University of Pretoria
Department of Economics Working Paper Series

Bear, Bull, Sidewalk, and Crash: The Evolution of the US Stock Market Using Over a Century of Daily Data

Shixuan Wang
University of Reading
Rangan Gupta
University of Pretoria
Yue-Jun Zhang
Hunan University
Working Paper: 2020-97
October 2020

Department of Economics
University of Pretoria
0002, Pretoria
South Africa
Tel: +27 12 420 2413

Bear, Bull, Sidewalk, and Crash: The Evolution of the US Stock Market Using Over a Century of Daily Data

Shixuan Wang[*], Rangan Gupta[**] and Yue-Jun Zhang[***]

Abstract

In this paper, we employ a four-state hidden semi-Markov model (HSMM), which outperforms a hidden Markov model (HMM), to identify market conditions of the Dow Jones Industrial stock market over the daily period of 16th of February, 1885 to 4th of June, 2020. Our results indicate that the four hidden states represent bear-, bull-, sidewalk-, and crash-markets, which in turn appropriately captures the various major historical events during the period of study. Our results have implications for investors and policymakers.

Keywords: Dow Jones Industrial Average, Stock Returns, Hidden (semi-)Markov Models

JEL Codes: C22, G10

1. Introduction

Historically, the stock market of the United States (US) has been identified as a leading indicator for major macroeconomic variables, like metrics of economic activity, inflation and interest rates (Stock and Watson, 2003; Simo-Kengne et al., 2016; Plakandaras, 2017; Pierdzioch and Gupta, 2020). Naturally, appropriate modeling of the states of the equity market is of paramount importance for policymakers, as well as investors, especially at high-frequency. This is because, one can then use this information in mixed frequency data sampling (MIDAS) models to produce nowcasts and real-time predictions of the variables that are sample at lower (monthly or quarterly) frequencies, such as output growth and inflation (Andreou et al., 2013; Breitung and Roling, 2015). Against this backdrop, we aim to identify the evolution of the market conditions of the daily returns Dow Jones Industrial Average (DJIA) over its entire available history covering the period of 16th of February, 1885 to 4th of June, 2020. The DJIA time series data is employed due to the high number of observations available, which in turn allows us to avoid incorrect inference due to sub-sample-specific characteristics or inefficiency of estimates obtained from small samples, as stressed by Gebka and Wohar (2019). Note that these authors analysed the predictive power of the DJIA index returns, measured at different quantiles of its distribution, for future return distribution for the period of 26th May, 1896 to 10th September, 2014, i.e., over a similar historical sample like ours.

Market conditions have been studied mostly with Markov-switching techniques. As an extension of the classical Hidden Markov model (HMM), the hidden semi-Markov model (HSMM) can arbitrarily specify the sojourn time distribution.[1] Given this, Bulla and Bulla (2006) show that HSMM outperforms the HMM in the reproduction of the stylized facts of daily financial returns. Since then, the HSMM has been a prevailing tool to quantitively identify the market conditions based on the distributional properties of the hidden states (see, Yue (2010), Lau et al., (2017), Liu and Wang (2017a, 2017b), Apergis et al., (2019) for detailed literature reviews, and alternative applications in this regard). In light of this, we also employ

[*] Corresponding author. Department of Economics, University of Reading, Reading, RG6 6AA, United Kingdom. Email: shixuan.wang@reading.ac.uk.

[**] Department of Economics, University of Pretoria, Pretoria, 0002, South Africa. Email: rangan.gupta@up.ac.za.

[***] Business School, Hunan University, Changsha 410082, China; Center for Resource and Environmental Management, Hunan University, Changsha 410082, China. Email: zyjmis@126.com.

[1] Sojourn time is the number of time periods that the hidden (semi-)Markov chain stays in a specific state.

the HSMM model, for the first time in the literature, to analyse and identify hidden states of the DJIA returns, since its inception spanning 136 years of daily data. The remainder of the paper is organized as follows: Section 2 outlines the methodology, while Section 3 discusses the data and the empirical results, with Section 4 concluding the paper.

2. Methodology

The HSMM is based on two coupled processes, the state process $\{S_t\}_{t=1}^{T}$ and the observation process $\{X_t\}_{t=1}^{T}$. The state process follows a semi-Markov chain[2], which is constructed by an embedded first-order Markov chain with sojourn time distribution. $\{S_t\}_{t=1}^{T}$ is hidden and unobservable, and can only take finite state space, i.e. $S_t \in \{1, 2, \dots, K\}$. The time series dependence of S_t is characterized by the transition probabilities defined as:

$$\gamma_{i,j} = P(S_{t+1} = j | S_{t+1} \neq i, S_t = i) \text{ with } \sum_{j \neq i} \gamma_{i,j} = 1 \ and \ \gamma_{i,i} = 0. \tag{1}$$

Arranging all possible transition probabilities together into a matrix produces the transition probability matrix (TPM) with $K \times K$ dimension. It should be noted that the diagonal entries in the TPM of HSMM are all zero. The hidden Markov model (HMM) has the geometric distribution for the sojourn time. Unlike HMM, the sojourn time in HSMM,

$$d_j(\tau) = P(S_{t+\tau+1} \neq j, S_{t+\tau-v} = j, v = 0, \dots, \tau - 2 | S_{t+1} = j, S_t \neq j), \tag{2}$$

can be controlled by any arbitrary distribution.

The observation process $\{X_t\}_{t=1}^{T}$ is observable, and is generated based on the state process $\{S_t\}_{t=1}^{T}$. Importantly, the observation at time t only depends on the state at time t via the component distribution,

$$b_j(x_t) = P(X_t = x_t | S_t = j). \tag{3}$$

Bulla and Bulla (2006) provides the likelihood function for the observations modelled by HSMM. Then the expectation-maximization (EM) algorithm (Baum et al., 1970) is used to estimate the model parameters in HSMM. Based on the estimated parameters, the unobservable state process can be globally decoded by the Viterbi algorithm (Viterbi, 1967), which enable us to reveal the timing and the evolvement of the states over the sample period. For practical settings, we follow Liu and Wang (2017a) in using the logarithmic distribution for the sojourn time and the normal distribution for the component distribution, since they are straightforward to interpret, and the fact that convergence in the EM algorithm can be generally reached.

3. Data and Empirical Results

Our analysis involves the log-returns of the DJIA over the daily period of 16th of February, 1885 to 4th of June, 2020, with the start and end dates being governed by the availability of data at the time of writing this paper. The data is sourced from MeasuringWorth: https://www.measuringworth.com/datasets/DJA/index.php. We first present the model comparison of HSMM and HMM with different number of hidden states in Table 1. The four-state HSMM provides the best fit in terms of log likelihood, and is also the best model according

[2] Only non-absorbing states are considered in this study. Additionally, we follow Bulla and Bulla (2006) to consider the right-censored HSMM, which does not require the assumption that last observation coincide with the exit of a state.

to the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). We will proceed with this best model, i.e., the four-state HSMM, for the remainder of our analysis.

[INSERT TABLE 1 HERE]

Next, Table 2 shows the estimation result of the four-state HSMM. State 1 can be interpreted as the crash market because it has an extreme negative mean (-0.320) and the largest standard deviation (3.423), which typically covers the left tail of the return distribution. State 2 is characterized by a negative mean (-0.069) and second largest standard deviation (1.467), and thus corresponds to the bear market. State 3 has a positive mean (0.072) and the lowest standard deviation (0.489), which meets distributional properties of the bull market. Lastly, State 4 has a mean insignificantly different from zero (t-stat: 1.147) and the second lowest standard deviation (0.872), thus capturing the sidewalk market. Figure 1 illustrates the decomposition of empirical (aggregated) distributions into the four component distributions. The information on sojourn time confirms our interpretation: the crash market is typically short-lived with average sojourn time of 31 days, and the bear market lasts slightly longer with 52 days on average, while the bull and the sidewalk markets tend to continue for over half a year (125 days). We have four observations by examining the TPM: 1) the bear-market is always after the crash; 2) the bear-market can transit to the sidewalk- (69%) and to the crash- (31%) markets, but it never directly evolves into the bull-market; 3) the bull-market is succeeded by the sidewalk- market with a probability of 99%, and; 4) the sidewalk-market is more likely to be followed by the bull- (81%), rather than the bear- (19%) markets.

[INSERT TABLE 2 HERE]

[INSERT FIGURE 1 HERE]

By employing the Viterbi algorithm (Viterbi, 1967), we can globally decode the timing of the four hidden states over the entire sample period, which in turn is displayed in Figure 2. To facilitate studying the evolution of the states of the market, we collect the information of the four hidden states into different sub-periods based on the well-known historical events, as shown in Table 3. In the 19th century and before World War I (WWI) period, the market was mainly in the bull or sidewalk phases, with some occasional crashes. In the WWI, the percentage of periods in the bull-market substantially decreased, while the percentage of other market conditions increased, and in particular the crash episodes. The economy started to recover after the WWI until the "Great Depression" in 1929, during which period there was no crash. Nevertheless, the market was in tremendous turmoil during the "Great Depression" with nine episodes of crashes, and no bull market at all. Intriguingly, the market experienced a "u-turn" since the start of World War II (WWII), with 1,263 days in the bull market (out of 1,782 days in total). This can be intuitively explained by the fact that WWII largely helped the US to boost its economy. Over the long-lasting Cold War, the market had relatively lower percentage of time spent in the bear phase, and modestly higher percentage in the bull-market. In the pre-Global Financial Crisis (GFC) period, the market seemed to be normal with similar pattern as observed in the 19th Century, pre-WWI, and post-WWI. Since then, the market encountered substantial turmoil during the period of the GFC in 2008 (and the associated "Great Recession"), and the European sovereign debt crisis (ESDC) in 2010, featured with 10.2% of total days in the crash- and 32.7% in the bear-markets. In the recent years, the US stock market had gradual recovery with 51.5% of total days in the bear phase, until another dramatic shock due to the outbreak of COVID-19.

[INSERT FIGURE 2 HERE]

[INSERT TABLE 3 HERE]

4. Conclusion

In this paper, we employ a HSMM on the DJIA returns covering the period of 16th of February, 1885 to 4th of June, 2020. We find that a four-state HSMM model, corresponding to bear, bull, sidewalk, and crash markets, fits the data the best when compared to the version of the model with two and three hidden states. This model also outperforms the HMM model estimated with the same number of hidden states. Finally, when we analyse the evolution of the returns, our model is appropriately able to associate the four states with the various historical events, such as WWI, the "Great Depression" WWII, the Cold War, global financial and European sovereign debt crises, and the recent outbreak of the Coronavirus. Given that, stock market movements act as the leading indicator for macroeconomic variables measured in lower frequencies, the high-frequency information contained in the hidden states identified by the HSMM model could be used by policymakers to nowcast the economy based on MIDAS models, and investors to conduct timely portfolio allocations.

Reference

Andreou, E., Ghysels, E., and Kourtellos, A. (2013). Should Macroeconomic Forecasters Use Daily Financial Data and How? Journal of Business & Economic Statistics, 31(2), 240-251.

Apergis, N., Gozgor, G., Lau, C.K.M., and Wang, S. (2019). Decoding the Australian electricity market: new evidence from three-regime hidden semi-Markov model. Energy Economics, 78, 129-142.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics, 41(1), 164-171.

Breitung, J., and Roling, C. (2015). Forecasting Inflation Rates Using Daily Data: A Nonparametric MIDAS Approach. Journal of Forecasting, 34(7), 588-603.

Bulla, J., and Bulla, I. (2006). Stylized facts of financial time series and hidden semi-Markov models. Computational Statistics & Data Analysis, 51(4), 2192-2209.

Gebka, B., and Wohar, M.E. (2019). Stock return distribution and predictability: Evidence from over a century of daily data on the DJIA index. International Review of Economics & Finance, 60, 1-25.

Lau, C.K.M., Vigne, S.A., Wang, S., and Yarovaya, L. (2017). Return spillovers between white precious metal ETFs: the role of oil, gold, and global equity. International Review of Financial Analysis, 52, 316-332.

Liu, Z., and Wang, S. (2017a). Decoding Chinese stock market returns: Three-state hidden semi-Markov model. Pacific-Basin Finance Journal, 44, 127-149.

Liu, Z., and Wang, S. (2017b). Understanding the Chinese stock market: international comparison and policy implications. Economic and Political Studies, 5(4), 441-455.

Pierdzioch C., and Gupta R. (2020). Uncertainty and Forecasts of U.S. Recessions. Studies in Nonlinear Dynamics & Econometrics, 24(4), 1-20.

Plakandaras, V., Cunado, J., Gupta, R., and Wohar, M.E. (2017). Do leading indicators forecast U.S. recessions? A nonlinear re - evaluation using historical data. International Finance, 20(3), 289-316.

Simo-Kengne, B.D., Miller, S.M., Gupta, R., and Balcilar, M. (2016). Evolution of the Monetary Transmission Mechanism in the US: the Role of Asset Returns. The Journal of Real Estate Finance and Economics, 52(3), 226-243.

Stock, J.H., and Watson, M.W. (2003). Forecasting Output and Inflation: The Role of Asset Prices. Journal of Economic Literature, 41(3), 788-829.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory, 13(2), 260-269.

Yu, S.Z. (2010). Hidden semi-Markov models. Artificial intelligence, 174(2), 215-243.

Table 1. Model Comparison

	Log Likelihood	AIC	BIC
2-State HSMM	-47747.15	95508.30	95567.93
3-State HSMM	-46666.82	93361.63	93480.91
4-State HSMM	**-46356.25**	**92758.49**	**92954.44**
2-State HMM	-47989.73	95993.45	96053.09
3-State HMM	-46739.19	93506.39	93625.66
4-State HMM	-46388.95	92823.90	93019.85

Note: The red bold number indicates the best metric for the criteria of model comparison.

Table 2. Estimation of the 4-State HSMM

		State 1	*State 2*	*State 3*	*State 4*
Conditional Distribution	*Mean*	-0.320	-0.069	0.072	0.028
	SD	3.423	1.467	0.489	0.872
	t-statistic	-3.183	-1.457	2.233	1.147
TPM	*From/To*				
	State 1 (Crash)	-	100.0%	0.0%	0.0%
	State 2 (Bear)	31.1%	-	0.0%	68.9%
	State 3 (Bull)	0.7%	0.0%	-	99.2%
	State 4 (Sidewalk)	0.0%	18.8%	81.2%	-
Sojourn Time	*No. of Days*	1161	5275	11191	19406
	No. of Times	38	102	89	153
	Average Sojourn Time	30.553	51.716	125.742	126.837

Figure 1. Empirical and Component Densities of the 4-State HSMM

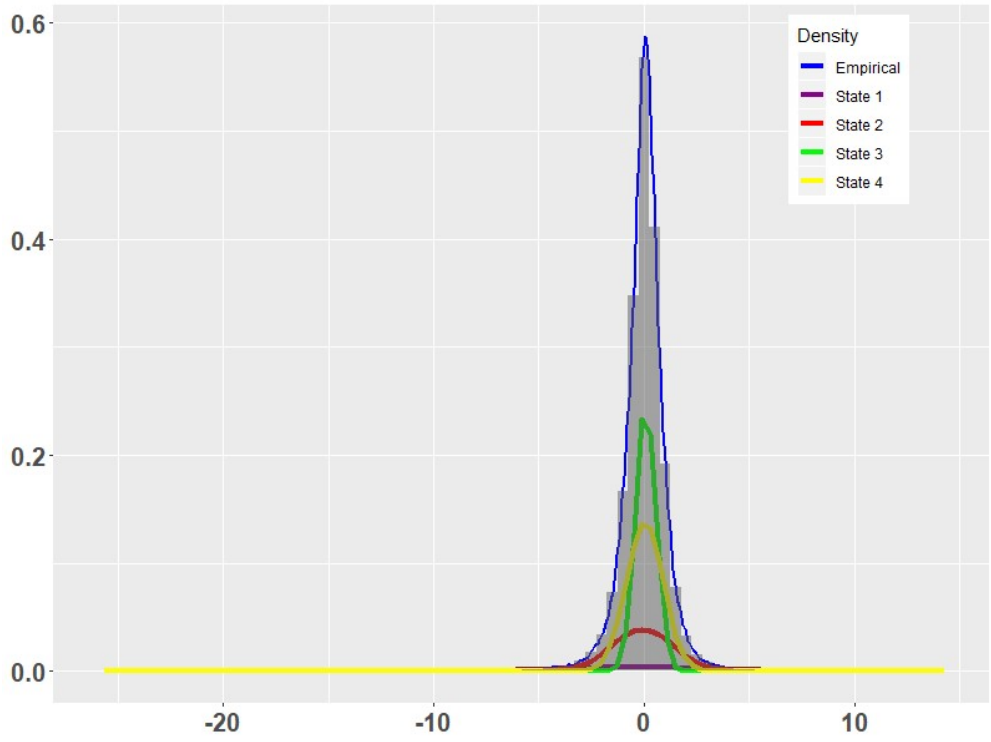
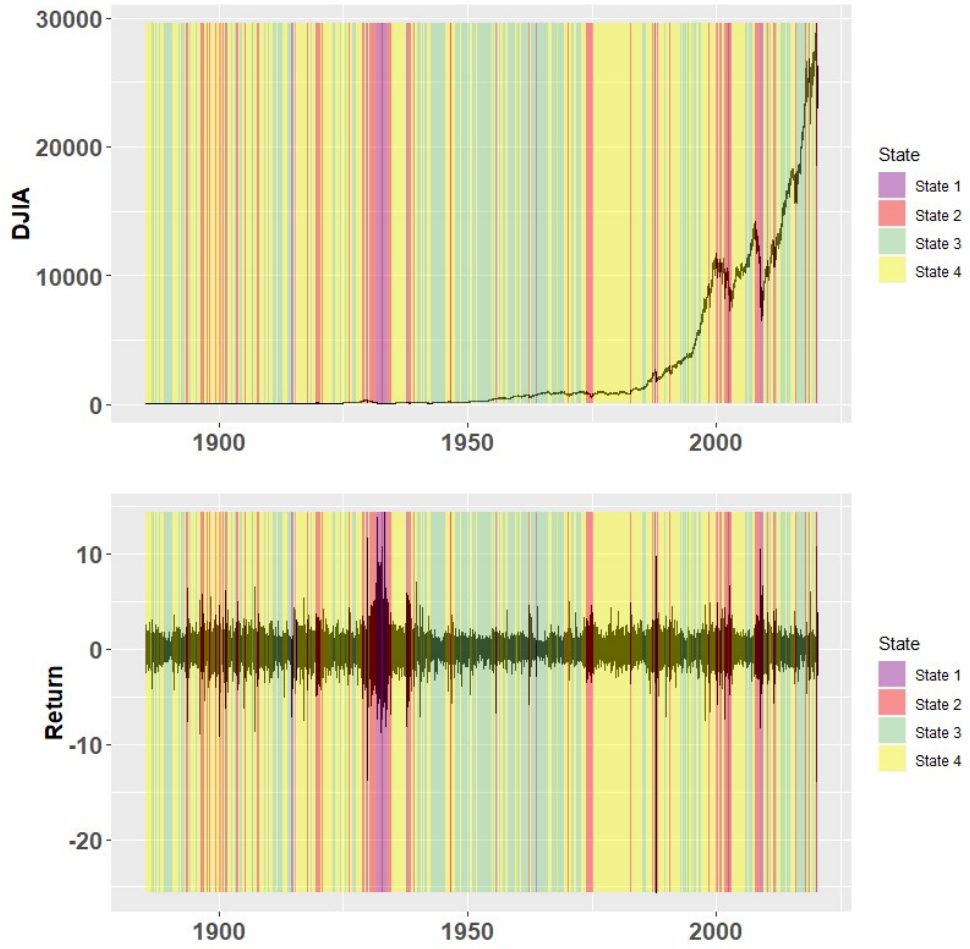


Figure 2. Global Decoding of the 4-State HSMM



Note: Upper panel: DIJA index along with the four identified states; Lower panel: returns of DIJA along with the four identified states.

Table 3. Information on the Four States in Different Sub-Periods

Label		**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
Famous Event		**19 Century**	**Pre-WWI**	**WWI**	**Post-WWI**	**Great Depression**	**WWII**	**Cold War**	**Pre-GFC**	**GFC and ESDC**	**Recent**
Start		1885-02-16	1900-01-01	1914-07-28	1918-11-12	1929-09-04	1939-09-01	1945-08-16	1991-12-27	2007-02-01	2012-08-01
End		1899-12-31	1914-07-27	1918-11-11	1929-09-03	1939-08-31	1945-08-15	1991-12-26	2007-01-31	2012-07-31	2020-06-04
Total Days		4488	4358	1176	3229	2981	1782	11878	3804	1386	1910
No. of Days	***State 1***	37	29	30	0	791	11	29	54	142	36
	State 2	512	675	217	563	1045	47	887	640	453	197
	State 3	1302	1146	34	515	0	1263	4802	941	205	983
	State 4	2637	2508	895	2151	1145	461	6160	2169	586	694
Percentage in Total Days	***State 1***	0.8%	0.7%	2.6%	0.0%	26.5%	0.6%	0.2%	1.4%	10.2%	1.9%
	State 2	11.4%	15.5%	18.5%	17.4%	35.1%	2.6%	7.5%	16.8%	32.7%	10.3%
	State 3	29.0%	26.3%	2.9%	15.9%	0.0%	70.9%	40.4%	24.7%	14.8%	51.5%
	State 4	58.8%	57.5%	76.1%	66.6%	38.4%	25.9%	51.9%	57.0%	42.3%	36.3%
No. of Times	***State 1***	3	3	4	0	9	2	8	5	2	2
	State 2	12	17	8	5	14	4	21	10	5	8
	State 3	9	14	1	4	0	8	29	9	5	11
	State 4	18	27	6	9	4	10	43	14	8	18

Note: WWI denotes the first World War; WWII represents the second World War; GFC is the global financial crisis in 2008; EDC stands for European sovereign debt crisis; State 1 is the crash market, State 2 is the bear market, State 3 is the bull market, and State 4 is the sidewalk market.