Discovering Technical Traders in the T-bond Futures Market

Andreas S. WEIGEND, Fei CHEN, Stephen FIGLEWSKI
Leonard N. Stern School of Business
New York University
New York, NY 10012
{aweigend|fchen|sfiglews}@stern.nyu.edu
www.stern.nyu.edu/~aweigend

Steven R. WATERHOUSE
Ultimode Systems
Data Mining Tools
2560 Bancroft Way #213
Berkeley, CA 94704
stevew@ultimode.com
www.ultimode.com

Abstract

This study uncovers trading styles in the transaction records of US Treasury bond futures. We use statistical clustering techniques to group together trades that are similar. Trade profit was held back in the clustering process. Results show that clusters differ significantly in their profit and risk characteristics. Some clusters uncover "technical" trading rules. Using the information about the individual accounts, we describe the assignments of accounts to clusters by entropy, and model the transitions of a given account through clusters by a first order Markov model.

Motivation

The Commodity Futures Trading Commission records every transaction of the bond futures market. The comprehensiveness of this data as well as the maturing of technologies to process large amounts of information is enabling researchers to investigate many aspects of the futures market that were not possible before. In this study, we adopt a data-driven clustering approach to address the following questions: What trading styles and hidden regularities can statistical methods uncover in bond futures? Do we find evidence consistent with technical trading? How consistently do small traders adhere to one trading style? Can we model transitions from one style to another? What are the profit and risk characteristics for these styles?

Raw Data and Feature Selection

The raw data set contains everyone transaction executed between 1989 and 1992. A transaction record consists of time of transaction, price, volume, buy or sell information, a specific account number with a clearing house identification, and a Customer Type Indicator (CTI).

The goal of uncovering trading style through clustering implies that the basic unit of analysis has to be an individual trade. We define a trade to begin when an account opens a long or short position, and to end when the account position size returns to zero.

Trading styles are characterized by five *trade-specific* variables:

- logarithm of maximum absolute cumulative position size of a trade;
- opening position size relative to the maximum cumulative position size;
- logarithmic exposure, the area under position size versus chronological time;
- buy or sell information of the first transaction of a trade;
- logarithmic length of a trade;

and five *market-specific* variables:

- logarithm of exponentially smoothed (by 30 minutes) volatility at the opening of a trade;
- exponentially smoothed (30 minutes) price at opening relative to the minimum and maximum of the last half hour bar, scaled to lie between -1 and 1;
- same exponentially smoothed price at first reversal;
- exponentially smoothed short term (30 minutes) market volume at opening;
- exponentially smoothed long term (3000 minutes) market volume at opening.

This study analyzes the transactions of the US Treasury bond futures September 1992 contract. We concentrate on non-exchange members trading off the floor (CTI type 4), who have no outstanding contracts at expiration and have executed at least 10 transactions. This filtering leaves us with a data set of 9271 trades, representing 1127 accounts. For a more detailed discussion of the data and variable selection process, see Chen et al. (1998a).

Clustering

The trades can be viewed as a set of unlabeled examples forming clusters in the space defined by our choice of trade-specific and market-specific variables. The goal is to find similar trades and categorize them as one cluster.

We assume that trades were generated by a finite mixture of distributions (Titterington, Smith and Makov 1985). Each distribution corresponds to a prototypical trade. Every observed trade is a noisy realization of one of these prototypes. The parameters of the mixture distribution are estimated by maximizing the likelihood of data given the model. Except for the long/short variable which is modeled as a binomial, all of the mixture distributions are Gaussians.

The probability of observing an input $\mathbf{x}$ is $P(\mathbf{x}) = \sum_{k=1}^{K} P(\mathbf{x} \mid k)P(k)$, where $\mathbf{x}$ is the input vector, and K the number of mixing Gaussians. We assume an independent Gaussian noise model:

$$P(\mathbf{x} \mid k) = \qquad (1)$$
$$\frac{1}{(2\pi)^{\frac{M}{2}} \mid \Sigma^k \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu^k)^T (\Sigma^k)^{-1} (\mathbf{x} - \mu^k)\right),$$

where M is the dimensionality of $\mathbf{x}$ and the parameters of kth Gaussian are mean μ^k and covariance Σ^k. The likelihood of data given model is

$$\mathcal{L} = \prod_{t=1}^{T} P(\mathbf{x}^{(t)}) = \prod_{t=1}^{T} \sum_{k=1}^{K} P(x^{(t)} \mid k)P(k). \qquad (2)$$

While a spherical structure of the covariance matrix (identity matrix scaled by a constant) is clearly too inflexible, full covariance matrices, allowing for correlations between input variables, are not desirable either. We restrict our model to diagonal matrices, with the understanding that the input variables are chosen and rendered as uncorrelated as possible.

Reasonable assumptions have to be made on the priors. We assume uniform priors on the mean and Jeffreys priors on the variance, see Cheeseman and Stutz (1995) for more details.

In estimating a mixture of distributions, we face two sets of unknowns: the posterior probability of a data point given the model, and the parameters of the model. We estimate them iteratively using the Expectation-Maximization algorithm (Dempster, Laird and Rubin 1977).

Results

This section analyzes the clustering results, assuming 7 clusters.

Trade Profit and Time until Expiration

Two pieces of information were withheld during clustering: the profit of each trade and its time until expiration.

Table 1 summarizes the profit of each cluster. While the overall sample considered here sustains an average loss per trade of USD 970, all clusters except 2 and 4 earn positive profits on average. The mean profits of cluster 2 and 3 are statistically significant, lying at least 2 standard deviations away from zero. Pronounced risk asymmetries exist: cluster 4 is a losing cluster with a great deal of variance in profit and a lot of downside risks. The same is true for cluster 7, except that the asymmetry is on the upside. These two clusters absorbed all the outliers. Cluster 3 is profitable statistically, and its variance is tight, suggesting risks associated with this cluster are quite small.

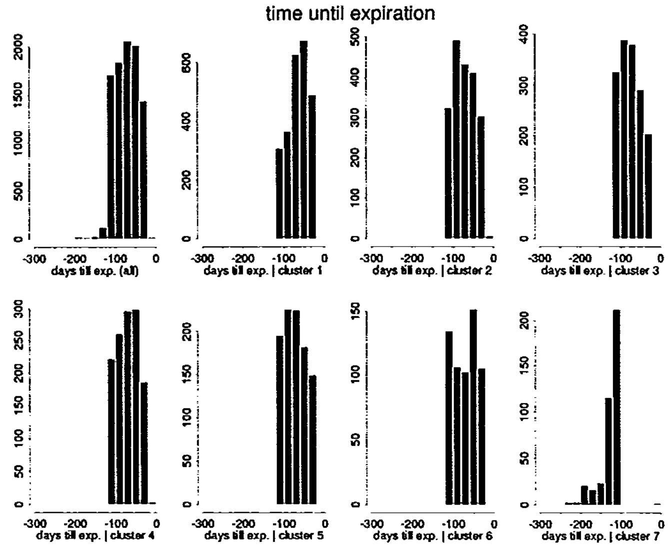

Figure 1: Histograms of trade opening time in days until expiration (zero is the time of expiration). The first histogram in the upper left corner shows the distribution for all trades. The rest are histograms conditional upon clusters.

Trading time until expiration is not a clustering variable, but it is useful for interpreting trading styles. The main result from Figure 1 is cluster 7, which has a trading style of opening early in the contract, in this case, at least 3 months prior to expiration.

Trade-Specific Variables

Trade length is a clear distinguishing feature among clusters. Figure 2 shows the conditional histograms. Although the unconditional histogram indicates that on average most trades are completed within a day, we see clear differences between clusters 1 and 7. Cluster 1 is a very quick trading style, with all of its trades finished in less than a day. Cluster 7, on the other hand, is a very long trading style, with most of its members holding trades longer than a day. Coupled with the fact that cluster 7 opens early in the contract, this finding means that early positions tend to stay put in the market. This is consistent with expectations.

Maximum position size is a measure of how big a trade is. Figure 3 shows the clustering result. Here the conditional plots are generated by dividing the conditional histogram by the unconditional. This provides a sense of how trades are distributed among clusters along the trade size dimension. The clear features are clusters 1 and 4. Cluster 1 contains day traders, and maintains very small position sizes (about 40% of all the smallest trades) but no large ones at all. Cluster 4 is

	all	1	2	3	4	5	6	7
N	9271	2465	1696	1592	1265	980	599	401
mean profit	-976	6	-400	390	-9260	550	160	6330
error of mean	990	40	80	180	6990	370	470	5930
ratio	1.0	0.17	5.37	2.17	1.37	1.50	0.34	1.06

Table 1: Summary table of profits per cluster, ordered in decreasing cluster size N. The rows are: cluster size N, mean of profit per trade (in dollars), error of the mean (std(profit)/$\sqrt{N}$) and ratio between mean profit and error of mean.

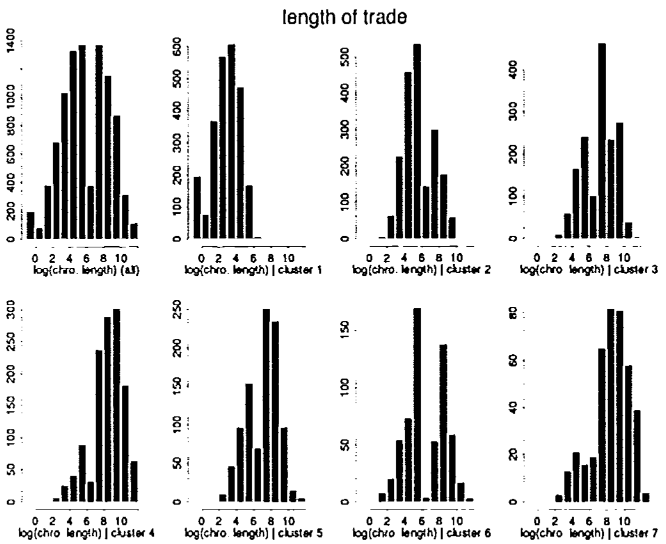

Figure 2: Histograms of logarithmic length of trade. The leftmost bar in the histograms corresponds to trades of length less than one minute. The gap between 6 and 8, common to all histograms, is due to the fact that most trades are either less than 8 hours or longer than 16 hours.

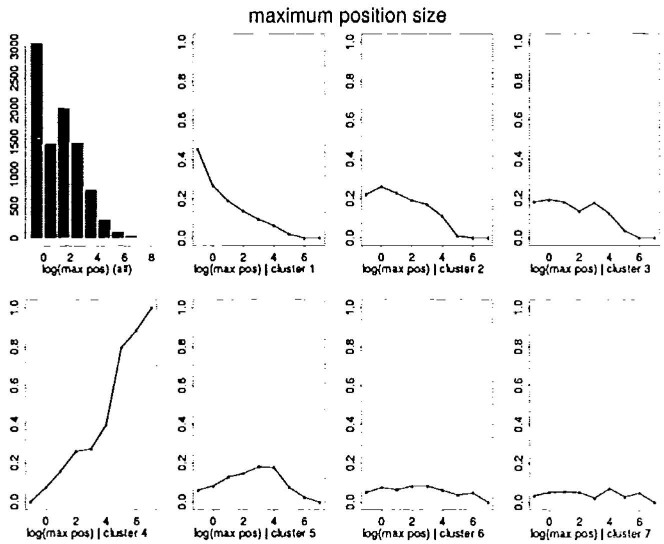

Figure 3: Relative histograms of logarithmic maximum position size. The upper left plot is the unconditional histogram. The rest are conditional plots showing the percentage of all trades each cluster contains for a given maximum position size.

just the opposite. This is a big trading style, containing 100% of the largest trades in the data while containing no small trades.

Market-Specific Variables

One interesting result comes from relative price, a variable chosen to help us uncover evidence of technical trading. This variable indicates the movement of prices at a particular point in time. Recall a value of -1 means the price has been going down in the past half hour, while +1 means the opposite.

Figure 4 shows distributions of relative prices in the half hour prior to opening a trade and first reversal of the position. Three most salient clusters are presented here. The results are best understood in the context of long/short information of a trade, which is displayed on top of each plot respectively.

Cluster 3 is a long style. This cluster opens a long position when the price has been going up, and reverses its position as the price has been going down. Such behavior is consistent with the technical trading style of trend following. Cluster 5 also contains mostly long trades, but it reacts to price movement exactly the opposite of cluster 3. Trades in this cluster initiate a long position when the price has been going down, and reverse to a short position when the price has been going up. This is consistent with the contrarian strategy. A mixing of these two trading strategies is observed in cluster 2. This cluster contains short trades (only 3% long) that both open and reverse when the price is at its peak with respect to the last half hour, not a strategy recommended by many.

Trade Entropy and Account Dynamics

The clustering results we have presented so far ignored account identification. However, one aspect of trading style analysis we are interested in is how trades within an account are related to various trading styles as defined by the clusters. We use *entropy* to answer this question. The entropy of an account j, H_j, is defined to be

$$H_j \equiv -E[\log p_k^j] \equiv -\sum_{k=1}^{K} p_k^j \log p_k^j \; , \qquad (3)$$

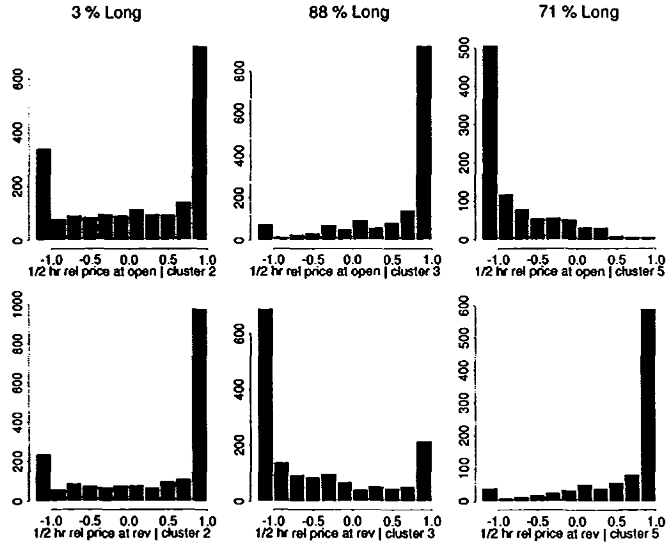

Figure 4: Histograms of relative price at opening a trade (first row) and at first reversal (second row). Only cluster 2, 3 and 5 are shown.

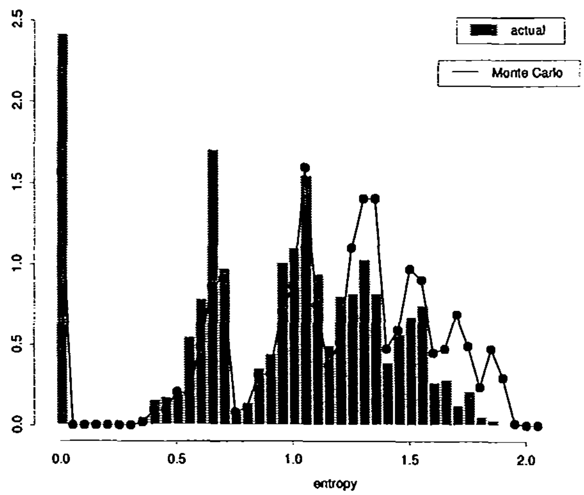

Figure 5: Histograms of trade entropy. The bar plot is the distribution of entropy results from clustering. The line plot shows the distribution of mean entropy values of 100 Monte Carlo sampling. The areas under the bar plot and the curve are both normalized to one.

where p_k^j is the probability that a trade in account j is in cluster k. The probabilities are obtained empirically. Entropy in this context is useful for checking model validity by ensuring that the same clustering results cannot be obtained through random assignments.

Figure 5 contains a histogram of the distribution of entropy over the 1127 accounts. To provide a sense of the baseline, we destroy the specific account and cluster information: clusters are randomly assigned to trades, while the probability of the clusters are conserved. The results of 100 Monte Carlo runs are summarized by the line plot. Compared with the actual distribution (with a mean entropy of 0.95), the rightward shift of the baseline (with a mean of 1.12) confirms that accounts in the dataset employ trading styles more often than just randomly.

Time order of trades within an account is another important piece of information that has also been ignored so far. Do trades tend to pursue one trading style, or do they often employ different ones? We describe the dynamics of trade movements by the probabilities with which an account moves from one style to another. Figure 6 summarizes this information in a Markov transition matrix. Each cell contains an empirical transition probability. The outstanding feature is cluster 7. In addition to being a cluster that opens early in the contract and holds lengthy trades, this is a trading style that almost always (with probability 0.93) stays in the same cluster.

Validating the Results

As a first step of validating the findings, we varied number of clusters and checked for spurious results. Specifying different numbers of clusters at the outset leads to the discovery of different sets of clusters. In order to compare them, we use hierarchical clustering. Figure 7 illustrates this process of clustering 13 centers: 7 clusters are described in this paper, the remaining 6 were obtained from another clustering run. We see that most clusters are similar, except cluster C. Looking closely at this cluster we discover that it splits into clusters 3 and 4 if we add one more cluster as part of the model specification. Trial and examination of different numbers of clusters lead us to believe that 7 clusters seem to be a reasonable assumption.

A second validating step involved clustering with several additional variables. These variables were taken from the same set of inputs, but with their ordering randomized in order to destroy any associations with specific trades. Clustering outcomes remained the same. The randomized variables did not show any relevant structure.

To compare the results further, we clustered with two more data generating models: Gaussians with spherical covariance matrices, and Gaussians with full covariance matrices. The spherical matrix helps us understand the impact of the diagonal assumption on clustering results, while the full covariance guides us to the choice of input variables. The latter generalization provided results consistent with those presented in this paper.

Conclusions and Outlook

This paper showed how cluster analysis can uncover some structure in T-bond futures trading data. We identified the cluster centers as trading styles and demonstrated that some styles correspond to conventional technical trading rules, whereas others–differing significantly in profit and risk–require new explana-

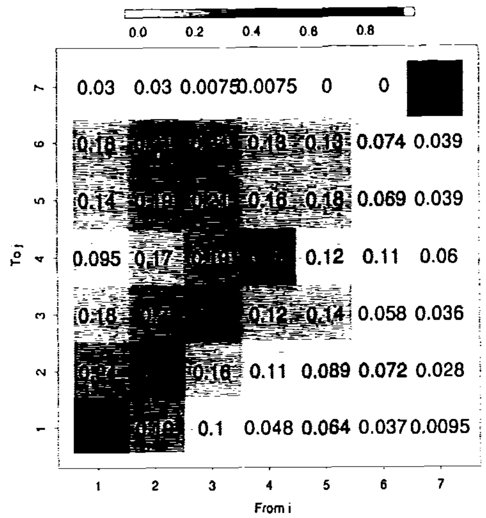

Figure 6: Image plot of a first order cluster Markov transition table. Each cell contains a gray scale representation of the empirical probability (also printed in the cell) that a trade is in cluster i (the "From" axis) and the next trade of the same account moves to cluster j (the "To" axis).

tions.

We view the work reported here as a starting point for several levels of investigation. The first level compares this model of trades as a finite mixture of distributions with a tree-based method and shows the strengths and weaknesses of these two approaches in predictive power, transparency and understandability (Chen et al. 1998b).

The second level focuses on the sequence of discovered styles for each account. Adding a set of variables that characterize the recent trading history of an account, we try to explain some of the style transitions. Contrasting profitable traders with non-profitable ones, we find behavioral asymmetries in variables characterizing market timing and risk taking.

The third level of analysis represents a significant change in the use of computational intelligence for trading. Traditionally, the core predictive part of a trading system extracts regularities from financial data. In contrast, we extract regularities from the actual behavior of traders. We model successful and unsuccessful traders separately. We adopt a two-stage architecture. We first learn when to act (timing), then how to act (buy or sell). To assess the knowledge represented in these two sets of models, we generate surrogate trades and evaluate their profits and risk performance on a held-out test set of market data.

Finally, we analyze the data in a richer context that includes information about external news events. While the impact of news events on prices and volatilities has been studied on the macro level for some markets, we are able to characterize reactions to such events on the micro level of individual traders. Employing the clustering methodology used in this paper, we can uncover distinct styles of response to an event. To bridge from the micro level to the macro level, we use concepts from mean field theory that allow us to suggest a mechanism for global price discovery from local trader-to-trader interactions.

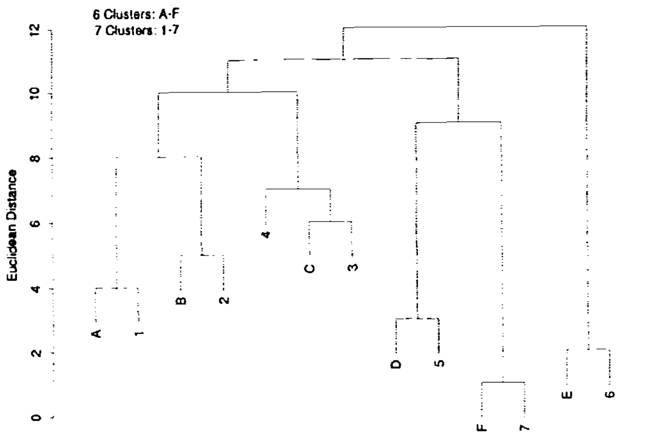

Figure 7: Results of hierarchical clustering on cluster centers obtained from two clustering runs, assuming 6 (labeled A-F) and 7 (labeled 1-7) clusters respectively. The y-axis reflects the Euclidean distance between the centers.

References

Cheeseman, P., and Stutz, J. 1995. Bayesian Classification (AutoClass): Theory and Results. *Knowledge Discovery in Data Bases II.* Menlo Park, Calif.: AAAI Press / The MIT Press.

Chen, F., Figlewski, S., Heisler, J., and Weigend, A. S. 1998a. Uncovering Hidden Structure In Bond Futures Trading, In Proceedings of the IEEE/IAFE/INFORMS 1998 Conference on Computational Intelligence for Financial Engineering, New York, NY.: IEEE/IAFE.

Chen, F., Figlewski, S., Waterhouse, S. R., and Weigend, A. S. 1998b. Modeling Financial Data Using Clustering and Tree-Based Approaches. In *International Conference on Data Mining*, Southampton, UK.: Computational Mechanics Publications.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B* 39:1–38.

Heisler, J. 1997. Loss Aversion Among Small Speculators in a Futures Market. In Proceedings of 1997 International Association of Financial Engineers (IAFE) Conference, Boston, Mass.: IAFE.

Lee, C., Shleifer, A. and Thaler R. 1991. Investor Sentiment and the Closed-End Fund Puzzle. *Journal of Finance* 46:75–109.

Titterington, D., Smith, A., and Makov U. 1985. *Statistical Analysis of Finite Mixture Distributions.* New York, Mass.: John Wiley.