Uncovering Hidden Structure in Bond Futures Trading

Fei Chen
Leonard N. Stern School of Business
New York University

Stephen Figlewski
Leonard N. Stern School of Business
New York University

Jeffrey Heisler
School of Management
Boston University

Andreas S. Weigend
Leonard N. Stern School of Business
New York University

January 1998

Working Paper Series
Stern #IS-98-1

Working Paper IS-98-01 (Information Systems)
Leonard N. Stern School of Business, New York University.
In: *Proceedings of the IEEE/IAFE/INFORMS Conference on Computational Intelligence for Financial Engineering (CIFEr'98, New York, March 1998)*
http://www.stern.nyu.edu/~aweigend/Research/Papers/TradeStyles

Uncovering Hidden Structure in Bond Futures Trading

Fei CHEN*, **Stephen FIGLEWSKI**‡,
Jeffrey HEISLER‡‡, **Andreas S. WEIGEND***

Abstract. This study uncovers trading styles in the transaction records of US Treasury bond futures. It uses transaction-by-transaction data from the Commodity Futures Trading Commissions' (CFTC) Computerized Trade Reconstruction (CTR) records. The data set consists of 30 million transaction—the complete US T-bond futures market for 3 years. Each transaction record consists of time (by the minute), price, volume, buy/sell, and an identifier of the specific account.

We use statistical clustering techniques to group together trades that are similar. Two sets of assumptions have to be made: (1) What is a trade? We define a trade to begin when an account opens a position, and to end when its position size returns to zero. We describe each trade by several trade-specific variables (e.g., length of trade, maximum position size, opening move, long or short) and several exogenous, market-specific variables (e.g., price, volatility, trading volume). (2) What process generated the data? We assume a mixture of Gaussians. An observed trade is interpreted as a noisy realization of one of the mixture components. This paper assumes identity covariance matrices. Furthermore, each trade is fully assigned to a single cluster. We compare this approach to diagonal and to full covariance structure with probabilistic assignments.

Trade profit was held back in the clustering process. It turns out that the clusters differ significantly in their profit and risk characteristics. Using conditional distributions, we summarize features of profitable trading styles and contrast them with losing strategies. We find that profitable styles tend to hold trades longer, trade at higher volatility, and trade earlier in the contracts. We also show how some clusters uncover "technical" traders. Using the information about the individual accounts, the assignments of accounts to clusters are described by entropy, and the transitions of a given account through clusters is modeled by a first order Markov model.

1 Motivation and Overview

The recent explosion of our ability to collect, store, and process information is offering us immense opportunities to discover complex hidden relationships in data. In this study, using the complete set of 3 million transaction records of the bond futures market, we seek to address the following questions: What trading styles and hidden regularities can statistical methods uncover in bond futures? How do trade-specific variables (opening move, position, exposure, etc.) interact with exogenous variables (price, volume, volatility, etc.)? How can machine learning techniques explain the rationales and behaviors of traders? Can the dynamics of speculators' movements among styles be modeled?

We adopt a data-driven clustering approach to find hidden properties characterizing different trading styles through systematic examinations of the data. The essential step in this process is the identification of meaningful input variables. The trades form "clouds" in the space defined by the set of trade-specific

*Information Systems Department, Leonard N. Stern School of Business, New York University, 44 West Fourth Street, New York, NY 10012, {fchen|aweigend}@stern.nyu.edu
‡New York University Salomon Center, 44 West Fourth Street, New York, NY 10012, sfiglews@stern.nyu.edu
‡‡School of Management, Boston University, 595 Commonwealth Avenue, Boston, MA 02215, jheisler@bu.edu

and exogenous variables. The task is to distinguish these clouds and interpret them as distinct trading styles. The choice of variables is thus crucial to revealing the hidden structures of the data. Clustering is a statistical methodology that groups trades into clusters based on probabilistic properties discovered only during the clustering process. The application of this technique in finance was first seen in the work by Gruber (1972).

The organization of this paper is the following: Section 2 discusses the raw data, the feature vectors used for clustering, and the data selection process; Section 3 provides an overview of the clustering algorithm; Section 4 interprets clustering results from three angles: variables not included in clustering, variables that are trade-specific, and variables that are exogenous to the trades; Section 5 considers the results on trade entropy and account dynamics; Section 6 presents the conclusions.

2 Data

2.1 Raw Data

The Commodity Futures Trading Commission's (CFTC) Computerized Trade Reconstruction (CTR) records transaction-by-transaction data of the bond futures market. This study analyzes the transactions of the US Treasury bond futures September 1992 contract. Each transaction record consists of a price, volume, buy or sell information, and a specific account number. Transaction time is also recorded. The market operates during two time periods each day. Daytime trading hours (during which more than 95% of trading volume occurs) are 7:20 am to 2:00 pm Chicago time. Night trading hours are either 5:20 pm to 8:05 pm standard time, or 6:20 pm to 9:05 pm daylight savings time. Information from buyers and sellers is combined in the CTR data so that each transaction can be reconstructed by the minute (Manaster and Mann 1996).

2.2 Feature Vectors

The goal here is to analyze and understand trading behavior by clustering similar trades. This implies that the base unit of analysis has to be an individual trade. A trade is defined to begin when an account opens a long or short position, and to end when the account position size returns to zero. Note that the term "trade" in this paper does not imply a single transaction between two counter parties, as is often used.

Trading styles are characterized by two groups of variables. The *trade-specific* variables reflect the state of each trade: how big the opening position size is, what the maximum position size is, how long the holding period is, and what the trade's exposure to the market is. The *exogenous* variables characterize the working environment in which a trade takes place: how prices have been moving when a trade opens, how volatile the market is at that time, and how heavy the overall trading activities are in the market.

Information at the opening of a trade is important. It is indicative of the market conditions when an account opens a trade. Information at the first reversal of position is equally as important. It tells us the circumstances under which a trader decides to behave differently than before. Thus trading styles are characterized exogenously by variables both at the time of opening and the time of first reversal.

First, we define notations used in the paper. Let t denote the chronological time. Let t_o be the time at which a trade opens, and t_e be the time a trade closes. The sign of a trade, long or short, is +1 or -1 respectively. Let x_t denote the signed transaction size at t, and p_t be the transaction price.

An "active minute" is defined to be a minute in which at least one transaction takes place in the full data set. Let $\min_{p_t}^{(30)}$ denote the minimum price of the price series of the 30 active minutes ending at t.

Similarly, $\max_{p_t}^{(30)}$ denotes the maximum price of the same price series.

The cumulative sum function of ξ_t, $\texttt{cumsum}(\xi_t)$, is

$$\texttt{cumsum}(\xi_t) \equiv \sum_{i=1}^{t} \xi_i \ . \tag{1}$$

When $\xi_t = x_t$, the cumulative sum becomes the position size at t, which is denoted by s_t.

Using these definitions, the trade-specific variables for clustering are:

- maximum absolute position size, $\max_{\text{duration of trade}} |s_t|$;
- opening position relative to the maximum position size, $\frac{|x_{t_o}|}{\max_{\text{duration of trade}} |s_t|}$;
- exposure, the area under position size versus chronological time;
- sign of trade, long $\equiv +1$ and short $\equiv -1$;
- logarithmic length of trade, $\log(\frac{t_e - t_o}{1440})$, where 1440 minutes correspond to a day.

The exogenous variables include information on prices, short and long term empirical volatilities, and trading volume. Prices are given at the time of opening a trade and at the time of the first reversal of the position. They are expressed as the relative location of the current price with respect to the minimum and maximum of the local bar of the last 30 minutes, and are scaled to lie between 1 and -1:[§]

$$sign(x_{t_o}) \left[2 \left(\frac{p_t - \min_{p_t}^{(30)}}{\max_{p_t}^{(30)} - \min_{p_t}^{(30)}} \right) - 1 \right] \ . \tag{2}$$

The normalized quantity is further multiplied by the sign of the trade, which is the direction of the first transaction. The rationale behind this is that trading styles of going long when prices have been rising should be treated the same as those of going short when prices have been falling. We fold these two scenarios into one.

Volatility and volume information at the time of opening a trade and first reversal of a position are computed with an exponential filter capturing information on the time scale of 30 active minutes (short term). In addition, volatility and market volume at opening are also computed with an exponential filter of 3000 active minutes, corresponding to one week (long term). The exponential filter for squared relative return is

$$z_t = (1 - \lambda) \left(\log(p_t) - \log(p_{t-1}) \right)^2 + \lambda z_{t-1} \ , \tag{3}$$

where the decay parameter, λ, is either 0.936 or 0.999, corresponding to 30 or 3000 active minutes respectively, z_t is the filtered series, and volatility $v_t = \sqrt{z_t}$. Analogously we filter the volume series on the same time scales.

2.3 Account Selection

The entire set of transaction data of the September 1992 contract consists of more than 3 million records. Different types of accounts (e.g., floor traders' accounts versus non-exchange members' accounts) manifest

[§]In cases when the denominator is zero, we set the relative price to zero.

different trading styles. Since we are currently not making comparisons between account types, we choose to analyze a subset of this data.

The first criterion for filtering the raw data is based on customer type. CFTC assigns a customer type indicator (CTI) code to each transaction. This study focuses on CTI code 4, corresponding to non-exchange members trading off the floor. This choice is based on the results that small speculators employ heuristic trading strategies that tend to exhibit more distinct trading styles than other groups (Lee, Shleifer and Thaler 1991, Heisler 1997).

Accounts with non-zero closing positions at expiration are excluded from the current analysis. Leakages may be caused by several factors, such as: actual delivery, exchange for physical, trades transferred from one account to another held by the same customer, or simply mis-entries. The data set does not give indications to such causes, so we choose to analyze only records with zero closings.

We divide the sample population into two groups: accounts with frequent trading and those with little. This paper reports results on accounts with more than 10 transactions. After filtering, we have a data set of 9271 trades, representing 1127 accounts.

3 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 exogenous variables. The goal is to find similar trades and categorize them as one cluster. Clustering is the appropriate data mining technique for this task. It groups trades that are "close" to each other into clusters, and at the same time maximizes the differences among the clusters (Banfield and Raftery 1992).

Viewed from a probabilistic framework, we assume that the 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. A Gaussian distribution is parameterized by its mean and its covariance matrix. We estimate these parameters by maximizing the likelihood of data given by the model.

The number of clusters is specified as part of the model, not estimated from the data. The probability density functions are currently assumed to be spherical Gaussians with identity covariance matrices. All input data are normalized to have a sample mean of 0 and a sample standard deviation of 1.

4 Results

This section analyzes the clustering results, assuming ten centers. Table 1 contains, by cluster, all mean values of the variables used during clustering. The 7 clusters listed in the table contain 99% of all trades. The remaining three have cluster sizes of 65, 5, and 3 respectively. We regard them as containing outliers and thus omit them from analysis. Figure 1 presents characteristic features of the clusters using Chernoff faces (Chernoff 1973).

Section 4.1 discusses two variables that are not included in the clustering: profit of the trade and the time until expiration. Section 4.2 discusses the trade-specific variables. Section 4.3 discusses the exogenous variables.

	all	*1*	*2*	*3*	*4*	*5*	*6*	*7*
N	9271	2815	2725	2257	744	407	180	75
length (log)	-1.35	-2.42	-1.89	-0.41	0.42	1.09	-0.89	-3.71
rel open	0.79	0.98	0.99	0.37	0.78	0.69	0.83	0.71
max position	16.02	6.82	6.92	25.31	9.41	11.39	6.66	9.37
exposure	133	10	10	134	52	316	18	9
% long	52%	5%	100%	53%	51%	68%	41%	24%
rel price at open	-0.011	-0.084	0.034	0.026	-0.001	0.100	-0.071	-0.545
rel price at rev	-0.047	-0.171	-0.020	-0.004	0.305	-0.096	-0.174	-0.496
*volat at open (30)**	5.2	5.2	4.2	4.2	11.0	6.7	6.4	10.0
*volat at rev (30)**	4.8	4.0	3.4	3.7	14.0	6.5	4.4	14.0
*volat at open (3000)**	5.2	4.9	4.8	4.9	5.0	12.1	5.1	5.3
volm at open (30)	3301	2633	2423	2321	2855	733	24930	55260
volm at rev (30)	3020	2487	2120	2206	3369	1613	1214	87872
volm at open (3000)	1824	1724	1738	1712	1693	277	8248	6419

* Volatility is in units of 10^{-4}.

Table 1: Summary of the clusters, ordered in decreasing cluster size N. The mean of all clustering variables is provided: logarithm of the length of the trade in minutes, relative opening move, maximum position size, trade exposure (in 1000s), percentage of long trades, relative price at open, relative price at reversal, volatility at open (short term, filtered at 30 minutes), volatility at reversal (short term), volatility at open (long term, filtered at 3000 minutes), volume at open (short term), volume at reversal (short term), and volume at open (long term). The second column contains information on the entire sample data set. Clusters 8 through 10 have sizes of 65, 5, and 3 respectively (not shown). They are omitted from analysis. The three large volume values (volume at open (30) for clusters 6 and 7, volume at reversal (30) for cluster 7) are caused by an outlier on August 27, 1992 in the CTR data set.

4.1 Trade Profit and Opening Time until Expiration

Two important pieces of information were deliberately withheld during the clustering: the profit of each trade and its time until expiration. This enables a clean interpretation and understanding of the trading styles.

Table 2 summarizes the variables for each cluster. In conjunction with the table, we interpret the results on profit with Figure 2, a box plot representation of the cluster profit distributions. This plot complements the summary table by providing better visualization of the outliers (Tukey 1990).

While the overall sample considered here sustains an average loss per trade of USD 970, clusters 2, 4, and 5 do earn positive profits on average. The mean profits of cluster 2 and 4 are statistically significant, lying at least 2 standard deviations away from zero. The profitable groups also prefer to buy (cluster 2 are all long), and tend to hold trades longer than a day (clusters 4 and 5 hold trades at least 1.5 day long).

Figure 2 reflects pronounced risk asymmetries. Cluster 1 is a losing cluster. It has a great deal of downside risks. The same is true for cluster 3. Cluster 4 is profitable on average, with outliers in both directions. Cluster 5 makes the largest profit on average, but it suffers greatly from a large variance in its

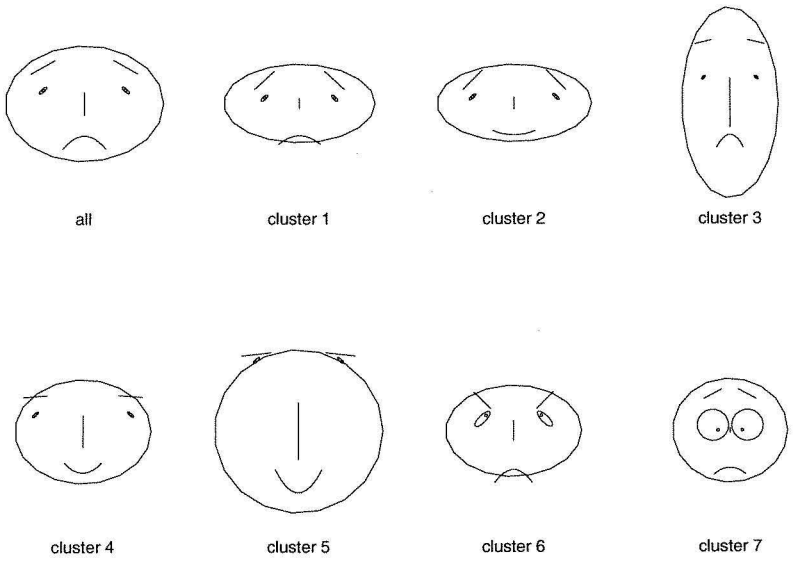

Figure 1: Chernoff faces of trading styles. The first face in the upper left corner characterizes all trades. The remaining are the clusters. Face mappings are:

- Curve of mouth: profit;
- Location of mouth: maximum position size;
- Area of face: exposure;
- Shape of face: relative opening position size;
- Length of nose: log length of trade;
- Location of eyes: volatility at opening;
- Shape of eyes: relative price at opening;
- Separation of eyes: relative price at first reversal;
- Width of eyes: volume at opening;
- Angle of eyes: volume at reversal;
- Location of pupil: opening time until expiration;
- Angle of eye brows: percentage of trades at night.

returns. Both cluster 6 and 7 are small in size. The spread of the distributions is tighter than that of the others.

Trading time until expiration is not a clustering variable, but it is useful for interpreting the trading styles. Figure 3 shows that clusters 5 through 7 have very different distributions from those of clusters 1 through 4. Cluster 5 is the earliest group to open trades, at an average of 129 days until expiration. Cluster 6 and 7 open the latest in the contract, with peaks at around 23 and 27 days respectively. Cluster 7 in particular has one large peak, indicating most of the trades in this cluster open at about the same time.

The percentage of trades at night is also not included in clustering. A night trade is defined to have at least one transaction occurring during evening trading hours. Table 2 shows that this variable is a separating feature for cluster 6, which trades at night more than twice as often as the other clusters.

The next two sections describe the results on trade-specific variables and exogenous variables.

	all	*1*	*2*	*3*	*4*	*5*	*6*	*7*
N	9271	2815	2725	2257	744	407	180	75
mean profit	-976	-628	491	-1686	1097	1949	-1137	-639
error of mean	990	150	100	850	550	2800	530	670
ratio	1.0	4.2	4.9	2.0	2.0	0.7	2.1	1.0
% profitable	48%	38%	51%	51%	61%	54%	39%	36%
days till exp.	71	69	72	71	70	129	23	28
% at night	4%	2%	2%	6%	7%	7%	13%	4%

Table 2: Summary of clusters, ordered in decreasing cluster size N. Information on variables not included in the clustering is provided: mean of profit per trade (in dollars), error of the mean (std(profit)/$\sqrt{N}$), ratio of mean profit and error of mean, percentage of profitable trades, mean opening time in days until expiration, and percentage of trades at night. The behavior of the trimmed mean profit is quite stable for all clusters. Trimming away 10% of the outliers does not reverse the sign of mean profit for any of the clusters (not shown).

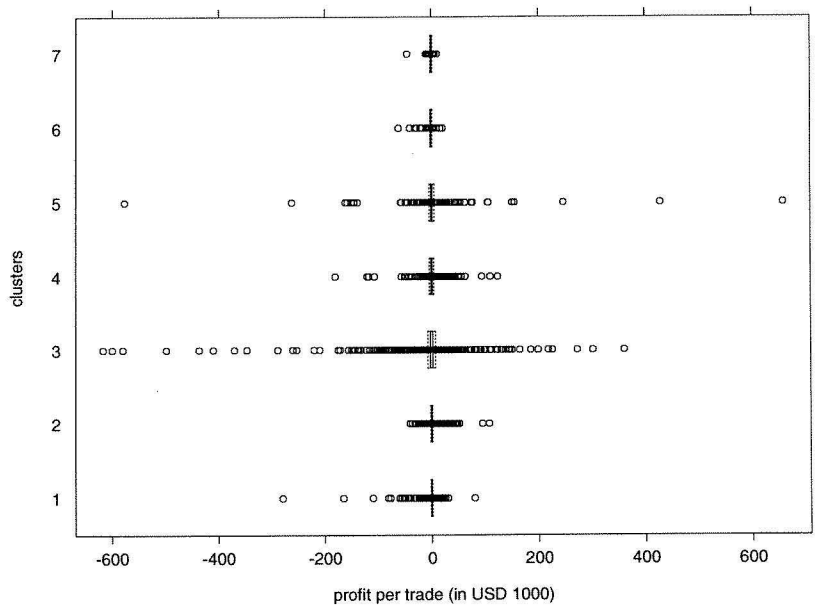

Figure 2: Box plots of profits. The width of the solid box (only visible in cluster 3) is equal to the interquartile distance, the difference between the third quartile of the data and the first quartile. The dotted lines extend to a distance of 1.5 times the interquartile distance from the median. Points beyond the dotted lines are plotted as small circles.

4.2 Trade-Specific Variables

Trade length is a clear distinguishing feature among clusters. We express length in logarithmic time. Figure 4 shows the conditional histograms. Table 1 indicates that although on average most trades are completed within a day, two of the three profitable clusters (4 and 5) stay in the market longer than a

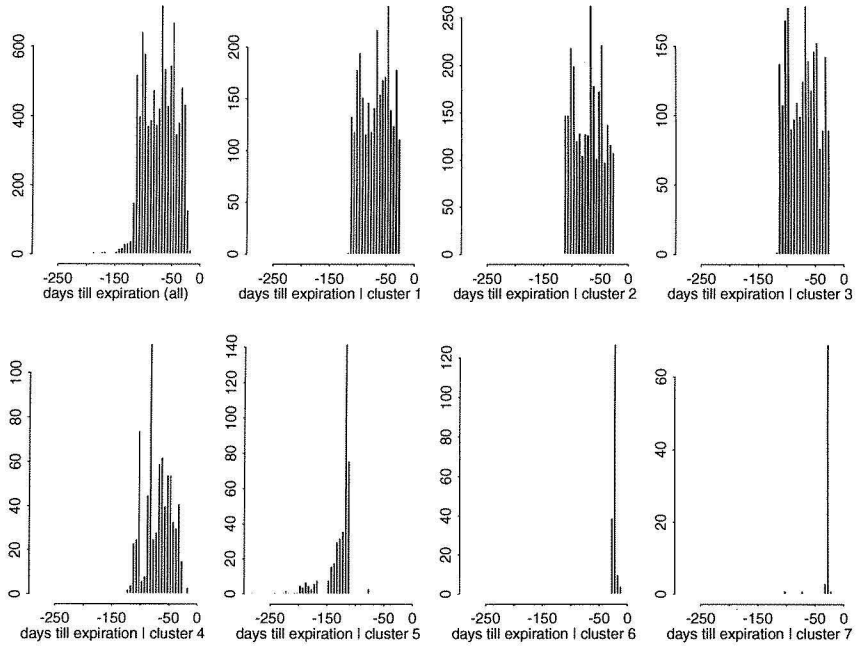

Figure 3: Histograms of trade opening time in days until expiration (with zero being the time of expiration). The first histogram in the upper left corner shows the distribution for all trades. The rest are histograms conditional upon clusters.

day. The conditional histograms clearly support this effect. Note that cluster 7 has a different distribution from all the other groups. This is a very quick trading style, with half of the trades carried out within a minute**.

Figure 5 shows that the relative opening position separates the clusters well. This variable computes the ratio of the opening position size relative to the maximum position size of a trade. Note that a value of 1 means the entire position was established in a single first transaction. Cluster 1 and 2 always open at full position size. Cluster 3, in contrast, only has one such case out of 2257 trades. The profitable clusters (2, 4, and 5), on the other hand, all have distributions peaking at 1.

4.3 Exogenous Variables

Figure 6 shows distributions of relative prices in the half hour prior to opening a trade. This variable indicates the movement of prices before a trade is opened. Cluster 7 has the most unique distribution from the remaining clusters. About 80% of trades in this cluster have a relative opening price of -1. Traders in this group adopt the strategy to either enter the market long when prices have been falling, or short when rising. Analyzing the relative price of first reversal of position with respect to the 30 minute local bar shows similar patterns (figure not shown).

The next three figures (Figure 7 through Figure 9) are histograms of volatilities. Figure 7 (1 week) shows that distributions for clusters 1 through 4 are similar to each other, but very different from clusters 5 through 7. For clusters 5, 6, and 7, trades do not occur at times with small price fluctuations. Note that

**Trade length of one minute could be an artifact of the CTR data, as the mechanics of CTI type 4 trading (non-exchange members trading off the floor) does not really permit such short a trade.

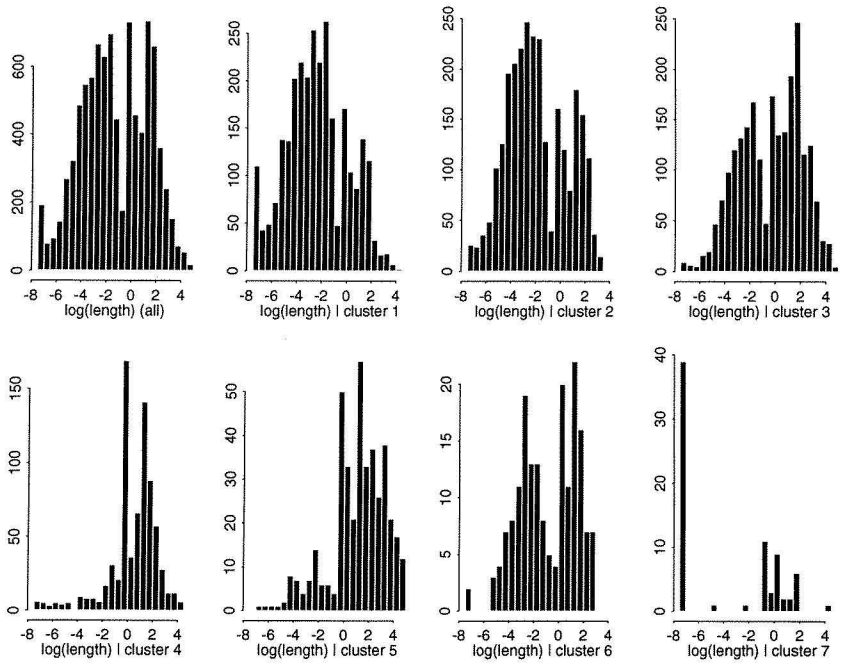

Figure 4: Histograms of logarithmic length of trade. The leftmost bar in the histograms corresponds to trades of length less than one minute, and zero corresponds to one day. Because of day and evening trading hours, the gap between -1 and 0, common to all histograms, is due to the fact that most trades are either less than 8 hours or longer than 16 hours.

cluster 5 has a much broader distribution than the others. This is the cluster with the largest expected returns. It is characterized on this dimension by trading only at high volatility. The distributions for cluster 6 and 7 are very peaked, a consequence of their tight distributions in trade opening time.

Figure 8 shows distributions of the same volatility variable, filtered on a much shorter time scale of 30 minutes. The shorter interval paints a very different picture for clusters 4 through 7. Cluster 4 is much more spread out, with more trades occurring at high volatility. Cluster 5 has a tighter distribution than before, and most of its trades take place at low volatility. The distributions of clusters 6 and 7 are separated further. More than half of the trades in cluster 6 are now seen to trade during low volatility times, an effect hidden with the long term filter.

Figure 9 shows that the main results from analyzing the distributions of short term volatility at reversal come from cluster 4. The hump of short term volatility at opening in Figure 8 gets distributed more evenly upon exiting. This cluster opens trades at low volatility but reverses positions at high volatility. The distribution for cluster 7 is even more peaked now, meaning this group exits at a higher volatility than the others. This is again the result of heavily concentrated trade opening time for cluster 7.

Figure 10 is the variable trade volume at opening. It tells an interesting story about clusters 5 and 6. A small volume coincides with an early transaction time relative to expiration, and a big volume implies the opposite. This distinction is clustered out between clusters 5 and 6. Cluster 5 is the most profitable. It trades very early on in the contract. Cluster 6 on average is the second largest loser, and it trades only during high volume times.

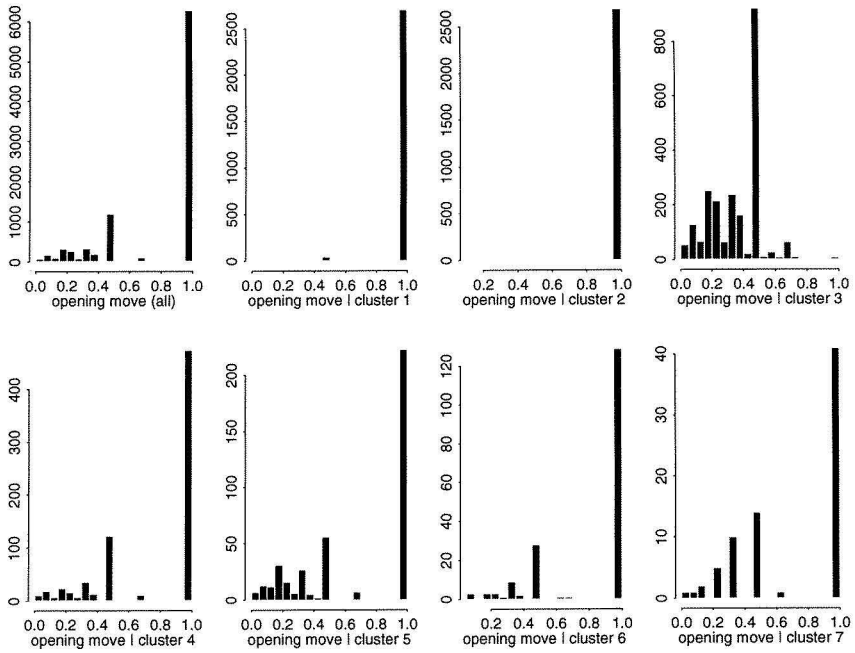

Figure 5: Histograms of relative position size at opening a trade. A value of 1 means the opening position is as large as the maximum position size of a trade. Note that except for cluster 3, all other clusters have a majority of their trades open at full.

5 Trade Entropy and Account Dynamics

One aspect of trading style analysis we are interested in is how trades within an account are related to trading styles. Do trades tend to pursue one trading style, or do they often employ different ones? We borrow an information theory concept, *entropy*, to answer this question. The entropy of an account j, H_j, is defined to be

$$H_j \equiv -E[\log p_i^j] \equiv -\sum_{i=1}^{K} p_i^j \log p_i^j \ , \tag{4}$$

where p_i^j is the probability that a trade in account j is in cluster i, and K is the number of clusters. The probabilities are obtained empirically. Entropy is an indication of how trades within an account are dispersed across clusters. On the one hand, if all trades in an account employed one single trading style, trade entropy would be zero. On the other hand, if no two trades employed same trading styles, entropy will be at its maximum. Entropy in this context is useful for checking model validity by ensuring that the same clustering results cannot be obtained through random assignments, and for supplying a gauge of success or failure for clustering algorithms.

Figure 11 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 0.74), the rightward shift of the baseline (with a mean of 0.89) confirms that accounts in the data set employ trading styles more often than just randomly.

The dynamics of trade movements indicate the relationship between trading styles and time. They are

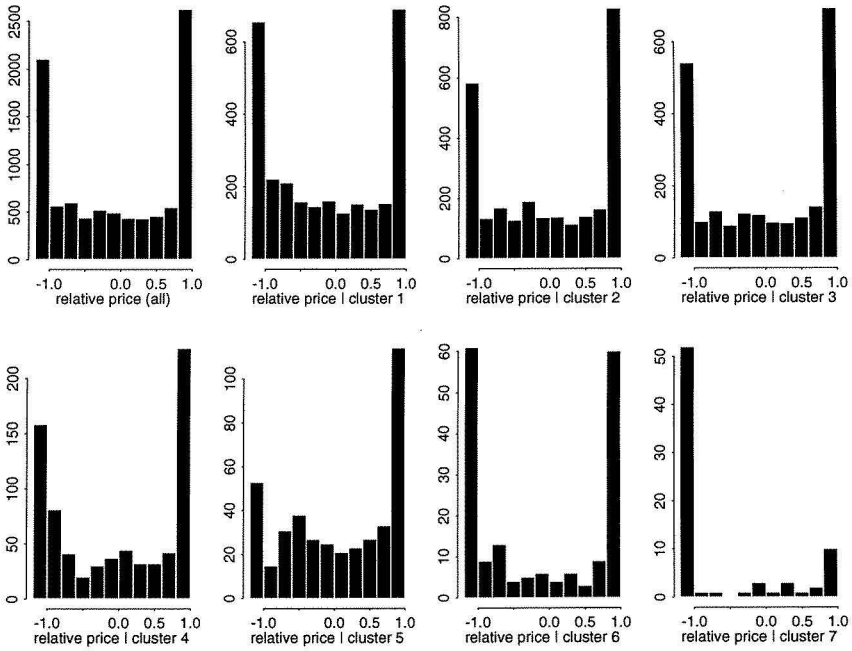

Figure 6: Histograms of relative price at opening a trade with respect to the minimum and maximum prices of the preceding 30 minutes. Because of the folding together long or short trades as discussed in Section 2.2, the rightmost bar contains two cases: open a long trade as prices have been rising, or open a short trade as prices have been falling.

To $j \Downarrow$ From $i \Rightarrow$	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	1040	740	522	188	4	31	22
2	773	1084	454	148	2	23	20
3	511	434	753	171	5	36	24
4	182	164	149	127	1	9	4
5	65	41	93	39	137	0	0
6	0	0	0	0	0	46	0
7	0	0	1	0	0	34	5
absorbing trades	236	258	274	71	258	0	0

Table 3: First order Markov transition table. Each cell contains the number of times a trade is in cluster i (row) and the next trade of the same account moves to cluster j (column). The last row contains for each cluster the number of trades that do not go to any other clusters, i.e. the last trade of each account. The same information is represented as a gray scale image plot in Figure 12.

described by the probabilities with which an account moves from one style to another. Table 3 summarizes this information in the Markov transition matrix. Each cell in the table counts how many times an account moves from cluster i to cluster j in sequential trades. Empirical probabilities are obtained through dividing each cell by its corresponding column sum. The outstanding feature is cluster 5. This is a profitable cluster, which tends to trade long, likes to hold, and prefers large position sizes. Figure 12 shows that trades in this

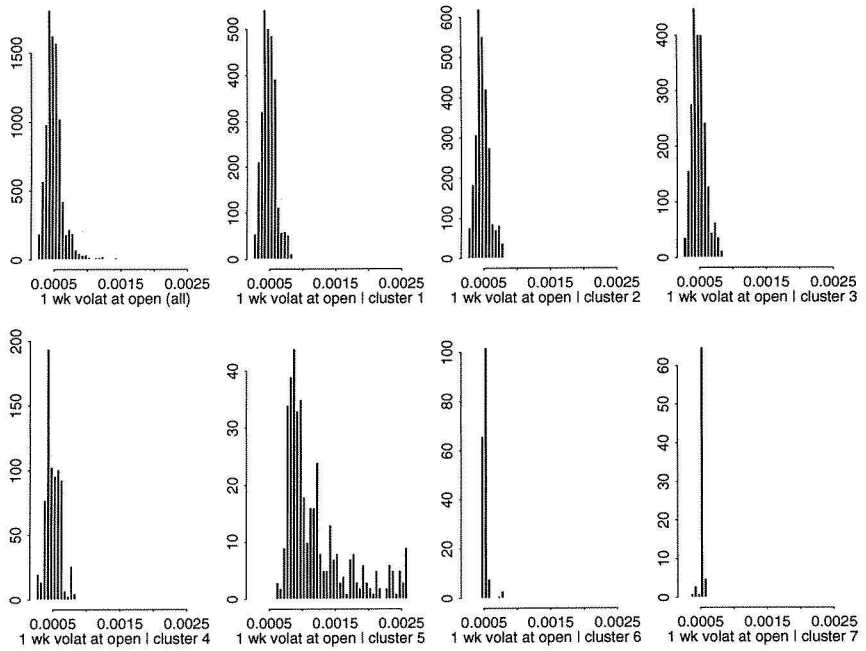

Figure 7: Histograms of volatility at opening a trade, exponentially filtered at 3000 active minutes (on average correspond to one week). The peaks of clusters 6 and 7 reflect the fact that most trades in these two clusters open at about the same time.

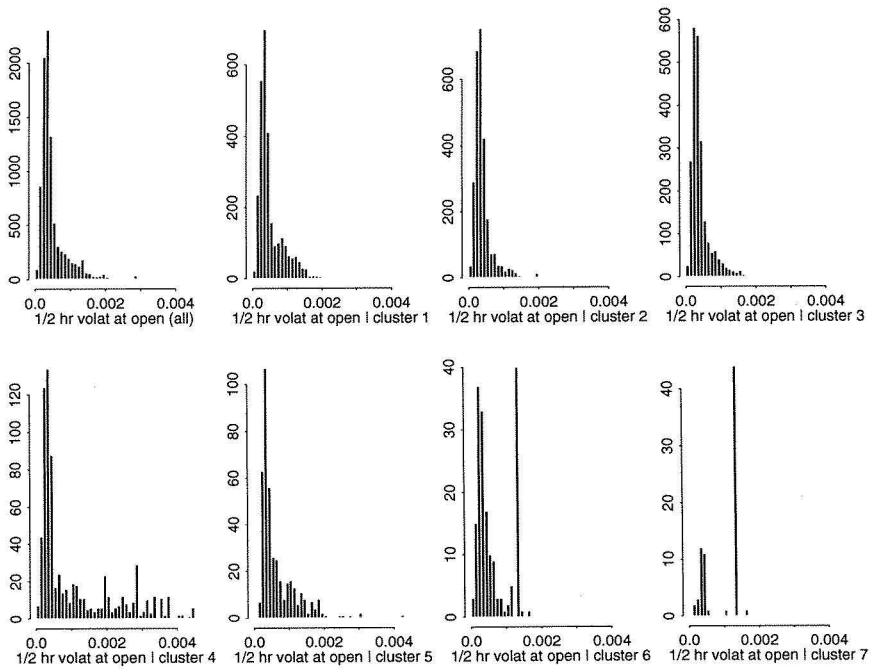

Figure 8: Histograms of instantaneous volatility at opening a trade, exponentially filtered at 30 active minutes. Peaks remain in clusters 6 and 7 for the same reason explained in Figure 7.

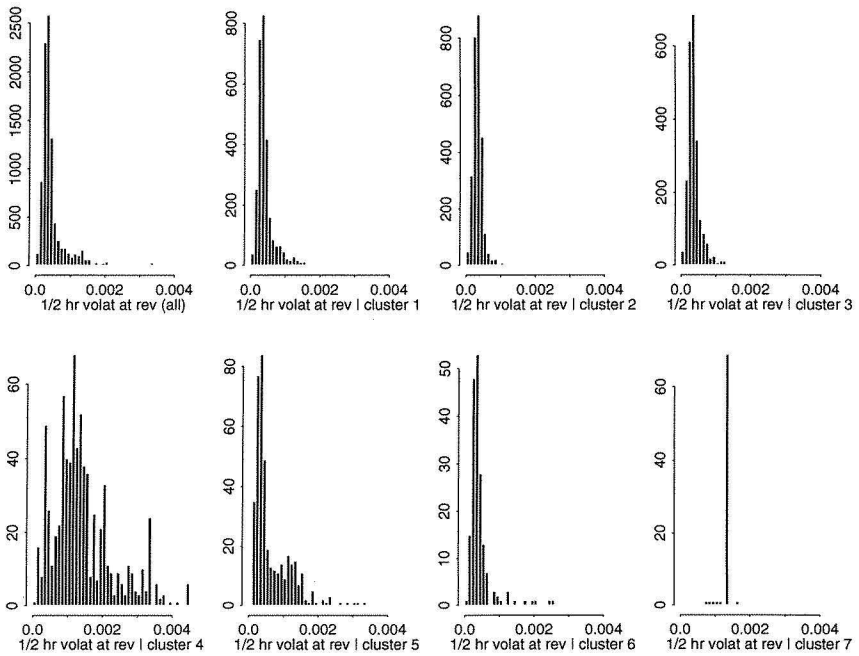

Figure 9: Histograms of instantaneous volatility at reversal, exponentially filtered at 30 active minutes.

cluster almost always (with probability 0.92) stay in the same cluster. Most trades in the other clusters, however, move into clusters 1, 2, and 3. Trades in clusters 1, 2, and 3 tend to stay in the same clusters too, but to a much lesser degree than cluster 5 (with probability 0.40, 0.44, and 0.38 respectively).

6 Conclusions

This paper takes a first step towards uncovering hidden structures in US Treasury bond futures. Our data set of the futures market consists of every single transaction executed with any clearing firm over a three year period. The enormous size of the data set enables us to analyze trading styles with the data-driven approach of clustering.

Two types of assumptions are made for the clustering: a data generating assumption as characterized by Gaussian prototypes, and a trading style assumption as defined by a set of trade-specific and exogenous variables. Clustering results revealed clear structures in the data. They confirmed the belief that the characterization of trading styles is closely related to the profitability aspect of trades, a variable that was not an input to clustering. In conclusion, results suggest that for CTI type 4 profitable trading styles

- hold trades longer than a day;
- trade at high volatility;
- trade early in the contract;
- persist in one trading style;
- open at full position as often as possible.

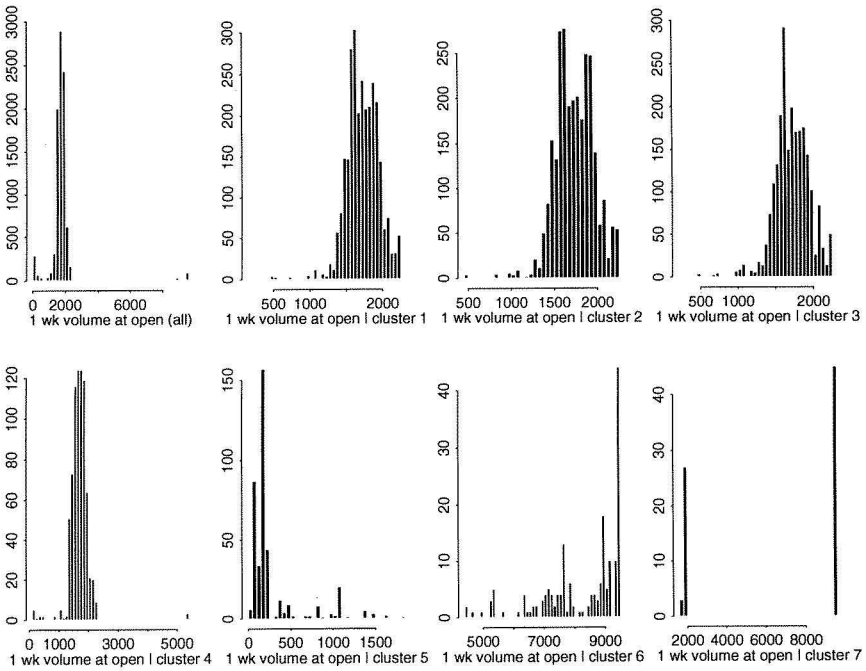

Figure 10: Histograms of exponentially filtered trade volume at opening a trade. Note that the x-axis varies in each histograms. It reflects the range of volume for each cluster. Cluster 7 has a very concentrated peak. Once again, this is because a majority of trades in this cluster open at about the same time.

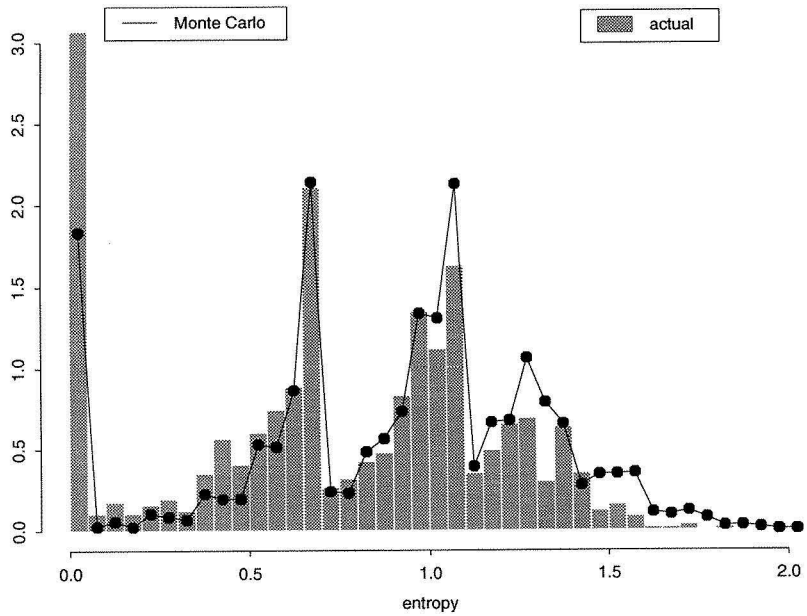

Figure 11: Histograms of trade entropy. The bar plot is the distribution of entropy results from clustering. The line plot shows the distribution of mean entropy values of 100 Monte Carlo sampling. The areas under the bar plot and the curve are both normalized to one.

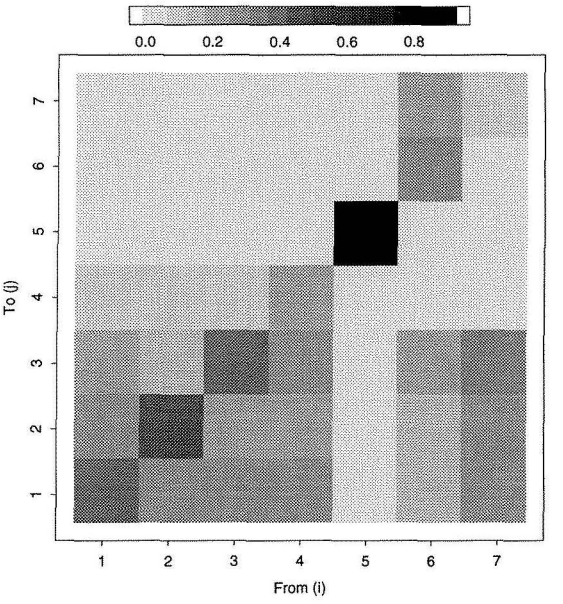

Figure 12: Image plot of the first order cluster Markov transition table. Each cell contains a gray scale representation of the empirical probability that a trade is in cluster i (the "From" axis) and the next trade of the same account moves to cluster j (the "To" axis).

In contrast, losing styles

- hold trades less than a day;
- trade at low volatility;
- trade late in the contract.

To validate the findings, we checked for spurious clusters. An outlier in volume (Table 1) was detected and replaced by a reasonable value. Clustering results on the cleaned data set indicate that this large shock in volume was responsible for the three outlier clusters of small size (cluster 8, 9, and 10).

A second validating step involved clustering with three 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 relaxed the assumption of an identity covariance matrix. We also transformed the input variables differently to suit them more closely to the Gaussian assumption. Two data generating models were used: Gaussians with diagonal covariance matrix, and Gaussians with full covariance matrix. The diagonal matrix helps us understand the impact of the spherical assumption on clustering results, while the full covariance guides us to the choice of input variables. These generalizations provided results consistent with those presented in this paper.

Taking into consideration the account identification information, we measured the trade entropy across the clusters, and examined the Markov transition properties of the accounts. Randomly assigning clusters to trades increased the entropy, confirming the existence of structure in the clusters. For all the clustering models and algorithms we tested, we found that the same Markov transition property, i.e., one cluster persistent in its trading style, remained.

This paper is a preliminary analysis towards an understanding of trading styles. Subsequent clustering will include a significantly larger portion of the data. We are also investigating the effects of additional variables on clustering, such as technical indicators and information related to the evolution path of trades. These additions will help us better interpret technical trading styles. We will also compare trading styles of various CTI types. Ultimately we hope to use the data set to examine the fundamental principles of behavioral finance and to enhance our understanding about how individual traders behave in the bond futures market.

References

Banfield, J. D. and Raftery, A. E. (1992). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**: 803–822.

Chernoff, H. (1973). The use of faces to represent points in k–dimensional space graphically, *Journal of the American Statistical Association* **68**: 361–368.

Gruber, M. (1972). Improved earnings forecasts through disaggregation of economic data, *in* S. Eilon and T. Fowkes (eds), *Applications of Management Science Models in Banking and Finance*, Gower Press, Ltd., London, England.

Heisler, J. (1997). Loss aversion among small speculators in a futures market, *Proceedings of 1997 International Association of Financial Engineers (IAFE) Conference, Boston.*

Lee, C., Shleifer, A. and Thaler, R. (1991). Investor sentiment and the closed–end fund puzzle, *Journal of Finance* **46**(1): 75–109.

Manaster, S. and Mann, S. C. (1996). Life in the pits: Competitive market making and inventory control, *The Review of Financial Studies* **9**(3): 953–975.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York.

Tukey, J. W. (1990). Data–based graphics: Visual display in the decades to come, *Statistical Science* **5**: 327–339.