposed by (Harvey, 1990), imposing the following restrictions

$$\tilde{Q} = \begin{bmatrix} I_q & 0 \\ 0 & Q_{short} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_{11} & 0 & 0 & \ldots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{K1} & \lambda_{K2} & \lambda_{K3} & \ldots & \lambda_{Kq} \end{bmatrix} \tag{2.12}$$

where Λ is the $K \times q$ sub-matrix in (2.8).

The specifications of $\tilde{\Lambda}$, $\tilde{\Phi}$, $\tilde{Q}$ and R in (2.8), together with (2.12), impose several constraints to the estimation. The EM procedure allows us to impose restrictions on the parameters in a closed-form. According to (Wu et al., 1996) and (Bork, 2009), we get the constrained $\tilde{\Phi}$, $\tilde{\Lambda}$, $\tilde{Q}$ and R as:

$$\mathsf{vec}(\tilde{\Phi}_r) = \mathsf{vec}(\tilde{\Phi}) + \left(A^{-1} \otimes \tilde{Q}\right) M(M(A^{-1} \otimes \tilde{Q})M')^{-1}(k_\Phi - M\mathsf{vec}(\Phi)) \tag{2.13}$$

where A is defined in (B.1), M is the $f \times 2K(r+K)$ matrix, f is the number of constraints, k_Φ is the f vector containing the constraints values such that $M\mathsf{vec}(\tilde{\Phi}) = k_\Phi$.
Equivalently, for the restricted Λ_r:

$$\mathsf{vec}(\Lambda_r) = \mathsf{vec}(\Lambda) + \left(E_1^{-1} \otimes R\right) G(G(E_1^{-1} \otimes R)G')^{-1}(k_\lambda - G\mathsf{vec}(\Lambda)) \tag{2.14}$$

where E_1 is defined in (B.1), G is the $s \times Kr$ matrix, s is the number of constraints, k_λ is the s vector containing the constraints values such that $G\mathsf{vec}(\Lambda) = k_\lambda$. $\tilde{Q}$ and R are evaluated using equations (B.3) and (B.4) and the restrictions, according to (Wu et al., 1996), can be imposed elementwise.

The final estimation scheme reads as follows:

1. Initialize $\tilde{\Lambda}(0)$, $\tilde{\Phi}(0)$, $\tilde{Q}(0)$ and $R(0)$
2. Perform the E-step using the estimations $\tilde{\Lambda}(j)$, $\tilde{\Phi}(j)$, $\tilde{Q}(j)$ and $R(j)$ and the Kalman Smoother (A.1).
3. Perform the M-step and evaluate the new estimators $\tilde{\Lambda}(j+1)$, $\tilde{\Phi}(j+1)$, $\tilde{Q}(j+1)$ and $R(j+1)$ using equations (B.2)-(B.4).
4. Use the unrestricted estimations and (2.13) and (2.14) to obtain the restricted ones.
5. Repeat 2, 3 and 4 above until the estimates and the log-likelihood reach convergence.

Finally, we select the number of long-term sentiment q using the AIC and BIC indicators.

3 Data

The approaches to sentiment analysis can be broadly classified into three categories. The first class is based on (mostly supervised) Machine Learning techniques. Three steps are typically considered. The first one is to collect textual data forming the training dataset. The second one is to select the text features for classification and to pre-process the data according to the selection. The final step is to apply a classification algorithm to the textual data. As an example, (Pang et al., 2002) compare the performance of Naive Bayes, support vector machines, and maximum entropy algorithm to classify positive or negative movie reviews. The second category is the lexicon-based approach. It also typically consists of three steps. The first step is the selection of a dictionary of N words which could be relevant for a specific topic (e.g. the word *great* is considered as a positive word to review a movie). The second one consists in tokenizing the textual data and, for each word in the dictionary, count how many times it appears in the text. This process can be visualized with a vector of length N where the i-th element represents the number of times the i-th word of the dictionary is mentioned in the text. Finally, a measure takes the vector of length N as an input and gives a quantitative score as an output. One can refer to (Loughran and McDonald, 2011) for a relevant example in the financial literature. The third and last approach is a combination

of methodologies coming from the first and second approach. For an overview of textual data treatments and computational techniques, we refer to the review paper (Vohra and Teraiya, 2013) and the book (Liu, 2015).

In this paper, we use a pre-classified index named The TRMI sentiment index. It is constructed using over 700 primary sources, divided into news and social media, and collects more than two million articles per day. For every article, a "bag-of-words" technique is used to create a sentiment score, which lies between -1 and $+1$, a buzz variable[3], and one or more asset codes, which in our case refer to companies. The time resolution of the sentiment data is one minute.

For any asset a, minute s, and day t we denote as $S^a_{t,s}$ the sentiment score and as $\text{Buzz}^a_{t,s}$ the buzz variable. The variable $S^a_{t,s}$ can be specific for *news* or *social* sentiment. In this paper, we use both the series, but analyze them separately. Since the following empirical analysis are performed using daily data, we need to aggregate the TRMI series on a daily basis. TRMI user guide suggests to use the following equation

$$S^a_t = \frac{\sum_{s=4:00\ \text{PM}^{t-1}}^{4:00\ \text{PM}^t} Buzz^a_{t,s} S^a_{t,s}}{\sum_{s=4:00\ \text{PM}^{t-1}}^{4:00\ \text{PM}^t} Buzz^a_{t,s}} \in [-1, 1]\,, \tag{3.1}$$

where S^a_t refers to the daily sentiment at day t, evaluated on a 24-hour window between the 4:00 PM of day $t-1$ (4:00 PM^{t-1}) and the 4:00 PM of day t (4:00 PM^t). Note that the TRMI server provides a daily frequency sentiment, where they use equation (3.1) with a 24-hours window from 3:30 PM to 3:30 PM of the day after. However, since we want to relate the sentiment series with close to close returns, we construct the daily sentiment series aggregating the high-frequency sentiment according to the trading closing hour of the NYSE (4:00 PM). For more details, please refer to (Peterson, 2016).

For the empirical analysis, we consider the TRMI sentiment index of 27 out of 30 stocks [4] of the Dow Jones Industrial Average (DJIA) over the period 03/01/2006 – 29/12/2017. The companies that we considered are the components of the DJIA as of December 29, 2017. Since the TRMI index divides the news sentiment from the social sentiment, we have a total of 54 time series. A description of tickers, sectors and summary statistics is available in Sector A of the online material. Finally, the MLSS model, in its current specification, does not manage missing values in data, while some of the sentiment time series present missing observations. The EM algorithm is naturally designed to handle missing observations. However, since the number of missing values is small [5], we fill them

[3] "The buzz field represents a sum of entity-specific words and phrases used in TRMI computations. It can be non-integer when any of the words/phrases are described with a minimizer, which reduces the intensity of the primary word or phrase. For example, in the phrase "less concerned" the score of the word "concerned is" mitigated by "less". Additionally, common words such as "new" may have a minor but significant contribution to the Innovation TRMI. As a result, the scores of common words/phrases with minor TRMI contributions can be minimized." See TRMI user guide.

[4] We only consider 27 assets because one is missing in the Thomson Reuters dataset and two have a high ratio of missing values at the beginning of the sample.

[5] 47 out of 54 sentiment series have less than 1% of missing observations. All the series have a percentage of missing which is smaller than 7.5%. For more information on the sentiment series, we refer to Section A of the online material.

using the rolling mean over the last 5 days.

4 Empirical analysis

In this section, we present the results of the estimation of the MLSS model for the investigated stocks, providing an economic interpretation for the long- and short-term component of the sentiment. In the analyses, we consider separately the case of news and social sentiment indicator.

The first quantity to fix is the number q of long-term sentiment factors. Using the Bayesian information criteria (BIC) we select $q_{\text{news}} = 2$ and $q_{\text{social}} = 2$. An additional criterion for the selection of the number of long-term sentiment is provided in Section C of the online material.

Table 1 reports the values of Φ and Λ with the estimation errors[6]. Bold values indicate parameters which are significantly different from 0 with a p-value smaller than 0.05. We notice that most of the estimated parameters are statistically significant.

As an illustrative example, the top panel of Figure 1 shows how the filter works for the Goldman Sachs news sentiment series. We observe that a high fraction of the sentiment daily variation is captured by the filter. In Section D of the online material we quantify more in detail the signal-to-noise ratio of the proposed filter. We find that the MLSS model has a signal-to-noise ratio approximately twenty times larger than the MLNSL. Moreover, the noise in social media is generally higher than the noise in newspapers. The bottom panel of Figure 1 reports separately the long-term and short-term components of the filtered sentiment signal.

The MLSS approach considers two new quantities extracted from the observed sentiment. The first novelty is the long-term sentiment which, by construction, represents the series of common trends in a particular basket of sentiment time series. The second novelty is the multivariate structure of sentiment, extracted using the symmetric matrix Q_{short}. In the next sections, we separately analyse the relation between these two quantities and the stock market prices. To this end, we extract the market factors from the stock prices of these assets. Denote as $r_t \in \mathbb{R}^{27}$ the vector of demeaned close-to-close log-returns and evaluate the unconditional covariance matrix Q_{ret} and the unconditional correlation matrix C_{ret}. We extract the factor loading matrix $\Lambda^{\text{mrk}} \in \mathbb{R}^{q_{\text{mrk}} \times 27}$ using the PCA on the matrix C_{ret} and define the return factors $R_t = \Lambda^{\text{mrk}} r_t \in \mathbb{R}^{q_{\text{mrk}}}$. We also define the market factors as $M_t^{\text{mrk}} = \Lambda^{\text{mrk}} p_t$, where $p_t \in \mathbb{R}^{27}$ is the vector of log-prices. In the following analysis, we consider $q_{\text{mrk}} = 1$ and name the first market factor Fixed Dow 27 [7].

4.1 Long-term Sentiment

We first investigate a possible economic interpretation of the long-term sentiment. Since we estimate model (2.6) separately for the news and social sentiment, the total number of long-term sentiment series is four. However, two of them are specific for the news sentiment while the other two are

[6] Note that the Λ matrices, as discussed in the supplementary material, have the upper triangular submatrix equal to zero.

[7] Since we are not re-balancing the index to track the real composition of the Dow Jones, we fix the composition as of December 29, 2017 and name the index Fixed Dow 27.

Tickers	$\Phi^{\mathbf{news}}$	$\Lambda^{\mathbf{news}}$		Signal to noise	
				MLSS	MLNSL
AXP	**0.464**	**1.177**		0.623	0.010
	(0.029)	(0.050)			
JPM	**0.732**	**−0.169**	**0.711**	0.326	0.023
	(0.016)	(0.035)	(0.058)		
VZ	**0.682**	**0.545**	−0.080	0.431	0.029
	(0.019)	(0.038)	(0.063)		
CVX	**0.545**	**0.103**	**0.894**	0.610	0.022
	(0.024)	(0.042)	(0.071)		
GS	**0.773**	**−0.239**	**0.718**	0.336	0.029
	(0.014)	(0.036)	(0.060)		
JNJ	**0.407**	**0.851**	**0.834**	0.788	0.010
	(0.030)	(0.039)	(0.065)		
MRK	**0.336**	**0.811**	**0.885**	0.832	0.008
	(0.033)	(0.036)	(0.059)		
PFE	**0.299**	**0.530**	**1.021**	1.185	0.007
	(0.029)	(0.031)	(0.052)		
UNH	**0.374**	**1.177**	**0.530**	0.574	0.009
	(0.037)	(0.056)	(0.093)		
BA	**0.585**	**0.376**	**0.742**	0.896	0.033
	(0.021)	(0.036)	(0.059)		
CAT	**0.633**	**0.309**	0.045	0.423	0.017
	(0.021)	(0.064)	(0.108)		
GE	**0.581**	**1.083**	**−0.196**	0.587	0.022
	(0.023)	(0.035)	(0.058)		
MMM	**0.295**	**0.958**	0.072	0.788	0.009
	(0.034)	(0.038)	(0.064)		
UTX	**0.331**	**0.422**	**−0.413**	0.690	0.011
	(0.035)	(0.057)	(0.094)		
XOM	**0.591**	−0.058	**1.025**	0.725	0.031
	(0.021)	(0.039)	(0.065)		
KO	**0.486**	**0.476**	**0.245**	0.620	0.015
	(0.028)	(0.033)	(0.055)		
PG	**0.337**	**0.838**	**−0.623**	0.929	0.008
	(0.031)	(0.041)	(0.068)		
AAPL	**0.593**	**0.221**	**0.160**	1.736	0.096
	(0.018)	(0.026)	(0.043)		
CSCO	**0.714**	**1.063**	**−1.094**	0.441	0.046
	(0.017)	(0.043)	(0.071)		
IBM	**0.603**	**0.754**	**−1.269**	0.853	0.040
	(0.020)	(0.038)	(0.063)		
INTC	**0.641**	**0.641**	**−0.299**	0.865	0.066
	(0.018)	(0.039)	(0.065)		
MSFT	**0.651**	**0.858**	−0.007	0.668	0.053
	(0.019)	(0.026)	(0.043)		
DIS	**0.439**	**0.454**	**−0.198**	1.074	0.013
	(0.025)	(0.028)	(0.046)		
HD	**0.611**	**1.137**	**0.232**	0.473	0.021
	(0.024)	(0.058)	(0.098)		
MCD	**0.404**	**−0.291**	0.020	1.401	0.013
	(0.024)	(0.034)	(0.057)		
NKE	**0.368**	**0.664**	**−0.285**	0.783	0.010
	(0.032)	(0.046)	(0.076)		
WMT	**0.516**	**0.147**	**0.619**	0.854	0.022
	(0.023)	(0.031)	(0.052)		

Table 1: Static parameters of model (2.6) for news sentiment. Values and standard errors of estimated Λ are multiplied by 10^3. In parenthesis we show the standard error of the estimated parameter. The last two columns show the signal to noise ratio for two competing models.

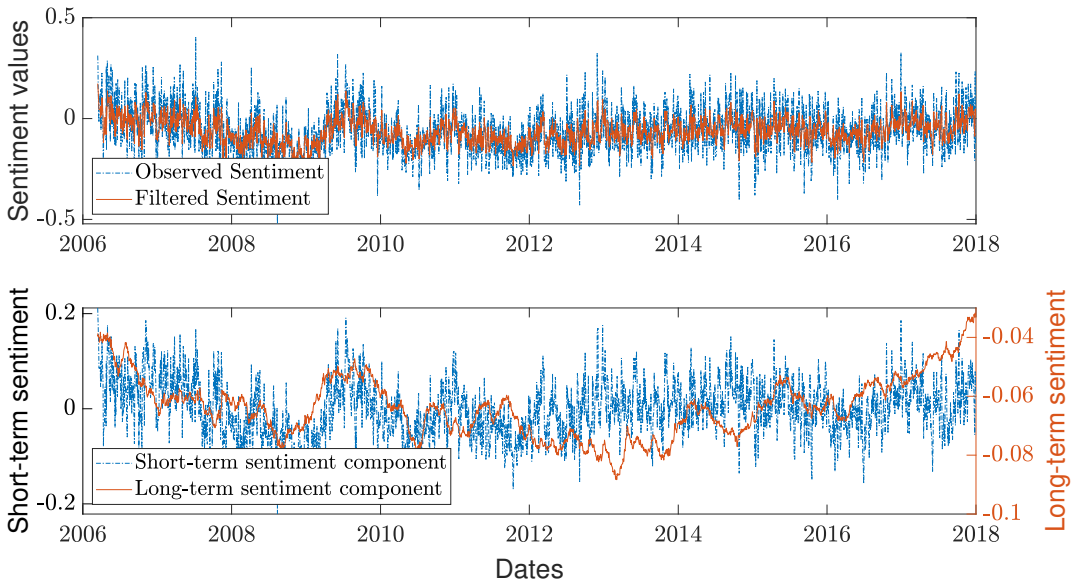

Figure 1: Goldman Sachs sentiment series. In the top panel, the blue line is the observed sentiment and the orange line is the filtered sentiment including both long-term and short-term component. In the bottom panel, we decompose the contribution of the long-term sentiment (in orange, RHS scale) and the short-term sentiment (in blue, LHS scale).

specific for the social sentiment. Using the Engle-Granger test (Engle and Granger, 1987), we observe that one of the factors of the news long-term sentiment is cointegrated with the Fixed Dow 27. Figure 2 shows the cointegration relation, pointing out that the main driver of the prices and the driver of the sentiment time series reflect the same common information. Since at daily scale the news from which the sentiment is extracted could in general reflect the performance of the market, the existence of such a co-integration relation sounds very reasonable and not so surprising. However, Figure 3 shows the standardized weights of the cointegrated factors. The weights of the market factor are very homogeneous across assets, as shown in the top panel, while the weights of the cointegrated factor of the long-term sentiment are very heterogeneous, as shown in the bottom panel. The values of the elements of the factor loading matrix Λ^{news} reported in Table 1 are either positive or negative [8]. Then, some firm's sentiment positively affects the common sentiment factors, while some other firm's sentiment negatively affects them. We checked whether the heterogeneity of weights were related with the number of news of a given asset, or with the buzz index, but we found no significant evidence. Unravelling the origin of the detected heterogeneity is an interesting research question, that could be probably answered by looking at the contents of the articles from which the sentiment was computed. Unfortunately, we do not have access to this kind of information.

[8] The elements of the factor loading matrix Λ^{social} are available upon request.

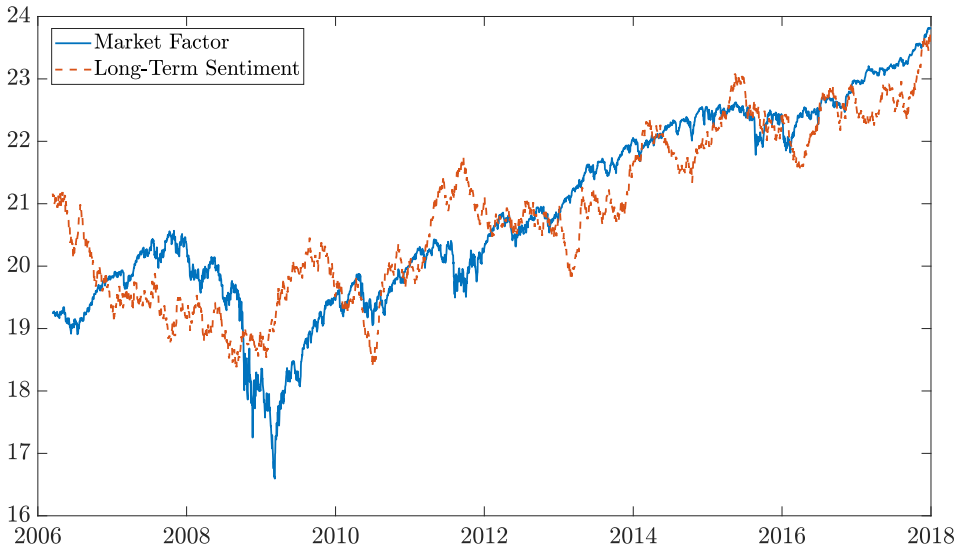

Figure 2: Co-integration between Fixed Dow 27, in blue, and the second factor of the news long-term sentiment, in orange. Time series are scaled. The Fixed Dow 27 is the first market factor built using log-prices.

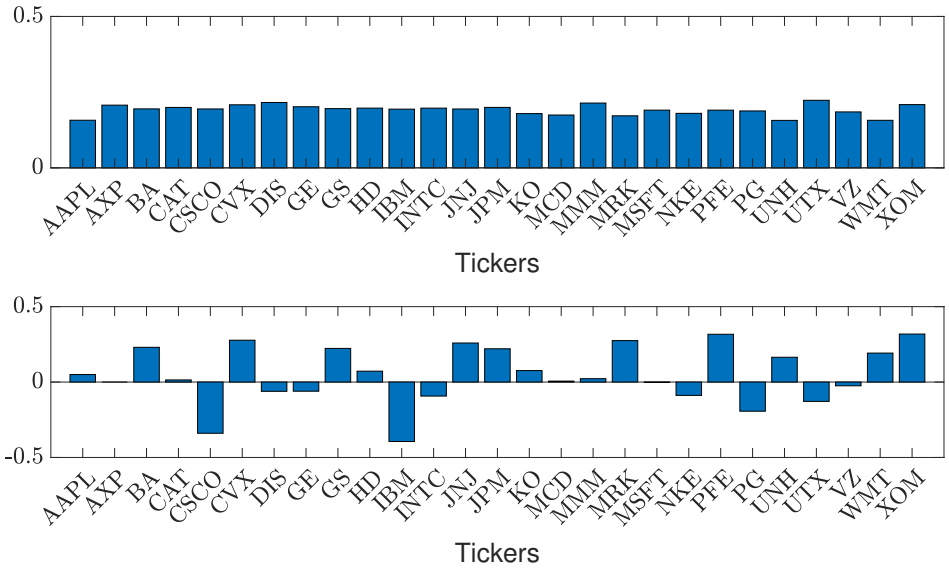

Figure 3: Values of the standardized factor loadings of the cointegrated series. Top panel: loadings of the Fixed Dow 27 index. Bottom panel: loadings of the second factor of the news long-term sentiment.

We do not have a clear interpretation for the other news long-term sentiment component. The loadings of this factor are reported in the second column of Table 1. We see that all the companies are positively related with the factor with the only exceptions of two financial stocks (GS and JPM). In addition, the two social long-term sentiment are not cointegrated with the Fixed Dow 27 and, even in this case, the loadings do not provide a clear association between the factors and the market or a specific sector.

4.2 Short-term Sentiment

The second novelty of the MLSS model is the multivariate structure of the short-term sentiment series. The question we want to address in this section is whether the correlation structure of the short-term sentiment is (linearly) related with the correlation structure of the daily returns. In the previous section, we observed that one of the factors of the long-term sentiment is cointegrated with the first market factor. We therefore expect the short-term sentiment to capture asset-specific features, i.e. we expect a close relation with the idiosyncratic dynamic of the returns[9]. To test this intuition for the correlation structure, we compare the results of the MLSS model with the results of the MLNSL model which, by construction, does not disentangle the factors from the sentiment series. If the intuition is correct, the correlation matrix of the sentiment extracted using the MLSS model should be linearly related with the return correlations and with the idiosyncratic return correlations. On the contrary, the correlation matrix of the sentiment extracted using the MLNSL model, which only captures the slowly changing dynamics of the sentiment series, and thus of the first market factor, should be linearly related with the returns correlation but mildly correlated with the idiosyncratic returns correlations. Finally, to test whether the filtering procedure is a crucial step in our approach, the correlation matrix of the observed sentiment is also considered.

We define C_{short} as the correlation matrix associated with the covariance matrix Q_{short}, C_{MLNSL} the correlation matrix associated with the covariance matrix Q of equation (2.2), $C_{\text{Obs}} = \text{Corr}(\Delta S_t)$ the unconditional correlation of the first difference of the observed sentiment, and C_{ret} the unconditional correlations matrix of the stock returns. We search for a linear element-wise relation between C_{ret} and C_{model}, where model is one of short, MLNSL, or Obs. The results are reported for the news case only, but the conclusions are similar for the social sentiment.

We perform a standard ordinary least squares estimation on the model

$$\mathsf{vechl}(C_{\text{ret}}) = \alpha + \beta^{\text{model}} \mathsf{vechl}(C_{\text{model}}), \tag{4.1}$$

where $\mathsf{vechl}(X)$ is the operator which collects the lower triangular elements of matrix X in a column vector. We compare the results obtained using the MLSS model ($C_{\text{model}} = C_{\text{short}}$), with the results obtained using the MLNSL model ($C_{\text{model}} = C_{\text{MLNSL}}$) and using the Observed sentiment ($C_{\text{model}} = C_{\text{Obs}}$). In addition, since the unconditional correlation between two assets is higher when

[9] We define idiosyncratic returns as the market returns where the first market factor is removed using the factor model (4.2)

Models	All assets			Same sector		
	R^2	F-statistic	p-value	R^2	F-statistic	p-value
MLSS	13.77 %	55.713	0.0000	37.89 %	23.182	0.0000
MLNSL	15.63 %	64.669	0.0000	28.78 %	15.359	0.0004
Obs	0.95 %	3.330	0.0689	4.19 %	1.662	0.2052
MLSS	11.34 %	44.659	0.0000	30.91 %	17.001	0.0002
MLNSL	4.31 %	15.700	0.0001	7.50 %	3.081	0.0873
Obs	1.01 %	3.554	0.0602	4.88 %	1.950	0.1707

Table 2: Top rows: Results from the linear regression (4.1). Bottom rows: Results from the linear regression (4.3). Left columns: OLS estimates when all the assets are considered; right columns: OLS estimates when only the correlations between stocks belonging to the same sector are considered. Obs rows: estimation based on the observed sentiment.

they belong to the same sector, we separately consider two cases. In the first case, we estimate model (4.1) considering all the pairs of assets. In the second case, we estimate model (4.1) considering only the pairs of assets belonging to the same economic sector according to Table A.1.

The top left panel of Table 2 shows the results with all the correlation pairs. In the first column we report the R^2 of the regression, in the second column we report the F-statistic and the relative p-value is reported in the third column. The regressions with C_{short} and C_{MLNSL} have high and significant p-values, while the regression with C_{obs} is not statistically different from the model with the intercept only. This finding has two implications. The first one is that the sentiment innovations have a similar correlation structure as that of the returns' innovations. In particular, if the returns of two assets are relatively highly correlated, then also the increment of the filtered sentiment of the news about these assets are relatively highly correlated. The second implication is that, if a filtering procedure is not applied on the observed sentiment data, the noise is too large to find significant results. In the top right panel of Table 2 we report the results of the model (4.1) applied to the pairs of assets belonging to the same sector. We observe that the R^2 increases for all models. This result is expected since it is well known that the return correlation is higher and more significant between two assets of the same sector. However, even if the R^2 increases, the number of pairs decreases. For this reason, the increment in the R^2 does not lead to an increment in the F-statistic, which fails to reject the null hypothesis for the C_{obs}. This result confirms that the C_{obs} matrix is not a significant regressor for C_{ret}.

Comparing the top panels of Table 2, we note that the increment in the R^2 is higher for the MLSS model rather than the MLNSL model. This evidence is consistent with the intuition that the short-term sentiment series, extracted using the MLSS model, are more related with the idiosyncratic returns. Indeed the correlation induced by the market factor is predominant in the first case, reported in the top left panel, where all the assets are considered, rather than the second case, reported in the top right panel, where the co-movements are not only driven by the first market factor, but they are also driven by sector-specific factors.

Now we extract the Fixed Dow 27 return from the asset returns using a one-factor model. We repeat the analysis comparing the matrices C_{short}, C_{MLNSL} and C_{Obs} with the unconditional correlation of the idiosyncratic returns. We extract the market factor R_t from the returns using

the factor model

$$r_t^i = \alpha^i + \beta^i R_t + z_t^i, \quad \forall i = 1, \ldots, 27 \tag{4.2}$$

where $z_t^i \sim N(0, \tilde{Q}_{\text{ret}})$. We then compute the cross-correlation matrix $\tilde{C}_{\text{ret}}$ from the covariance matrix $\tilde{Q}_{\text{ret}}$ and estimate the following model

$$\mathsf{vechl}(\tilde{C}_{\text{ret}}) = \alpha + \beta^{\text{model}} \mathsf{vechl}(C_{\text{model}}). \tag{4.3}$$

The bottom panels of Table 2 report the results. In the bottom left panel we show the results for the model (4.3) where all the correlation pairs are considered. The first evidence is that the MLNSL R^2 dramatically decreases, while the MLSS R^2 remains almost the same. This finding suggests that almost all the return correlations explained by the C_{MLNSL} matrix are associated with the market factor R_t, while the matrix C_{short}, which represents the fast trends on the sentiment data, also captures different dynamics.
In the bottom right panel, we show the results for the model (4.3) where we consider only the correlation pairs for assets belonging to the same sector. In this case the differences between the MLSS and MLNSL are more severe. Indeed, the MLSS model still has a high and highly significant R^2, while the F-statistic for the MLNSL model fails to reject the null that β^{MLNSL}, defined in equation (4.3), is equal to 0. Again, the model with the observed sentiment has not significant p-values.

As a last observation, we see the different behavior of the sectors in this regression exercise. Figure 4 reports the scatter plot of the elements of C_{short} versus the corresponding values of C_{ret} when the two stocks belong to the same economic sector, characterized by a specific marker. We also superimpose the regression line obtained from equation (4.1). Note that the behavior is different among sectors. The financial sector, marked with blue dots, is the one with highest linear relation and the three assets belonging to this sector have all high returns and sentiment correlations. On the contrary, the consumer cyclical sector, marked with garnet-red triangles, has a high dispersion among the correlations of the 5 assets.

In summary, Sections 4.1 and 4.2 support the intuition behind the MLSS model. Indeed, the slowly changing components of the sentiment are effectively captured by the long-term sentiment. We successfully confirmed this hypothesis in Section 4.1. At the same time, the short-term sentiment effectively describes the firm-specific behavior of the returns. Section 4.2 shows that the MLSS model can capture different features of the returns, while the MLNSL mainly captures the sentiment component associated with the market.

5 Contemporaneous and lagged relations

The goal of this section is to assess the explanatory power of the sentiment with respect to the market returns using the different filters presented in the previous sections. In particular, we show that both the extraction of long-term and short-term sentiment components and the multivariate

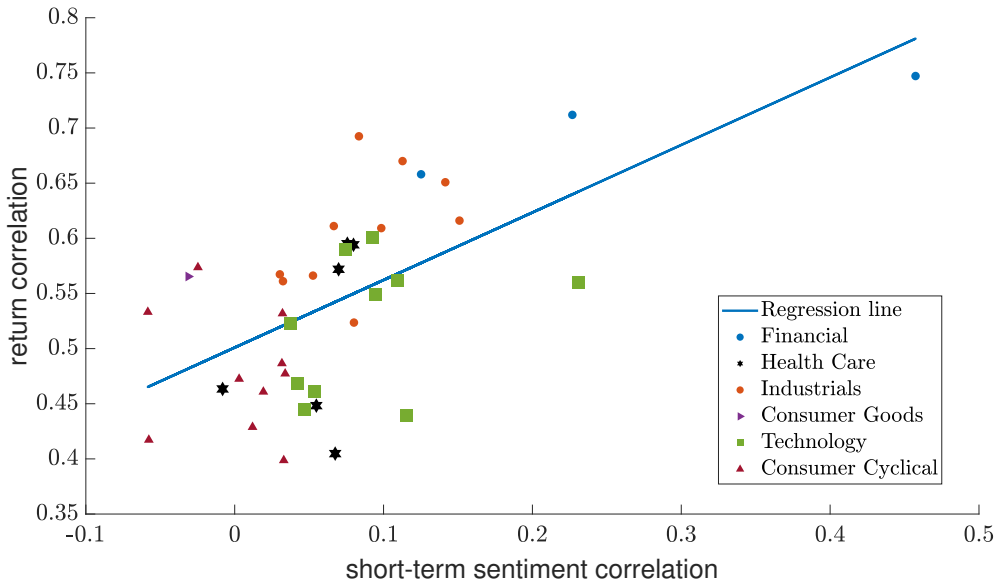

Figure 4: Scatter plot of the news short-term sentiment correlations and the return correlations for pairs of assets in the same sector. The line corresponds to the regression (4.1).

specification of the model are crucial ingredients to capture the synchronous and lagged effects.

We consider the asset prices P_t^i of the 27 stocks of the Fixed Dow 27 and construct the equally weighted portfolio

$$M_t = \frac{1}{27}\sum_{i=1}^{27} P_t^i \tag{5.1}$$

as a representative portfolio and denote with r_t^m its log-returns. We consider a representative portfolio for two reasons. Firstly, (Beckers, 2018) shows that the returns predictability using sentiment indicators is higher when using market indexes rather than single stocks. Secondly, using a representative portfolio we can compare different filtering techniques which do or do not consider the multivariate structure.

We define $\bar{S}_t^{\text{news}} = \frac{1}{27}\sum_{i=1}^{27} S_t^{i,news}$ and $\bar{S}_t^{\text{social}} = \frac{1}{27}\sum_{i=1}^{27} S_t^{i,social}$ as the sentiment associated to the representative portfolio. We consider six different filtering techniques defined as follow:

1. S_t^{MLSS} is the filtered signal obtained using the MLSS model in equation (2.6). The resulting filtered quantities are 4 long-term sentiment factors F_t^{MLSS}, 2 for the news and 2 for the social sentiment, and 54 short-term sentiment series Ψ_t^{MLSS}, 27 for the news and 27 for the social sentiment. We compute the cross-sectional average for the news short-term sentiment $\bar{\Psi}_t^{MLSS,news}$ and social short-term sentiment $\bar{\Psi}_t^{MLSS,social}$. As a final result, we define

$$S_t^{MLSS} = \left[\Delta F_t^{MLSS,news}, \Delta F_t^{MLSS,social}, \bar{\Psi}_t^{MLSS,news}, \bar{\Psi}_t^{MLSS,social}\right]' \in \mathbb{R}^6.$$

2. S_t^{LSS} is the filtered signal obtained applying the MLSS model directly to the univariate series $\bar{S}_t^{\text{news}}$ and $\bar{S}_t^{\text{social}}$. For identifiability reasons, the number of factors is one. The motivation behind this model is to test whether a simple cross-sectional average of sentiment time series can be an effective proxy of the sentiment of the representative asset. This approach intentionally neglects the multivariate structure of the sentiment and treats it as a non relevant feature. A similar reasoning has been used in (Borovkova et al., 2017). The resulting filtered quantities are 2 long-term sentiment factors F_t^{LSS}, one for the news and one for the social sentiment, and 2 short-term sentiment series $\bar{\Psi}_t^{LSS}$, one for the news and one for the social sentiment. The final model reads

$$S_t^{LSS} = \left[\Delta F_t^{LSS,news}, \Delta F_t^{LSS,social}, \bar{\Psi}_t^{LSS,news}, \bar{\Psi}_t^{LSS,social}\right]' \in \mathbb{R}^4.$$

3. S_t^{MLNSL} is the filtered signal obtained using the MLNSL model in equation (2.2) from the 54 observed sentiment time series. The resulting filtered quantities are 54 filtered sentiment series F_t^{MLNSL}, 27 for the news and 27 for the social sentiment. We compute the cross-sectional average for the news sentiment $\bar{F}_t^{MLNSL,news}$ and social sentiment $\bar{F}_t^{MLNSL,social}$. As a final result, we define

$$S_t^{MLNSL} = \left[\Delta \bar{F}_t^{MLNSL,news}, \Delta \bar{F}_t^{MLNSL,social}\right]' \in \mathbb{R}^2.$$

4. S_t^{LNSL} is the filtered signal obtained applying the LNSL model, introduced by (Borovkova and Mahakena, 2015) and presented in equation (2.1), to $\bar{S}_t^{\text{news}}$ and $\bar{S}_t^{\text{social}}$. As for the LSS model, the motivation behind this choice is to test whether the multivariate structure of sentiment is a relevant feature or not. We obtain two filtered sentiment series $\bar{F}_t^{LNSL}$, one for the news and one for the social sentiment. We then define

$$S_t^{LNSL} = \left[\Delta \bar{F}_t^{LNSL,news}, \Delta \bar{F}_t^{LNSL,social}\right]' \in \mathbb{R}^2.$$

5. S_t^{SDFM} is the signal obtained applying a standard dynamic factor model estimated on the sentiment series with Principal Component Analysis. In order to be consistent with the MLSS model, we extract two principal factors for the news sentiment and two principal factors for the social sentiment. As for the Sentiment Dynamic Factor Model (SFDM), the motivation behind this choice is to test whether the multivariate specification of the MLSS model provides any additional advantage. We obtain four filtered sentiment series $\bar{F}_t^{SDFM}$, two for the news and two for the social sentiment. We then define

$$S_t^{SDFM} = \left[\Delta \bar{F}_t^{SDFM,news}, \Delta \bar{F}_t^{SDFM,social}\right]' \in \mathbb{R}^4.$$

6. S_t^{obs} only considers the observed sentiment $\bar{S}_t^{\text{news}}$ and $\bar{S}_t^{\text{social}}$

$$S_t^{Obs} = \left[\Delta \bar{S}_t^{\text{news}}, \Delta \bar{S}_t^{\text{social}}\right]' \in \mathbb{R}^2.$$

In summary, the six models allow us to separate the effect of the different components. The MLSS model exploits all the possible information from the multivariate time series and all the relevant factors are considered. The average across assets is computed at a later stage on the short-term sentiment. For this reason, it does not affect the long-term components. The LSS model computes the cross-sectional average as a first step and does not exploit the multivariate structure. Then, both the short-term and long-term components are different from the one of the MLSS model. The MLNSL and LNSL models differ only on the step of the aggregation. The first model applies the filter on the multivariate time series, while the second model applies the filter on the aggregated time series. The SDFM model uses a non-parametric PCA to extract the common features and ignores the possible presence of a short-term sentiment component. The advantage of this model is that, similarly to the MLSS model, it extracts the main factors starting from the row data without aggregating them. Finally, the Obs model works as a benchmark.

5.1 Quantile regression

In this section, we investigate the lagged relation between sentiment and market returns. The recent literature for the DJIA (Garcia, 2013) and for the gold futures (Smales, 2014) found that the reaction to news is more pronounced during recessions. For this reason, we use the quantile regression in place of a simple linear regression to obtain a more comprehensive analysis of the relationship between variables. In Section E of the online material we report the investigation on the contemporaneous relation between sentiment and returns.

5.1.1 Lagged relations

We consider the following quantile regression

$$r_t^m(\tau) = \alpha^0(\tau) + \alpha^1(\tau)\, r_{t-h}^m(\tau) + \beta^{\text{model}}(\tau)\, S_{t-h}^{\text{model}}, \tag{5.2}$$

where *model* denotes one of the six filtering models presented above and $\alpha_1(\tau) \neq 0$ for $h \neq 0$. According to (Koenker and Machado, 1999), we can compare the explanatory power of a selected model according to the R^1 measure. In particular, if we consider the functional expression for the quantile regression

$$\hat{V}(\tau) = \min_{(\alpha^0, \alpha^1, \beta)} \sum_{t=1}^{T} \rho_\tau \left(r_t^m - \alpha^0 - \alpha^1 r_{t-h}^m - \beta S_{t-h}\right), \tag{5.3}$$

where $\rho_\tau(u) = u(\tau - I_{u<0})$, we can define the quantile R^1 measure as

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)}, \tag{5.4}$$

τ **quantiles**	$R^1(\tau)$ **measure**					
	MLSS	LSS	MLNSL	LNSL	SFM	Obs
0.01	12.7***%	4.5**%	0.3%	0.2%	0.5%	0.1%
0.05	3.2**%	1.3%	0.1%	0.0%	0.2%	0.1%
0.10	1.7**%	1.2**%	0.0%	0.0%	0.3%	0.1%
0.33	0.2%	0.1%	0.0%	0.0%	0.1%	0.0%
0.50	0.2**%	0.1**%	0.1**%	0.1**%	0.0*%	0.0%
0.66	0.4***%	0.2**%	0.1*%	0.1*%	0.1*%	0.0%
0.90	2.8***%	1.0**%	0.2***%	0.1%	0.1%	0.1%
0.95	5.3***%	1.6**%	0.3*%	0.1%	0.1%	0.2%
0.99	11.9***%	3.4**%	0.0%	0.5%	1.4**%	1.0**%

Table 3: The R^1 measure across the value τ for the one-lag quantile regression. We denote with *** the significance at 1%, ** the significance at 5% and * the significance at 10%

where $\tilde{V}(\tau)$ is evaluated restricting equation (5.3) with the intercept parameter and the market return at time $t-h$. In contrast with the R^2 measure of the linear models, $R^1(\tau)$ is a local measure of goodness of fit and only applies to a particular quantile. In addition, (Koenker and Machado, 1999) show that using $\hat{V}$ we can test the significance of the β^{model} parameters. However, the likelihood ratio test introduced in (Koenker and Machado, 1999) assumes that the residuals of the quantile regression are i.i.d. while the observed residuals show heteroskedasicity. For this reason, we perform a multivariate Wald test where the covariance structure of the estimators is estimated using a block bootstrap. For an introduction of confidence intervals estimations with bootstrap methods refer to MacKinnon (2006). To maintain the autocorrelation structure of the data, the block bootstrap length is set to 22 days (one month in financial time).

As a first step, we consider the lag $h = 1$. We evaluate the $R^1(\tau)$ statistic and test the significance using the χ^2 Wald-test. Table 3 reports the values and significance of the R^1 measure. A finding is common among all models: the values of R^1 are higher in the tails and lower close to the median. In addition, what we observe is extremely promising for the Long-Short modeling approach. The significance of the noisy sentiment is zero for almost all quantile levels. Filtering the time series is essential to recover predictability. However, filtering alone is not sufficient. Indeed, neither the predictability of the LSNL model nor of the multivariate extension MLSNL is statistically significant except for some case. Significance is recovered only when the filtered sentiment is decomposed into the short-run and long-run components. This is true for extreme returns, both positive and negative. The result is stronger when the LSS model is replaced by the MLSS, meaning that the cross-sectional dependence is an important ingredient to enhance predictability. It is worth noticing that the values of R^1 presented in Table 3 come from models with different number of regressors, therefore an adjustment with respect to the number of regressors should be applied. However, given that the number of observations used to evaluate the measures is $T = 3020$, a correction in the spirit of the adjusted R^2 would be of order 0.2% and would not affect the results.

A further advantage of the long-short decomposition is that we can properly asses the relative contribution of the two components. In particular, we use a bootstrap Wald test to assess the

τ **quantiles**	p-**values**	
	L_{t-1}^{ST}	L_{t-1}^{LT}
0.01	0.000***%	0.019***%
0.05	0.251***%	35.900%
0.10	0.170***%	5.700*%
0.33	3.300**%	28.500%
0.50	0.169***%	2.100**%
0.66	2.000**%	0.195***%
0.90	0.404***%	0.021***%
0.95	0.001***%	25.200%
0.99	0.003***%	20.800%

Table 4: p-values for the Wald test statistics. The statistics are asymptotically χ^2 with degrees of freedoms equal to the number of tested parameters.

significance of the parameters in the MLSS model. Considering the $S^{MLSS} = \left[\Delta F_t^{MLSS}, \bar{\Psi}_t^{MLSS}\right]$, the significance of the parameter $\beta^{LT} \in \mathbb{R}^4$ and $\beta^{ST} \in \mathbb{R}^2$ can be tested using

$$\tilde{V}^{LT}(\tau) = \min_{(\alpha^0, \alpha^1, \beta^{LT})} \sum_{t=1}^{T} \rho_\tau \left(r_t^m - \alpha^0 - \alpha^1 r_{t-h}^m - \beta^{LT} \Delta F_{t-h}^{MLSS}\right) \tag{5.5}$$

and

$$\tilde{V}^{ST}(\tau) = \min_{(\alpha^0, \alpha^1, \beta^{ST})} \sum_{t=1}^{T} \rho_\tau \left(r_t^m - \alpha^0 - \alpha^1 r_{t-h}^m - \beta^{ST} \bar{\Psi}_{t-h}^{MLSS}\right), \tag{5.6}$$

As before, we use a Wald test based on block bootstrap resampling to assess the significance of β^{ST} and β^{LT}.

We report the p-values of the Wald test statistics in Table 4. The contribution given by the short-term sentiment is strongly significant, in particular for extreme quantiles. On the contrary, the long-term sentiment is not significant in 6 out of 9 quantiles. The results support the intuition that, if today a very high or very low return appears, it can be partially explained by the yesterday's rapidly changing mood, while the permanent trend in the sentiment series have almost no impact.

The experiments performed in the contemporaneous (see Section E in the online material) and one-lag cases show that the MLSS model is the best model to capture the return variations. For this reason, for the multi-period analysis we will only consider the MLSS model.
Considering a general h, we wonder if extra lags can add explanatory power to the regression exercise. Using the functional form

$$\hat{V}^{h,\text{MLSS}}(\tau) = \min_{\left(\alpha^0, \alpha^1 \beta^1 \in \mathbb{R}^6, \beta^2 \in \mathbb{R}^{6(h-1)}\right)} \sum_{t=h+1}^{T} \rho_\tau \left(r_t^m - \alpha^0 - \alpha^1 r_{t-1}^m - \beta^1 S_{t-1}^{\text{MLSS}} - \beta^2 \mathcal{L}_h(S_{t-1}^{\text{MLSS}})\right), \tag{5.7}$$

we separate the contributions given by the first and higher order lags. The bootstrap Wald test can be used to assess the null hypothesis that $\beta^2 = 0$.

Following (Tetlock, 2007; Garcia, 2013), we fix a maximum number of $h = 5$ and Table 5 reports the p-values for the different values of h. The h-lagged sentiment series are uninformative

τ	$h=2$	$h=3$	$h=4$	$h=5$
0.01	18.133%	29.136%	57.652%	72.784%
0.05	0.618%*	0.946%*	4.317%*	3.009%*
0.10	0.907%*	0.773%*	4.341%*	1.968%*
0.33	65.530%	47.389%	74.932%	74.071%
0.50	62.489%	70.078%	80.725%	90.581%
0.66	43.722%	53.518%	52.853%	74.962%
0.90	4.831%*	0.662%*	0.063%*	0.208%*
0.95	12.800%	2.504%*	0.628%*	2.468%*
0.99	38.580%	71.448%	81.945%	87.196%

Table 5: p-values for the Wald statistics for different values of h. We denote with * the values where β^2 is significantly different from zero.

in the median region, where the one lag sentiment have less explanatory power too. However, in agreement with (Garcia, 2013), the lagged sentiment remains informative for few days and, in our case, this is true for the 5%, 10%, 90%, and 95% quantile levels. It is worth noticing that the 1% and 99% quantiles are unaffected by higher-order lags. This shows that, in case of very good or very bad days, the returns are strongly driven by very fresh news ($h = 1$) while the older news have no informative power.

6 Portfolio allocation with sentiment data

This section details an economic application of the MLSS model in portfolio selection and benckmarks the results against a buy-and-hold strategy. We consider the equally weighted portfolio in equation (5.1) and the six filtered signals S_t^{MLSS}, S_t^{LSS}, S_t^{MLNSL}, S_t^{LNSL}, S_t^{SDFM} and S_t^{Obs} introduced in the previous section. It is worth noticing that (Beckers, 2018) and (Garcia, 2013) showed that the predictability power of the sentiment series declined after 2007. For this reason, we want to challenge the filtering techniques to predict the future daily returns in the time window 2007-2019.

In the first part of this section, we use the sentiment signals as exogenous variables to build a simple classifier and we introduce six trading strategies based on the six sentiment time series. Then, we test these strategies on the February 2007 - June 2017 window. This period offers a large series with different economic conditions. The sentiment models are estimated in the same time window. The estimation of multivariate models (MLSS and MLNLS) employs a backward looking technique based on smoothing recursions. Then, one may argue that for the multivariate case the estimation technique may introduce some sort of forward looking bias. We test that this bias, if any, is not likely to be the dominant effect. We perform a robustness check where we use the parameter values from February 2007 - June 2017 period to filter the TRMI sentiment series from July 2017 to December 2019. In this way, the trading signals cannot be affected by any forward looking bias. The results in the out-of-sample period confirm those from February 2007 - June 2017, showing that the trading strategies built on the MLSS model are the best performers. The details of the robustness check can be found in Section G of the online material.

6.1 Trading strategies

In the financial literature, several papers support the strong out-of-sample performance of the equally weighted portfolio (e.g. DeMiguel et al. 2009). The $1/n$ portfolio without rebalancing is used as a baseline for our trading strategies and the long passive position in this portfolio is called *buy-and-hold* strategy. Given that the buy-and-hold portfolio offers a good out-of-sample performance, we assume an investor who only deviates from the baseline strategy if a strong signal which predicts a negative return arrives from the sentiment series. For this reason, the criterion variable needs to capture the behavior of the left tail of returns distribution. We define the criterion binary variable as

$$Y_t = \begin{cases} 1, & \text{for } \tilde{r}_t^m < z_{1/3} \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

where $z_{1/3}$ is the 1/3 Gaussian quantile and $\tilde{r}_t^m = r_t^m/\sqrt{RV_t}$ are the standardized market returns with the realized variance, RV_t, evaluated by means of 5-minute intraday returns. The standardization of the returns is crucial to eliminate possible effects due to the persistence of volatility. The choice of the 33% quantile is consistent with the findings of Section 5.1.1. Moreover, it is a balance between a more conservative choice – a smaller quantile only sensitive to more extreme and predictive events – and a larger quantile, which provides a larger number of selling signals but less predictive power.

Since the goal of this paper is to show that the choice of the filtering procedure is essential, a simple classification technique is used. As a classifier, we consider the following conditional logit model

$$P\left(Y_{t+1} = 1 | X_t\right) = \text{logit}\left(X_t^{\text{mod}}\theta\right), \tag{6.2}$$

where $\text{logit}(X_t\theta) = \frac{e^{X_t\theta}}{1+e^{X_t\theta}}$ and $X_t^{\text{mod}} = \left[1, \tilde{r}_t^m, S_t^{\text{mod}}\right]$. We recall that S_t^{mod} is a vector whose dimension depends on the filtering model. For further details see the first part of Section 5.
The predicted binary value is defined as

$$\hat{Y}_{t+1}^{\text{mod}} = \begin{cases} 1, & \text{for } \text{logit}(X_t^{\text{mod}}\theta) > 0.5 \\ 0, & \text{otherwise} \end{cases}. \tag{6.3}$$

The main advantages of the conditional logit model are twofold. On one hand, the conditional logit model can be easily estimated using MLE. On the other hand, we can easily assess the fitness of the model on the data using the Mc Fadden's R^2 measure defined in (McFadden et al., 1973) as

$$R^2 = 1 - \frac{\log(L_m)}{\log(L_0)} \in [0, 1].$$

L_m represents the maximum likelihood of the complete model (6.2) and L_0 is the maximum likelihood of the bare model based only on the intercept. The models are estimated using overlapping rolling windows of 6 months (126 observations). We verified that this choice is sufficient to capture the time-varying nature of the explanatory power of the sentiment series. Figure 5 shows the value

of R^2 over the February 2007 - June 2017 period. The MLSS model has the highest R^2 w.r.t the other models, which typically translates in a higher predictive power. In addition, the MLSS R^2 has a high variability, suggesting that the predictive power changes through time. This latter finding suggests that the sentiment signal can be a good returns predictor in certain periods and a poor predictor in others. This intuition will be exploited later to generate trading strategies based on the R^2. In equation (6.3) we have used a naive threshold of 0.5 to generate a trading signal. The choice of the optimal or a dynamic threshold is discussed in Section I of the online material.

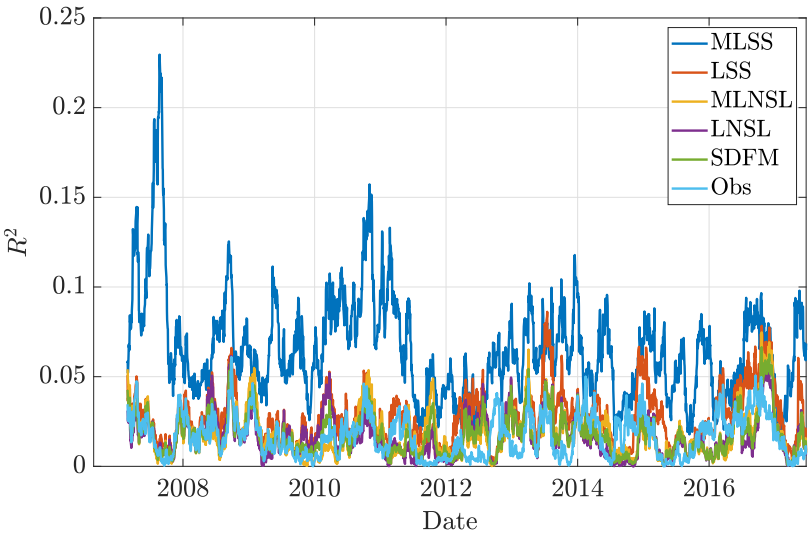

Figure 5: McFadden's R^2 for the different filtering methods using negative abnormal returns.

The estimated $\bar{Y}_t^{\text{mod}}$ defined in (6.3) translates in the trading signal

$$s_{t+1}^{\text{mod}} = \begin{cases} 1, & \text{if } \hat{Y}_{t+1}^{\text{mod}} = 0 \\ -1, & \text{if } \hat{Y}_{t+1}^{\text{mod}} = 1 \end{cases} \tag{6.4}$$

where $s_{t+1}^{\text{mod}} = 1$ ($s_{t+1}^{\text{mod}} = -1$) represents a buy (sell) signal in the equally weighted portfolio (5.1). At any day t, at the closing time of the trading day, the investor uses the sentiment signal S_t^{mod} and the standardized realized daily returns $\tilde{r}_t^m$ to forecast the binary variable $\hat{Y}_{t+1}^{\text{mod}}$ and the relative trading signal. Naming c_0 the number of shares bought or sold in any transaction, there are three possible scenarios

1. $s_t^{\text{mod}} = s_{t+1}^{\text{mod}}$: In this case the prediction on the future realization does not change and the investor does not re-balance the portfolio.

2. $s_t^{\text{mod}} = +1$ and $s_{t+1}^{\text{mod}} = -1$: The investor had a long position in the equally weighted portfolio at time t but the prediction changed. She sells the current position and short sells c_0 shares of the same portfolio.

3. $s_t^{\text{mod}} = -1$ and $s_{t+1}^{\text{mod}} = +1$: The investor had a short position in the equally weighted portfolio at time t but the prediction changed. She buys $2c_0$ shares of the portfolio.

Please notice that the only exception is for s_1^{mod} because we initialized $s_0^{\text{mod}} = 0$. In this case, the equally weighted portfolio is bought when $s_1^{\text{mod}} = 1$ and it is short sold when $s_1^{\text{mod}} = -1$. The investor's portfolio is then built as

$$\begin{cases} P_{t+1}^{\text{mod}} = s_{t+1}^{\text{mod}} c_0 M_{t+1} + \text{cash}_{t+1}, \\ \text{cash}_{t+1} = \text{cash}_t - (s_{t+1}^{\text{mod}} - s_t^{\text{mod}}) c_0 M_{t+1} - |s_{t+1}^{\text{mod}} - s_t^{\text{mod}}| c_0 M_{t+1} \frac{\text{cost}}{2}, \end{cases} \tag{6.5}$$

where *cost* is the percentage trading cost and M_t is defined in (5.1). The first equation in (6.5) shows that the value of the portfolio is composed by the value of the invested amount $s_{t+1}^{\text{mod}} c_0 M_{t+1}$ plus the cash position. The latter increases when $s_{t+1}^{\text{mod}} < s_t^{\text{mod}}$, meaning that the investor sells the portfolio and receives cash, and decreases when $s_{t+1}^{\text{mod}} > s_t^{\text{mod}}$, meaning that the investor buys and erodes the cash position. The second equation includes the impact of the transaction costs. Specifically, every time that a transaction happens, i.e. $s_{t+1}^{\text{mod}} \neq s_t^{\text{mod}}$, the investor pays an extra cost proportional to the current value of the equally weighted portfolio M_{t+1}.

We fix the starting point $s_0^{\text{mod}} = 0$, $\text{cash}_0 = 100,000\$$ and the parameter $c_0 = 100,000\$/M_0$. In the paper we only report the results for the case with trading costs, while the results with zero trading costs are reported in Section F of the online material. From now on, we refer to *without trading costs* when the portfolio in equation (6.5) is evaluated with cost = 0 and to *with trading costs* when costs = 0.1% as in (Gilli and Schumann, 2009) and (Avellaneda and Lee, 2010). In the following sections, the number of transactions is evaluated as $Tr^{\text{mod}} = \sum_{i=0}^{T-1} |s_{i+1}^{\text{mod}} - s_i^{\text{mod}}|$ and the transaction costs are evaluated as $Tc^{\text{mod}} = \sum_{i=0}^{T-1} |s_{i+1}^{\text{mod}} - s_i^{\text{mod}}| c_0 M_{i+1} \frac{\text{cost}}{2}$. It is worth noticing that the change of signal effectively produces two transactions. For instance, if the signal moves from $s_t = 1$ to $s_{t+1} = -1$, the first transaction is the liquidation of the long position and the second transaction is the short position on the asset. In addition, most of the time the selling signal appears for only one day and disappears the day after. Then, the typical path of a selling signal is given by $s_t = 1$, $s_{t+1} = -1$ and $s_{t+2} = 1$ producing a total of four transactions.

The transaction costs can strongly depress the overall performance of the portfolio. To partially mitigate this drawback, we can decrease the number of transactions using the McFadden's R^2 as a measure of the reliability of the signal $\hat{Y}_t^{\text{mod}}$. We compute the empirical quantile $z_\alpha^{1,t}\left(R^2\right)$ of the McFadden R^2 over the time window $(1, \cdots, t)$. The quantile $z_\alpha^{1,t}\left(R^2\right)$ is $\mathcal{F}_t$-measurable and does not introduce a forward looking bias. We can reduce the number of trades conditioning the selling signal at time t on the level of the McFadden's R^2 evaluated in the previous 6 months. The R^2 adjusted trading signal is then defined as follows

$$\bar{s}_t^{\text{mod}} = \begin{cases} -1, & \text{if } \hat{Y}_{t+1}^{\text{mod}} = 1 \text{ and } R_t^{2,\text{mod}} \geqslant z_\alpha^{1,t}\left(R^{2,\text{mod}}\right) \\ 1, & \text{otherwise .} \end{cases} \tag{6.6}$$

The value α determines the reduction in the number of trades. The higher α is, the smaller is the

Measures	BH	MLSS	LSS	MLNSL	LNSL	SDFM	Obs
A. return (%)	**8.975**	7.891	6.977	8.882	8.143	6.827	6.986
A. volatility (%)	19.132	**15.209**	18.136	17.952	19.431	17.672	19.955
A. neg. volatility (%)	15.523	**11.767**	14.339	14.474	15.374	13.767	16.055
A. Sharpe ratio	0.469	**0.519**	0.385	0.495	0.419	0.386	0.35
A. Sortino ratio	0.578	**0.671**	0.487	0.614	0.53	0.496	0.435
MDD ($)	59377	57235	50335	49773	63595	**45016**	62785
Number of trades	1	553	161	81	73	297	93
Transaction costs ($)	50	37974	14866	5565	4085	20335	7544

Table 6: Performances of the seven strategies with transaction cost for the period February 2007 - June 2017. In bold, the best performance per row. BH is the buy-and-hold portfolio, while MLSS, LSS, MLNSL, LNSL, SDFM, and Obs correspond to portfolios built from the corresponding model for the sentiment time series.

number of transactions. The parameter α should be considered as a subjective transaction cost of the investor. The six strategies, together with the buy-and-hold strategy itself, are evaluated according to six measures, the *annual return*, the *annual volatility*, the *annual negative volatility*, the *Sharpe ratio*, the *Sortino ratio*, and the *maximum drawdown* (MDD). In the next section, in a first step, the portfolios with the trading signals (6.4) with and without trading cost are analysed. Then, we assess the impact and the performance of the trade reduction strategy based on (6.6).

6.2 Empirical application: February 2007 - June 2017

The 2007-2009 crisis and the 2009-2017 bull market are good backtesting periods for the sentiment portfolios because we can test the return predictability during different market conditions.

Table 6 reports the performances of the six sentiment strategies together with the buy-and-hold portfolio with trading costs. The sentiment-based strategies have, excluding the LNSL and the Obs, a smaller volatility and MDD than the buy-and-hold portfolio. In addition, the MLSS portfolio produces returns similar to the buy-and-hold strategy, lower negative volatility, and consequently higher Sharpe and Sortino ratios than all the other strategies [10]. The lower performance for the annual returns is due to the higher transaction costs. Indeed in Section F of the online material we show that, when the trading costs are not considered, the MLSS strategy produces higher annual returns than all the other strategies. In addition, when we compare *without trading costs* experiment with the *with trading costs* experiment, the excessive number of transactions for the MLSS strategy reduces the Sharpe ratio gain with respect to the buy-and-hold portfolio from 40% to 10% and the Sortino ratio gain from 48% to 16%. In Section H of the online material we show that the selling signal generated by the MLSS sentiment series corresponds to statistically significant returns predictability. The transaction costs incurred by the MLSS portfolio throughout the nine years amount in total to 38% of the starting capital. For this reason, we employ the trading signal $\bar{s}^{\text{MLSS}}$ defined in equation (6.6), which penalizes signals with moderate McFadden's R^2. Table 7

[10] We have also tested a simple strategy where we remove the first factor, extracted with PCA, from the observed sentiment. However, this procedure is not sufficient to remove all the observation noise and to extract an effective trading signal. The performances of this strategy are Annual return 8.79%, Annual volatility 19.17%, and Sharpe ratio 0.459 (against 0.469 for the BH strategy and 0.519 for the MLSS based strategy).

Measures	BH	$\alpha = 0\%$	$\alpha = 20\%$	$\alpha = 35\%$	$\alpha = 50\%$	$\alpha = 65\%$	$\alpha = 80\%$
A. return (%)	8.975	7.891	8.225	9.575	9.84	**10.248**	9.184
A. volatility (%)	19.132	15.209	15.679	14.083	13.601	**13.538**	17.201
A. neg. volatility (%)	13.601	10.888	11.196	9.901	9.511	**9.443**	12.216
A. Sharpe ratio	0.469	0.519	0.525	0.680	0.723	**0.757**	0.534
A. Sortino ratio	0.660	0.725	0.735	0.967	1.035	**1.085**	0.752
MDD ($)	59377	57235	63522	49160	**33264**	35600	59486
Number of trades	1	553	437	349	273	169	57
Transaction costs ($)	50	37974	30626	25283	20074	12007	4127

Table 7: Performances of the MLSS based strategies built from equation (6.6) for different values of α $\times 100$. BH is the buy-and-hold portfolio. In bold, the best performance per row.

reports the performances of the strategies based on the penalized signal for different values of α. As expected, the higher the value of α and the lower the number of transactions is. In addition, the R^2-based signal produces higher quality signal and effectively increases the performance of the portfolios. The number of transactions decreases almost linearly but the Sharpe and Sortino ratios strongly increase. They reach a maximum value when $\alpha = 0.65$. These findings further corroborate the intuition that the MLSS sentiment strongly anticipates future returns during the financial crisis, given that the R^2 values in figure 5 are higher than the unconditional average during the $2007-2009$ period. Again this feature is peculiar for the MLSS filter while no evidence of return predictability is reported for the other filtering techniques. Again, the statistical significance of these strategies is reported in Section H of the online material.

7 Conclusions

In this paper, we presented a novel way to filter multivariate sentiment time series. The approach is very general and encompasses previous models discussed in the literature. Using a dynamic factor model, we were able to identify two different sentiment components. The first one, named long-term sentiment and modeled as a random walk, captures the common trends which drive the long-term dynamics. The second component, dubbed short-term sentiment and modeled as a VAR(1) process, captures short-term swings of market mood. An extensive empirical section investigates the different features of the two sentiment components. In a first analysis, we pointed out that one of the long-term sentiment factors co-integrates with the first principal component of the market. Quite surprisingly, the structure of the sentiment factor loadings does not mimic the typical uniform profile of the market factor. Some assets are over-expressed and contribute to the factor with a positive or negative sign, while others are under-expressed. Concerning the short-term sentiment, its multivariate dependence structure explains a sizable fraction of the residual covariance in a single factor market model. This result suggests that the short-term component captures transient and rapidly changing trends associated with the idiosyncratic components of the market. In a second analysis, based on quantile regression, we showed that the Multivariate Long-Short Sentiment model provides the highest explanatory power of lagged and contemporaneous returns. Essential to achieve statistical significance are the multivariate nature of the approach and

the separation of the sentiment signal in a long and a short component. In particular, disentangling the short-term sentiment is crucial to capture the behavior of extreme returns. In a further analysis, we observed that newspapers and social media differently react to negative and positive returns. Specifically, they can effectively explain abnormal returns from one to five days in advance, but they almost immediately digest the positive market realizations while they echo negative realizations for several days to come.

It is worth noting that (Tetlock, 2007) and (Garcia, 2013) reported results similar to ours for the unfiltered sentiment focusing on period before 2007. Using the TRMI dataset, (Beckers, 2018) showed that the forecasting power on returns of the sentiment dropped dramatically after 2007. Our results suggest that the filtering procedures are more important nowdays than in the past. Consistently, in a final investigation, we performed an asset allocation exercise where the selling signal are based on the sentiment series. In line with results from the quantile regression, the portfolio based on the MLSS filter significantly outperforms the benchmark buy-and-hold strategy and the other strategies based on different filtering techniques.

Supplementary materials

The supplementary materials include additional analysis to support the evidences found in the paper. Section A provides an overview of the stocks used in the empirical analysis and some descriptive statistics of the sentiment series. Section B introduces a possible alternative for the modeling of the long-term sentiment. Section C provides additional evidences on the choice of $q = 2$ long-term sentiment series in model (2.6). Section D compares the different signal-to-noise ratios filtered by the MLSS and MLNSL model. Section E investigates the contemporaneous relation between the sentiment and return series. Appendices F, G, H and I report the *without trading costs* analysis, the robustness check, the statistical significance of the portfolio allocation exercise and the selection of the optimal classification threshold, respectively.

References

Algaba, A., Ardia, D., Bluteau, K., Borms, S., Boudt, K., 2020. Econometrics meets sentiment: An overview of methodology and applications. Journal of Economic Surveys .

Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. Journal of Economic Perspectives 31, 211–36.

Allen, D., McAleer, M., Singh, A., 2015. Machine news and volatility: The Dow Jones Industrial Average and the TRNA real-time high-frequency sentiment series. Handbook of High Frequency Trading , 327–344.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance 59, 1259–1294.

Appel, G., 2003. Become your own technical analyst: How to identify significant market turning points

using the moving average convergence-divergence indicator or macd. The Journal of Wealth Management 6, 27–36.

Audrino, F., Tetereva, A., 2019. Sentiment spillover effects for US and European companies. Journal of Banking & Finance 106, 542 – 567.

Avellaneda, M., Lee, J.H., 2010. Statistical arbitrage in the US equities market. Quantitative Finance 10, 761–782.

Banbura, M., Modugno, M., 2014. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics 29, 133–160.

Beckers, S., 2018. Do social media trump news? The relative importance of social media and news based sentiment for market timing. The Journal of Portfolio Management 45, 58–67.

Bork, L., 2009. Estimating US monetary policy shocks using a factor-augmented vector autoregression: an EM algorithm approach. Available at SSRN 1358876 .

Borovkova, S., 2015. The role of news in commodity markets. Available at SSRN 2587285 .

Borovkova, S., Garmaev, E., Lammers, P., Rustige, J., 2017. SenSR: A sentiment-based systemic risk indicator. Technical Report 553. De Nederlandsche Bank. .

Borovkova, S., Mahakena, D., 2015. News, volatility and jumps: the case of natural gas futures. Quantitative Finance 15, 1217–1242.

Calomiris, C.W., Mamaysky, H., 2019. How news and its context drive risk and returns around the world. Journal of Financial Economics 133, 299–336.

Cavaliere, G., Xu, F., 2014. Testing for unit roots in bounded time series. Journal of Econometrics 178, 259–272. Recent Advances in Time Series Econometrics.

Chen, I., Kelkar, Y.D., Gu, Y., Zhou, J., Qiu, X., Wu, H., 2017. High-dimensional linear state space models for dynamic microbial interaction networks. PloS one 12, e0187822–e0187822.

Corsi, F., Peluso, S., Audrino, F., 2015. Missing in asynchronicity: A Kalman-EM approach for multivariate realized covariance estimation. Journal of Applied Econometrics 30, 377–397.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. The Journal of Finance 66, 1461–1499.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? The Review of Financial Studies 22, 1915–1953.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) , 1–38.

Durbin, J., Koopman, S.J., 2012. Time series analysis by state space methods. Oxford University Press.

Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: Representation, estimation, and testing. Econometrica 55, 251–276.

Feuerriegel, S., Gordon, J., 2019. News-based forecasts of macroeconomic indicators: A semantic path model for interpretable predictions. European Journal of Operational Research 272, 162–175.

Garcia, D., 2013. Sentiment during recessions. The Journal of Finance 68, 1267–1300.

Gerber, A.S., Gimpel, J.G., Green, D.P., Shaw, D.R., 2011. How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. American Political Science Review 105, 135–150.

Gilli, M., Schumann, E., 2009. An empirical analysis of alternative portfolio selection criteria. Swiss Finance Institute, Swiss Finance Institute Research Paper Series .

Groß-Klußman, A., Hautsch, N., 2011. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. Journal of Empirical Finance 18, 321–340.

Harvey, A.C., 1990. Estimation, prediction and smoothing for univariate structural time series models. Cambridge University Press. pp. 168 – 233.

Hill, S.J., Lo, J., Vavreck, L., Zaller, J., 2013. How quickly we forget: The duration of persuasion effects from mass communication. Political Communication 30, 521–547.

Huynh, T.L.D., Foglia, M., Nasir, M.A., Angelini, E., 2021. Feverish sentiment and global equity markets during the covid-19 pandemic. Journal of Economic Behavior & Organization 188, 1088–1108.

Jiao, P., Veiga, A., Walther, A., 2020. Social media, news media and the stock market. Journal of Economic Behavior & Organization 176, 63–90.

Jungbacker, B., Koopman, S.J., 2008. Likelihood-based analysis for dynamic factor models. Tinbergen Institute, Tinbergen Institute Discussion Papers .

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering 82, 35–45.

Kim, S.H., Kim, D., 2014. Investor sentiment from internet message postings and the predictability of stock returns. Journal of Economic Behavior & Organization 107, 708–729. Empirical Behavioral Finance.

Koenker, R., Machado, J.A.F., 1999. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94, 1296–1310.

Lillo, F., Micciché, S., Tumminello, M., Piilo, J., Mantegna, R.N., 2015. How news affect the trading behavior of different categories of investors in a financial market. Quantitative Finance 15, 213–229.

Liu, B., 2015. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. The Journal of Finance 66, 35–65.

MacKinnon, J.G., 2006. Bootstrap methods in econometrics. Economic Record 82, S2–S18.

Mai, F., Tian, S., Lee, C., Ma, L., 2019. Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research 274, 743–758.

McFadden, D., et al., 1973. Conditional logit analysis of qualitative choice behavior .

Nyman, R., Kapadia, S., Tuckett, D., 2021. News and narratives in financial systems: Exploiting big data for systemic risk assessment. Journal of Economic Dynamics and Control 127, 104119.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, pp. 79–86.

Peterson, R., 2016. Trading on Sentiment: The Power of Minds Over Markets. John Wiley & Sons, Ltd.

Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M., Mozetic, I., 2015. The effects of Twitter sentiment on stock price returns. PLOS ONE 10, e0138441.

Ranco, G., Bordino, I., Bormetti, G., Caldarelli, G., Lillo, F., Treccani, M., 2016. Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics. PLOS ONE 11, e0146576.

Schnaubelt, M., Fischer, T.G., Krauss, C., 2020. Separating the signal from the noise - financial machine learning for twitter. Journal of Economic Dynamics and Control 114, 103895.

Shumway, R.H., Stoffer, D.S., 1982. An approach to time series smoothing and forecasting using the EM algorithm. Journal of Time Series Analysis 3, 253–264.

Smales, L.A., 2014. News sentiment in the gold futures market. Journal of Banking & Finance 49, 275 – 286.

Smales, L.A., 2015. Asymmetric volatility response to news sentiment in gold futures. Journal of International Financial Markets, Institutions and Money 34, 161–172.

Sun, L., Najand, M., Shen, J., 2016. Stock return predictability and investor sentiment: A high-frequency perspective. Journal of Banking & Finance 73, 147 – 164.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance 62, 1139–1168.

Thorsrud, L.A., 2018. Words are the new numbers: A newsy coincident index of the business cycle. Journal of Business & Economic Statistics , 1–17.

Vohra, S., Teraiya, J., 2013. A comparative study of sentiment analysis techniques. Journal JIKRCE 2, 313–317.

Wu, L.S.Y., Pai, J.S., Hosking, J., 1996. An algorithm for estimating parameters of state-space models. Statistics & Probability Letters 28, 99–106.

Appendix

A Filter and Smoother recursions

In this section, we report Kalman Filter and Smoother recursions ancillary to the EM algorithm. The derivation of the formulas which follow can be found in (Shumway and Stoffer, 1982).
Starting from system (2.7(, we calculate recursively the Kalman Filter as:

$$\begin{aligned}
\tilde{F}_{t|t-1} &= E\left[\tilde{F}_t|S_1,\ldots,S_{t-1}\right] = \tilde{\Phi}\tilde{F}_{t-1|t-1} \\
P_{t|t-1} &= E\left[\left(\tilde{F}_t-\tilde{F}_{t|t-1}\right)\left(\tilde{F}_t-\tilde{F}_{t|t-1}\right)'|S_1,\ldots,S_{t-1}\right] = \tilde{\Phi}P_{t-1|t-1}\tilde{\Phi}'+Q \\
K_t &= P_{t|t-1}\tilde{\Lambda}'\left(\tilde{\Lambda}P_{t|t-1}\tilde{\Lambda}'+R\right)^{-1} \\
\tilde{F}_{t|t} &= \tilde{F}_{t|t-1}+K_t\left(S_t-\tilde{\Lambda}\tilde{F}_{t|t-1}\right) \\
P_{t|t} &= P_{t|t-1}-K_t\tilde{\Lambda}P_{t|t-1}
\end{aligned}$$

where we take $\tilde{F}_{0|0}=\mu$ and $P_{0|0}=\Sigma$. Now, using backward recursions $t=T,\ldots,1$ we derive the Smoother as

$$\begin{aligned}
J_{t-1} &= P_{t-1|t-1}\tilde{\Phi}'\left(P_{t|t-1}\right)^{-1} \\
\tilde{F}_{t-1|T} &= \tilde{F}_{t-1|t-1}+J_{t-1}\left(\tilde{F}_{t|T}-\tilde{\Phi}\tilde{F}_{t-1|t-1}\right) \\
P_{t-1|T} &= P_{t-1|t-1}+J_{t-1}\left(P_{t|T}-P_{t|t-1}\right)J_{t-1}' \\
P_{t-1,t-2|T} &= P_{t-1|t-1}J_{t-2}'+J_{t-1}\left(P_{t,t-1|T}-\tilde{\Phi}P_{t-1|t-1}\right)J_{t-2}'
\end{aligned} \tag{A.1}$$

where $P_{T,T-1|T}=\left(I-K_T\tilde{\Lambda}\right)\tilde{\Phi}P_{T-1|T-1}$.

B Expectation Maximization

The log-likelihood of the model (2.7) is

$$\begin{aligned}
l\left(S_t,\tilde{F}_t,\theta\left(j\right)\right) =& \log f(\tilde{F}_0)+\sum_{t=1}^{T}\log f(\tilde{F}_t|S_{t-1})+\sum_{t=1}^{T}\log f(S_t|\tilde{F}_t) \\
=& -\frac{1}{2}\log|\Sigma|-\frac{1}{2}\left(\tilde{F}_0-a\right)\Sigma^{-1}\left(\tilde{F}_0-a\right)' \\
&-\frac{T}{2}\log|\tilde{Q}|-\frac{1}{2}\sum_{t=1}^{T}\left(\tilde{F}_t-\tilde{\Phi}\tilde{F}_{t-1}\right)\tilde{Q}^{-1}\left(\tilde{F}_t-\tilde{\Phi}\tilde{F}_{t-1}\right)' \\
&-\frac{T}{2}\log|R|-\frac{1}{2}\sum_{t=1}^{T}\left(S_t-\tilde{\Lambda}\tilde{F}_t\right)R^{-1}\left(S_t-\tilde{\Lambda}\tilde{F}_t\right)'
\end{aligned}$$

where a and Σ are the parameters s.t. $\tilde{F}_0\sim\mathcal{N}(a,\Sigma)$.

E-step

The objective function to maximize is, from Shumway and Stoffer (1982),

$$G\left(a,\Sigma,R,\tilde{Q},\tilde{\Lambda},\tilde{\Phi}\right)=E_m\left[\log f|S_1,\ldots,S_T\right],$$

where E_m denotes the conditional expectation relative to a density containing the mth iterate values $a(m),\Sigma(m),R(m),\tilde{Q}(m),\tilde{\Lambda}(m)$ and $\tilde{\Phi}(m)$. Using now the Kalman smoother (A.1) we can derive

$$E\left[\left(S_t-\tilde{\Lambda}\tilde{F}_t\right)\left(S_t-\tilde{\Lambda}\tilde{F}_t\right)'|S_1,\ldots S_T\right]=\left(S_t-\tilde{\Lambda}\tilde{F}_{t|T}\right)\left(S_t-\tilde{\Lambda}\tilde{F}_{t|T}\right)'+\tilde{\Lambda}P_{t|T}\tilde{\Lambda}',$$

$$\begin{aligned}E\left[\left(\tilde{F}_t-\tilde{\Phi}\tilde{F}_{t-1}\right)\left(\tilde{F}_t-\tilde{\Phi}\tilde{F}_{t-1}\right)'|S_1,\ldots,S_T\right]=&P_{t|T}+\tilde{F}_{t|T}\tilde{F}'_{t|T}+\tilde{\Phi}P_{t-1|T}\tilde{\Phi}'\\&+\tilde{\Phi}\tilde{F}_{t-1|T}\tilde{F}'_{t-1|T}\tilde{\Phi}'-P_{t,t-1|T}\tilde{\Phi}'\\&-\tilde{F}_{t|T}\tilde{F}'_{t-1|T}\tilde{\Phi}'-\tilde{\Phi}P_{t,t-1|T}-\tilde{\Phi}\tilde{F}_{t-1|T}\tilde{F}'_{t|T},\end{aligned}$$

lead to

$$\begin{aligned}G\left(a,\Sigma,R,\tilde{Q},\tilde{\Lambda},\tilde{\Phi}\right)=&-\frac{1}{2}\log|\Sigma|-\frac{1}{2}tr\{\Sigma^{-1}\left[P_{0|T}+\left(\tilde{F}_0-a\right)\left(\tilde{F}_0-a\right)'\right]\}\\&-\frac{T}{2}\log|\tilde{Q}|-\frac{1}{2}tr\{\tilde{Q}^{-1}\left(C-B\tilde{\Phi}'-\tilde{\Phi}B'+\tilde{\Phi}A\tilde{\Phi}'\right)\}\\&-\frac{T}{2}\log|R|-\frac{1}{2}tr\{R^{-1}\left(E_3-\tilde{\Lambda}E_2'-E_2\tilde{\Lambda}'+\tilde{\Lambda}E_1\tilde{\Lambda}'\right)\},\end{aligned}$$

where

$$\begin{aligned}&A=\sum_{t=1}^{T}\left(\tilde{F}_{t-1|T}\tilde{F}'_{t-1|T}+P_{t-1|T}\right),\quad B=\sum_{t=1}^{T}\left(\tilde{F}_{t|T}\tilde{F}'_{t-1|T}+P_{t,t-1|T}\right),\quad C=\sum_{t=1}^{T}\left(\tilde{F}_{t|T}\tilde{F}'_{t|T}+P_{t|T}\right),\\&E_1=\sum_{t=1}^{T}P_{t|T}+\tilde{F}_{t|T}\tilde{F}'_{t|T},\quad E_2=\sum_{t=1}^{T}S_t\tilde{F}'_{t|T},\quad E_3=\sum_{t=1}^{T}S_tS_t'.\end{aligned}\tag{B.1}$$

M-step

The resulting update equations are

$$\Lambda(m+1)=E_2E_1^{-1},\qquad\tilde{\Phi}(m+1)=BA^{-1}\tag{B.2}$$

$$\tilde{Q}(m+1)=\frac{1}{T}\left(C-B\tilde{\Phi}\left(m+1\right)'-\tilde{\Phi}\left(m+1\right)B'+\tilde{\Phi}\left(m+1\right)A\tilde{\Phi}\left(m+1\right)'\right)\tag{B.3}$$

$$R(m+1)=\frac{1}{T}\left(E_3-\tilde{\Lambda}(m+1)E_2'-E_2\tilde{\Lambda}(m+1)'+\tilde{\Lambda}(m+1)E_1\tilde{\Lambda}(m+1)'\right)\tag{B.4}$$

$$a(m+1) = \tilde{F}_{0|T}, \qquad \Sigma(m+1) = P_{0|T}\,. \tag{B.5}$$

For simplicity, in our estimations we impose $\tilde{F}_0 = 0$.