Empir Econ (2011) 41:663–679
DOI 10.1007/s00181-010-0397-0

Sentiment dynamics and stock returns: the case of the German stock market

Thomas Lux

Received: 9 July 2009 / Accepted: 22 May 2010 / Published online: 1 October 2010

Abstract We use weekly survey data on short-term and medium-term sentiment of German investors in order to study the causal relationship between investors' mood and subsequent stock price changes. In contrast to extant literature for other countries, a trivariate vector autoregression for short-run sentiment, medium-run sentiment, and stock index returns allows to reject exogeneity of returns. Depending on the chosen VAR specification, returns are found to either follow a feedback process caused by medium-run sentiment, or returns form a simultaneous systems together with the two sentiment measures. An out-of-sample forecasting experiment on the base of estimated subset VAR models shows significant exploitable linear structure. However, trading experiments do not yield convincing evidence of significant economic gains from the VAR forecasts, and it appears that predictability of returns from sentiment decreases during the recent market gyrations.

Keywords Investor sentiment · Opinion dynamics · Return predictability

JEL Classification G12 · G14 · C22

1 Introduction

Many market participants and financial practitioners seem to believe in the informational content of various measures of "sentiment." Under an efficient market

T. Lux (✉)
Department of Economics, University of Kiel, Olshausen Str. 40, 24118 Kiel, Germany
e-mail: lux@bwl.uni-kiel.de

T. Lux
Kiel Institute for the World Economy, Kiel, Germany

perspective, however, publicly available information such as the balances of surveys of traders' expectations should already be incorporated in current prices. Excess profits on the base of measures of sentiment would contradict this view. Such an argument, however, neglects the additional source of risk introduced by non-fundamental traders via their potentially erratic changes of mood. This source of "noise trader risk" has been investigated by Black (1986) and DeLong et al. (1990) who demonstrate that noise trader risk imposes limits-to-the-arbitrage activities of rational traders. As a consequence of this additional form of market risk non-rational traders could "create their own space," i.e., a kind of ecological niche in which they can survive despite their erratic beliefs.

In DeLong et al. (1990) the mood of noise traders is a stochastic variable with independent realizations in each trading period. Noise trader risk is, therefore, unpredictable, in its direction and the extent of its changing moods of euphoria or pessimism. One could argue that if rational traders could get reliable signals on the contemporaneous and future disposition of noise traders, they should find it more profitable to built up contrarian positions. However, as long as future sentiment is not perfectly predictable uncertainty about the persistence and expected reversal time of a wave of positive or negative sentiment might still limit the extent of rational arbitrage (Shleifer and Vishny 1997). Therefore, noise traders might have a significant impact on asset prices without necessarily triggering sufficient arbitrage activity against their distortive influence.

The proximity of certain popular measures of "sentiment" to noise traders' mood in theoretical models has spawned a sizeable empirical literature on the predictive performance of sentiment. Most studies in this area have used one of the widely publicized indices of investors' sentiment in the U.S. or indirect measures like the ratio of equity put to call trading volume, number of advancing issues to declining issues and others. Examples of this literature include Brown and Cliff (2005) who find that returns at 1–3 year horizons are negatively related to sentiment. However, the same authors (Brown and Cliff 2004) find no predictive power of sentiment for market-wide near-term returns. They do find, however, a significantly negative effect from sentiment to the returns of a portfolio of small stocks. They also identify a causal influence from market returns to sentiment. Verma et al. (2008), in contrast, find some predictive power after decomposing sentiment into rational and irrational components. Rational components are identified by a regression of the raw sentiment data on a set of fundamental factors while the residuals of this regression are interpreted as the irrational part. Both components have a positive impact on near-term returns with a larger effect of the "rational" term.

More subtle channels of predictive power of sentiment variables are explored by Baker and Wurgler (2006) and Baker et al. (2009): The former use a composite proxy for sentiment based on the first principal component of a collection of statistics that may reflect sentiment (including the closed-end fund discount, NYSE share turnover, number and first-day returns of IPOs, and the dividend premium) and also consider an orthogonalized index that eliminates potential systematic risk factors due to macroeconomic fundamentals. They find significant predictive capacity of both indices for selections of stocks that should be particularly sentiment-prone (such as young or small stocks, distressed stocks, non-dividend-paying stocks, etc.) while the sentiment proxies display no predictive power for the aggregate market. Baker et al. (2009) extend this approach to five additional countries besides the U.S. and find similar cross-sectional patterns of predictability for stocks that are difficult to arbitrage in various countries. In a somewhat similar vein, Lemmon and Portniaguina (2006) find contrarian predictability of the U.S. small stock returns from the residual component of consumer confidence unrelated to macroeconomic fundamentals.

Apart from the U.S. market, the relationship between sentiment and returns has also been investigated for the Shanghai stock market by Kling and Gao (2008). Their results closely resemble those of Brown and Cliff (2004) in that they identify causation from market returns on sentiment but not vice versa.

Evidence on the German stock market can be found in recent articles by Schmeling (2007, 2009) and Hengelbrock et al. (2009). Schmeling (2007) reports significant predictability in long-horizon regressions plus some indication of profitability of sentiment-based trading. Using consumer sentiment as a proxy of individual investor sentiment, Schmeling (2009) also finds significant negative correlations to market-wide returns for a sample of 18 industrialized countries. He also shows that predictability is more pronounced for countries with less developed market institutions, and confirms the findings of Baker and Wurgler (2006) that hard-to-arbitrage stocks are affected more strongly by sentiment. Hengelbrock et al. (2009) also find market-wide predictability at medium horizons for the U.S. and Germany. However, while sentiment is again a contrarian indicator for the U.S. market (and its predictability disappears over the 1990s), German sentiment is positively related to future returns which is harder to explain even under a noise trader and limits-to-arbitrage perspective.

Both Schmeling (2007) and Hengelbrock et al. (2009) use survey data from *sentix* as the proxy for German investor sentiment. This study is based on another survey, which has been launched in 2004 under the name of *animusX-Investors sentiment*. Rather than investigating single-equation regressions, we adopt the perspective of Brown and Cliff (2004) and Kling and Gao (2008) and study the joint dynamics of sentiment and market returns. In contrast to their results, we find strong evidence for a causal influence from sentiment to returns. Bivariate and trivariate Granger causality tests speak unanimously for medium-run sentiment as the main driver in our simultaneous system. Testing down from complete VAR models with full coefficient matrices to more parsimonious subset VAR models also shows a tendency of ending up with unidirectional causation from medium-run sentiment to both returns and short-run sentiment.

Out-of-sample forecasting experiments indicate significant gains in forecast accuracy of VAR-based predictions compared to the benchmark of a random walk with drift for the most parsimonious subset VAR models. Any attempt at exploiting more information—either via richer VAR structures or via consideration of fundamental factors—deteriorates results. We also find that the influence of sentiment tends to fade away during the recent period of abnormally large market fluctuations. Trading experiments on the base of estimated subset VAR models do not provide convincing evidence of easy exploitation of the predictive capacity of sentiment.

Our study proceeds as follows: Sect. 2 provides information on our data set of sentiment measures. While these seem to be quite popular among practitioners the data we use have apparently not been the subject of academic studies so far. Section 3 investigates sentiment and returns as a simultaneous system, exploring the causality structure, model specification issues, and forecasting performance of VAR models. Section 4 continues with a "market timing" experiment, and Sect. 5 concludes.

2 Data

Since mid 2004, *animusX-Investors sentiment*,[1] a provider of technical services and information for German investors, provides weekly surveys on the prospects of the German stock market over short-run and medium-run horizons. These surveys are conducted via e-mail among about 2,000 registered private and institutional investors. According to the organizer of the survey, response rates are about 25%. The incoming categorial responses are recorded from Thursday, 6 p.m. to Saturday, 12 p.m. each week, and the result is communicated in the form of a diffusion index (balance between optimistic and pessimistic responses) on the following Sunday at about 8 p.m. Our data set covers these diffusion indices for short-run and medium-run sentiment for the time horizon from the 29th calendar week of 2004 to the 46th week of 2009 (a total of 280 observations).[2] Short-run sentiment (S-Sent) concerns expectations for the following week while medium-run sentiment (M-Sent) covers respondents' expectations over the next 3 months. While the *animusX* sentiment data receive relatively wide coverage in the financial press, we are not aware of any previous scientific study using these data.

Figure 1 exhibits the time development of the two indices together with the weekly returns (defined as continuously compounded returns, i.e., differences between the logs of the index) of the German share price index DAX. Returns are obtained from *Datastream* and are defined as log differences between weekly closing notations of the DAX.[3] As one might have expected, the short-run index exhibits much higher volatility than the medium-run index. While the short-run index flips often very dramatically from predominant optimism to pessimism and vice versa, the medium-run index rather seems to perform long swings between more moderate perceptions.

Table 1 provides some sample statistics of our data. Since we will split the entire record into an in-sample of 150 observations and an out-of-sample part of the remaining 130 observations in our subsequent analysis, we give results for both the full sample and the reduced in-sample. The statistics confirm our impression that the short-run index has more variability and less temporal dependence than its medium-run counterpart. The ADF statistics indicate that we have little reason to doubt the stationarity of

[1] Information on the services of *animusX* is available at http://www.animusx.de. The company offers a whole range of technical tools as well as various sentiment data collected via weekly surveys among its subscribers.

[2] Survey participants actually have the choice between five categorial answers. These are essentially a strongly pessimistic (expected price drop), mildly pessimistic (bottom formation), neutral, mildly optimistic (ceiling formation), and strongly optimistic view (expected price increase). The weights attached to the different categories are not made public by the owners. However, the company assures that these weights have been kept constant throughout the history of this survey. As is typical of diffusion indices, the admissible range is $[-1, 1]$.

[3] In our trading experiment below we will also consider weekly opening quotations as the information received on Sunday evening could only be exploited by investment or withdrawal from the stock market on Monday morning with prices differing from last week's closing prices.

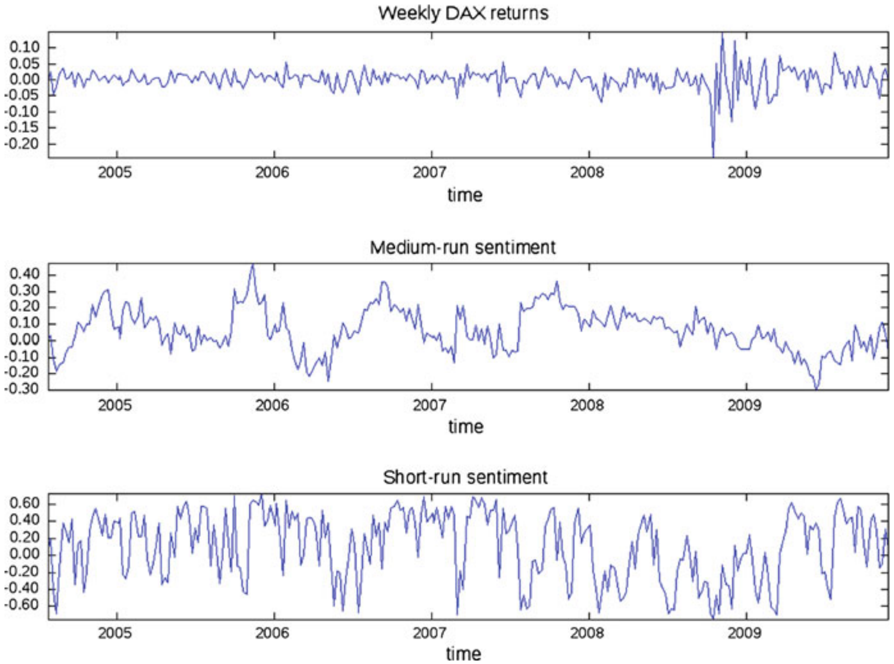

Fig. 1 Sentiment and stock market returns. The time horizon is from the 29th calendar week of 2004 to the 46th week of 2009

Table 1 Summary statistics

	Mean	SD	Skewness	Kurtosis	ρ_1	ADF
Panel A: Full sample (280 observations)						
S-Sent	0.110	0.393	−0.414	−1.007	0.598	−6.447
M-Sent	0.063	0.131	0.089	−0.165	0.807	−3.689
Returns	0.001	0.034	−1.287	10.862	−0.114	−11.384
Panel B: In-sample (150 observations)						
S-Sent	0.222	0.354	−0.732	−0.464	0.436	−4.632
M-Sent	0.073	0.136	0.204	−0.223	0.777	−2.856
Returns	0.005	0.019	−0.412	0.301	−0.097	−9.632

Notes: S-Sent and M-Sent denote the short-run and medium-run sentiment index, respectively. The ADF test statistics have been computed with one lag, but inclusion of further lags did not change the results qualitatively. The one-sided 5 and 1% critical values are −1.957 (−1.942) and −2.590 (−2.602), for samples with 280 (150) entries, respectively

all three time series. Since both sentiment series are bounded between −1 and 1, strict non-stationarity seems practically impossible anyway. Our subsequent VAR analyses confirm stationarity of the trivariate system since the estimated coefficient matrices have only stable eigenvalues throughout.

Table 2 Lag selection for trivariate VAR model

Lags	Information criteria		
	BIC	AIC	HQ
0	−14.811	−14.811	−14.811
1	**−16.686**	−16.867	−16.794
2	−16.601	−16.965	**−16.817**
3	−16.392	−16.941	−16.718
4	−16.261	−16.997	−16.698
5	−16.095	**−17.019**	−16.643
6	−15.823	−16.937	−16.484
7	−15.540	−16.845	−16.314
8	−15.213	−16.712	−16.103
9	−14.960	−16.654	−15.966
10	−14.669	−16.560	−15.792

Notes: The minimum value of each criterium is indicated by bold letters

3 VAR analysis of sentiment and returns

Since the two sentiment indices as well as returns appear to be stationary, we can explore their dynamic relationship by estimating trivariate VAR models.[4] Denoting the triples of observations on S-Sent, M-Sent, and returns at time t by a vector $Y_t = (Y_{1t}, Y_{2t}, Y_{3t})'$, we assume a driving process of the form

$$Y_t = V + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + U_t. \tag{1}$$

In (1), V is a 3×1 vector of intercept terms, the A_i are 3×3 coefficient matrices with entries $a^k_{i,j}$ (i the row, j the column number, and k the lag order) and U_t is a vector of disturbances. We start by examining the appropriate order of the VAR models to be estimated. As can be seen from Table 2, different selection criteria favor different lag orders: the Bayesian criterion (BIC) opts for only one lag to be included while the Hannan–Quinn criterion (HQ) favors two lags and the Akaike Criterion (AIC) assumes its minimum at five lags. It is, therefore, impossible to judge on a priori grounds which of the three models would have to be preferred. A sequence of likelihood ratio tests (available in the supplementary material for the paper at the journal homepage) yields similarly diverse behavior: proceeding from larger to smaller lags, we find that the null hypothesis $A_4 = 0$ (i.e., no significant entries in the coefficient matrices of order four or higher) is the first one rejected. However, subsequently, we see that the null hypotheses $A_3 = 0$ and $A_2 = 0$ can both not be rejected and disregarding the test results for higher lags, the results for lower dimensions would favor a parsimonious VAR(1) model.

[4] Cf. Lütkepohl (2005) for the econometrics of simultaneous systems. All computations reported in this article have been performed in GAUSS 8.0.

Table 3 Coefficient estimates of VAR(2) model

	Lag	S-Sent	M-Sent	Ret
S-Sent	1	0.343	0.813*	2.106
M-Sent	1	−0.028	0.642**	0.306
Ret	1	0.004	0.056**	−0.120
S-Sent	2	0.209	−0.394	−3.010
M-Sent	2	−0.006	0.184*	0.094
Ret	2	0.000	−0.036	−0.203

Notes: This shows parameter estimates obtained via maximum likelihood. * and ** denote significant parameters at the 5 and 1% significance level, respectively

Owing to this ambiguity, we proceed by estimating VAR models up to order 5 and investigate their implied causal structures and forecasting capabilities. Table 3 shows the estimated parameters for the VAR(2) model as a benchmark case. Most interestingly, we find a strongly significant effect from sentiment on returns, but not the other way around. This indicates that investors' sentiment causes returns and should be exploitable to predict future market movements. Note that the effect is restricted to medium-run sentiment while short-run sentiment itself also appears to be caused unidirectionally by medium-run sentiment. Since only the second coefficient is significant in all three equations, our estimated VAR(2) model identifies M-Sent as an exogenous variable whose dynamics drives both S-Sent and returns. This is in total contrast to previous VARs of sentiment and stock market returns for the U.S. (Brown and Cliff 2004) and China (Kling and Gao 2008) that identify causality from returns on sentiment, but not the other way around. The estimated covariance matrix indicates significant correlation of innovations of short-run sentiment and returns as well as similarly significant correlation between both sentiment measures. In contrast, medium-run sentiment and returns appear not to be instantaneously correlated. Specification tests show that residuals suffer from significant skewness, but do not display excess kurtosis or temporal dependence.

The results for the VAR(1) and VAR(5) models preferred by the BIC and AIC criteria are very similar. For VAR(1) we obtain significant parameters again for the central column, i.e., causal influence of M-Sent on both S-Sent and returns plus a strong relation of S-Sent to its own past which was absent in the VAR(2) model. Although the VAR(5) model allows for a total of 45 autoregressive parameters, not too many of them appear significant. These are again the entries of the central column of the lag 1 coefficient matrix plus two more significant parameters at lag 5 which entail feedback effects from both S-Sent and returns on M-Sent which consequently is not exogenous anymore in the VAR(5) model.[5]

[5] The negative coefficient from S-Sent to M-Sent at higher lags also bears some similarity to previous findings. Namely, such an inverse relationship between different sentiment measures has also been found by Brown and Cliff (2005) for the U.S., and Schmeling (2007) for a different set of German sentiment data. In particular, the latter paper found a negative effect from individual investors' sentiment on that of institutional investors. Identifying short-run sentiment with individual investors' disposition and medium-run sentiment with that of institutional investors, our finding of a negative effect at lag 5 is similar to his results. Nevertheless, the interpretation is cumbersome: Following the above authors, individual (short-term) euphoria would drive prices up which makes institutional (medium-term) investors expect a reversal of the

Table 4 reports Wald test statistics for Granger causality tests between returns and sentiment. Since we are mainly interested in the endogeneity or exogeneity of returns, we first test jointly for Granger causality from both indices on returns and vice versa in the trivariate model (panel A). We confirm the impression from the VAR estimates: The sentiment variables appear to be exogenous to the system at least in the VAR(1) and VAR(2) specifications. Panel B of Table 4 looks at causality tests for bivariate models of each pair of our three variables. As can be seen the only significant entries are indeed those for causality from M-Sent to S-Sent as well as returns.

As we have seen above, only few of the coefficients of the chosen VAR specifications appear to be significant under any information criterion. Using estimates for many coefficients whose true parameter might be zero adds estimation noise that might be detrimental to our attempted use of the resulting models for forecasting. It, thus, seems natural to proceed from model selection in terms of the number of lags to a broader concept of model selection allowing for an arbitrary number of zero coefficients in our VAR structures. In principle, in the absence of any prior knowledge on the structure of dependencies between variables, any subset VAR model is a potential candidate for the data-generating process. It would, therefore, be an obvious idea to apply information criteria to select the optimal subset model. However, the number of potential models to be considered is tremendous. While a case-by-case comparison of the $2^{12} = 4046$ models in subset VAR(1) specifications (including the nine coefficients of the 3×3 matrix A_1 plus three constants) appears feasible, the $2^{48} \approx 3 \times 10^{14}$ possible cases of VAR(5) submodels can certainly not all be evaluated individually.

A number of approaches have been developed in order to test down from a general model with full coefficient matrices to a preferred subset specification (cf. Penm and Terrell 1984; Brüggemann 2004; Lütkepohl 2005).[6] A comparative study of a variety of proposed algorithms by Brüggemann (2004) found two methods to be most reliable. One is the so-called *topdown* algorithm (denoted TD in the following): It starts equation-wise from the most general specification and sequentially eliminates components. If a certain selection criterion (i.e., BIC, AIC, or HQ) improves upon elimination of an entry, its coefficient is set equal to zero while it is maintained in subsequent estimations if the criterion deteriorates without this entry. A possible shortcoming of this approach is that it neglects interdependencies between the single equations. An alternative algorithm that tests down from the most general specification at the system level (denoted SY in the following) consists in sequentially eliminating all regressors with t-ratios below a certain threshold η. As shown in Brüggemann (2004), the choice:

$$\eta = \{(\mathrm{e}^{c_T/T} - 1)(T - N + j - 1)\}^{\frac{1}{2}}$$

Footnote 5 continued
trend in the medium run. Besides the problem that these considerations would coincide in the simultaneous assessment of the near-term and medium-term prospects by the same participants in our survey, we also lack an indication of the direct influence of S-Sent on returns (all pertinent coefficients are insignificant).

[6] Practical applications of subset or restricted VAR models are surprisingly rare. A cursory search only brought forward a study by Lin et al. (1994) on exchange rate forecasts based on VAR models of their fundamental determinants. Although the selection criterion used in this study is slightly different from those applied here, the results are quite similar: While full VAR models performed worse than the random walk framework, the restricted models had significant informational content for some exchange rates.

Table 4 Causality tests

	Sent → Ret	Ret → Sent	
Panel A: Trivariate model			
Var(1)			
Wald	7.371	1.236	
p-value	0.026	0.539	
Var(2)			
Wald	10.034	3.146	
p-value	0.040	0.534	
Var(5)			
Wald	16.670	15.828	
p-value	0.082	0.105	
	S-Sent → Ret	Ret → S-Sent	M-Sent → Ret
Panel B: Bivariate models			
Var(1)			
Wald	0.907	0.008	6.162
p-value	0.341	0.927	0.013
Var(2)			
Wald	0.145	1.862	9.802
p-value	0.930	0.394	0.007
Var(5)			
Wald	4.720	5.092	10.821
p-value	0.451	0.405	0.055
	Ret → M-Sent	S-Sent → M-Sent	M-Sent → S-Sent
Panel B: Bivariate models			
Var(1)			
Wald	0.378	0.014	6.820
p-value	0.539	0.907	0.009
Var(2)			
Wald	0.317	0.965	7.853
p-value	0.854	0.617	0.020
Var(5)			
Wald	1.108	8.644	11.336
p-value	0.953	0.124	0.045

Notes: Granger causality tests report the p-values of the null hypotheses that all coefficients of the first variable(s) at all lags are jointly insignificant in the equation(s) of the second variable(s). p-values are computed from the underlying Chi-square distribution under the null

with T the sample size, N the overall number of coefficients, and j the elimination round is equivalent to selection based on AIC, HQ, and BIC criteria for the choices $c_T = 2, c_T = 2 \ln \ln T$, and $c_T = \ln T$, respectively. While SY takes the systemic dependency between variables into account, it did not uniformly turn out to be superior to TD in the comprehensive simulations conducted by Brüggemann.

In order to shed light on the out-of-sample forecasting capacity of a VAR model of sentiment and returns, we adopt both TD and SY in relation with the AIC, BIC, and

HQ criterion, respectively, taking as the pertinent starting point the full VAR models preferred by these criteria.

Since specification tests were not unambiguously in favor of Normality of residuals, we also turn to Least Squares (LS) instead of ML estimation. In order to sequentially estimate restricted VAR models we adopt the feasible (or estimated) generalized least square estimator (FGLS) which uses a two-stage procedure by first estimating consistently the covariance matrix and then using this estimate to obtain consistent parameter estimates (Lütkepohl 2005). Table 5 provides results of out-of-sample forecasting exercises on the base of various subset model selection schemes. We consider both single returns as well as cumulative returns over longer horizons. All entries are relative mean squared errors (RMSEs) of out-of-sample forecasts which are standardized by dividing the MSE of the VAR-based forecasts by the MSE of the random walk with drift as a benchmark model.[7] The table also indicates significance under the adjusted Diebold–Mariano test for nested alternatives (Clark and West 2007). The second row of each panel shows the number of non-zero coefficients of each specification (in the case of panel B: the average number of non-zero coefficients). Panel A shows the performance of forecasts with a frozen subset VAR structure, the one preferred for the in-sample data by the pertinent selection algorithm. Parameter estimates have been updated over the course of the out-of-sample forecasting exercise by taking into account the new information when changing the forecast origin. Since updating might also be applied to the subset VAR selection itself, we also consider forecasts on the base of a more flexible approach in panel B. Here, the subset VAR selection algorithms are also evoked again with every change of the forecast origin in the out-of-sample period.[8]

Overall, the complete set of results indicates some forecastability of returns from an appropriately determined subset VAR specification. In general, forecasting performance increases with added flexibility of model estimation procedures (i.e., from panel A to B) while it deteriorates when moving from the strict BIC to the more liberal HQ and AIC criteria. Inspecting more closely the results, it becomes quite obvious that it is the second column of A_1 (i.e., dependencies of all variables on M-Sent) that provides forecasting power while capturing potential higher order effects turns out to deteriorate the results. The hope that we might improve results via more subtle independencies between M-Sent and S-Sent, therefore, did not materialize itself. Note also that TD is somewhat better able then SY to filter out the most promising subset model which is in harmony with similar results in the Monte Carlo simulations of Brüggemann (2004). Interestingly, inspecting the development of the updated subset model selections reported in panel B, one observes that all three TD runs as well as BIC-SY dispense with the dependency of returns on M-sentiment at some point during the out-of-sample period. In two cases, this happens at the same point in time (October 2008)

[7] From a more general autoregressive benchmark specification, all subset specification algorithms would choose the random walk with drift as the preferred submodel, i.e., eliminate all autoregressive terms.

[8] Results for unrestricted VAR models and models with constant (in-sample) parameters have also been computed and are available in the supplementary material on the journal website. The former shows a very dismal performance with MSEs throughout worse than the random walk, while the latter are very similar to the more refined estimates exhibited in Table 5, albeit with somewhat more modest forecasting results.

Table 5 RMSEs of out-of-sample forecasts. Forecasts from models estimated via constrained FGLS with (a) Updating of parameter estimates only and (b) Updating of entire model-selection algorithm

Horizon # Coeff.	BIC-TD 5	HQ-TD 6	AIC-TD 19	BIC-SY 6	HQ-SY 10	AIC-SY 21
Panel A: Updating of parameter estimates						
Forecasts of single returns						
1	0.994	0.994	1.047	0.994	1.006	1.053
2	0.990	0.992	1.044	0.992	1.000	1.052
3	0.994	0.995	0.981	0.996	1.003	1.016
4	0.985*	0.987*	0.971	0.990*	0.996	1.000
5	0.987*	0.990	0.986	0.992	0.998	0.998
6	0.988*	0.990	0.987	0.994	0.999	0.996
7	0.983*	0.986*	1.000	0.991	0.997	1.002
8	0.984*	0.987*	0.997	0.992	0.997	1.002
Forecasts of cumulative returns						
1	0.994	0.994	1.047	0.994	1.006	1.053
2	0.983*	0.984*	1.107	0.985*	1.006	1.095
3	0.978*	0.980*	1.091	0.983	1.008	1.079
4	0.958*	0.963*	1.076	0.968*	1.003	1.047
5	0.944**	0.951*	1.052	0.960*	1.003	1.030
6	0.934**	0.943**	1.049	0.957*	1.004	1.031
7	0.910**	0.931**	1.046	0.949*	1.000	1.027
8	0.904**	0.918**	1.035	0.942*	0.997	1.024

Horizon # Coeff.	BIC-TD 4.29	HQ-TD 6.05	AIC-TD 14.24	BIC-SY 5.02	HQ-SY 8.50	AIC-SY 19.85
Panel B: Updating of model-selection algorithm						
Forecasts of single returns						
1	0.989*	1.018	1.172	1.010	1.015	1.153
2	0.988*	1.033	1.192	0.996	1.038	1.239
3	0.981**	0.959	1.037	0.995	0.968	1.236
4	0.977**	0.972	1.013	0.990	0.965	1.125
5	0.979**	0.984*	1.083	0.992	0.982*	1.067
6	0.980**	0.988*	1.033	0.993	0.991	1.029
7	0.981**	0.993	1.043	0.992	0.995	1.015
8	0.980**	0.991	1.067	0.993	0.996	1.006
Forecasts of cumulative returns						
1	0.989*	1.018	1.172	1.010	1.015	1.153
2	0.958**	1.021	1.061	0.983*	1.031	1.130
3	0.932**	0.983	0.984*	0.972*	0.988	1.120
4	0.904**	0.988	0.957*	0.961*	0.984	1.040
5	0.883**	0.975	0.974*	0.957*	0.970	1.027
6	0.864**	0.971	0.966	0.952*	0.973	1.005

Table 5 Continued

Horizon # Coeff.	BIC-TD 4.29	HQ-TD 6.05	AIC-TD 14.24	BIC-SY 5.02	HQ-SY 8.50	AIC-SY 19.85
7	0.853**	0.964	0.986	0.945*	0.964	1.031
8	0.837**	0.950	0.978	0.937*	0.959	1.038

Notes: For each model, this shows the relative MSE of the forecasts from the pertinent VAR (i.e., original MSE divided by that of the random walk model with drift). * and ** identify cases of significantly better predictive accuracy of the VAR forecasts using one-sided 10 and 5% significance levels under the adjusted Diebold–Mariano test. In the pertinent test statistics, we used the Newey–West estimator with automatic lag selection by Andrew's method to compute the standard errors of MSE differences. Constrained models have been estimated with updating of parameters only (panel A) and updating of the subset VAR selection plus parameter estimation (panel B) during their out-of-sample application. The subset model selections have been performed both with a topdown criterion (TD) as well as with a system-wide elimination strategy (SY), both based on either the BIC, HQ or the AIC information criteria (cf. Brüggemann 2004)

while the two remaining cases happen before or after that period, respectively. AIC-TD prefers dependency on S-Sent after this break whereas all other algorithms support independence of returns from both sentiment variables for the last part of the sample.

We also conducted a number of sensitivity checks for the findings reported in Table 5. In order to see whether sentiment is just a proxy for neglected fundamental factors, we have added the following list of typical fundamental determinants to our VAR framework: the dividend yield, the price-to-earnings ratio, the short-term interest rate (1 week EURIBOR rates), and the term spread between EURIBOR and 10-year government bonds. We have introduced this additional information in two ways: first, by adding the four fundamental variables as exogenous factors to the trivariate VAR structure of Eq. 1, and second by estimating large seven-variate VAR models of sentiment indices, returns, and fundamental factors. A detailed record of the results along the structure of Table 5 can be found in the supplementary material on the journal website. The outcome is that under both alternatives, the quality of forecasts virtually always deteriorates. Unrestricted models now have mean squared errors up to almost ten times those of the random walk at longer forecasting horizons. Restricted models show no particular pattern concerning the preferred fundamental factors. If fundamentals are introduced as exogenous variables, almost none of the restricted models show any significant forecasting improvements. If we allow for dynamic interaction between all seven variables, the results are somewhat better than with exogenous fundamentals but still worse throughout than those of our trivariate system without fundamental factors. It, thus, appears, that sentiment has more informational content than typical fundamental factors. Adding those factors on top of our more parsimonious model only seems to magnify estimation noise and does not provide any benefit for out-of-sample forecasting.

We have also repeated our previous exercise with an alternative pair of sentiment indices from *sentix behavioral indices* (www.sentix.de),[9] an index that has been used in some previous analyses of the German stock market. In the *sentix* survey, the short-run

[9] We are grateful to one reviewer for suggesting to compare the results for *animusX* to those of the very similar *sentix* series and for providing the pertinent data for this exercise.

horizon is 1 month, while the medium-run is the 6-months horizon. Visual inspection shows that both the two short-run indices and the two medium-run indices do, nevertheless, march very much in lockstep with each other. Interestingly, despite the higher volatility of the short-run indices (and the relatively sizable relative difference in 1 week vs. 1 month horizon), their comovements are even closer (correlation coefficient: 0.86) than those of the two medium-run indices (correlation coefficient: 0.52). When replicating the steps of our analysis with the *sentex* data, we also found a strong influence from its medium-run component to returns and highly significant evidence for Granger causality. However, here there is a bidirectional influence as the null hypothesis of absence of causality from returns to sentiment can also be rejected at standard levels of significance. Results of out-of-sample forecasting exercises nevertheless resemble closely those of the *animus* indices. One difference is that with this data set, the BIC-TD algorithm (that performed best in Table 5) tends to completely eliminate any dependence of returns on sentiment. The best forecasting results are obtained based upon the slightly less stringent HQ criteria and SY selection algorithms with all three alternatives HQ-TD, BIC-SY, and HQ-SY achieving about the same performance (detailed results are again available in the supplementary material to this article). While the performance of both *animusX* and *sentex* indices is about the same, it certainly would be interesting to explore whether there are potential forecast gains from combining the data of both sources. We leave this question for future research.

4 Profitability of VAR-based trading strategies

Our forecasts of future returns could easily be used as input in the design of a straightforward trading strategy. Since our simultaneous system provides us with forecasts of price changes for the stock market index, we attempted to test the "market timing" potential of our VAR models. We, therefore, allow our "VAR trader" to switch between a long and a short position depending on her/his expectation of near-term returns. As an alternative, we also considered switching between the stock market and a safe investment in bonds. Results were in general similar. Note that while short-selling of the index as such is technically impossible, a number of derivative instruments exist that are inversely linked to major stock market indices, so that the long–short strategy can at least be closely approximated in practice. As it turned out, the results of these trading experiments were in general unsatisfactory, and where the VAR strategies seemed to yield excess profits, these were not significant under a bootstrap test. We, therefore, confine ourselves to reporting only a small selection of results.

Given the expectations for the next period[10] produced by our VAR models, a risk-neutral myopic investor would switch between a long and a short position if the alternative provides a higher expected return net of transaction costs. Denoting by r_t^e the continuously discounted expected return of the stock market, an investor who is currently invested in stocks will switch to a short position if $r_t^e < -c$. Vice versa, a move from a short to a long position will happen if $r_t^e > c$ holds. If the reverse of one of these inequalities holds, the trader will simply keep on to her current position.

[10] Since forecastability is more pronounced over longer horizons, we also tried strategies based on forecasts over multiple periods. However, results were not better than those reported here.

For computing returns of the VAR-based strategy, we have taken into account that a signal obtained from the VAR model with the new sentiment information on Sunday evening could only be used for trading once the stock market opens again on Monday morning. We have, therefore, used the weekly opening prices as transaction prices if a signal was received during the weekend. However, as a comparison with the closing quotation of the previous week shows, the impact of the weekend return is small. We have run all strategies with transaction costs of $c = 0, 0.01, 0.025$, and 0.05, but only report results for the lower alternatives. A buy-and-hold strategy for our complete out-of-sample record from June 2007 to mid-November 2009 would have generated a negative return of -31.2%.

We confine ourselves to exhibiting results for the "topdown" (TD) model-selection algorithms in Table 6. In line with the forecasting performance, the results for the systemic (SY) selection approach are either similar, or worse (these details are available in the supplementary material). As can be seen from Table 6, positive returns are obtained only for the most parsimonious model (selected by BIC-TD) and only in the absence of transaction costs. While the contrast between the return of this active strategies and the buy-and-hold benchmark appears spectacular, the performance worsens drastically once transaction costs are introduced. Interestingly, this is not because of the raw costs incurred by trading activities, but because of relatively weak signals issued by the VAR models. Keeping the pattern of transactions recommended in the absence of transaction costs, but deducing fees of $0.1, 0.25$, or 0.5% leads to only very slight changes of overall profits (or losses).

Of course, the question remains whether any of these results is significant under the distribution of profitability of a pertinent strategy in the absence of exploitable linear structure. In order to assess the significance of excess profits, we perform two bootstrap tests: We resample the out-of-sample observations either with or without replacement and apply our trading strategy for 1,000 bootstrapped resamples of both types. To be precise: Bootstrapping means here that we destroy the temporal order of the triples of observations consisting of S-Sent, M-Sent, and returns at time t. Our random draws from the trivariate series are, then, subjected to the same sequence of operations like the "true" ordering. That means, we either keep the selected subset VAR of the in-sample record fixed and recursively update parameter estimates, or we apply recursively the pertinent subset VAR selection algorithm itself to the random sample. Since we have destroyed all structural dependencies, parameter estimates and "identified" VAR structures should be spurious and should not have any forecasting value. Table 6 also displays the 90, 95, and 99% quantiles of the bootstrapped distributions. As it turns out, the bootstrap test (whether performed with or without replacement) shows that not even the impressive performance of the BIC-TD with recursive model selection is significant at typical confidence levels.[11] Essentially, the large fluctuations of the market in 2008/2009 lead to a very broad distribution of bootstrapped profits in which not even the almost 50% increase from the buy-and-hold strategy to the actively managed BIC-TD strategy stands out.

[11] Note also that Jensen's alpha does not indicate a positive risk-adjusted excess return despite the huge difference to the buy-and-hold return.

Table 6 Profitability of trading based on VARs: full out-of-sample record

	B & H	Active	Bootstrapped quantiles			SR	JA
			90%	95%	99%		
Panel A: No transaction costs							
TD with updating of model selection							
BIC	−0.312	0.171	0.413	0.496	0.762	0.016	0.000
			(0.443)	(0.535)	(0.782)		
HQ	−0.312	−0.189	0.422	0.551	0.901	−0.045	0.000
			(0.460)	(0.662)	(0.890)		
AIC	−0.312	−0.015	0.516	0.690	1.023	−0.015	0.000
			(0.526)	(0.728)	(1.082)		
TD with frozen subset VAR structure and updated parameters							
BIC	−0.312	−0.445	0.564	0.714	1.040	−0.085	−0.003
			(0.553)	(0.695)	(1.033)		
HQ	−0.312	−0.654	0.362	0.547	0.861	−0.118	−0.004
			(0.338)	(0.555)	(0.845)		
AIC	−0.312	−0.452	0.456	0.675	1.002	−0.091	−0.003
			(0.445)	(0.636)	(0.960)		
Panel B: Transaction costs of 0.1%							
TD with updating of model selection							
BIC	−0.312	−0.398	0.236	0.368	0.721	−0.080	−0.002
			(0.190)	(0.357)	(0.748)		
HQ	−0.312	−0.531	0.344	0.521	0.896	−0.102	−0.003
			(0.379)	(0.574)	(0.961)		
AIC	−0.312	0.140	0.439	0.584	0.955	0.011	0.001
			(0.442)	(0.640)	(0.927)		
TD with frozen subset VAR structure and updated parameters							
BIC	−0.312	−0.548	0.478	0.640	0.943	−0.106	−0.003
			(0.440)	(0.607)	(0.923)		
HQ	−0.312	−0.592	0.319	0.494	0.869	−0.113	−0.004
			(0.309)	(0.488)	(0.869)		
AIC	−0.312	−0.632	0.391	0.565	1.041	−0.119	−0.005
			(0.354)	(0.550)	(0.879)		

Notes: Forecasts are computed on the base of VAR models with either updating of parameters or repeated application of the subsample selection algorithm. The bootstrapped quantiles are obtained by either scrambling the trivariate out-of-sample data or by drawing with replacement from this sample (those in brackets). SR and JA denote the Sharpe ratio and Jensen's alpha of the active strategies (the Sharpe ratio of the market portfolio is −0.110)

5 Conclusion

Using a new set of short-run and medium-run sentiment data for the German stock market, we investigated the structural properties of VAR models including short-run and medium-run sentiment measures as well as returns of the stock index DAX at the weekly frequency. In striking contrast to similar studies for the U.S. and Shanghai markets, we found that, depending on the specification of our VAR model, either sentiment is exogenous and drives returns, or returns and sentiment define a simultaneous system with mutual causation. Given the large number of insignificant entries in the full VAR models, we estimate restricted models using the "topdown" and "systemwide" elimination algorithms in conjunction with the BIC, HQ and AIC information criteria. As it turns out, estimated models are the more successful in out-of-sample forecasting, the more rigorously they cut down on the number of estimated parameters of the VAR structure. More liberal algorithms apparently accumulate estimation noise through additional parameter estimates that are not set equal to zero. Comparing six different algorithms we find an almost completely monotonic relationship between the number of parameters admitted to the VAR model and the deterioration of the quality of forecasts. Overall, it appears that only the direct effect from M-Sent to returns contributes to successful forecasts, and more subtle effects of multi-directional interaction at higher lags are either not present or cannot be captured by the linear VAR model.

We also found that fundamental factors like interest rates, price-to-earnings ratio, and dividend-yield did not provide any additional explanatory power beyond that of M-Sent. Indeed, adding such factors only led to deterioration of forecasts, so that M-Sent seems to completely encapsulate this information. However, since M-Sent itself is highly predictable, it can certainly not be interpreted as a measure of *new* fundamental information. It rather appears like a slowly moving basic mood of the market that changes in reaction to both fundamental factors as well as other, perhaps psychological influences. In contrast, short-run sentiment seems more in harmony with the structural perception of sentiment in noise trader models, in that it performs wild, short-lived swings between euphoria and depression. However, it also seems to be a largely autistic component: Although it gets itself a strong impetus from M-Sent, its feedback on M-Sent and returns is tenuous.

Despite the apparent causal relationship from M-Sent to returns and the significant out-of-sample forecasts, exploitation of these structural dependencies seems not easy: The results from our trading experiments turned out quite diffuse and disappointing. We also found a breakdown of the causal link from M-Sent to returns (as indicated by the selection of coefficients of the subset selection algorithms) during the major period of market turmoil in 2008, which could be either due to evolutionary adaptation of market forces extracting the predictive component of M-Sent or to inadequacy of the linear model during the period of the financial crisis.

Acknowledgments The author is grateful for helpful comments to Jonas Dovern, Carsten-Patrick Meier, Adriano Rampini, and participants at the staff seminar at the Kiel Institute. Very detailed and helpful comments by two referees are also gratefully acknowledged. I also thank Thomas Theuerzeit of *animusX* for providing the sentiment data and offering useful insights into the structure of the survey. This research has been supported by the Volkswagen Foundation through its grant for a research project on *Financial Markets as Complex Networks.*

References

Baker M, Wurgler J (2006) Investor sentiment and the cross-section of stock returns. J Financ 61:1645–1680
Baker M, Wurgler J, Yuan Y (2009) Global, local and contagious investor sentiment. Working paper, Harvard Business School

Black F (1986) Noise. J Financ 41:529–543
Brown G, Cliff M (2004) Investors sentiment and the near-term stock market. J Empir Financ 11:1–27
Brown G, Cliff M (2005) Investor sentiment and asset valuation. J Bus 78:405–440
Brüggemann R (2004) Model reduction methods for vector autoregressive processes. Springer, Berlin
Clark T, West K (2007) Approximately normal tests for equal predictive accuracy in nested models. J Econom 138:291–311
DeLong J, Shleifer A, Summers L, Waldman R (1990) Noise trader risk in financial markets. J Political Econ 98:703–738
Hengelbrock J, Theissen E, Westheide C (2009) Market response to investor sentiment. Working paper, University Bonn
Kling G, Gao L (2008) Chinese institutional investors' sentiment. Int Financ Mark Inst Money 18:374–387
Lemmon M, Portniaguina E (2006) Consumer confidence and asset prices: some empirical evidence. Rev Financ Stud 19:1499–1529
Liu T-R, Geslow M, Irwin S (1994) The performance of alternative VAR models in forecasting exchange rates. Int J Forecast 10:419–433
Lütkepohl H (2005) New introduction to multiple time series. Springer, Berlin
Penm J, Terrell R (1984) Multivariate subset autoregressive modelling with zero constraints for detecting 'overall causality'. J Econom 24:311–330
Schmeling M (2007) Institutional and individual sentiment: smart money and noise trader risk? Int J Forecast 23:127–145
Schmeling M (2009) Investor sentiment and stock returns: some international evidence. J Empir Financ 16:394–408
Shleifer A, Vishny R (1997) The limits of arbitrage. J Financ 52:35–55
Verma R, Baklaci H, Soydemir G (2008) The impact of rational and irrational sentiments of individual and institutional investors on DIJA and S&P500 index returns. Appl Financ Econ 18:1303–1317