2

Did Twitter Mood Really Predict the DJIA? Misadventures in Big Data for Finance

Michael Lachanski

Princeton University (Class of 2015)

Did Twitter Mood Really Predict the DJIA?

Misadventures in Big Data for Finance

Abstract

In "Twitter mood predicts the stock market", Bollen et al. [2011a] claim to be able to predict whether the DJIA will close with a higher or lower value than the previous day's value 86.7% of the time. We replicate and extend results by Pav [2012a] which suggests that these results constitute a Type I error. We show that the results presented by Bollen et al. [2011a] imply that there exists a simple market timing strategy which typically achieves a Sharpe Ratio of at least 5.8, nearly an order of magnitude greater than any published result we have found on market timing ability. We show that *any* out-of-the-box correction for multiple comparison bias eliminates all of the statistically significant results from Bollen et al.'s linear time series analysis. We document several other errors and potential errors, concluding that the evidence presented in Bollen et al. is not sufficient to show that Twitter mood actually predicted the DJIA.[1]

CLASSIFICATION: G100, G12, Y8, C390

Acknowledgments: The author would like to thank Burton Malkiel, Thomas Espenshade, Stephan Luck, German Rodriguez, and Rene Carmona for their continued support. This essay has benefited considerably from helpful comments by Frank Fabozzi, Steven Pav, Francine Loza, Modibo Camara, Yuliy Sannikov, and an anonymous referee. All errors are my own.

[1] All code for this essay is available at www.github.com/mlachans/DTMRPTDJIA

2.1 Introduction

In this essay, we look closely at the results presented in "Twitter mood predicts the stock market" (TMP hereafter). Excluding the DJIA, 3-month Treasury, and S&P 500 data, we use no data in the analyses of this essay, limiting ourselves to an ex-post analysis of the results presented in TMP.[2] We validate and extend arguments presented in Pav [2012a,b]. First, we show that the 86.7% DJIA up-down predictability reported in TMP almost certainly violates the efficient market hypothesis (EMH hereafter). Second, we demonstrate that the significant linear hypothesis tests motivating the non-linear analysis results arise completely from multiple comparison bias. Finally, we show that the normalization technique used in TMP's visualizations and potentially used in the linear hypothesis testing can induce false positives.

2.1.1 Review of TMP and Outline

TMP's headline result is that by feeding previous days' stock prices and their measures of Twitter mood into a Self-Organizing Fuzzy Neural Net Classifier (SOFNN hereafter) the authors were able to predict 13 out of 15 days' directional DJIA movements from December 1, 2008 to December 19, 2008. Bollen et al. (BMZ hereafter) find that the mood they label CALM is most predictive of stock prices. Pav [2012a] presents a single Monte Carlo simulation in which the ability to predict whether the market will be up or down $\frac{13}{15}$ days leads to an annualized SR of 9. Given that in his experience, realized Sharpe Ratios (SRs hereafter) greater than 3 would be "the stuff of legend", he concludes that the result is likely to be misstated. In §2.2 we show that, if BMZ's result held ex-sample, it would still be the greatest trading strategy ever discovered in that trading based on the SOFNN classifier would lead to large SRs (> 5) even after making adjustments for the risk-free rate and realistic transactions costs, features left out of the original analysis by Pav [2012a].

The statistical content of TMP, as it relates to the DJIA, consists of two parts. The first is a set of 49 hypothesis tests on linear autoregressive distributed lag (ARDL hereafter) models and

[2] In other words, at no point in this essay do we *directly* analyze the *joint* distribution of Twitter mood and DJIA prices.

the second is a set of out-of-sample tests using non-linear models. Using standard t-statistics we expect 5% of all hypothesis tests to show significance even if the null result is true. BMZ conduct so many hypothesis tests using standard t-statistics that they cannot rule out the possibility that their most significant results are in fact null. In §2.3, we review the theory of multiple hypothesis testing, present the Bonferroni adjusted coefficients referred to by Kuleshov [2011], Pav [2012a] and extend their criticisms. Even under the most powerful currently available adjustments for multiple comparison bias, none of the p-values reported in the linear hypothesis testing section of TMP are significant at any standard level.[3]

It is likely that BMZ use a normalization procedure on their mood time series which, while not uncommon in content analysis, prevents them from being able to claim that the *non-normalized* CALM time series Granger cause anything.[4] §2.4 constructs an example by simulation in which X and Y are coincident, Y does not Granger cause X yet locally normalized Y Granger causes X. §2.5 concludes.

2.2 Implied Sharpe Ratio Exceeds That of Any Previously Discovered Strategy

> To put these figures into context, an achieved (i.e. in real trading, not backtesting) Sharpe ratio of $1\text{yr}^{-1/2}$ is considered 'good'; an achieved value of $2\text{yr}^{-1/2}$ is considered 'excellent'; anything north of $3\text{yr}^{-1/2}$ is the stuff of legend.
>
> ---
>
> *sellthenews.tumblr.com*
>
> STEVEN PAV on "Twitter mood predicts the stock market"

Pav [2012a] conducts one Monte Carlo simulation over DJIA data from 1970 to 2012 to obtain an SR of 9.2. Since, in his experience in quantitative finance, an SR of 2.5 has never been realized over any significant length of time for a market timing strategy Pav concludes that the results in

[3] Powerful in the sense that they are valid and reject the null hypothesis in situations where less powerful adjustments accept the null hypothesis.

[4] See the equation on page 452 of Bollen et al. [2011b] for an example of the type of normalization under analysis.

TMP are likely to be overstated.

We do not disagree with his conclusions, however his quantitative analysis is biased towards reporting a higher SR because he sets the risk-free rate and transaction costs to zero.[5] Pav conducted only a single simulation and it is unclear if his estimate of the SR is representative of those we would obtain if we attempted to exploit BMZ's findings.[6] His simulations may also not be useful for our purposes because he uses the DJIA time series from 1970 and 1930 respectively. Until recently, there were no instruments replicating cash flows from holding the DJIA. [7] Because the DJIA is an untradable price-weighted index, simulated returns from trading the index before the development of these instruments are not necessarily subject to the EMH.[8] In this section, we use estimates for the risk-free rate from the daily yield of the 1 month constant maturity U.S. Treasury daily yield and transaction costs to assess the SR that this strategy might reasonably achieve.[9] Our transaction cost schedule is taken from Tetlock et al. [2008] who used transaction costs ranging from 0 to 10 basis points round-trip as realistic baselines to assess the viability of text analytics based trading strategies.[10] Since our hypothetical trader trades daily, each cost is applied daily.

Our strategy follows the set-up in Lachanski [2015] and trades at unit leverage. κ is defined as in Pav [2012b] and Lachanski [2015]. Thirteen out of fifteen times, our portfolio gains the absolute value of DJIA's daily return. Two out of fifteen times, our strategist guesses incorrectly and loses the absolute value of the DJIA's daily return. We conduct two sets of simulations. Our first simulation assumes, as Pav [2012a] does, that arbitrage identified in TMP has always existed.

[5] We learned this from correspondence with Pav.

[6] It can be proven that if the stochastic process generating strategy returns is stationary then a performance simulation will yield the correct expected SR if the time series is long enough. For a typical buy-and-hold strategy, time varying volatility lets us reject the stationarity assumptions at the outset. We are aware of a number of alternative conditions in which a single performance simulation on a long enough non-stationary time series will generate correct estimates of the SR, but these conditions are technical and verifying them would take us beyond the scope of this essay. Instead of attempting to verify these conditions, we defer to standard practice as explicated in Carmona [2014] and conduct a large number of simulations (1000) instead.

[7] http://www.investopedia.com/articles/investing/010815/etfs-tracking-dow.asp, accessed 2015/03/19.

[8] This a purely technical objection. In practice, the DJIA and S&P 500 track each other with such a high correlation coefficient that a timing strategy which works in one will almost certainly work in the other.

[9] The 1 month constant maturity U.S. Treasury daily yield is calculated from 3 month Treasury bill yields by FRED.

[10] The DJIA index does not reflect dividend payouts of the underlying firms (except to the extent that dividend payouts announcements lower the price of the equities) and so its "adjusted close", used for calculating returns, is its nominal price.

We conduct the first simulation from 7/31/2000[11] to 12/31/2014 using S&P500 volatility data for calibrating the SR bound. We use DJIA data from Yahoo! Finance via the quantmod package for our simulations. Our second simulation assumes that the arbitrage identified in TMP only became possible after Twitter was opened to the public and runs from 07/15/2006[12] to 12/31/2014.

For formalizing the violation of the EMH, we follow Lachanski [2015] and calculate a theoretical "good deal" upper bound on SR. We take:

$$R^f = \max\left(1+\text{Daily Yield of 3 month Treasury}\right)^{252} \approx 1.0622$$

and relative risk aversion as 6.4. S&P 500 excess return volatility over periods starting from 2000 and 2007 are 19.98% and 21.25% respectively. Following Lachanski [2015], these parameters translate to "good deal" bounds on predictable annualized SR of 1.36 and 1.44 respectively. Table 2.1 contains our simulation results that assume, like Pav does, that the arbitrage identified in TMP has always existed. Table 2.2 contains our simulations based on the idea that only after the introduction of Twitter, by allowing one to cheaply estimate collective mood states, did the trading opportunity identified by BMZ come into existence. Our simulations replicate Pav's results and validate his intuition: 86.7% DJIA predictability leads to such profitable trades that adding realistic trading costs and the risk-free rate to our simulations changes nothing qualitatively. The SR bound we constructed was violated in every single simulation. Over the 11,000 simulations conducted for this section, none produced an SR lower than 5.8. Not only does TMP's strategy present a "good deal", but if the 86.7% DJIA predictability finding is correct then it presents the greatest deal ever published in the "good deal" literature (as surveyed in Lachanski [2015]).

The point estimate of 86.7% in TMP almost certainly violates the EMH. Not only this, but even minimum value in the 95% confidence interval for DJIA predictability given in TMP would violate the bound we have constructed. Concretely, we can use the lower bound of TMP's 95% confidence

[11] We chose this as the start date because it was after the development of DJIA tracking financial instruments and because this corresponds with the start of the Federal Reserve Bank of St. Louis (FRED hereafter) 1-month Treasury Constant Maturity Rate time series.

[12] http://latimesblogs.latimes.com/technology/2011/07/twitter-delivers-350-billion-tweets-a-day.html, accessed 2015/04/01.

interval calculated by Pav, $g = 7/30$. In this case, our theoretical SR with $r_f = 0$ is 5.2. In general, even small abilities to time the market at the daily frequency leads to violation of the "good deal" bound. We can see that the minimal g violating the bound, for $\hat{\kappa}_{r_f=0}$ calculated from 7/15/2006 to 2014/12/31, is approximately 0.0676. In other words, a daily edge greater than 6.76% would likely violate the bound we have constructed.[13] In other words, the result in TMP is, given the realized distribution of SR documented above and the theoretical upper limit on SR for a strategy that trades daily, very likely to be overstated.

2.3 Multiple Comparison Bias

> ..it is clear that nothing limits...the number of features according to which one can distribute [natural events or social facts] into several groups or distinct categories...One could distinguish first of all legitimate births from those occurring out of wedlock, one can also classify births according to the age, profession, wealth, or religion of the parents...usually these attempts through which the experimenter passed don't leave any traces; the public will only know the result that has been found worth pointing out; and as a consequence, someone unfamiliar with the attempts which have led to this result completely lacks a clear rule for deciding whether the result can or can not be attributed to chance.
>
> ANTOINE COURNOT (1843) ON THE MULTIPLE HYPOTHESIS PROBLEM AS QUOTED IN SHAFFER (1995)

When conducting a single hypothesis test, if our null hypothesis is true, then our p-value will be distributed uniformly in $[0, 1]$. We set the probability of a false positive, or a rejection of the null given that the null is true, at level α. In most studies, the probability of a false positive is fixed

[13] This is why, as noted in Lachanski [2015], we are also skeptical about many of the results summarized by Nassirtoussi et al. [2014]. As the distribution of returns becomes more heavy-tailed, $\kappa \to \infty$ and so the edge necessary to achieve a fixed SR increases. Emerging markets and FX markets, on which many of the predictive exercises in Nassirtoussi et al. [2014] are based, generally have heavier tails than the S&P 500. However, simulations not presented in this essay suggest that, even accounting for larger κ values in these markets, one still obtains SRs from the aforementioned predictability results that exceed the "good deal" bounds constructed in this section.

Theory: $r_f = 0$	$r_f = 0$	r_f =1 mo. Treasury daily yield	Trading Cost
9.01	8.95	8.82	0 bp
	8.79	8.67	1 bp
Theory r_f =1 mo. Treasury daily yield	8.64	8.53	2 bp
8.96*	8.49	8.36	3 bp
	8.33	8.21	4 bp
$\hat{\kappa}_{r_f=0}$	8.18	8.06	5 bp
0.56	8.03	7.91	6 bp
	7.87	7.75	7 bp
$\hat{\kappa}_{r_f=}$ 1 month Treasury daily yield	7.72	7.60	8 bp
0.56	7.57	7.45	9 bp
	7.42	7.29	10 bp

Table 2.1: Our first column contains theoretical SRs using a calculation in Lachanski [2015] and estimated from the raw returns and excess returns respectively. The second and third columns come from the simulation. Our upper bound for average annualized SR over this time period is 1.36. Even with the addition of the risk-free rate and the inclusion of daily transaction costs, the predictive power of the strategy in TMP far exceeds our upper bound on ex-ante SR over this time period. ∗This value is biased upwards because TMP does not predict excess returns.

Theory: $r_f = 0$	$r_f = 0$	r_f =1 mo. Treasury daily yield	Transaction Costs
8.45	8.43	8.38	0 bp
	8.28	8.24	1 bp
Theory r_f =1 mo. Treasury daily yield	8.14	8.09	2 bp
8.45**	8.00	7.94	3 bp
	7.85	7.80	4 bp
$\hat{\kappa}_{r_f=0}$	7.71	7.66	5 bp
0.61	7.56	7.52	6 bp
	7.42	7.37	7 bp
$\hat{\kappa}_{r_f=}$ 1 month Treasury daily yield	7.28	7.23	8 bp
0.61	7.13	7.08	9 bp
	6.99	6.94	10 bp

Table 2.2: This table presents the mean annualized SR under the assumption that the arbitrage opportunity identified in TMP has existed since 7/15/2006. Our first column contains theoretical SRs using a calculation in Lachanski [2015] and $\hat{\kappa}$ estimated from the DJIA returns, corresponding with the case in which $r_f = 0$, and excess DJIA returns respectively. The second and third columns come from the simulation. Our upper bound for average annualized SR over this time period is 1.44. Even with the addition of the risk-free rate and the inclusion of daily transaction costs, the predictive power of the strategy in TMP far exceeds our upper bound on ex-ante SR over this time period. $**$This value is biased upward because TMP does not predict excess returns.

at level $\alpha = 0.05$. This is typically called a significance level. We can see that in this situation:

$$\mathbb{P}\{\text{False Positive}\} = 1 - \mathbb{P}\{\text{No False Positive}\} = 1 - (1 - \alpha) = \alpha = 0.05$$

It is a straightforward conclusion that if many people individually test different, independent hypotheses with true nulls, then we expect between two and three researchers to have statistically significant findings in their sample.[14] We call a rejection of the null hypothesis when the null is true a Type I error. For instance, suppose 49 researchers conduct hypothesis tests in which the null hypothesis is true. Then, the probability of making a Type I error is:

$$\mathbb{P}\{\text{Type I error}\} = 1 - \mathbb{P}\{\text{No Type I error}\} = 1 - (1 - \alpha)^{49} \tag{2.1}$$

In the case of TMP, BMZ conduct 49 hypothesis in a single study in which they use both significance levels $\alpha = 0.1$ and $\alpha = 0.05$. Using (31), we can see that the probability of at least one Type I error, assuming the null, is:

$$1 - (1 - 0.1)^{49} = 0.994 \gg \alpha = 0.1 \text{ and } 1 - (1 - 0.05)^{49} = 0.919 \gg \alpha = 0.05 \tag{2.2}$$

if we do not somehow adjust for the number of hypotheses we are testing. We can see that, across the whole family of hypotheses tested by BMZ, it is a near certainty under the null that one of them is a false positive. We describe the situation in which, because of the sheer number of hypotheses being tested, the probability of a false positive diverges "meaningfully"[15] from the pre-specified α as one in which our statistics suffer from multiple comparison bias. To see this problem in TMP visually, Pav [2012a] presents a quantile-quantile (Q-Q hereafter) plot, which plots the quantile

[14] In the context of empirical asset pricing, Harvey et al. [2015] for instance, use variants on the multiple comparison adjustments presented in this section to show that most published factor-pricing models are false.

[15] The importance of the divergence, of course, depends on the application. Even testing two independent hypotheses will raise the probability of a false positive to 0.0975 if we specify $\alpha = 0.05$, but for some commercial applications (e.g. A/B testing for advertising) these are acceptable increases in the probability of Type I error and multiple comparison adjustments are rarely performed. On the other hand, in genetic association studies in which there are thousands of covariates and false positives would be especially deleterious to medical or ethical decision making, unnecessarily low-power multiple hypothesis tests (e.g. Bonferroni) are the norm; see Rietveld et al. [2015] for a recent example.

function of the uniform distribution against the quantile function of the empirical distribution of the p-values in TMP's Table 2.2. We replicate his results in our own Q-Q plot, displayed on the left panel of Figure 2.1.

The null hypothesis is that all 49 p-values are drawn from a uniform distribution. Carmona [2014] notes that if, in fact, these p-values are drawn from a distribution other than the uniform distribution (i.e. the null is false), we should expect points in the lower left corner of the Q-Q plots in Figure 2.1 to be significantly below the line $y = x$. For TMP's p-values, we can see no such thing: the lowest p-values fall about where we would expect them to if the entire family of p-values was drawn from a uniform distribution. On the other hand, in our right-hand panel of Figure 2.1, in which CALM is a significant predictor of the stock market, we observe that many points in the lower left-hand corner of the Q-Q plot fall below the line $y = x$.

Unfortunately, visualizations like these can be misleading. The lowest p-values in TMP's Q-Q plot do appear to be slightly below the line $y = x$. Worse, it is possible to simulate examples in which the expected and observed quantiles are drawn from different distributions and yet visually, the Q-Q plots indicate that the expected and observed distributions appear to be a perfect match in the sense that all points appear to fall on the line $y = x$.[16] The only way to rigorously determine whether or not the results in TMP are the result of multiple comparison bias is to correct for the number of hypothesis tests that BMZ conduct. Since these corrected p-values will always correspond with significant statistics, we refer to these ex-post p-value adjustments as multiple hypothesis tests or multiple comparison adjustments. Our presentation of multiple hypothesis testing follows a combination of Shaffer [1995] who covers multiple hypothesis testing generally, and Harvey and Liu [2014] who covers it for financial applications. All other works are cited as they are used.

[16] An example of misleading Q-Q plots is available here: http://stats.stackexchange.com/questions/2492/is-normality-testing-essentially-useless, accessed last 2015/03/15.

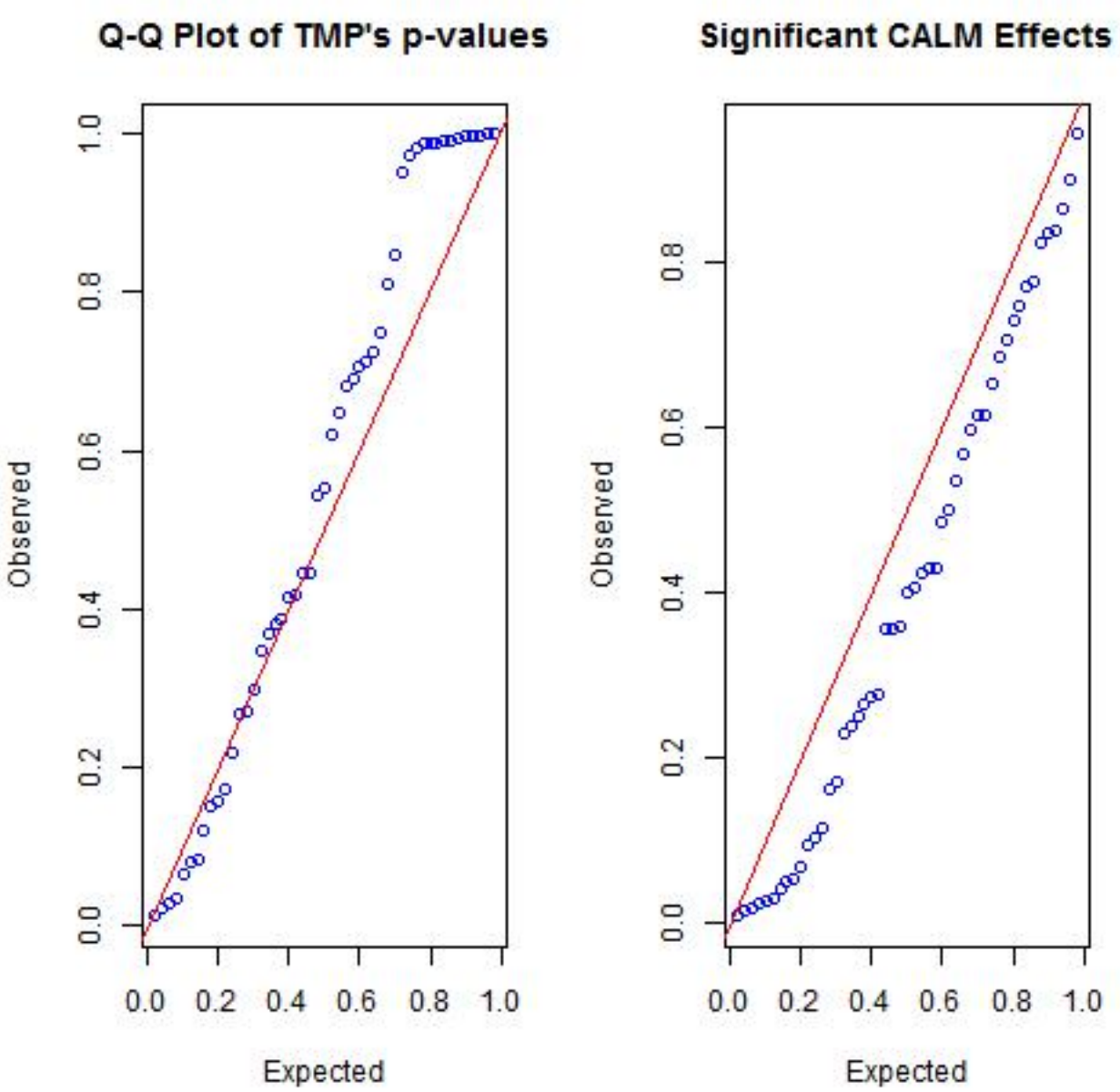

Figure 2.1: Pav [2012a] presents a Q-Q plot equivalent to our left hand plot. The horizontal axis of our plots presents the quantiles of the theoretical uniform distribution while the y-axis presents the empirical distribution of TMP's p-values on the left-hand panel and with one case in which the null hypothesis is false on the right-hand panel. For the right hand panel, our 7 CALM hypothesis test p-values are drawn from a $[0, 0.1]$ uniform distribution while the rest are drawn from the [0, 1]. Notice that, visually, BMZ's results do not collectively appear to diverge from the p-values we would expect to find if the null hypothesis was true while our simulated results on the right plot do diverge.

2.3.1 Controlling the Familywise Error Rate

We begin by presenting Table 2.3, which contains the uncorrected p-values reported in Table 2.2 of TMP.[17]

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	**0.085**	0.272	0.952	0.648	0.120	0.848	0.388
2 Days	0.268	**0.013***	0.973	0.811	0.369	0.991	0.7061[18]
3 Days	0.446	**0.022***	0.981	0.349	0.991	0.991	0.723
4 Days	0.218	**0.030***	0.998	0.415	0.989	0.989	0.750
5 Days	0.300	**0.036***	0.989	0.544	0.553	0.996	0.173
6 Days	0.446	**0.065**	0.996	0.691	0.682	0.994	**0.081**
7 Days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

Table 2.3: These p-values are taken from Table 2.2 in TMP. Throughout this section, we will be applying multiple hypothesis adjustments to this table of p-values.

There are two philosophies to multiple hypothesis testing. In the first philosophy, we aim to control the family-wise error rate (FWER hereafter). The FWER is the probability of one false positive in our entire family of hypotheses if null is true. An obvious correction is to simply multiply our p-values by the number of hypotheses we are testing. This is called the Bonferroni correction and we derive it below with some additional notation that will be used for specifying the other tests. First, let's give an example in which the Bonferroni correction gets the job done using the same situation described in (2):

$$1 - \left(1 - \frac{0.1}{49}\right)^{49} = 0.095 < \alpha = 0.1 \text{ and } 1 - \left(1 - \frac{0.05}{49}\right)^{49} = 0.048 < \alpha = 0.05 \tag{2.3}$$

so that, in this case, the probability of a Type I error is actually less than the the pre-specified significance level. We will prove that the Bonferroni correction always achieves this outcome. Let M be the number of hypotheses we are testing and let us suppose that our p-values have been ordered from least to greatest. Let α_1 be the first (minimum) unadjusted p-value and α_k be the k-th

[17] If BMZ have truncated rather than rounded their p-values, as we suspect they have, any adjustments we make to those p-values will be biased towards significance. We report all units to three decimal places.

smallest element in our list so that:

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_k \leq \cdots \leq \alpha_{M-1} \leq \alpha_M \tag{2.4}$$

Finally, let p_k be the "adjusted" p-value. Then, our Bonferroni correction is:

$$p_k = \begin{cases} \alpha_k M & \text{if } \alpha_k M < 1 \\ 1 & \text{otherwise} \end{cases} \tag{2.5}$$

Let m be the number of true null hypotheses and all true null hypotheses comprise set $I_0 \in I$ where I is the set of all hypotheses tested. Note that $m \leq M$. Using Boole's inequality over the set of events $\left\{A_i\right\}_{i=1}^{n}$:

$$\underbrace{\mathbb{P}\left\{\bigcup_{i=1}^{n} A_i\right\} \leq \sum_{i=1}^{n} \mathbb{P}\left\{A_i\right\}}_{\text{Boole's Inequality}}$$

we show that (5) will always reduce the probability of Type I error below α:

$$\begin{aligned} FWER &= \mathbb{P}\left\{\bigcup_{I_0}\left(p_k \leq \frac{\alpha}{M}\right)\right\} \text{ definition of } FWER \\ &\leq \sum_{I_0} \mathbb{P}\left\{p_k \leq \frac{\alpha}{M}\right\} \text{ Boole's Inequality} \\ &\leq m\frac{\alpha}{M} \text{ uniformity of p-values under null and } \sum_{i=1}^{n} \frac{\alpha}{n} = \alpha \\ &\leq \frac{M\alpha}{M} \text{ since } M \geq m \\ &= \alpha \end{aligned}$$

Thus, Bonferroni corrected p-values less than α will guarantee that our results are significant at level α.

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 Days	1.000	0.637	1.000	1.000	1.000	1.000	1.000
3 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.4: This table presents Bonferroni corrected p-values for TMP's linear models. These p-values are referred to, but not actually presented in Kuleshov [2011], Pav [2012a]. All remaining tables in this section are original.

Bonferroni Results

Using (5), the Bonferroni corrected p-values are in Table 2.4. Pav [2012a] and Kuleshov [2011] suggest that under Bonferroni testing none of the p-values presented by BMZ are significant at any standard level, but neither produces the Bonferroni adjusted p-values.

Unfortunately, as we can see from our example in (3) our Bonferroni test is conservative in the sense that it is possible that the true significance of the results will be less than α.[19] It can be shown that the Bonferroni test unnecessarily increases the number of Type II errors (i.e. false negatives).[20] This motivates the Holm [1979] test we present below.

Holm Results

The Holm test uses the order of the test statistics to increase the power of the test and, like the Bonferroni test, requires no assumptions on our data. The Holm adjustment sets the k-th smallest p-value to:

[19]This is particularly easy to see when all of the underlying covariates being tested have correlation equal to one. In this case, we should make no adjustment at all, yet the Bonferroni adjustment will increase the hurdle our p-values must pass by a factor of M. As one might expect, most of the less stringent tests we will use assume positive association (not necessarily linear association) among all hypotheses being tested.

[20]A false negative is a failure to reject the null when it is false.

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 Days	1.000	0.637	1.000	1.000	1.000	1.000	1.000
3 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.5: This table presents Holm corrected p-values for TMP's linear models. Even with the additional information provided by the order of the test statistics, none are significant. Even the second smallest p-value, Holm-corrected, is 1.000.

$$p_k = \begin{cases} \alpha_k (M + 1 - k) & \text{if } \alpha_k (M + 1 - k) < 1 \\ 1 & \text{otherwise} \end{cases} \tag{2.6}$$

but adds the additional criterion: for the minimum k such that $p_k > \alpha$ (being careful not to reorder the p-values after adjustment), accept all null hypotheses associated with p-values $\{\alpha_{k+1}, \ldots, \alpha_M\}$ at level α. We can see that for α_1, the Holm adjustment is equivalent to the Bonferroni adjustment but becomes less stringent thereafter. Procedures that require an ordering of $\alpha_{i \in \{1,\ldots,M\}}$ to conduct a hypothesis test are called "sequential-rejection" or "multistage" methods. Because the Holm method starts with α_1 and continues sequentially until α_M, it is called a step-down method.

Using (6) on the p-values in TMP gives us the results presented in Table 2.5. Unfortunately, without making additional assumptions on the moods we do not have access to more powerful tests that control the FWER. Nonetheless, we make these assumptions below to show how robust the insignificance of TMP's results are.

Hochberg, Hommel and Sidak Adjustments Tests

Tetlock [2007] and Loughran and McDonald [2015] both report that mood state measurements one might expect to move in opposite directions (positive and negative word counts in the *Wall Street Journal*, for instance) tend to positively correlate. Inspection of Figure 2.2 in TMP suggests that

several of seven tested time series might be correlated (in particular, look at OF, CALM, ALERT, SURE and HAPPY before and after the 2008 election and Thanksgiving).[21] Neither of these facts necessarily imply that the underlying mood time series in TMP will have a positive correlation, but both of them could be consistent with such a correlation. If we were willing to suppose that all of our moods are independent or have positive correlation, then the hypothesis tests presented in Table 2.2 of TMP would have either zero or positive association in the sense that the significance of one p-value would either decrease or have no impact on the other p-values in that table. Making either assumption of independence between our hypothesis tests or "no negative association" between the hypothesis tests of our moods allows us to use the Hochberg and Hommel tests. The Hochberg test is a sequential rejection step-up test and is defined as follows:

$$\begin{aligned}
p_M &= \alpha_M \\
p_{M-1} &= \min\left\{p_M, 2\alpha_{M-1}\right\} \\
p_{M-2} &= \min\left\{p_{M-1}, 3\alpha_{M-2}\right\} \\
&\vdots \\
p_{M-k+1} &= \min\left\{p_{M-k+2}, k\alpha_{M-k+1}\right\} \\
p_{M-k} &= \min\left\{p_{M-k+1}, (k+1)\,\alpha_{M-k}\right\} \\
p_{M-k-1} &= \min\left\{p_{M-k}, (k+2)\,\alpha_{M-k-1}\right\} \\
&\vdots \\
p_2 &= \min\left\{p_3, (M-1)\,\alpha_2\right\} \\
p_1 &= \min\left\{p_2, M\alpha_1\right\}
\end{aligned}$$

where p_k are computed from top to bottom. Hommel's method is more powerful than Hochberg's

[21] Table 2.1 of TMP reports that all of the moods generated by GPOMS have positive partial correlation with the mood generated by OF in a multiple regression framework. In fact, if each individual GPOMS mood had a large enough Pearson correlation with the OF mood we could establish that all six moods were positively correlated; unfortunately we only have access to *partial* correlations from Table 2.1 and several of those appear to be statistically insignificant from zero.

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	0.999	0.999	0.999	0.999	0.999	0.999	0.999
2 Days	0.999	0.637	0.999	0.999	0.999	0.999	0.999
3 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
4 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
5 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
6 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
7 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Table 2.6: This table presents Hochberg corrected p-values for TMP's linear models. Even with the additional information provided by the order of the test statistics and the assumption of independence or positive association, none are significant.

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	0.999	0.999	0.999	0.999	0.999	0.999	0.999
2 Days	0.999	0.611	0.999	0.999	0.999	0.999	0.999
3 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
4 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
5 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
6 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999
7 Days	0.999	0.999	0.999	0.999	0.999	0.999	0.999

Table 2.7: This table presents Hommel corrected p-values for TMP's linear models. Even with the additional information provided by the order of the test statistics and the assumption of independence or positive association, none are significant.

but also more computationally complex and we do not present the method here. Both tests are more powerful than both the aforementioned Bonferroni and Holm tests. Because we cannot rule out positive correlation or independence between the mood time series, we present the tests in order of power in Tables 2.6 and 2.7.

Finally, if we assume independence between the mood time series, which is extremely unlikely to be the case given the evidence discussed in the previous paragraph and in TMP, we have access

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	0.978	0.999	1.000	1.000	0.999	1.000	0.999
2 Days	0.999	0.473	1.000	0.999	0.999	1.000	1.000
3 Days	0.999	0.656	1.000	0.999	1.000	1.000	1.000
4 Days	0.999	0.761	1.000	1.000	0.999	1.000	1.000
5 Days	0.999	0.815	1.000	1.000	1.000	1.000	0.999
6 Days	0.999	0.951	1.000	0.999	1.000	1.000	0.975
7 Days	1.000	0.998	1.000	0.995	1.000	1.000	0.998

Table 2.8: This table presents Sidak sequential rejection step-down corrected p-values for TMP's linear models. Even with the additional information provided by the order of the test statistics, none are significant at any standard level.

to the Sidak sequential rejection step-down adjustment.[22] It is defined as follows:

$$\begin{aligned}
p_1 &= 1-\left(1-\alpha_1\right)^M \\
p_2 &= \max\left\{p_1, 1-\left(1-\alpha_2\right)^{M-1}\right\} \\
p_3 &= \max\left\{p_2, 1-\left(1-\alpha_3\right)^{M-2}\right\} \\
&\vdots \\
p_{k-1} &= \max\left\{p_{k-2}, 1-\left(1-\alpha_{k-1}\right)^{M-k}\right\} \\
p_k &= \max\left\{p_{k-1}, 1-\left(1-\alpha_k\right)^{M+1-k}\right\} \\
&\vdots \\
p_{M-1} &= \max\left\{p_{M-2}, 1-\left(1-\alpha_{M-1}\right)^{2}\right\} \\
p_M &= \max\left\{p_{M-1}, \alpha_M\right\}
\end{aligned}$$

where, as usual, p_k is computed as ordered above. Our survey of the literature suggests that, for TMP, this is the strongest ex-post FWER multiple comparison adjustment we have access to and we present the Sidak sequential rejection step down results in Table 2.8.

None of the results for any of our FWER results are significant at any standard level. More recent work in this area, especially in financial econometrics, assumes access to underlying data or

[22] v8doc.sas.com/sashtml/stat/chap43/sect14.htm, accessed 2015/03/15. In fact, the Sidak adjustment (which results from assuming independence) can be combined with other tests in this section in various ways. None of the FWER-restricting combinations we have tried resulted in any of the p-values being close to significant at any standard level.

technical knowledge about the stochastic process generating the data that we do not have access to Romano and Wolf [2005], White [2000]. Given the results of our Sidak step down procedure and the sheer number of hypotheses being tested, more sophisticated methods are unlikely to reject the null as long as we restrict FWER. These tests indicate it is extremely likely that there is at least one false positive in TMP's results.

2.3.2 Controlling False Discovery Rate

> Evaluating a trading strategy is not a mission to Mars. Being wrong could cost you your job and money will be lost - but it is unlikely to be a matter of life and death. However, reasonable people may disagree with this view.
>
> ---
>
> *Evaluating Trading Strategies*
>
> HARVEY AND LIU (2014)

The second approach to multiple comparison bias instead aims to control the false discovery rate (FDR hereafter). Many presentations[23] of the properties of FDR versus the FWER use a variation on Table 2.9.

	Null True	Alternative True	Total
Not called significant	TN	FN	$M - R$
Called significant	FP	TP	R
Total	m	$M - m$	M

Table 2.9: For cases in which all values of the table are known with certainty, this table is called a confusion matrix. In statistics this table is commonly called a contingency table. Our setup is the same as in §2.3.1 with additional notation: let R be the number of null hypotheses rejected by our tests (i.e. significant findings or "discoveries"). TN is the number of true negatives (a negative is a failure to reject the null). FN is the number of false negatives. FP is the number of false positives, i.e. Type I errors. TP is the number of true positives.

[23] http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture10.pdf, accessed 2015/03/29.

If the values of Table 2.9 are known with certainty our FDR is:

$$FDR = \frac{FP}{FP + TP} = \frac{FP}{R} \tag{2.7}$$

If the values of Table 2.9 are unknown, we can define the FDR using only the statistical properties of our null hypothesis Storey [2001]:

$$FDR = \mathbb{E}\left\{\frac{FP}{R} | R > 0\right\} \mathbb{P}\{R > 0\} \tag{2.8}$$

where $\mathbb{P}\{R > 0\}$ can be read as the "the probability of at least one discovery". Using our table, we can recast our FWER tests as fixing the probability:

$$\mathbb{P}\{FP \geq 1\} \leq \alpha$$

so that a multiple comparison adjustment over FDR controls fixes the expected fraction of false positives at level α while a multiple comparison adjustment over FWER fixes the probability of a single false positive at level α. We can see that, since BMZ's findings suggest a family of strategies based around their CALM measure, our FWER control methods may be too stringent to evaluate the commercial or academic value of the findings in Table 2.2 in TMP. Because FDR tests are more powerful than FWER tests, and the p-values generated by multiple comparison adjustments that control FDR correspond with statistical tests of the fraction of false positives over all positives, FDR methods are both preferred in financial applications in general and are more appropriate for assessing TMP in particular.

We start with the most common FDR test. No special assumptions are required for the Benjamini, Hochberg and Yekuteli test (BHY hereafter). It is a step-up method, meant to be evaluated from $k = M$ to $k = 1$ according to the following formula:

$$p_k = \frac{M\left(\sum_{i=1}^{M} i^{-1}\right)}{k}\alpha_k \tag{2.9}$$

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
4 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7 Days	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2.10: This table presents the BHY sequential rejection step-up corrected p-values for TMP's linear models. Given the form we have presented it in, the BHY test is valid for arbitrary underlying time series.

Lag	Opinion Finder	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	0.595	0.950	0.999	0.959	0.735	0.999	0.950
2 Days	0.950	0.441	0.999	0.999	0.950	0.999	0.999
3 Days	0.950	0.441	0.999	0.950	0.950	0.999	0.999
4 Days	0.890	0.441	0.999	0.950	0.999	0.999	0.999
5 Days	0.999	0.441	0.999	0.999	0.999	0.999	0.770
6 Days	0.950	0.595	0.999	0.999	0.999	0.999	0.595
7 Days	0.999	0.769	0.999	0.950	0.999	0.999	0.769

Table 2.11: This table presents the Benjamini & Hochberg sequential rejection method p-values for TMP's linear time series tests. None of our p-values are significant at any standard level.

for the maximal k such that $\alpha_k < \alpha$, we declare $\left\{\alpha_1, \ldots, \alpha_k\right\}$ to be significant. We present the adjusted p-values in Table 2.10.

Under the approach that Harvey and Liu [2014] deem "most appropriate for evaluating trading strategies", the BHY correction, none of the p-values TMP present are less than 1. Nonetheless, we use the reasoning in §2.3.1 to justify the use of stronger tests with positive association or independence assumptions on the underlying mood time series. In Table 2.11, we present the Benjamini and Hochberg corrected p-values, which are valid under the assumption of independence or positive association.

Surprisingly, the initial results of Kuleshov [2011], Pav [2012a] are robust to generically more powerful multiple comparison tests like the Holm adjustment, tests derived from stronger assumptions on the underlying time series (Hochberg adjustment, Hommel adjustment and Sidak adjustment), and techniques that adopt the much looser standard of limiting the FDR rather than the

FWER (Benjamini and Hochberg adjustment and BHY). Under no multiple comparison adjustments we have found, including several left out of this section for brevity but available in the mutoss R package and present in our GitHub repository, are any of the p-values reported by BMZ close to significant at any standard level. All of these tests suggest that $\frac{FP}{R} = 1$.[24] We have not tested the universe of FDR adjustments and there may exist an adjustment in the literature under which the p-values in TMP are significant, but we are doubtful that the assumptions required by such an FDR adjustment would be met by the collective mood time series of TMP. We conclude that BMZ present no statistical evidence in their linear time series analysis that Twitter mood predicts the stock market.

2.3.3 Multiple Hypothesis Testing for Non-linear Machine Learning Algorithms

When evaluating their SOFNN classifier, Bollen et al. [2011a] treat the random walk hypothesis (RWH) as the null hypothesis. This hypothesis, discretized, treats every day's market movement as the outcome of a binomial tree in which the probability of an up movement is 0.5. BMZ (incorrectly) assert that the EMH claims that the probability of forecasting whether the market will be up or down given all available information is 0.5.[25] While this is not correct under the EMH, this is correct under the RWH and it provides one with a natural benchmark to assess the statistical significance of an algorithm's ability to forecast the up-down pattern of the market. We can formalize this hypothesis test using the binomial distribution, which we present below:

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i} \tag{2.10}$$

[24] Some tests do not return adjusted p-values but simply their estimate of FDR.

[25] BMZ's conflation of the EMH and RWH is incorrect; in general, the RWH is not equivalent to the EMH. EMH does not imply the RWH as LeRoy [1973], Lucas [1978] construct economies in which the EMH is true but the RWH is false. See LeRoy [1989] for additional information. Market price patterns are typically better fit by a random walk with trend than the RWH. Also note that in a test we do not present here, we can use the up-down pattern of DJIA values to reject the hypothesis that the daily DJIA difference follows a random walk without trend with 1% statistical significance.

where in the above:

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} \tag{2.11}$$

equals the number of different ways groups of i objects can be chosen from a set of n objects and i can be thought of as the number of failed predictions if we fail to predict correctly with probability p. We can ask the question: if the RWH is correct, with what probability would our algorithm forecast thirteen out of fifteen days or more? In this case, our null is that $p = 0.5$, $n = 15$ and $i = 2$. A hypothesis test of the SOFNN evaluates:

$$\mathbb{P}\left\{2 \text{ or fewer failures}\right\} = 0.003 = \sum_{j=0}^{i=2} \binom{15}{j} 0.5^j \left(1-p\right)^{15-j} \leq \alpha$$

where we reject the null at significance level α if the inequality above is true. If Bollen et al. [2011a] only tested one time series with their SOFNN, it would be significant at the 1% level. However, they test eight such time series. Assuming independence, we can construct the following FWER-restricting null:

$$0.029 = 1 - \left(1 - \mathbb{P}\left\{2 \text{ or fewer failures}\right\}\right)^8 \leq \alpha$$

In this case, the non-linear hypothesis test result is still significant at the 5% level.[26]

We note that over the time period tested in TMP the market suffered 8 down moves. In a machine learning context, a more typical evaluation of the SOFNN would be against a baseline that always predicted "down". This is equivalent to setting a different null hypothesis in which the market suffers a down move with probability $\frac{8}{15}$. In this case, our probability of a failure to correctly predict the market's up-down outcome on a given day drops to 0.467. Our hypothesis test, because of the small number of days, is sensitive to small changes in the baseline expectation of the probability:

$$0.059 = 1 - \left(1 - \mathbb{P}\left\{2 \text{ or fewer failures}\right\}\right)^8 \leq \alpha \text{ if } p = 0.467$$

[26] Pav [2012a] makes, but does not explain, this calculation.

and the result is no longer significant at the 5% level.

On the other other hand, one could argue that a more economically meaningful test is against the long-run daily up-down probabilities. Using the DJIA time series from 2015/02/13 to 1999/03/24 we estimate that the long run daily probability of an up move is 0.52.[27] Under these circumstances, BMZ's results look more impressive, because our learned baseline would mislead us in December 2008. Our FWER adjusted significance is:

$$0.018 = 1 - (1 - \mathbb{P}\{2 \text{ or fewer failures}\})^8 \leq \alpha \text{ if } p = 0.52$$

which is significant at the 5% level. As we can see from these examples the hypothesis tests of the SOFNN are sensitive to parameters specified in the null hypothesis.[28] Nonetheless, before taking into account the bias induced by the global normalization procedure (BMZ appear to normalize the entire time series, including the test set, for their SOFNN evaluation), it is certainly plausible that TMP presents statistically significant evidence that the SOFNN algorithm, with BMZ's CALM time series as an input, predicts the stock market in its non-linear hypothesis section.

2.3.4 Conclusions

If the effect discovered in TMP was real, then by the central limit theorem, one would obtain p-values low enough in their "CALM" time series to overcome the hurdles set by multiple comparison adjustments simply by sufficiently extending the length of time series. Since the authors of Bollen et al. [2011a] had access to the entire Twitter feed across time, we strongly suspect the reason that they did not do this is because doing so causes the Twitter mood effect to vanish. The authors of TMP have not presented evidence that any of the p-values in TMP's linear time series section are significant, when one adjusts for the multiple comparison bias inherent in testing a large number of hypotheses. Fortunately, Bollen et al. [2011a] provide enough evidence, in the form of their

[27] This is consistent with RWH with trend.

[28] If BMZ had assigned more days to their test set, the confidence bounds would be tighter and the tests would more sharply discriminate between the RWH and the RWH with trend.

reported p-values, to conclude that one *cannot* reject that all Twitter mood data is independent of future DJIA movements.

2.4 A Second Potential Error

Bommarito[29], was first to point out the local normalization in TMP takes in information from the future and to suspect that this normalization was used in TMP's statistical tests. In a series of emails[30], Bommarito asked Bollen:

> Are the z-scores [i.e. locally normalized mood time series] used in any assessment of prediction, or are these z-scores only used for plotting the signals?

Bollen responded:

> The Z-scores are not used in our assessment of prediction accuracy. Please have a look at the equations on page 6 (and related discussion) vs. equation (2.1).

The equations on page 6 relate to the SOFNN, but not the linear time series results.[31] In this section, we show that there exist a priori reasonable joint distributions of equity prices and mood time series such that the local normalization procedure can reverse Granger causality. More precisely, given an underlying time series in which X_t does not Granger cause D_t but is Granger caused by X_t, locally normalized $\tilde{X}_t$ can Granger cause D_t. Thus, if the locally normalized time series appears to Granger cause the stock market, that does not imply that the underlying CALM time series Granger causes the stock market. In two examples below, we show how TMP's p-values

[29] http://etf-central.com/2010/10/21/update-j-bollen-h-mao-x-j-zeng-twitter-mood-predicts-the-stock-market, accessed 2015/02/14.

[30] Obtained from personal correspondence with Professor Bommarito.

[31] It is possible Bollen meant to refer to the fact that the equations use the term X_t rather than $\mathbb{Z}_{X_t}$ but we note even in the final, published version of TMP, the authors only specifically disavow using the locally normalized time series for Section 2.5, the SOFNN section of their paper and not Section 2.4, the linear statistics section. Furthermore, Bollen did not explicitly answer Professor Bommarito's question of whether the local normalization is only used for visualization purposes in TMP. Throughout §2.4, we *assume* that $\mathbb{Z}_{X_t}$ rather than X_t was used as the explanatory variable in the linear time series section of TMP. We wish to emphasize that, while this could explain the results in TMP, we are uncertain that this is actually the case. We reached out to BMZ several times seeking clarification; they have not responded to any emails regarding TMP.

could result entirely from the use of this inappropriate normalization and from the DJIA Granger causing or being contemporaneously determined with Twitter mood.

2.4.1 Monte Carlo Simulation 1: Fatal Inference

It turns out to be trivial to construct an example in which collective mood is contemporaneously caused with the stock market, but our locally normalized time series, which takes in k days of information from the future, appears to Granger cause stock prices. We present such an example using Monte Carlo simulation below. In all examples, we use 1000 iterations and a time series of 174 periods, which means 173 differenced periods. This corresponds with what the authors claim to test in TMP.

In our first and simplest example, we draw day t's DJIA difference D_t from an independent and identically distributed (i.i.d. hereafter) standard normal distribution each period. Then, our mood is given by:

$$X_t = \beta D_t + \epsilon_t \tag{2.12}$$

with β strictly positive so that higher differences in the DJIA cause one to feel calmer the same day.[32] Even if stock market differences increase calmness, if calmness reacts rapidly, a functional form like (12) will adequately capture the mood-stock market relationship. As usual, $\mathbb{E}\left[\epsilon_t\right] = 0$. For now we will drop ϵ_t but our simulations suggest that if ϵ_t is normal or t with 4 degrees of freedom, there is no qualitative difference to our results. Applying our local normalization to the

[32]This condition is also necessary to avoid imaginary solutions in calculations below.

underlying mood time series gives us:

$$
\begin{aligned}
\mathbb{Z}_{X_t} &= \frac{X_t - \frac{1}{2k+1}\sum_{i=t-k}^{t+k} X_i}{\sqrt{\frac{1}{2k}\left(\sum_{i=t-k}^{t+k}\left(X_i - \frac{1}{2k+1}\left(\sum_{i=t-k}^{t+k} X_i\right)^2\right)\right)}} \\
&= \frac{\beta D_t - \frac{1}{2k+1}\sum_{i=t-k}^{t+k}\left(\beta D_i\right)}{\sqrt{\frac{1}{2k}\left(\sum_{i=t-k}^{t+k}\left(\beta D_i - \frac{1}{2k+1}\left(\sum_{i=t-k}^{t+k}\beta D_i\right)\right)^2\right)}} \text{ using (12)} \\
&= \frac{\left(\frac{2k}{2k+1}\right) D_t - \left(\sum_{i=t-k}^{t-1} D_i\right) - \underbrace{\sum_{i=t+1}^{t+k} D_i}_{\text{from the future}}}{\sqrt{\frac{1}{2k}\left(\sum_{i=t-k}^{t+k}\left(D_i - \frac{1}{2k+1}\left(\sum_{i=t-k}^{t+k} D_i\right)\right)^2\right)}} \\
&= \frac{\left[\left(\frac{2k}{2k+1}\right) D_t - \left(\sum_{i=t-k}^{t-1} D_i\right) - \underbrace{\sum_{i=t+1}^{t+k} D_i}_{\text{from the future}}\right](2k+1)}{\sqrt{2k\left(\sum_{i=t-k}^{t+k} D_i^2\right)}} \\
&= \left(\frac{2k+1}{\sqrt{2k}}\right)\left(\frac{\left[\left(\frac{2k}{2k+1}\right) D_t - \left(\sum_{i=t-k}^{t-1} D_i\right) - \underbrace{\sum_{i=t+1}^{t+k} D_i}_{\text{from the future}}\right]}{\sqrt{\sum_{i=t-k}^{t} D_i^2 + \underbrace{\sum_{i=t+1}^{t+k} D_i^2}_{\text{from the future}}}}\right)
\end{aligned}
$$

For "small" values of D_i, however, supposing that $k = 1$ and using the MacLaurin expansion, we can see that the negative linear effect dominates and so increases in D_{t+1} decrease the value of $\mathbb{Z}_{X_t}$. While the effect on our normalized time series is evident for either D_i large or small,

the overall effect of future information is unclear. Under the assumption of a linear coincident relationship between moods and markets, a change in our future information has a non-linear effect on our locally normalized mood time series. To see if our linear ARDL models can detect this, we run a Monte Carlo simulation in which: $\beta \sim Poisson\{10\}$, $k = 3$ and $\epsilon_t \sim N(0, 1)$. We generate D_t and X_t 1000 times according to the above assumptions. Then, we normalize our mood time series X_t according to the formula specified in Bollen et al. [2011a]. We use the Granger causality test on the first and second lags of mood as in TMP. We should find that our p-values are uniformly distributed in [0, 1] for the underlying mood time series Granger causality tests. However, if our local normalization takes in information from the future in a way detectable by our ARDL regression models, then our p-values for Granger causality tests on the normalized mood time series should not be uniformly distributed.

In the simulation, shown in Figure 2.3, our locally normalized time series have a significant first lag with the DJIA differences; the average significance is less than 0.01% and nearly all of the p-values are significant at 5%. Less than 1% of the time we expect to achieve results insignificant at the 5% level despite the fact that the underlying time series has no ability to predict the DJIA. We can also see that our statistical tests correctly find that our underlying mood time series, which does not cause the differenced DJIA series by construction, does not Granger cause market price increases.

We can see that while this simulation provides an example in which our local normalization is fatal for inference on the underlying time series, the relationship we have assumed between markets and mood does not, on average, lead to Granger causality tests replicating TMP's p-value pattern. On average, our tests find that both the first and second lag are significant, rather than just the second lag. Is it possible for the underlying relationship between moods and stocks to be such that, when locally normalized, the information content of mood appears to "skip a day" as it does in TMP's 2nd plot?

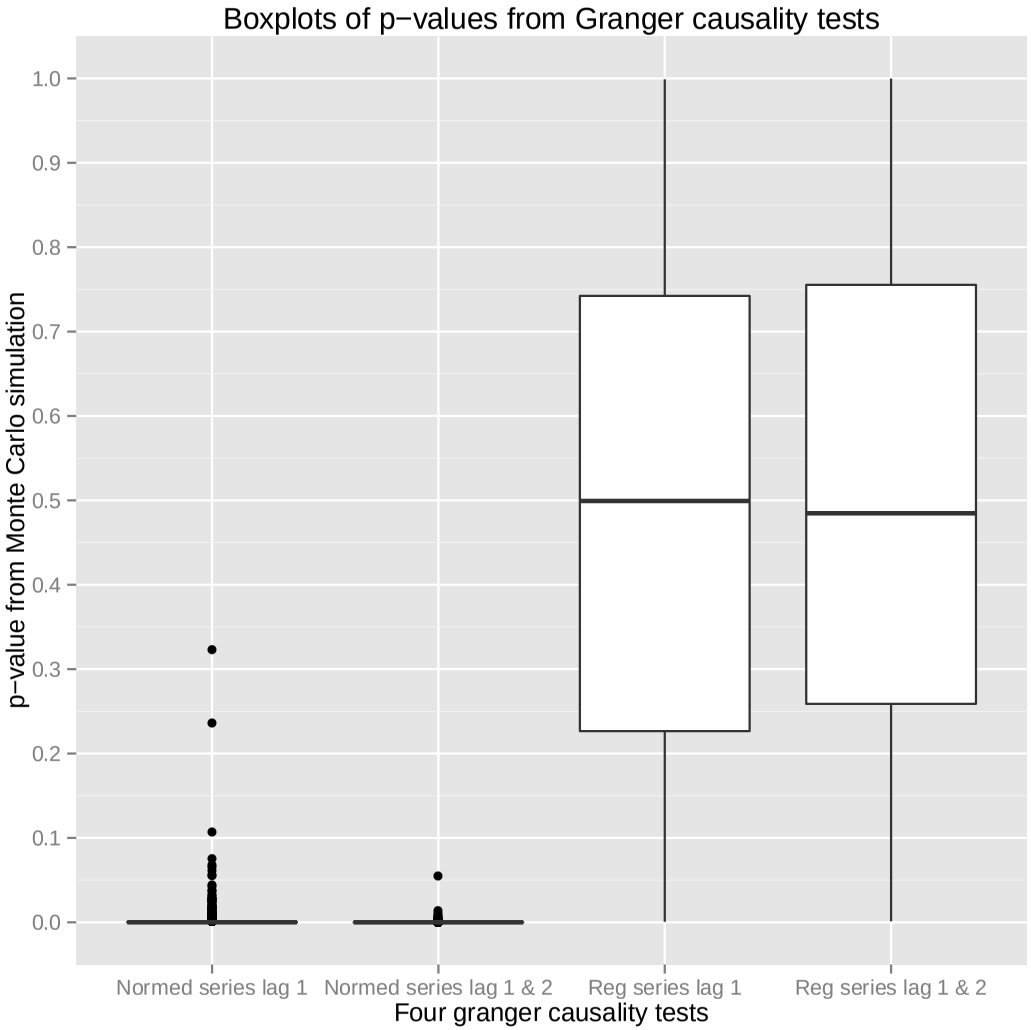

Figure 2.2: This figure contains the box plot of p-values derived from 1000 Granger causality tests on the first and first two lags of locally normalized mood times (the two leftmost plots) and the underlying mood time series (the two rightmost plots) simulated according to (12). Notice that while our Granger causality tests correctly discover that the p-values of the underlying mood time series are uniformly distributed, our tests find that the locally normalized mood time series Granger cause stock market increases. While this shows that the local normalization procedure in TMP can lead to misidentification of Granger causality, because our Granger causality tests finds only the first lag significant and TMP does not report a significant first lag, collective mood is unlikely to have come from a stochastic process resembling (12).

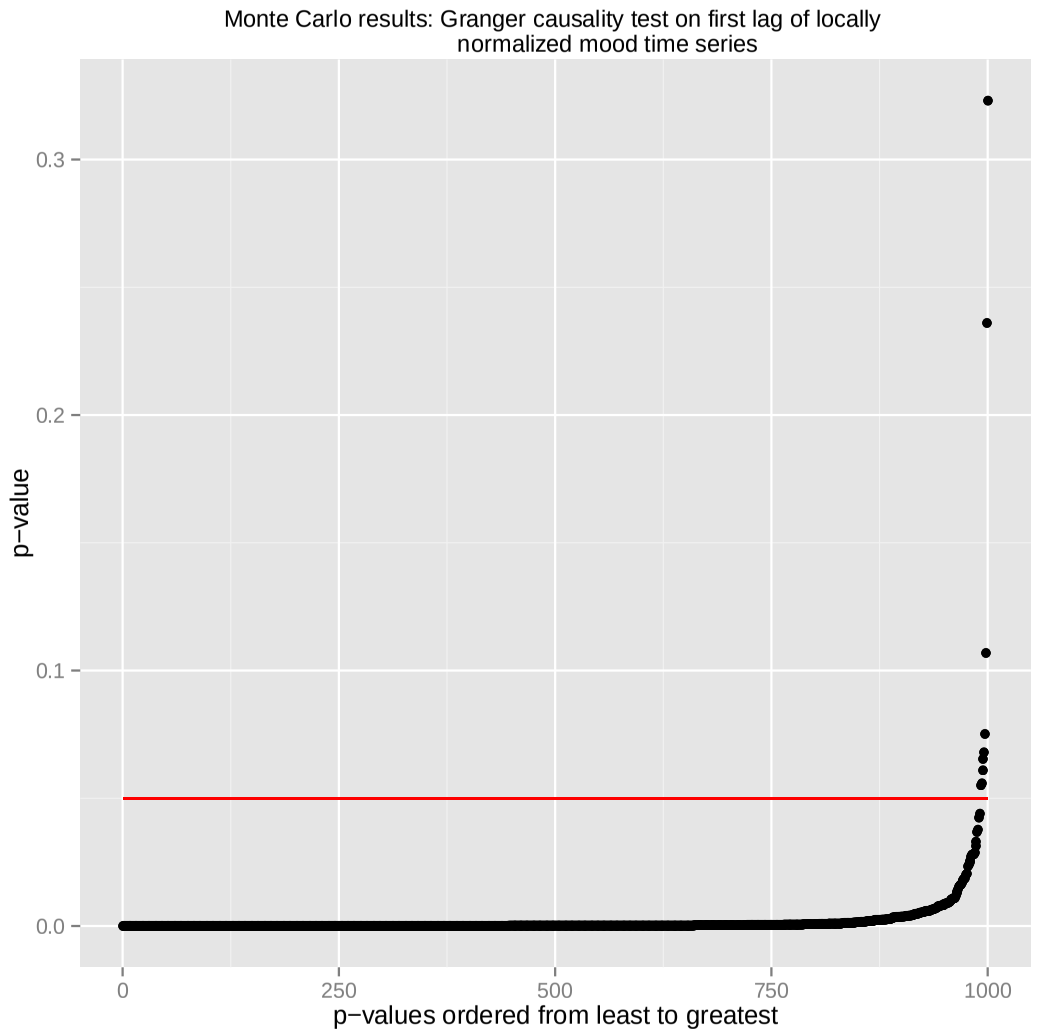

Figure 2.3: This figure contains the p-values of 1000 Granger causality tests on the first lag of our locally normalized mood time series simulated from (12) and ordered from least to greatest. All of the points below the red line are statistically significant at the 5% level. The vast majority of Granger tests report that locally normalized mood Granger causes the stock market even though the underlying time series, by construction, does not.

2.4.2 Monte Carlo Simulation 2: Hedonic Treadmill

We turn to theory to help narrow our search for a functional relationship between equity market behavior and collective mood that could explain the p-values presented in TMP. Kahneman [2000] points out that most people's individual mood is on a "hedonic treadmill." This means that one's subjective mood state is determined not by our absolute level of wealth or even by recent changes in wealth but by deviations above or below our expectations of wealth. We assume that the composed-anxious axis CALM measurements function similarly to self-reported measures of individual subjective well-being referenced by Kahneman. In this case, since the stock market has an expected change of near zero each day, Gilbert and Karahalios [2010] suggest that significant deviations above or below the expected daily change should affect the stock market. Based on their work, we conjecture the following functional form for the relationship between the stock market and mood:

$$X_t = f\left(\nabla D_t\right) + \epsilon_t$$

where f is increasing in its argument and ∇, as usual, is our difference operator. Recall that: $\nabla D_t = D_t - D_{t-1} = DJIA_t - 2DJIA_{t-1} + DJIA_{t-2}$, or a measure of what Gilbert and Karahalios [2010] describe as stock market "acceleration". In this essay we assume that f is analytic so that we can always linearize it for small changes in ∇D_t:

$$X_t = \beta \nabla D_t + \epsilon_t \tag{2.13}$$

where $\beta \in \mathbb{R}^+$. For simplicity we drop our ϵ_t for the rest of our calculations. We have normalized time series:

$$\begin{aligned}
\mathbb{Z}_{X_t} &= \frac{X_t - \frac{1}{2k+1}\left(\sum_{t=-k}^{k} X_t\right)}{\sigma\left(X_{-k}, X_{-k+1}, \ldots, X_{k-1}, X_k\right)} \\
&= \frac{\beta\nabla D_t - \frac{1}{2k+1}\left(\sum_{t=-k}^{k} \beta\nabla D_t\right)}{\sqrt{\frac{1}{2k}\left(\sum_{i=t-k}^{t+k}\left(\beta\nabla D_i - \frac{1}{2k+1}\left(\sum_{i=t-k}^{t+k}\beta\nabla D_t\right)\right)^2\right)}} \text{ using (3)} \\
&= \frac{\nabla D_t - \frac{1}{2k+1}\left(D_{t+k} - D_{t-k-1}\right)}{\sqrt{\frac{1}{2k}\left(\sum_{i=t-k}^{t+k}\left(\nabla D_i - \frac{\left(D_{t+k} - D_{t-k-1}\right)}{2k+1}\right)^2\right)}} \text{ using telescoping sum} \\
&= \frac{\nabla D_t - \frac{1}{2k+1}\left(D_{t+k} - D_{t-k-1}\right)}{\sqrt{\frac{1}{2k}\left(\sum_{i=t-k}^{t+k}\left(D_i - D_{i-1} - \frac{\left(D_{t+k} - D_{t-k-1}\right)}{2k+1}\right)^2\right)}}
\end{aligned}$$

We can see that, under this specification, information from *only* k days in the future enters the normalization in the numerator, which will systematically affect $\mathbb{Z}_{X_t}$ for small values of D_i. If BMZ chose $k = 2$, then we would obtain:

$$
\begin{aligned}
\mathbb{Z}_{X_t} &= \left(2 \frac{D_t - D_{t-1} - \frac{1}{5}\left(D_{t+2} - D_{t-3}\right)}{\sqrt{\left(\sum_{i=t-2}^{t+2}\left(D_i - D_{i-1} - \frac{\left(D_{t+k} - D_{t-k-1}\right)}{2k+1}\right)^2\right)}} \right) \\
&= 2 \frac{D_t - D_{t-1} - \frac{1}{5}\left(D_{t+2} - D_{t-3}\right)}{\sqrt{\left(\sum_{i=t-2}^{t+2}\left(D_i^2 - D_i D_{i-1} + D_{i-1}^2 + \frac{\left(D_{t+2} - D_{t-3}\right)}{5} D_{i-1} - D_i \frac{\left(D_{t+2} - D_{t-3}\right)}{5} + \left(\frac{D_{t+2} - D_{t-3}}{5}\right)^2\right)\right)}} \\
&= 2 \frac{D_t - D_{t-1} - \frac{1}{5}\left(D_{t+2} - D_{t-3}\right)}{\sqrt{\left(\sum_{i=t-2}^{t+2}\left(D_i^2 - D_i D_{i-1} + D_{i-1}^2\right)\right)}} \\
&\qquad \text{linear effects from the future dominate other future terms} \\
&= 2 \frac{D_t - D_{t-1} - \overbrace{\frac{1}{5}\left(D_{t+2} - D_{t-3}\right)}}{\sqrt{\left(D_{t+2}^2 + 2\left(D_{t+1}^2 + D_t^2 + D_{t-1}^2 + D_{t-2}^2\right) + D_{t-1}^2 - \left(\sum_{i=t-2}^{t+2} D_i D_{i-1}\right)\right)}}
\end{aligned}
\tag{2.14}
$$

Our analysis of the D_{t+2} term is the same as our future terms in the previous section, small increases in D_{t+2} will increase $\mathbb{Z}_{X_t}$ (fixing all terms $D_{i \neq t+1} \approx 0$). On the other hand, for small changes D_{t+1}, fixing all other terms $D_{i \neq t+1} \approx 0$, our Taylor expansion gives no linear term; thus the magnitude of linear changes in D_{t+2} on $\mathbb{Z}_{X_t}$ is larger in this case than for D_{t+1}. This means it is possible, for an appropriate underlying distribution of D_t and specification of f for information to appear to "skip a day" in our Granger causality analysis.

This inspires the Monte Carlo simulation in which, letting $\beta \sim Poisson\{10\}$,[33] $k = 2$ and $\epsilon_t \sim N(0, 1)$, we generate X_t 1000 times according to the above assumptions. As usual, D_t is from an i.i.d. normal. Then, we normalize our mood time series X_t as in TMP. We use the Granger causality test on the first and second lags of mood as in TMP. Again, we should find that our p-values are uniformly distributed in $[0, 1]$ for the underlying mood time series Granger causality tests. However, if our local normalization takes in information from the future in a way detectable

[33]The calculations in (14) suggest that β does not matter. Without noise, it does not. However we must use (13) without dropping our ϵ_t term to avoid multicollinearity in our post-simulation Granger causality analysis. In this case, we cannot factor β out of the denominator. Simulations suggest that larger β terms make the second lag easier to detect.

by our ARDL regression models, then our p-values for Granger causality tests on the normalized mood time series should not be uniformly distributed. Our results are shown in Figure 2.4.

As our calculations in (14) suggest, the locally normalized time series can detect Granger causality for the first and second lag collectively, but not for the first lag by itself. In this case, 64% of the time the Granger causality test indicates our first and second lag are collectively significant at the 5% level, compared with 5% of the time for the first lag. More than half of the time our normalized mood time series will have the properties ascribed to it by BMZ without the underlying time series predicting the stock market. Thus, one explanation for the linear test results in TMP, in which information in Tweets appears to skip a day, could be that BMZ used the local normalization and that their CALM time series is a function of stock market "acceleration".lying mood time series Granger causality tests. However, if our local normalization takes in information from the future in a way detectable by our ARDL regression models, then our p-values for Granger causality tests on the normalized mood time series should not be uniformly distributed. Our results are shown in Figure 2.5.

2.5 Lots of Computation but Not Much Science: Future Work

> If you torture the data long enough, it will confess.
>
> ATTRIBUTED TO NOBEL LAUREATE RONALD COASE

BMZ conclude their essay:

> ...future research may need to take into account social and cognitive effects in which individual agents are endowed with the ability to learn from past experiences and can adjust their behavior accordingly. The investigation of such phenomena in online social networking environments is part of an exciting new research front commonly referred to as "computational social science".

Bollen et al. are researchers in fields in which the number of labeled and unlabeled data points can range into the trillions. In this context, a mood model that estimates 964 co-occurrence weights between n-grams and the POMS-Bi before scoring a corpus is sparse. Unfortunately, empirical

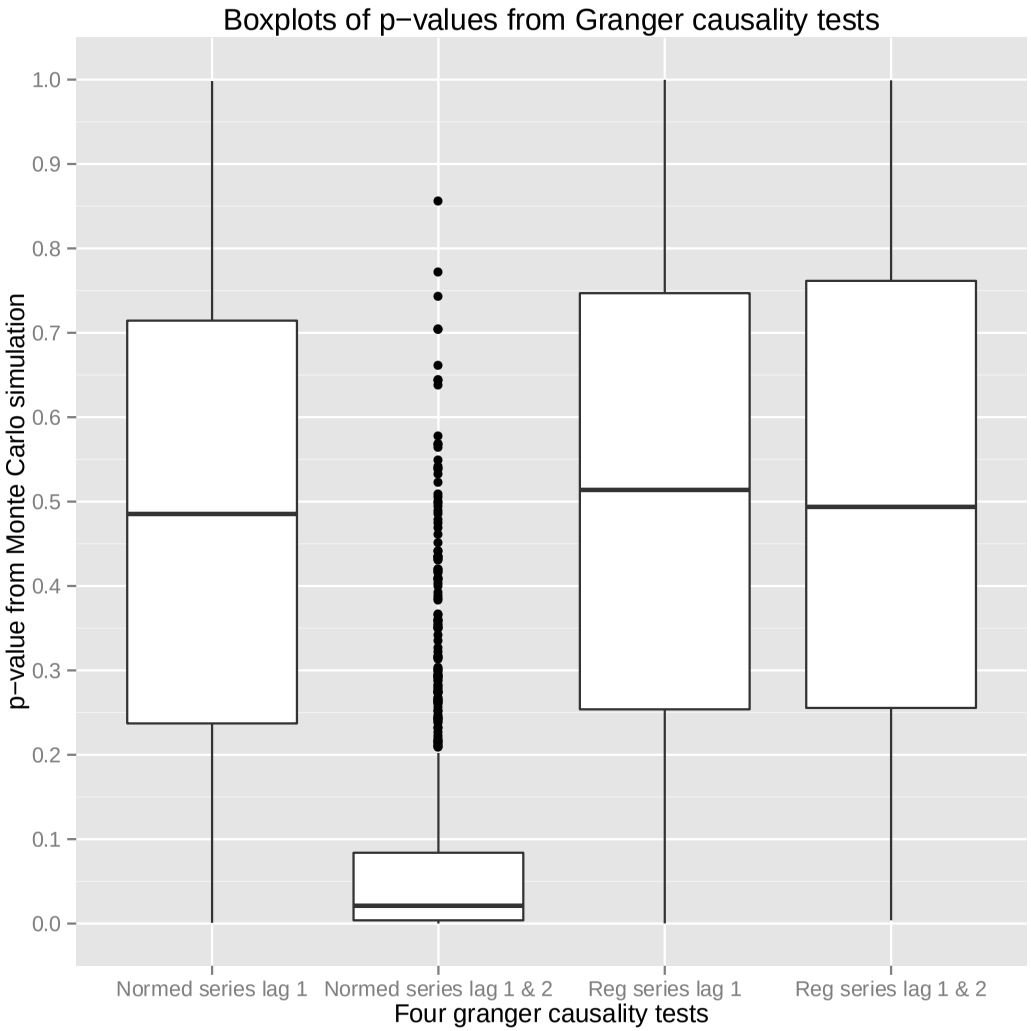

Figure 2.4: This figure contains the box plot of p-values derived from 1000 Granger causality tests on the first and first two lags of locally normalized mood times (the two leftmost plots) and the underlying mood time series (the two rightmost plots) simulated according to (13). Notice that while our Granger causality tests correctly discover that the p-values of the underlying mood time series are uniformly distributed, our tests find that the second lag of the locally normalized mood time series Granger causes stock market increases. The average p-value pattern generated in this figure matches the p-value pattern reported in TMP in the sense that the first lag is insignificant but the second lag is significant, suggesting that Twitter mood information somehow skips a day before being reflected in stock prices.

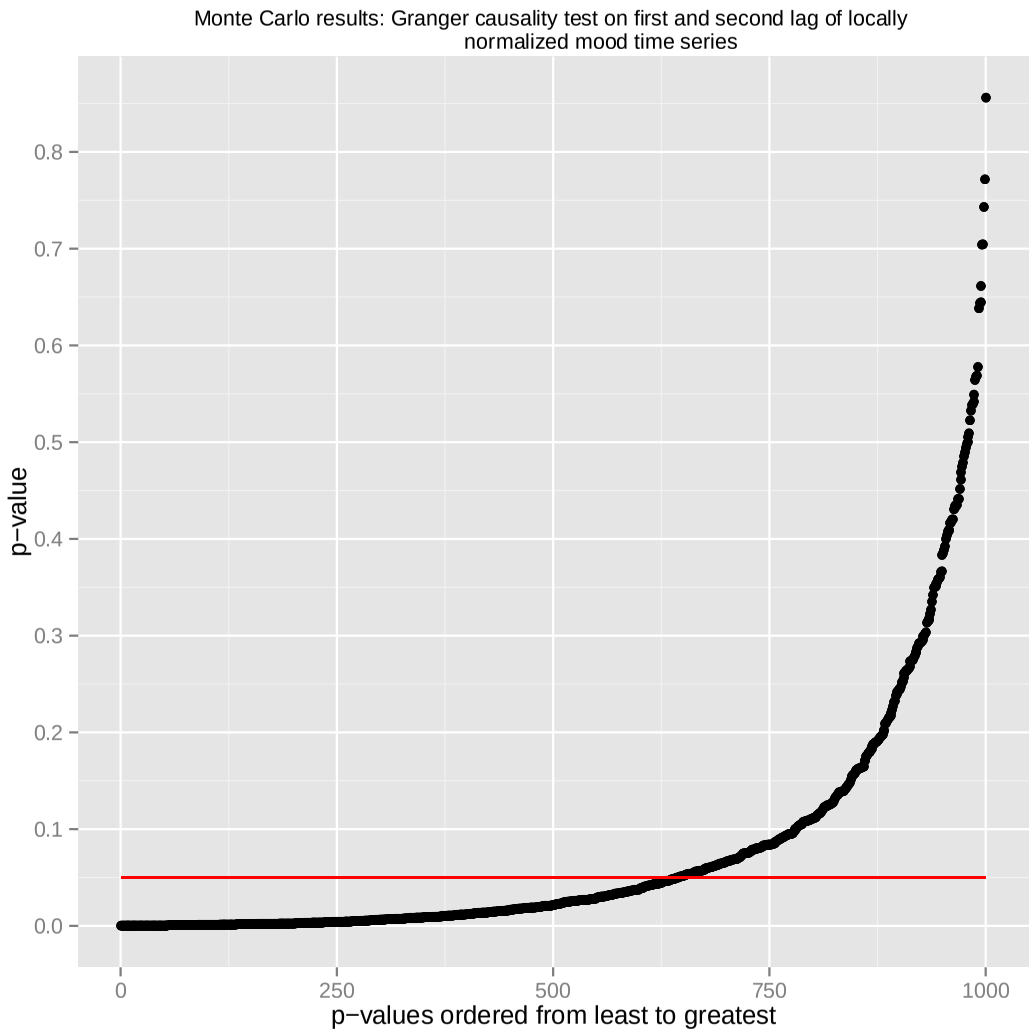

Figure 2.5: This figure shows the p-values from 1000 Granger causality tests on the first two lags of the locally normalized collective mood time series. Points below the red line are statistically significant at the 5% level. In other words, even when collective mood has no predictive power for stock prices, if the stochastic process generating mood is given by (13) and collective mood is normalized according to the procedure shown in TMP, more than half of the time our simulation will generate p-values similar to those shown in TMP.

asset pricing at daily timescales is still a "small" data problem for which the risk of overfitting complex text analysis algorithms is high. BMZ conduct their entire time series analysis over a period of less than 200 days. If the number of tokens in the text they analyze varies sufficiently, then by an elementary theorem in linear algebra they could choose these co-occurrence weights to match the in-sample data exactly. Given this possibility, the extremely high market predictability BMZ discover is unsurprising.

The groundwork by Pav [2012a] motivated a closer look at the inferential strategy used in TMP. In the big data regime, multiple comparison adjustments are a must. Without them, we can expect the number of spurious correlations reported in the literature to skyrocket. We found that the market predictability results in TMP almost certainly yield an EMH-violating market timing strategy under realistic assumptions on the risk-free rate and transaction costs. We found that any reasonable adjustment for the multiple comparison bias in TMP eliminated all of the significant results in the linear time series section. We found that the local normalization strategy used in TMP is another potential source of error in inference.

We conclude our investigation of the Twitter mood effect by echoing Kuleshov [2011], who writes:

> The methodology problems of Bollen et al., and the fact that several groups were unable to replicate their accuracy raise serious concerns about the validity of these authors' results. Given the boldness of their claims, I believe they ought to either publish their methods and their code, or withdraw these claims.

In future work, we will attempt to replicate the Twitter "calm"-ness effect in-sample.

Bibliography

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, mar 2011a. ISSN 18777503. doi: 10.1016/j.jocs.2010.12.007. URL http://linkinghub.elsevier.com/retrieve/pii/S187775031100007X.

Johan Bollen, Alberto Pepe, and Huina Mao. Modeling Public Mood and Emotion : Twitter Sentiment and Socio-Economic Phenomena. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Modeling*, pages 450–453, 2011b.

Rene Carmona. *Statistical Analysis of Financial Data in R, 2nd edition*. 2014. ISBN 9781461487876. doi: 10.1007/978-1-4614-8788-3.

Eric Gilbert and Karrie Karahalios. Widespread Worry and the Stock Market. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 58–65, 2010.

Campbell R. Harvey and Yan Liu. Evaluating Trading Strategies. 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2474755. URL http://papers.ssrn.com/abstract=2474755.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the Cross-Section of Expected Returns. 2015.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL http://www.jstor.org/stable/10.2307/4615733.

Daniel Kahneman. Experienced Utility and Objective Happiness: A Moment-Based Approach.

In *Choices, Values and Frames*, chapter 37. Cambridge University Press and the Russell Sage Foundation, New York, 2000.

Volodymyr Kuleshov. Can Twitter predict the stock market? 2011.

Michael Lachanski. Not Another Market Timing Scheme! : Detecting Type I Errors with 'Good Deal' Bounds. *Journal of Undergraduate Research in Finance*, 1(1), 2015.

Stephen LeRoy. Risk Aversion and the Martingale Property of Stock Prices. *International Economic Review*, 14(2):436–446, 1973.

Stephen LeRoy. Efficient Capital Markets and Martingales. *Journal of Economic Literature*, 27 (4):1583–1621, 1989. URL http://www.jstor.org/stable/2727024.

Tim Loughran and Bill McDonald. Textual Analysis in Accounting and Finance : A Survey. 2015.

Robert Lucas. Asset Prices in an Exchange Economy. *Econometrica: Journal of the Econometric Society*, 46(6):1429–1445, 1978. URL http://www.jstor.org/stable/1913837.

Arman Khadjeh Nassirtoussi, Teh Ying Wah, Saeed Reza Aghabozorgi, and David Ngo Chek Ling. Text Mining for Market Prediction: A Systematic Review. *Expert Systems with Applications*, 41(16):7653–7670, jun 2014. ISSN 09574174. doi: 10.1016/j.eswa.2014.06.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0957417414003455.

Steven Pav. The junk science behind the 'Twitter Hedge Fund', 2012a. URL http://sellthenews.tumblr.com/post/21067996377/noitdoesnot.

Steven Pav. Converting Timing Edge to Sharpe, 2012b.

Cornelius A Rietveld, Tõnu Esko, Gail Davies, Tune H Pers, Patrick Turley, Beben Benyamin, F Christopher, Valur Emilsson, Andrew D Johnson, James J Lee, Christiaan De Leeuw, Riccardo E Marioni, Sarah E Medland, Michael B Miller, Olga Rostapshova, Sven J Van Der Lee, Anna A E Vinkhuyzen, Najaf Amin, Dalton Conley, Cornelia M Van Duijn, Rudolf Fehrmann,

Lude Franke, Edward L Glaeser, Narelle K Hansell, Caroline Hayward, William G Iacono, Carla Ibrahim-verbaas, Vincent Jaddoe, David Laibson, Paul Lichtenstein, C David, Patrik K E Magnusson, Nicholas G Martin, George Mcmahon, Nancy L Pedersen, Steven Pinker, David J Porteous, Danielle Posthuma, Fernando Rivadeneira, Blair H Smith, John M Starr, Henning Tiemeier, J Nicholas, Maciej Trzaskowski, André G Uitterlinden, C Frank, Mary E Ward, Margaret J Wright, George Davey, Ian J Deary, Magnus Johannesson, Robert Plomin, M Peter, Daniel J Benjamin, David Cesarini, and D Philipp. Correction for Rietveld et al., Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences*, 112(4):E380–E380, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1424631112. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1424631112.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005. ISSN 00129682. doi: 10.1111/j.1468-0262.2005.00615.x.

Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

John D Storey. A direct approachto false discovery rates. *Journal of the Royal Statistical Society B*, 64(3):479–498, 2001.

Paul C. Tetlock. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 63(3):1139–1168, 2007.

Paul C. Tetlock, Maytal Saar-tsechansky, and Sofus Macskassy. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

Halbert White. A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126, 2000. ISSN 0012-9682. doi: 10.1111/1468-0262.00152.