Trade-based Information and Sentiment

Xing Huang Asaf Manela Jimmy Wu

Washington University in St. Louis, Olin Business School

Email {xing.huang, amanela, jimmywu}@wustl.edu

April 2019

1 Introduction

One important motive for trade is information—informed investors trade on valuable information, whereas uninformed investors speculate on perceived information (or sentiment). Recent literature on active investment management uses fund portfolio holdings to evaluate stock selection skills possessed by fund managers (Cremers and Petajisto (2009); Kacperczyk et al. (2005)). If a subset of fund managers are informed and possess stock selection skills, there would be valuable information embedded in their holdings which could predict stock returns. Moreover, uninformed investors may have systematic bias in their belief formation (e.g. extrapolative bias) which may have implications for asset prices in equilibrium and lead to return predictability as well (Greenwood and Shleifer (2014); Da et al. (2019)).

Given the high-dimensional nature of trading data, previous work mainly relies on theoretical guidance to identify potential informed investors and applies simple aggregation across the trades of these investors (for example, skilled institutional investors (Wermers et al. (2012)), insiders (Lakonishok and Lee (2015)), short sellers (Boehmer et al. (2008)), etc.), or to identify specific types of systematic bias of investors (for example, extrapolative bias (Odean (1999); Grinblatt and Keloharju (2002)).

We propose instead a more data-driven approach and use machine-learning to extract valuable information and predictive sentiment from investors' trades to predict future returns. Our method allows for flexible heterogeneity across investors, and exploits both the information embedded in informed investors' trades and the sentiment driven by systematic bias in uninformed investors's trades. Assessing the improvement in out-of-sample predictive power relative to simpler aggregators used in past literature, we hope to quantify the economic gains from using machine learning to analyze trading data, and to better understand asset prices. The predictors that contribute to the improved predictive power of trading data may shed light on various topics such as investor trading behavior, investor skill measures, and the joint behavior of stock prices and trading volume. This study could also inform the debate about active money management skill and on the effects of information and sentiment on asset prices.

2 Research strategy

2.1 Multinomial logistic regression

Our main approach is to model the endogenous variable as the number of shares or dollar holdings. Assume there are M mutual funds and N unique stocks in the market. Let α_{it} denote abnormal return of asset i

from t to $t+1$. We consider a shares demand model as follows:

$$s_{ijt} = \lambda\left(\alpha_{1t}, \ldots, \alpha_{Nt}, p_{1t}, \ldots, p_{Nt}, m_{it}\right) + \varepsilon_{ijt} \tag{1}$$

where s_{ijt} is the number of shares in asset i held by investor j at time t, and $m_{it} \equiv \sum_j s_{ijt}$ is total shares held by investors in our sample. We model the demand as a multinomial logitistic regression:

$$p\left(\boldsymbol{s}_{it}|\boldsymbol{v}_{it}\right) = Multinomial\left(\boldsymbol{s}_{it}|\boldsymbol{q}_{it}, m_{it}\right) \tag{2}$$

where the probability that a given share of stock i is held by fund j at time t is modeling with a logit $q_{ijt} = \frac{e^{\eta_{ijt}}}{\sum_j e^{\eta_{ijt}}}$ with $\eta_{ijt} = \varphi_{0j} + \boldsymbol{\varphi}_j' v_{it}$, and $\boldsymbol{v}_{it}$ is a vector of covariates that determine $j's$ demand for i at time t, and that includes abnormal returns.

The need for dimensionality reduction is that suppose we want to predict α_{it} with investor demand s_{ijt}'s. We could potentially run the regression $\boldsymbol{\alpha}_t = F\left(\boldsymbol{s}_t\right) + \boldsymbol{e}_t$, which has $N \times M$ rhs variables for each stock on the left. But after fitting the (backward / demand) regression (2), we can form a *sufficient reduction projection* in the direction of $\boldsymbol{\alpha}_t$, $\boldsymbol{z}_t = \boldsymbol{\varphi} s_t \in \mathcal{R}^N$ with elements $z_{it} = \sum_j \boldsymbol{\varphi}_j' s_{ijt}$, which summarizes all the information in investor demand that is relevant for predicting abnormal returns. We can then use it in the forward regression $\boldsymbol{\alpha}_t = F\left(\boldsymbol{z}_t, \boldsymbol{m}_t\right) + \boldsymbol{e}_t$, which has (only) $N+1$ rhs variables for each stock on the left.

2.2 Hurdle distributed multiple regression

Given investor's bounded rationality, investors are likely to break down the investment decision into two steps (Gabaix (2014); Nieuwerburgh and Veldkamp (2010)): they first select the stocks they want to acquire information and then decide how they want to trade on those stocks. To take into account the two-step decision making process, we will break up (1) into a two-part hurdle distributed multiple regression (HDMR) as in Kelly et al. (2019):

$$\begin{aligned} h_{ijt}^* &= h_{ij}\kappa_{jt} + w_{it}'\delta_{jt} + v_{ijt}, \quad h_{ijt} = 1(h_{ijt}^* > 0) \\ s_{ijt}^* &= s_{ijt}\lambda(\varphi_{0j} + \boldsymbol{\varphi}_j' v_{it}) + \varepsilon_{ijt} > 0, \quad s_{ijt} = h_{ijt} s_{ijt}^* \end{aligned}$$

The first two equations describe the choice to include stock $i$into the choice set or not, while the next two equations capture the number of shares investors want to hold on the asset.

2.3 Predicting returns by extracted information/sentiment

We will evaluate the improvement in return prediction using the machine learning technology compared to traditional simple aggregator. Through decomposing the improvement into different components, we can further explore whether and how much the additional predictive power comes from simple machine learning using multinominal logistic regression alone, or augmented with modeling the choice set selection by the hurdle rate; we can also examine how much additional predictive power is gained by extracting information compared with extracting sentiment, or how much predictive information/sentiment is contained in investor's purchase decisions vs. sale decisions.

2.4 Interpreting the information/sentiment extraction

To better understand the process of extracting information or sentiment from trades, it is useful to examine loadings on the predictors, which will shed lights on a couple of other interesting topics. For example, what drives trade? Among the potential relevant factors (e.g. past, current, future returns, news coverage, analyst forecasts/revisions, firm characteristics, etc.), what matter more for buys? What matter more for sales? How does that challenge our thinking about modern portfolio theories?

We could also infer the skill/informedness or the systematic bias of investors from the coefficient φ_j', and link the measure to the type of investors (institution vs individual), the characteristics of institutions (style, fund manager characteristics, skill measure, etc.) and the characteristics of individuals (personal information, behavioral bias/sophistication measure, etc.) Given the sparse feature of investor's portfolio, using the two-part HDMR may improve the performance relative to multinomial logistic regression. The estimates from HDMR could clarify which factors matter more for the intensive margin (how much to trade) vs extensive margin (whether to enter the choice set)?

3 Computing language

We plan to use Julia and Python to conduct the analysis.

References

Boehmer, Ekkehart, Charles M. Jones, and Xiaoyan Zhang, "Which Shorts are Informed?," *Journal of Finance*, 2008, *63*, 491–527.

Cremers, Martijn and Antti Petajisto, "How Active Is Your Fund Manager? A New Measure That Predicts Performance," *Review of Financial Studies*, 2009, *22*, 3329–3365.

Da, Zhi, Xing Huang, and Lawrence Jin, "Extrapolative Beliefs in the Cross-Section: What Can We Learn from the Crowds?," *Working Paper*, 2019.

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129*, 1661–1710.

Greenwood, Robin and Andrei Shleifer, "Expectations of returns and expected returns," *Review of Financial Studies*, 2014, *27* (714-746).

Grinblatt, Mark and Matti Keloharju, "What Makes Investors Trade?," *Journal of Finance*, 2002, *56*, 589–616.

Kacperczyk, Marcin, Clemens Sialm, and Lu Zheng, "On the Industry Concentration of Actively Managed Equity Mutual Funds," *Journal of Finance*, 2005, *60*, 1983–2011.

Kelly, Bryan, Asaf Manela, and Alan Moreira, "Text Selection," *Working Paper*, 2019.

Lakonishok, Josef and Inmoo Lee, "Are Insider Trades Informative?," *Review of Financial Studies*, 2015, *14*, 79–111.

Nieuwerburgh, Stijn Van and Laura Veldkamp, "Information Acquisition and Under-Diversification," *Review of Economic Studies*, 2010, *77*, 779–805.

Odean, Terrance, "Do Investors Trade Too Much?," *American Economic Review*, 1999, *89*, 1279–1298.

Wermers, Russ, Tong yao, and Jane Zhao, "Forecasting Stock Returns Through an Efficient Aggregation of Mutual Fund Holdings," *Review of Financial Studies*, 2012, *25*, 3490–3529.