Some Sources of Trading Volume

Tom Hyer
Akuna Capital

April 7, 2018

Abstract

A long-standing puzzle in financial economics is that of the volume of trading; financial models invariably predict volumes which are orders of magnitude lower than those observed in reality. We discuss some features of financial markets, which tend to increase the volume of trading for a given dislocation in price; we also provide reasons why these features can persist in real markets.

1 Introduction

This paper was inspired by a sort of challenge from John Cochrane [**1**]: "Volume is The Great Unsolved Problem of Financial Economics." In mulling this over, I discovered that I was holding two puzzle pieces that might contribute to the solution of this Great Problem.

First, I had the good fortune to work for a while with Tom Rubio, one of the giants of the Chicago Treasury market, in a role that allowed wide-ranging discussions of practical market making. He had a fondness for the specialized relative-value traders in the corners of the Treasury market, who make modest but steady income streams through mastery of a particular detail of the market. I realized that the operation of these traders would tend to increase trading volume, possibly by a substantial factor. Section 2 discusses this effect, and provides a toy model which can be solved analytically, to more formally illustrate the mechanism.

Second, I had been aware for some time of the efforts undertaken by large quant funds to obtain reliable non-market information sources; it occurred to me that, due to market incompleteness, trading based on price information inevitably increases volumes. Section 3 explores this effect, and again exhibits an analytic solution for a (different) toy model.

These two sections are largely independent, and perhaps should have been published separately. On the other hand, they both bear on the same problem; and it is frankly unlikely that I will have anything further to contribute to this question. So it seems best to bundle them together, and leave others to judge their value.

2 Mechanics of Relative-Value Trading

2.1 Motivating example

In the U.S. Treasury market, relatively few participants expect to profit from outright interest rate risk. Instead, the market is populated by a wide variety of detail-oriented traders, each of whom trades some bonds against others based on *relative* value considerations:

- Curve shape: sometimes through spreads, more often through calendar butterflies or more complex trades
- Issue date: trading of rolls, or of on-the-run against 1-old bonds, and so forth
- Deliverability against futures
- And so on.

These traders – or, more precisely, their high-frequency co-located software agents – form the bulk of the Treasury market. If we think of the sources of market randomness as a set of uncorrelated eigenmodes, the vast majority of traders are attempting to remain neutral in the first and even second mode, while working to profit from dislocations in less volatile modes.

What happens when a true directional dislocation, such as a rise in 10-year yields caused by a large bond issue, reaches the market? Each analyst sees it through his own lens, and each takes advantage of it by buying the cheap bond as part of a larger portfolio trade. These trades themselves cause smaller dislocations elsewhere, which collectively ripple through the market until a new equilibrium is reached.

2.2 A toy one-period model

Consider a market of M securities with prices p_i. We postulate the existence of market makers who make markets in each security, without reference to other prices; we assume their aggregate effect is such that any RV trader can trade a quantity v_i of the i^{th} security with a price slippage of v_i/ζ (thus executing at an average price of $p_i + v_i/2\zeta$ where p_i is the initial price) for some market-wide liquidity parameter ζ. This is not a realistic model, since the market makers charge no spread and always lose money; we use it only to quantify the actions of *relative value* (RV) traders in the same securities.

Suppose that RV traders possess some market information: the j^{th} RV trader knows the unique equilibrium value of some linear combination $h_j \cdot p$ of security prices. In our model, this knowledge is precisely correct and all RV trader estimates are consistent with an equilibrium price state $\vec{Q}$; but each RV trader will own only the portfolio for which he knows the outcome. We assume there are $N \gg M$ RV traders, and that their known portfolios h_j are uniformly distributed on the unit M-sphere. It follows that they span the space of possible portfolios.

Now suppose we start from an initial state $\vec{p} \neq \vec{Q}$, such as might immediately follow an exogenous market dislocation; and that each RV trader trades in sequence. Let $g = h_1$ be the known portfolio of the first RV trader. To maximize his returns, he will buy a quantity δ_1 such that

$$\sum_k (Q_k - p_k - \delta_1 g_k/\zeta) g_k = 0;$$

i.e., up to the point where his marginal return is no longer positive. This yields

$$\delta_1 = g \cdot (Q - p)\zeta/g^2 = \zeta g \cdot (Q - p)$$

.

This RV trader's actions serve to flatten out price errors parallel to g, or equivalently to project the price errors $p - Q$ onto the subspace orthogonal to g. Thus $(p - Q)^2$, a measure of the total market dislocation, decreases by an expected factor of $1 - M^{-1}$. In plain language, because of the breadth of the market, a single trader can only reduce the dislocation by a fraction of $O(M^{-1})$; the rest is outside his scope (as, *e.g.*, TED spreads are to Treasury traders specializing in liquidity premiums).

The expected absolute value of δ_1 is $\zeta u(M)|Q - p|$, where

$$u(M) = \Big(2\Gamma(1 + M/2)\Big) / \Big(M\sqrt{\pi}\Gamma(1/2 + M/2)\Big) \simeq \sqrt{2/\pi M}.$$

However, the resulting trading volume (summed over securities) is

$$\mathcal{V} = \zeta |g \cdot (Q - p)| \sum_k |g_k|;$$

because this is minimized when $Q - p$ is a basis vector,

$$E[\mathcal{V}] \geq \zeta |Q - p| E\Big[|g_1| \sum_k |g_k|\Big] = \zeta |Q - p| (2/\pi + (1 - 2/\pi)M^{-1}).$$

Thus this trader's expected trading volume $\mathcal{V} \simeq 2\zeta|Q - p|/\pi$. Crucially, this is largely independent of M – even though the reduction in the market dislocation is proportional to M^{-1}. As M grows, and each trader sees an insignificant fraction of a larger market, his own trading volume *does not decrease*.

This reasoning can be repeated for subsequent RV traders, whose collective action will cause $p \to Q$; but in the process the total trading volume generated will be *proportional to* M. This is the motivation for our choice of this toy model: it illustrates the general principle that cross-asset relative value trading multiplies the total market volume, possibly by a large factor.

2.3 Variant of the Toy Model

We have assumed for simplicity that the RV portfolios h_j are distributed so that each RV trader takes some position in every security. This simplifies the analytical treatment by making the problem isotropic in the space of prices, but means that every trader manages a large and diverse portfolio, which seems unrealistic.

To quantify this effect, consider an alternative model where each $h_i = e_k$, the unit basis vector for some $k \leq M$. This is a variant of our model where the RV traders become "absolute value" traders, each with complete knowledge of the equilibrium value Q_k of a single security.

If the initial price dislocation is isotropically distributed, then M relative value trades – each eliminating the dislocation in one price p_k – will occur, each with expected size $M^{-1/2}\zeta|Q - p|$. Thus the total trading volume is now proportional to $\sqrt{M}$, rather than to M.

2.4 Effect in real markets

In reality, of course, RV traders form portfolios with more than one security but far less than the entire market; we denote by K the typical portfolio size, where $1 < K \ll M$. We assert that the amount of trading volume generated by an exogenous price dislocation is approximately proportional to $\sqrt{KM}$, though we are able to offer proofs only for the extremes $K = 1$ and $K = M$.

An initial price distortion in the bond market diffuses through a variety of calendar spreads, liquidity spreads, treasury/swap and treasury/eurodollar basis trades, and other such strategies, until its effect is spread over a wide range of instruments. We believe $K \sim 3$ and $M \sim 15$ are reasonable approximations to the connectedness of this market. Based on this, we expect that trading volumes might be increased by a factor of roughly seven, compared to what would be expected in a world where risk is immediately allocated to its final owner.

In the equity markets the situation is different, and not so closely related to our simple model. Many statistical arbitrageurs rely on mean-reversion relations between pairs or within small groups of stocks; however, each might have a different idea of the relevant "control group" for a given stock. In addition, block trading techniques which hedge a position with a large basket can spread a localized price distortion widely. Thus estimation of K and of M become much more difficult. We must ask ourselves:

- What is the typical portfolio size used by value-aware traders (as opposed to single-stock specialist market makers)? This is our K.
- How large is the universe of prices which will respond to a dislocation? This is our M.

We are willing to suggest that $K \geq 2$ and $M \sim 20$ are defensible choices; if this is correct, equity market volumes will be increased by a similar factor to that in rates markets.

2.5 Against internalization

To either an academic or an ambitious practitioner, this trading appears to be mostly waste. If trading volumes are increased by this mechanism, then a mature financial industry should find ways to remove the wasted trading, reducing volumes to a much lower level.

One might expect that traders would seek to amalgamate multiple strategies into a combined strategy which, by eliminating wash trades in its execution, would reduce transaction costs (and trading volumes). It is worth considering how practitioners address this very issue.

The largest unified statistical arbitrage firms strive for maximal amalgamation, with an "optimizer" which chooses which trades to execute in the market by considering the combined pricing signals from many internal strategies. The enterprise-level intellectual efforts at these firms are not focused on their optimizers, but on the "simulators" which attempt to test the validity and incremental contribution of new strategies, and on other attempts to accurately measure the performance of the strategies.

The difficulty of this problem appears to increase with scale. Bear in mind that market impact is the hardest thing to test in simulation, as it is not about what the market did but what it *would have done* in the presence of a different trading strategy. And strategies are necessarily chosen based on performance in simulation, rather than in a real market.

As the breadth and sophistication of simulations increases, so does that of potentially parasitic strategies. To suppose that the arc of history bends toward perfect simulation is overconfidence that some favored methods are perfectible; there is no factual basis for this belief. To date, the real-world record shows that the largest quantitative funds have grown more through diversification than by perfecting their original mission.

In the bond trading market, the unified approach is also used, but an alternative is a looser consortium: a *house* with internal traders who can execute at the *house bid/ask* which is artificially tight compared to that available in the external market. The house profits when internal trades cancel, but must pay to execute those that do not. The problem of measuring a strategy's value is not solved at the house level, but delegated to the internal traders who must decide whether their strategies are worth trading.

In the house model, trading itself supplies the feedback which allows strategies to be evaluated, improved, and eventually discarded. This business model is still not completely scaleable, because the actions of the internal traders are transparent to the house, and the most successful traders will fear that their strategies may be reverse-engineered. Those whose strategies are robust enough to remain profitable at market spreads will prefer the relative anonymity of exchange trading. Conversely, some strategies may prosper by producing (or closely following) toxic order flows and inflicting losses on the house.

This is far from a complete discussion of internalization. But there seems to be a widespread assumption that maximal internalization is an inevitable feature of mature markets. I believe this is a "Second Foundation" error [**2**], an unmerited faith that everything can be made orderly. The empirical sample, while small, does not lead to this conclusion.

3 Price Signals in Incomplete Markets

We have already alluded to the central role of the price signal, not just to be compared to some trader's estimate of "fair value" but as a unique source of trustworthy (because of its resistance to manipulation) information about the state of the world. Neal Stephenson [**3**] provides an analogy from the natural world: "... insects here see you as a big slab of animated but not very well defended food. The ability to move, far from being a deterrent, serves as an unforgeable guarantee of freshness." In finance, price signals from trading are valuable precisely *because* they are involuntary.

However, by their nature, prices combine information from many sources, and are also undeniably subject to short-term fluctuations. Many trading strategies make trading decisions which are informed by the prices of securities not traded by that strategy; such inputs induce fluctuations in the trading decision, which will generally cause increased trading.

The situation is further complicated by the incompleteness of markets. We are in the habit of saying that fundamental analysis can state when a company is undervalued; but this is a somewhat lazy approximation based on the pretense that every price-relevant factor has been analyzed. In practice, all such analysis is incomplete: it cannot cover all possible sources, or even all idiosyncratic sources, of change in the stock's value. Thus even a single price incorporates effects outside the scope of any given analysis.

3.1 A toy diffusion model

Let S_t be the price of some tradable instrument which in a risk-neutral world would follow a Martingale process (*e.g.*, a futures contract). We assume that the risk-neutral price process for S_t is a Brownian motion with unit volatility. Suppose that at time T_0, thanks to careful fundamental analysis, one trader knows that with probability $p_0 < 1$, the evolution of S_t over the time interval $[T_0, T_1]$ will have some nonzero drift μ. However, the binary fact of whether this scenario is realized will not be knowable to the trader until $t > T_1$ (if ever).

We also suppose that the trader seeks to maximize the figure of merit $E[X] - \lambda \mathrm{Var}[X]$ for his profit X, where $\lambda > 0$ represents his risk aversion. At time T_0, the resulting optimal holding Δ does not depend on S.

However, for $t \in (T_0, T_1)$, the arrival of new price information will allow the trader to update his estimate of p, and thus to adjust his optimal portfolio. The Bayesian posterior estimate

$$\hat{p} \equiv E[p; S_t] = p_0 / \Big(p_0 + (1 - p_0) \exp\{\mu^2 t/2 - \mu(S_t - S_0)\} \Big)$$

drives a new optimal position of $\mu\hat{p}/2\lambda\phi$, where

$$\phi \equiv 1 + \hat{p}(1 - \hat{p})\mu^2(T_1 - t)$$

incorporates variance contributions from the Brownian driver and from the uncertain drift.

In this model, the trader demands rather than supplies liquidity, since his confidence (as measured by $\hat{p}$) increases as the price moves in his favor. The total variation of $\hat{p}$ is unbounded, but the expected increase in profits quickly approaches a finite limit. In Table 1, for a few values of p_0 and $\mu\sqrt{T_1 - T_0}$, we show numerical results for the trading volume and (in parentheses) the expected profit, as a function of the rebalancing interval δ_t. In each case, the displayed results are divided by the corresponding quantity when $\delta_t = 1$; that is, we show the ratio by which trading volume (and profit) increase due to rebalancing.

Table 1:

$p_0,\ \mu\sqrt{T_1 - T_0}$	$\delta_t = 0.25$	$\delta_t = 0.1$	$\delta_t = 0.03125$	$\delta_t = 0.01$
0.5, 0.1	1.04 (1.002)	1.06 (1.002)	1.11 (1.002)	1.20 (1.002)
0.5, 1.0	1.40 (1.190)	1.63 (1.228)	2.12 (1.245)	2.98 (1.251)
0.1, 1.0	1.64 (1.360)	2.02 (1.440)	2.83 (1.478)	4.25 (1.489)
0.9, 1.0	1.12 (1.041)	1.20 (1.049)	1.36 (1.052)	1.65 (1.053)

The gain from trading is greatly increased only when p_0 is fairly small – *i.e.*, when the trader must rely on the market to verify his estimate. In such a case, the result of optimization is a *trend following* strategy. Conversely, when $p_0 \sim 1$, the main value of the new market information is in exiting a trade whose premise has been falsified by later market events.

Trading volumes, of course, increase without limit as the update frequency increases. However, even for fairly large update intervals there is already a significant increase in the traded volume. Thus our toy model, though simple, illustrates an important mechanism: by using the market as a source or validator of trading signals, we create signals which are themselves volatile.

4 Conclusions

To understand the volume of trading, we must first understand in more detail the mechanisms of price discovery in markets which are simultaneously *incomplete* and *highly coupled*. Coupling between markets occurs when relationships between prices are better known than any individual prices (usually because they are more stable). Even a simple model shows that the propagation of information via these coupling relationships generates a manyfold increase in volumes.

Incompleteness of markets means that market information, while rapid and precise, is inevitably adulterated. A trader may gain fundamental knowledge of only part of any market price, and must accept the additional components of that price as a source of unpredictable noise. This requires portfolio adjustment, with a concomitant increase in trading volume, even when treating a single security in isolation.

A rational trader cannot know the correct price for even one instrument; this is a calculation problem beyond the capability of any single market participant. Instead, traders are the equivalent of computational nodes in a distributed system of value calculation; and trading is the *involuntary and unforgeable* signal by which trustworthy information propagates between nodes. To assume that these nodes need not exist – that some more perfect equilibrium would somehow allow risk to flow directly to its final owner – is exactly equivalent to assuming away the problem of volume.

To quote Cochrane again: "Information seems to need trades to percolate into prices. We just don't understand why." I would offer an answer in two parts:

1. Trading reliably (because involuntarily) leads prices toward equilibrium. The price stops moving exactly when those with superior information no longer have an incentive to trade. Nothing else can work, because any other signal could be abused to enable profitable trading.

2. As we have just shown, the process of price discovery entails trading whose volume can far exceed that exchanged by any single participant. Trading based on price signals further increases volume, because such signals are inherently noisy.

Understanding the price calculation process is perhaps not sufficient to fully understand trading volumes in the real world; but it is a necessary and substantive step in the right direction.

5 Acknowledgements and References

References

[1] John Cochrane, *Volume and Information.* At https://tinyurl.com/yax72fje; October 2016.

[2] Isaac Asimov, *The Foundation Trilogy.* Gnome Press, 1951-53.

[3] Neal Stephenson, *Cryptonomicon.* Avon, 1999.

Roger Lee, of the University of Chicago, offered helpful comments and corrections on a draft of this work. I am also indebted to Nick Dunbar and an anonymous reviewer for help in refining the arguments presented here.