The Empirical Behavior of Commodity Prices at High Frequencies

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Joseph Dennis Mahon

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Glenn D. Pederson, Adviser

December 2016

Acknowledgements

This thesis went through many drafts and took many hours of work. I benefitted from a lot of help along the road. I'm grateful to my adviser, Glenn Pederson, for his guidance and patience. It would never have been possible in the first place without the generosity of Kent Horsager, who provided me with the data that I analyzed. And I'm thankful to my colleagues at the Federal Reserve Bank of Minneapolis for providing a stable and supportive workplace, and particularly to Dave Fettig and Toby Madden, both of whom retired before I finished.

I never would have gotten it done without the encouragement and inspiration of my wonderful wife, Emily. Proud of me?

Dedication

This thesis is dedicated to Dennis Mahon, who taught me that nothing in life is free, but we're free to be free.

Abstract

I analyze the intraday prices and volatility of three exchange-traded agricultural commodities at high-frequencies. The dataset spans only a few trading days, but includes every transaction on the exchanges down to the microsecond. After reviewing some of the empirical issues with high-frequency data and the methods in the literature devoted to estimating volatility from them, I develop an approach that can disentangle the continuous or integrated volatility from discrete jumps in the price process. I illustrate applications to estimating spot volatility and in volatility forecasting, and conclude that these methods provide useful refinements to existing approaches.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

This thesis attempts to fill some of the gaps in our knowledge about commodity market behavior at high frequencies. It deals with the application of time-series econometrics to newly-available intra-day financial data, with a particular focus on agricultural commodities. The data are of interest to exchange operators, traders, regulators and academics for a variety of reasons, to be explained in this introductory chapter. However, analyzing of these data generates new challenges, which are surveyed here as well, along with a description of the methods I use to deal with them. I conclude by discussing several applications for which this analysis may be useful.

1.1 The advent of electronic trading

With a few exceptions, the open-outcry system of pit trading on financial and commodities exchanges is extinct. Over time, exchanges have moved to electronic trading systems in which traders submit offers and orders from (often remote) computer terminals. The orders are processed by the exchanges using algorithms depending on when they are received, the price offered, and the trading rules established by exchanges.

This development has in turn given rise to new trading schemes that operate on very short timescales. Investors, always looking for speculative profits, understood that having direct market access and the ability to act on new information nearly instantaneously could open up short-term arbitrage opportunities. Discrepancies between the prices of assets and their derivatives (between spot, futures and options prices in the case of commodities) could, at least in theory, be exploited through rapid portfolio management and hedging. Likewise, correlations may exist between different asset prices or with other market indicators (e.g., bid-ask spread or order arrivals) that allow acting on new information before it is fully assimilated in the market price—the equivalent of technical analysis at light speed. There is now a large and growing algorithmic trading industry making use of high-speed computing and machine learning methods to profit from such schemes (Johnson, 2010, provides an overview).

These developments have attracted critical attention from regulators, market participants and financial journalists. One concern is that, rather than improving liquidity, high-speed trading has made these markets more volatile (see e.g., Lowenstein, 2012). This could translate to greater risk borne by non-financial users: farmers and food processors for agricultural commodities; transportation firms, electric utilities or manufacturers in the case of energy or metals; sovereign reserve funds in the case of currencies. Incidents such as the "Flash Crash" of 2010, when a cascade of similar algorithms executing orders in the same direction led to a large correction in equities markets, have also raised concerns about the systemic risks this activity may present.

Further, electronic trading creates new possibilities for the manipulation of market prices. A simple example is for a firm to "paint the order book" by submitting several orders near the exchange maximum to cause others to anticipate a large trading volume, thus bidding the price up or down, and then cancelling the order prior to its execution, taking advantage of the change in price to liquidate previously-held positions. This kind of simple example may be easy to spot, leading to an arms race of trading strategies. The fact that by some estimates 90 percent of orders from high-frequency traders are cancelled prior to execution (See Dodd, 2010) has been cited as evidence that this kind of manipulation is widespread (though high cancellation rates may be due in part to traders revising their trading strategies in the wake of unfavorable price movements). Commodities offer another angle on this issue. Commodities and their derivatives are traded for production-risk management purposes in addition to their role in the portfolio diversification and speculation roles played by other assets. These markets are also "thin" relative to other financial markets—much of commodity and commodity derivative trading occurs over the counter, and the exchange traded contract volumes are much lower than for other assets, making them potentially more prone to manipulation. The movement of speculative finance into commodities markets thus has consequences well beyond those markets. So it is an interesting question how their fluctuations compare to other financial markets.

This activity is of obvious interest both to regulators and exchange operators who may wish to improve the market rules, as well as to traders hoping to profit by detecting

such strategies in action. Fortunately, electronic trading has also made available a vast amount of data that may shed light on the problem.

1.2 The challenge of high-frequency financial data

With the advent of electronic trading, an enormous amount of high-frequency, intraday data on prices, offers and orders is now available. These data present new challenges to researchers, to be described in this and the next section. The standard tools of time-series analysis used in financial econometrics to study daily price fluctuations (e.g., ARIMA and GARCH models) are of limited use in analyzing high-frequency data. A large and growing literature in financial econometrics has sprung up around this topic (see Engle and Roberts, 2009, for one review). However, much of the applied research in this area has focused on equities and exchange rates, which is another motive for my focus on commodities.

Intraday financial data has grown in availability as computers have become more widespread. Prior to electronic trading, it was typically collected by "sampling" the market at regular intervals, recording prices and other variables at a given frequency. One problem this immediately introduces is that actual market transactions are spaced irregularly throughout the day. For example, it is well known that transactions exhibit a u-shaped, diurnal pattern, with the volume of transactions higher near the start and close of trading than at mid-day. Also, transactions are known to "cluster," with many transactions centered around an event (the arrival of news or a large order) and extended lulls with little activity.

This clustering creates a dilemma for data collectors. Using shorter timer intervals allows more detail to be captured during periods of high activity, but also records more null observations (zero values when using the returns series) during low-activity periods. With longer time intervals there are fewer null observations, but the time series is "rougher," containing more sharp jumps (Anderson, Bolerslev and Diebold, 2003). In either case, the observed shocks follow a heavy-tailed distribution.

It has been recognized since Mandelbrot's (1963) path-breaking research that daily price fluctuations, returns and trading volumes in asset markets follow heavy-tailed

"power-law" distributions, with a strikingly consistent set of parameter values. This poses a challenge to traditional asset pricing theories that assume pricing errors are normally or log-normally distributed.

Because these assumptions appear at odds with observed volatilities, numerous approaches have been developed that attempt to model and forecast returns and volatility in a fashion that captures these features, including ARCH/GARCH, stochastic volatility and jump-diffusion models. These models allow shocks to vary over time and cluster together, which also causes the time-aggregated variability of a series to be heavy-tailed, even though the moment-by-moment shocks (the "integrated volatility") are normally distributed.

But at high frequencies, shocks appear even more heavy-tailed (see chapter 2), and the above-mentioned methods become less useful. Further, while these models have successfully been applied to analyze day-to-day returns and volatilities, for intra-day trading they miss important details to be discussed in the next section.

Alternatively, electronic trading makes it possible to collect data in real time. Instead of sampling the market at regular intervals, prices and other variables can be recorded as transactions come in. Such data capture the trading process in real time, but the observations are irregularly spaced in time. Since standard time-series models are based around regular time intervals, methods must be developed that take this into account. Engle and Roberts (2009) survey some of these "tick-time" models.

Irregularly-spaced data thus come with their own set of difficulties, but they also can shed light on some of the other problems presented by high-frequency data and electronic trading. This is the subject of the next section.

1.3 Market microstructure

The time-series methods typically used for modeling day-to-day fluctuations in financial markets have a strong justification: they are rooted in economic theory. In an efficient market, an asset's price reflects all available information, and the arrival of new information that may alter prices is by definition unpredictable. An autoregressive model (such as a random-walk) in which future prices are represented as a function of current

prices plus some stochastic shock is thus a natural framework, and by fitting such a model to the data I can get estimates of an asset's expected return and volatility useful in portfolio analysis.

However, equilibrium asset-pricing models with a no-arbitrage condition may provide a poor guide to understanding the intra-day data from real-world financial markets, even if those models are an accurate description of the underlying behavior of market participants. One reason was already mentioned in the previous section: data from markets are often sampled at regular intervals, but trading occurs at irregular intervals. Another reason is that real-world exchanges price assets in discrete ticks (cents, eights of a dollar, etc.). If underlying asset prices follow a diffusion process, these *nonsynchronous trading* and *discrete price change* effects will introduce noise to estimates of volatility.

The real-world trading process introduces numerous other sources of noise: order limits and the strategic aspects of orders discussed above, asymmetric information between traders, the *bid-ask bounce* that causes apparent transient price oscillate between bids and asks for trades involving market makers, etc. A large literature on market microstructure consists of attempts to model and measure such effects (See Hasbrouck, 2007 for an overview).

High-frequency data offer the equivalent of a microscope, allowing us to observe the trading process in real time. For example, it is possible, at least in principle, to observe the effect of new information on prices, which is especially interesting if participants are asymmetrically informed. However, the problem of microstructure noise becomes more acute as the frequency of observation increases.

Moreover, as I will demonstrate in Chapter 2, the tick-by-tick data often contain instances of sudden substantial changes in price or "jumps." These jumps can contribute to observed volatility, though they are distinct from the continuous or "integrated volatility" embedded in the traditional Brownian motion model of asset prices.

For these reasons, Ait-Sahalia and Yu (2008) note that the approach to estimating volatility at high frequencies, "consisted in ignoring the data sampled at the highest frequencies out of concern for the noise that they might harbor, and sample instead once every 15 or 30 minutes" (p. 2).

But such subsampling possibly discards much potentially useful information. Therefore, another approach has been to develop volatility estimators that are robust to microstructure noise.

1.4 Objectives

While high-frequency data present new promises they also pose new challenges. Researchers have recently begun to tackle these econometric challenges, but much of the existing research is focused on equities and exchange rates. Commodity markets provide a novel application of these methods.

Therefore, my primary objective in this study is to *analyze volatility in high frequency commodity exchange data* by adapting methods in the existing literature.

The general question I wish to explore in this thesis can be phrased: Is there information in high-frequency commodity price movements that can be used to estimate intra-day volatility and forecast future volatility? These are of interest for several practical purposes, and a large literature exists on various estimators and their properties, but most of the published research focuses on equities and exchange rates.

Answering this question involves dealing with the econometric challenges described above. How do the issues of market microstructure noise and the appearance of discontinuous "jumps" complicate attempts to uncover these relationships?

Thus a second objective is to *identify jumps in the data* and to disentangle the jump component in order to *analyze volatility in separately from jumps*, by developing methods that allow inference in the presence of noise and jumps. It is also of interest to *quantify jumps and noise* as well, since that is possibly using these methods.

Because this is a thesis in applied economics, I want to show how to make use of these estimates. My third objective is to *measure instantaneous volatility and forecast future volatility* using my estimates. These "spot volatility" measures and volatility forecasts are of interest to market participants because of the relation to portfolio management and option pricing, as I will illustrate in Chapter 5.

Financial econometrics is typically parametric, that is, it fits financial data to a statistical model to obtain estimates of the parameters of interest, e.g., lag coefficients,

shock variance, etc. The choice of a model may or may not be suggested by some economic theory. The virtue of such methods is that they make use of asymptotic theory to obtain estimators that are unbiased and efficient in the mean squared error sense.

Due to the complexity of the trading process at high frequencies, I am less confident in the choice of models generating the data, for the reasons outlined in the preceding sections.

1.5 The approach of this study

I therefore would like to use non-parametric methods that make up for what they sacrifice in efficiency with robustness to model misspecification. A variety of nonparametric methods for estimating volatility of an asset at high frequencies known as realized volatility (RV) estimators involve summing intraday squared returns. In theory, as the length of the sampling interval goes to zero (i.e., as the data become higher frequency), these estimators converge to the actual volatility. (Anderson *et al.*, 2003).

However, this result requires the observed data are the "true" underlying asset returns, i.e., they ignore microstructure noise. Actual RV estimates of volatility increase rapidly as the frequency of sampling increases, indicating microstructure noise is significant (Ait-Sahalia *et al.*, 2005). One way this problem has been treated is by subsampling the data, as mentioned above, but this potentially throws away information that could shed light on the high-frequency trading process.

Further, Fan and Wang (2007) show that RV methods also become increasingly error-prone when the data contain discontinuous jumps, as is common in financial time series. Disentangling jumps and volatility is another dimension of the problem. It is of interest both to improve volatility estimates and to estimate the size and variability of jumps, in order to determine their pricing implications. A recent survey of this literature is Ait-Sahalia and Jacod (2012).

In this paper I follow a method similar to that of Fan and Wang. This method can analyze both jumps and the underlying volatility, and can be used to estimate market microstructure noise. This model will be discussed in detail in the next chapter but I will describe it briefly here.

The first step in the analysis is to identify jumps in the data. This is accomplished through the use of a discrete wavelet transform. Wavelets are a class of functions that have found use in many areas of applied mathematics such as data compression, image analysis, bioinformatics and seismology. In the time series context they can be used in filtering, similarly to Fourier analysis.

But unlike Fourier analysis, in which an entire stationary time series is decomposed into frequency components, wavelets can analyze the variation in a signal at multiple time and frequency scales or "resolutions." This makes them useful for describing non-stationary processes, such as those with jumps and time-varying volatility that I are interested in. Using a higher resolution allows us to easily locate the jumps in a large data set.

Once the jumps are identified, I can study them in isolation, estimating the jump sizes and variability. Also, the jumps can be removed from the remaining data and I can use existing RV estimators (including a wavelet-based RV estimator proposed by Fan and Wang) to model the integrated volatility of the commodity price series.

I apply the foregoing methods to a set of high-frequency commodity market observations obtained from a private source. The data cover intra-day trades on the Chicago Board of Trade for corn, wheat and soybeans. The dataset includes prices as well as bids and offers several levels deep.

With direct electronic access to the exchange, the data were recorded with every transaction, rather than at regularly-spaced intervals. Such data are sometimes referred to as ultra-high-frequency, and provide rich details on the trading process that can be used to estimate market microstructure noise. The data are thus subject to the modeling challenges of irregularly spaced data, which my methods can accommodate. They can also be subsampled to conduct analysis using constant interval methods. The data are discussed in greater detail, along with descriptive summary statistics, in the beginning of the next chapter.

Volatility analysis occupies a central role in asset (derivative) pricing and in portfolio selection. Therefore, my results have a number of applications, two of which I demonstrate.

The remainder of this thesis proceeds as follows. In the next chapter I provide a description of my data set, along with some exploratory time series analysis. In the following chapter, after reviewing some of approaches developed to estimating volatility from high-frequency data, I detail the methods used in my analysis. The fourth chapter contains my statistical results and interprets them in the context of the model and the markets in which I am interested. I illustrate some applications in the fifth chapter. In a brief concluding chapter I summarize and discusses future directions for research.

Chapter II

Data

In this chapter I describe my data set. After detailing the source of the data, I provide an initial exploratory data analysis based on graphics and summary statistics to convey the stylized facts.

2.1 The data set

2.1.1 Commodities exchanges

High-frequency data from the Chicago Board of Trade was obtained for three commodities: corn, soybeans, and wheat. The data series include: prices, quantities traded, bid and ask quotes, and order book information. For each commodity the asset being tracked is the first expiring futures contract. Therefore, the price is close to the spot price of the underlying commodity. The time-series for each commodity covers six trading days: Friday, August 26, 2011 through Friday, September 2, 2011. The exchange operated from 9:30 a.m. to 1:15 p.m. central time.

Table 2.1: Example of high-frequency data

Time	Commodity	BidPrice	AskPrice	LastPrice	LastQty	TotalQty	BidBook	AskBook
2011-09-02T09:30:00.030771	CBL ZC 1112	7.4	7.4	7.4	10	10	1982	1210
2011-09-02T09:30:00.041446	CBL ZC 1112	7.4	7.4	7.4	3	13	1982	1210
2011-09-02T09:30:00.055453	CBL ZC 1112	7.4	7.405	7.4	5	18	177	63
2011-09-02T09:30:00.090912	CBL ZC 1112	7.4	7.405	7.4075	1	19	194	70
2011-09-02T09:30:00.107487	CBL ZC 1112	7.4	7.405	7.405	1	20	194	70
2011-09-02T09:30:00.116119	CBL ZC 1112	7.405	7.4075	7.405	1	21	108	68
2011-09-02T09:30:00.163522	CBL ZC 1112	7.41	7.4125	7.4125	1	22	125	114
2011-09-02T09:30:00.283866	CBL ZC 1112	7.4075	7.41	7.41	3	25	110	77
2011-09-02T09:30:00.342490	CBL ZC 1112	7.4075	7.41	7.4075	1	26	124	72
2011-09-02T09:30:00.416437	CBL ZC 1112	7.405	7.4125	7.405	1	27	113	76
2011-09-02T09:30:00.634498	CBL ZC 1112	7.4025	7.4075	7.4	1	28	150	35
2011-09-02T09:30:00.671130	CBL ZC 1112	7.405	7.4075	7.405	1	29	88	33
2011-09-02T09:30:00.771404	CBL ZC 1112	7.405	7.41	7.41	1	30	88	48
2011-09-02T09:30:00.893891	CBL ZC 1112	7.4125	7.415	7.415	1	31	31	117
2011-09-02T09:30:00.920036	CBL ZC 1112	7.4125	7.415	7.4125	1	32	30	117
2011-09-02T09:30:00.986763	CBL ZC 1112	7.4125	7.415	7.4125	1	33	28	120

In Table 2.1, I provide an excerpt of the data for corn on September 2, 2011, reflecting trades in the first second of the trading day. These are known as "tick-by-tick" data. That is, rather than recording information at regular time intervals (say, every hour), they are recorded every time an order is executed. The data capture the trading process at its finest level. The time increments of the data go down to the microsecond.

2.1.2 Electronic trading and direct market access

Exchanges operating on an electronic trading platform offer members direct market access: members have a direct connection to the exchange through which they can submit orders. Orders are filled via a matching algorithm run by the exchange. Traders with direct market access also can access data from the exchange order book. In principle, the exchange has the complete order book and could make it available in its entirety. However, limitations on memory and bandwidth prevent the exchanges from providing more than a few levels of order book information. Further, exchanges can charge members substantial fees for access to this data, indicating it is of some usefulness to traders.

Electronic trading has also opened the door to statistical arbitrage. Because trades can be executed in tiny fractions of a second, the correlations between asset prices implied by standard financial theory (between futures and options on the same underlying asset) can break down at such small time scales, even if they generally hold at time scales that would make arbitrage by humans impossible. A great deal of resources have gone into developing algorithmic trading strategies to exploit these short term arbitrage opportunities, as well as investment in physical capital to achieve faster connections to exchanges (Budish, *et al*., 2015).

2.1.3 High-frequency financial time series

The data sets for each day include from ten to several thousand observations. These are much smaller in number than for tick-by-tick data for equities or foreign exchange, where

trades can reach several million observations in a day. This is due to the commodities markets being somewhat more thinly traded.

Tick-by-tick data present a number of challenges to the analyst. They frequently contain errors due to mistakes in reporting by the exchanges, and the data are irregularly spaced in time. They are also quite "noisy" due to the market microstructure issues discussed in the first chapter. These challenges have given rise to a number of new techniques for analyzing the data (to be discussed later). First, I will survey some stylized facts about the data.

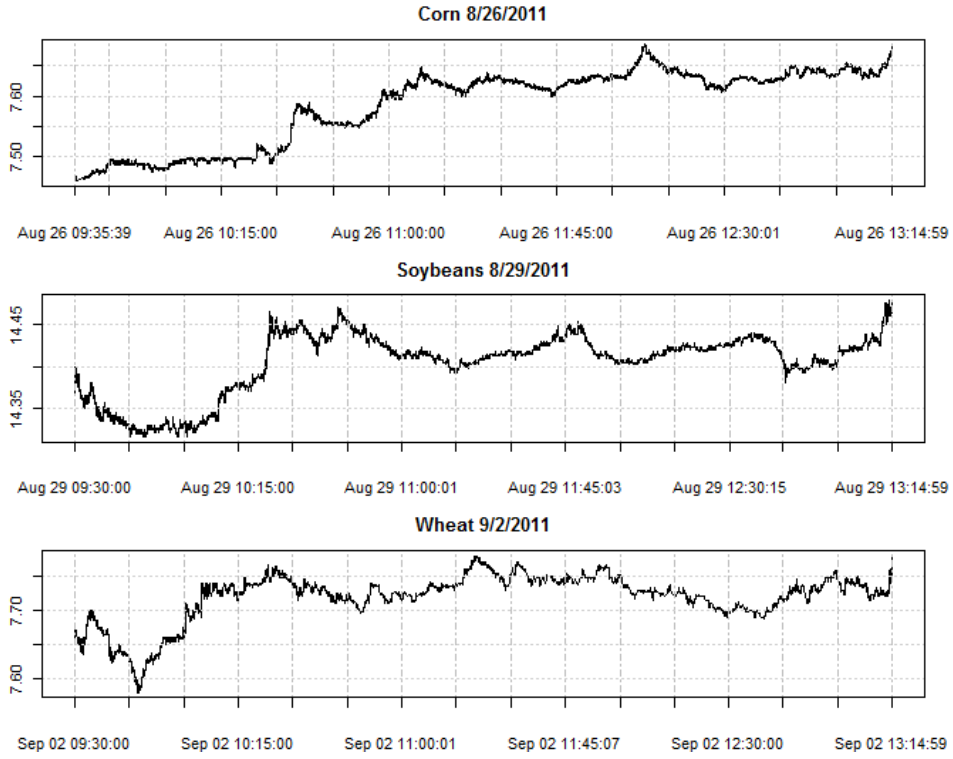
2.2 Exploratory data analysis

The peculiar features of high-frequency data can best be illustrated by plotting some select examples. These examples bring out the usefulness of sub-sampling or time aggregating the data. After that, I analyze the empirical distribution of the data and conduct some exploratory time series analysis on both autocorrelation and volatility using standard time series methods.

2.2.1 Graphical analysis

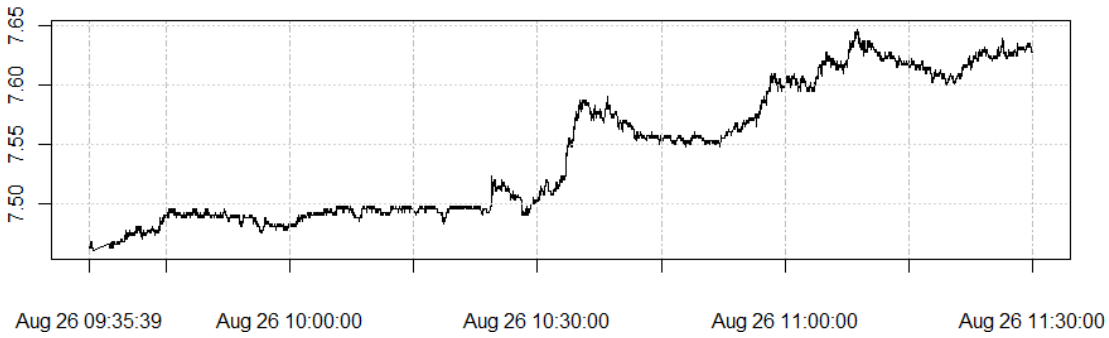
In Figure 2.1, I portray a day's worth of tick-by-tick values of the last traded price for corn, soybeans and wheat on August 26, August 29, and September 2, 2011, respectively.

Figure 2.1: Intraday tick-by-tick prices for three commodities



These series have been selected to illustrate a number of key features of the data. First they are jagged, with many small instantaneous jumps. Second, they contain occasional large jumps over slightly longer intervals. Third, the tick-by-tick sequencing illustrates how market activity tends to happen in clusters, there are many orders executed in a short time, and then no activity at all for extended periods.

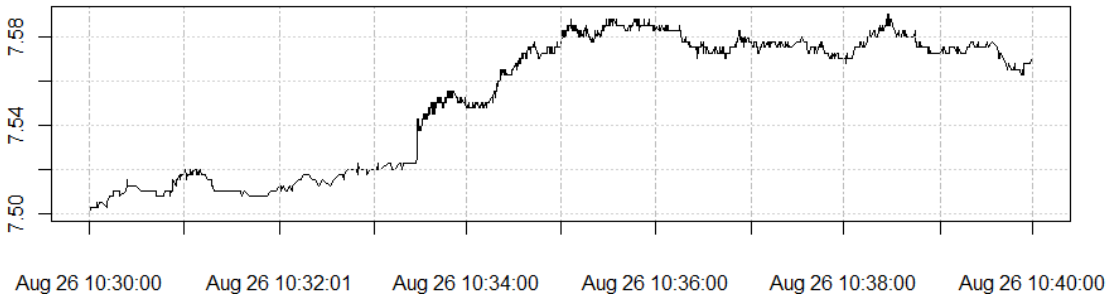
Figure 2.2: Two hours of corn prices, 8/26/2011



These properties are better illustrated by zooming in on a shorter time span. Consider, for example, in Figure 2.2, I report corn prices on August 26 for the first two hours of the trading day. One apparent observation is that the series does not appear any

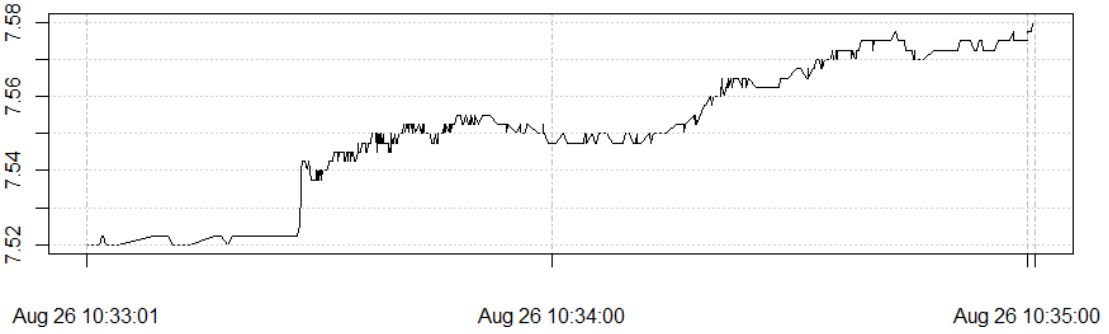
less jagged. Most of the fluctuations are small, but there is a substantial price jump between approximately 10:30 and 10:40 a.m. Zooming in further between those endpoints reveals further detail.

Figure 2.3: Ten minutes of corn prices, 8/26/2011



In Figure 2.3, more than in the previous graph, the price increments of the exchange, in this case one quarter of a cent, become apparent. Bold segments on the series represent periods where the trading price is oscillating between two price points, potentially the bid and ask prices offered by market makers. Therefore, there are two important sources of market microstructure noise evident: discrete prices and bid-ask bounce. In a hypothetical market where price variation could be arbitrarily small, the series at this resolution would appear smoother because prices wouldn't jump between exchange limits but adjust continuously to clear the market. (However, bid-ask effects would still be present if market making worked in the same fashion.)

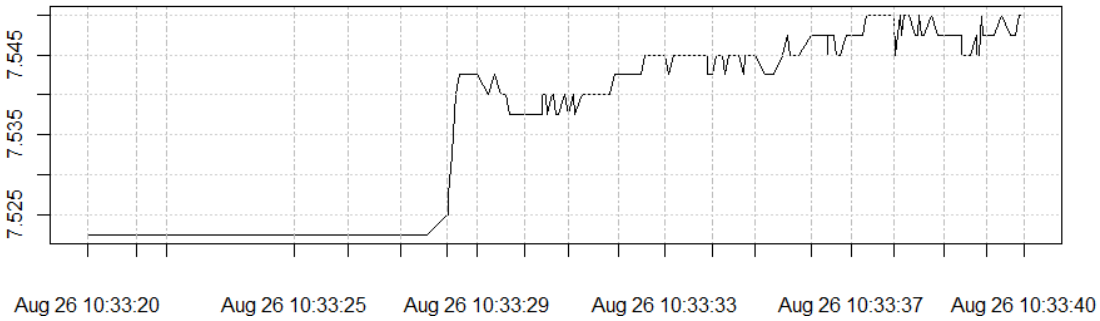
Figure 2.4: Two minutes of corn prices, 8/26/2011



As revealed in Figure 2.4, the large price movement noted earlier happened between 10:33 and 10:35 a.m. Magnifying once again brings out more structure. At this level of resolution the price increments become clearer, as does the irregular temporal spacing of the data. There are periods of heavy trading activity and stretches where no

trading happens. The latter are represented by flat segments and especially short diagonal segments on the line—the chart interpolates data points, so any movement up or down by one price increment that is horizontal represents a period with no trading.

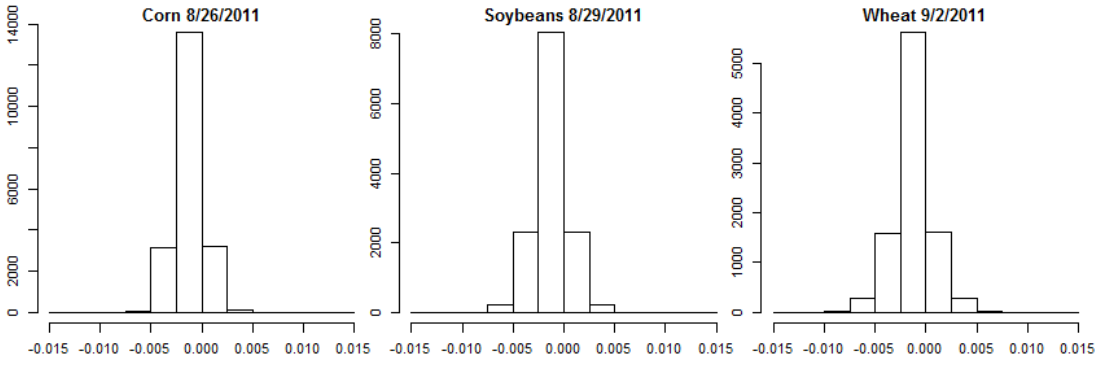
Figure 2.5: Twenty seconds of corn prices, 8/26/2011



In Figure 2.5 it is evident that that much of the sharp price increase that is visible in the longer-time axis chart is actually a gradual climb over a short period where prices were changing by the minimum exchange increment. The glaring exception is a large jump that occurs midway between 10:33 and 10:34 a.m. I zoom in one more time over a 20-second horizon to get a closer look.

The previously mentioned features, irregular clustering of activity in time and price changes mostly at the exchange minimum, are more evident in Figure 2.5. It reveals that the large jump is indeed a discrete event and not one accounted for by a succession of larger changes. Indeed, the price jumped by four increments in one trade. Given the size of the usual price jump, this event sticks out.

Figure 2.6: Distributions of tick-by-tick price changes for three commodities



Such events will be a primary focus of my analysis. It's not surprising that when looking at a trade-by-trade frequency that most price changes are at the exchange minimum. This makes jump events even more interesting than volatility. One might begin by asking how frequently they occur. In Figure 2.6, I display the distribution of tick-by-tick trades for the three commodities on the days shown in Table 2.1.

The histograms illustrate that the vast majority of ticks involved no change in price. These represent either a trade at the last price or a change in the order book (the bid or ask values). A substantial number of the remaining transactions involved price changes at the minimum of a quarter of a cent, and while it is possible to see that there is a small number of larger price changes, the volume of zero-price-change transactions is so large that they are hardly visible.

As a more concrete example, the corn market on August 26 had more than 20,000 transactions, 13,591 of which involved no change in prices. There were 3,192 ticks with an increase of 0.25 cents, and a roughly equal number, 3,170, where the price fell by that much. Only 92 transactions involved a price increase of half a cent, and more than 68 involved a half cent drop. Only one transaction was a decline of 0.75 cents and three were increases by that amount, and one transaction, the jump discussed above, was an increase of a cent.

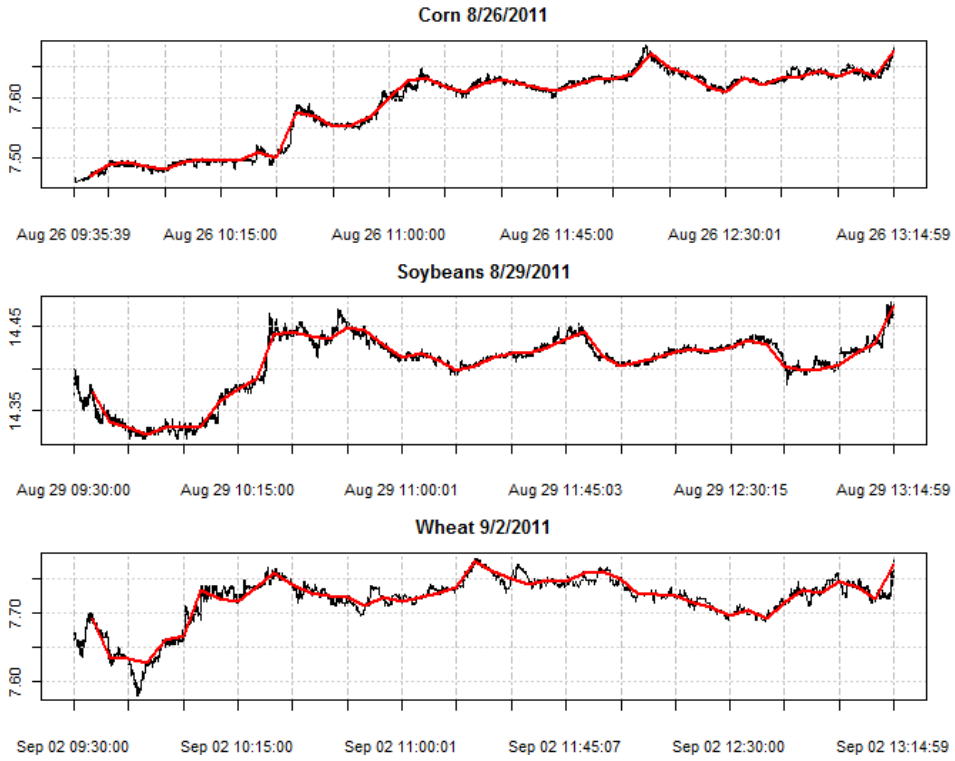
This distribution is typical of the tick-by-tick commodity market data. It poses a few challenges when conducting an analysis. First, trying to estimate a probability density of the returns on such an extremely peaked data set will deliver a nonsensical result. Further, if the jumps are of interest, there are few when looking at the truly ultra-high frequency of tick-by-tick data.

Consider again the corn example. Taking a one-minute window around the jump (between 10:33 a.m. and 10:34 a.m.) would reveal a price increase of 3 cents, compared to one for the instantaneous jump. This is one reason why in practice researchers and traders sample the data at lower frequencies, since it reduces noise, makes volatility more apparent, and jumps are more substantial.

2.2.2 Time aggregation

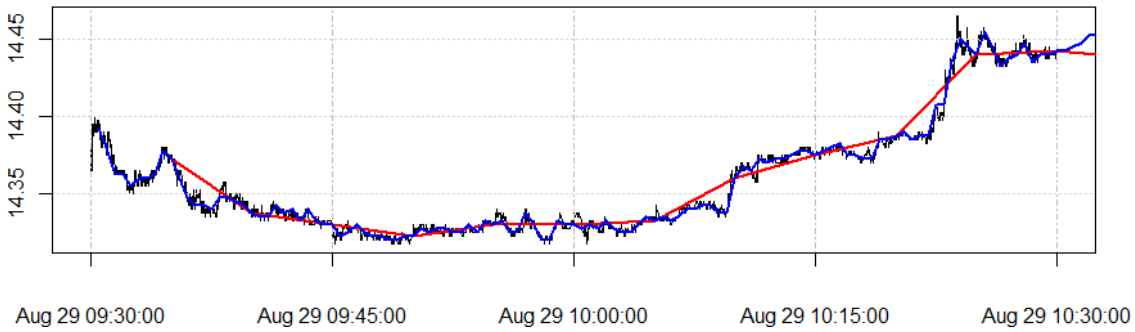
Sampling the tick-by-tick time series at lower frequencies has the immediate benefit of taking irregularly-spaced data on regular time increments, making available standard time series methods (e.g. ARIMA models). It makes large data sets smaller and reduces computer time for complex calculations. These benefits come at the cost of potentially losing some information in the higher-frequency data. [These issues are discussed in detail in Chapter 3, but here I illustrate with graphics.]

Figure 2.7: Raw and time-aggregated data



A recently-developed package for the R statistical software called "highfrequency" makes time aggregation and data cleaning straightforward. Using its time-aggregation functionality, series can be sampled at, for example, the five-minute frequency. In Figure 2.7, I have added these series as red lines on the three earlier charts. At this sampling frequency and the resolution of these charts, sampling appears to simply smooth out the series.

Figure 2.8: Two sampling frequencies, Soybeans 8/29/2011



To appreciate this more fully, I choose the first hour of trading for soybeans on August 29, during which the price swung down substantially followed by a recovery and a sharp increase. In Figure 2.8, I plot aggregations at the five-minute and 30-second frequencies, in red and blue lines, respectively.

Figure 2.9: A closer look at two frequencies, Soybeans 8/29/2011

It is not surprising that the five-minute series is smoother than the 30-second one, but the comparison is instructive. While the lower-frequency both eliminates much of the noise and all of the zero-price-change observations in the original series, it also misses some detail, such as the volatile period early in the day. Progressively zooming in on the price jump between 10:15 and 10:30 is instructive (see Figure 2.9).

While lower-frequency sampling reduces noise to levels where volatility can be meaningfully measured, choosing too low a frequency misses some of the action. It is for this reason that Ait-Sahalia, *et al.* (2005) recommend sampling as frequently as possible and specifying a model for microstructure noise.

In the earlier literature, sampling at too low of a frequency was also cautioned against because it made jumps more pronounced, which would cause realized volatility measures to fail. However, recent advances in modelling have made this less of a concern, and indeed the jumps are a primary interest in my analysis, as I will discuss later in this chapter. First, however, I will explore further how sampling frequency changes the distribution of measured returns.

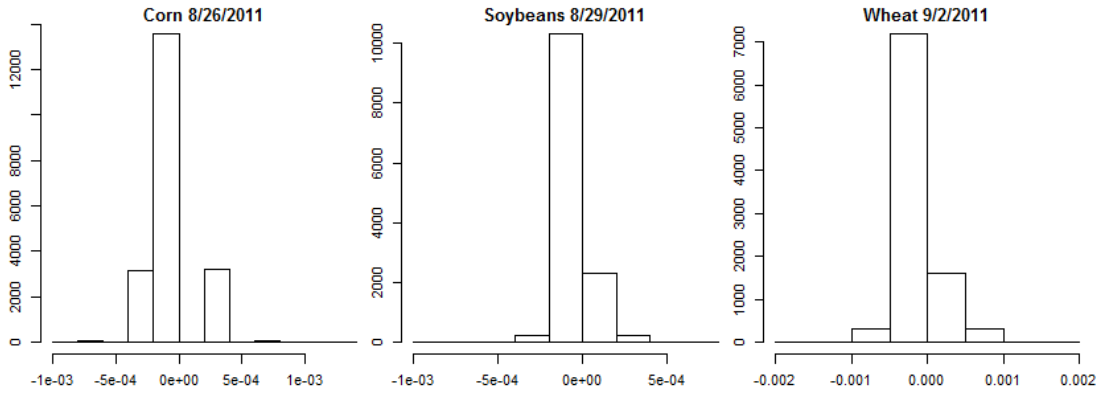
2.2.3 Empirical distributions

I have been focusing on prices, but returns are usually of more interest in finance. I define the (one-period) return at time t as in (1).

$$r_t = P_t/P_{t-1} - 1 = \log(P_t/P_{t-1}) = \log P_t - \log P_{t-1} = p_t - p_{t-1} \quad (1)$$

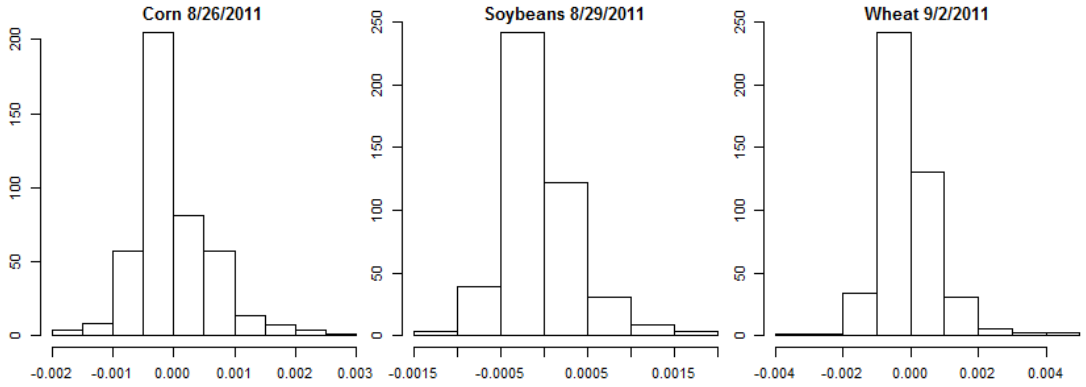
where P_t is the price of an asset at time *t*. It is conventional to denote the log price by the lowercase p_t, which I adopt for the remainder of this study. The primary reasons for this treatment are to place the price of an asset in terms of the change in value or return on investment, to obtain stationarity through differencing in series that are integrated or contain unit roots, and to avoid the problem of non-normal shocks (for which reason many models assume price movements are log-normal). Campbell, Lo and MacKinlay (1997, sec. 1.4) and Tsay (2010 sec. 1.2) elaborate on the reasons for focusing on returns.

Figure 2.10: Distributions of tick-by-tick returns for three commodities



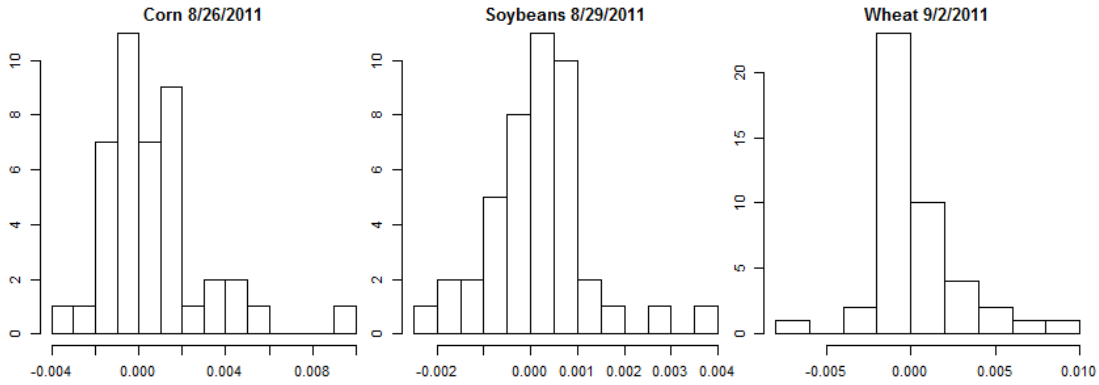
In an earlier section I illustrated how the vast majority of tick-by-tick price movements were within the exchange minimum or were zero. Simply transforming the series to returns does not change this, as displayed in Figure 2.10.

Figure 2.11: Return distributions with 30-second aggregation



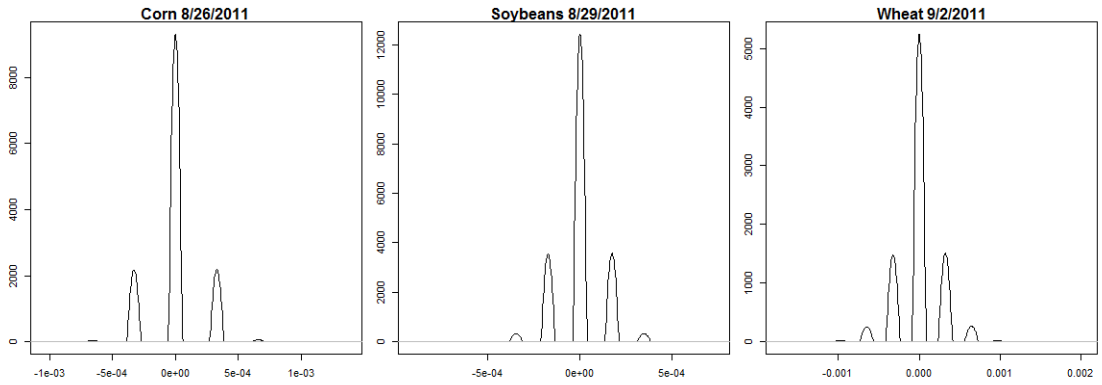
When I plot histograms with the data aggregated at 30 seconds, there is a greater spread in the frequency distribution of returns (see Figure 2.11).

Figure 2.12: Return Distributions with 5-minute aggregation



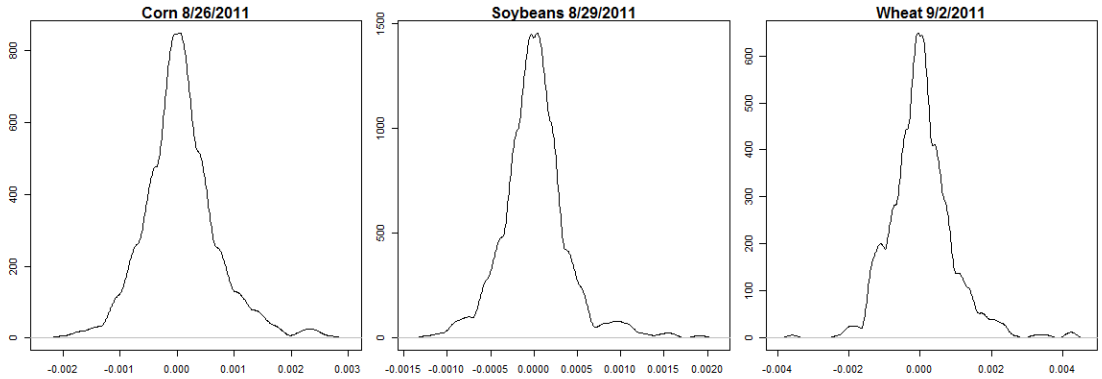
The frequency distribution is even more widespread when the data are aggregated at 5 minute intervals (see Figure 2.12).

Figure 2.13: Estimated density of tick-by-tick returns



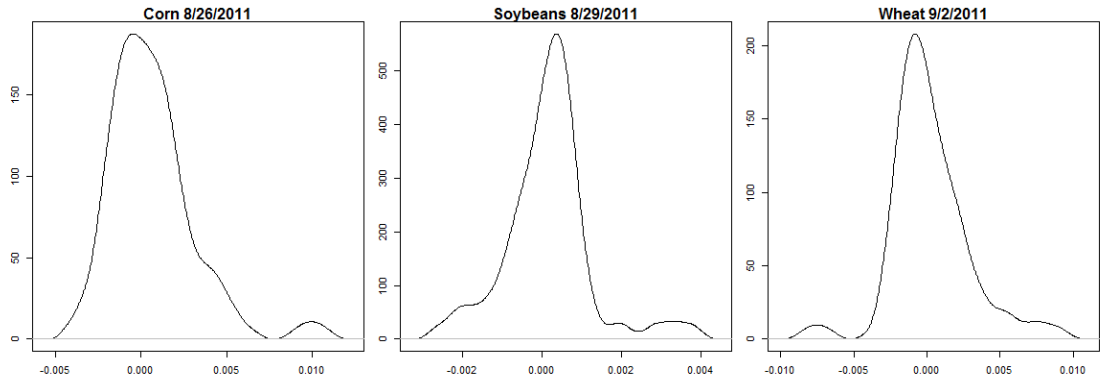
At this point, one may ask about the empirical probability distribution of returns, estimated for example via kernel smoothing [see Tsay (2010, sec.4.1.5) or Wasserman (2005 chs. 4 and 6) for more on kernel and other smoothing methods]. Unsurprisingly, the plots for the tick-by-tick returns appear identical to their histograms (see Figure 2.13).

Figure 2.14: Return density at 30-second aggregation



Fitting densities for returns at the 30-second and five-minute levels of aggregation (shown respectively in Figure 2.14 and 2.15) produces plots that appear more bell-shaped.

Figure 2.15: Return density at 5-minute aggregation



In the preceding section, I demonstrated with an example how using longer intervals of aggregation can lead to the appearance of larger jump events. Comparing the densities plotted above bears this out. Once again, I confront the tradeoff of whether to sample the market less frequently, filtering out noise but possibly discarding information versus sampling more frequently, studying the market microstructure at the cost of ignoring interesting patterns at larger-scales. In the wavelet analysis used in this paper, while not eliminating the trade-off, it is possible to take both perspectives.

Up to this point, I have been simply treating the price movements essentially as a sequence of independent shocks. However, it is informative to apply some standard time-series models to explore the correlations of price movements over time.

2.2.4 Autocorrelation

A useful way to visualize the correlation of a time-series variable with its own previous values is to plot the autocorrelation function (ACF) at different lags (this section follows Shumway and Stoffer, 2010, sec. 1.6 and 2.3). To define the ACF, I begin with the (sample) autocovariance function for lag h, as written in (2).

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t-1}^{n-h}(x_{t-h} - \bar{x})(x_t - \bar{x}) \quad (2)$$

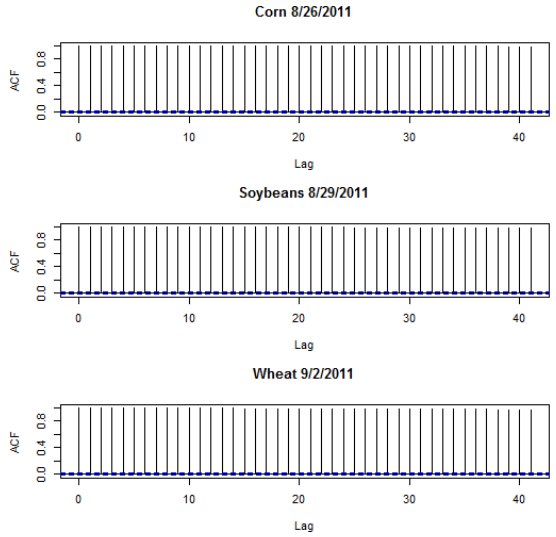
For h = 0, 1, 2,…, where $\hat{\gamma}(0)$ is the sample variance of x_t. The ACF for lag h is written,

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0) \quad (3)$$

These estimators, while informative for exploratory data analysis, have a couple of limitations. First, they assume stationarity because using the sample mean and variance

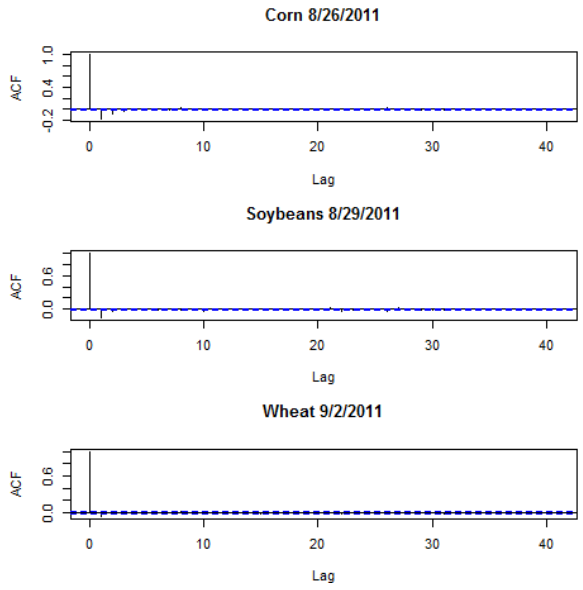
as estimators of the expectation and variance of x_t assumes they are constant. Second, they require imposing constant time intervals because otherwise plotting at different lag lengths becomes a challenge.

Figure 2.16: ACFs for prices, 1-second aggregation



As a proxy for the tick-by-tick data, I can use the series aggregated at one-second intervals and compare those sampled less-frequently. Obviously, plotting ACFs for the price series will show persistent high autocorrelations at all lag lengths. In Figure 2.16, I illustrate the three examples used in previous sections before moving on to returns.

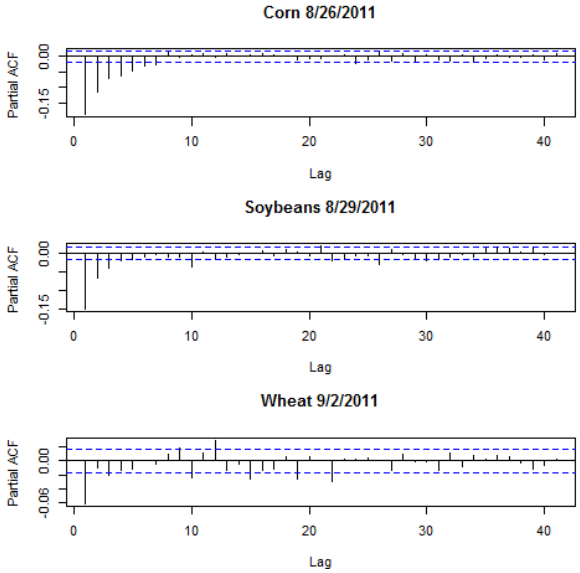
Figure 2.17: ACFs for returns, 1-second aggregation



In contrast, returns at the one-second frequency show only a slight negative autocorrelation at the one-period lag, as shown in Figure 2.17. This lack of correlation at higher-order lags (at least at the 1-second frequency) suggests returns may follow a random walk. This opens the door to volatility estimation using simple ARIMA models.

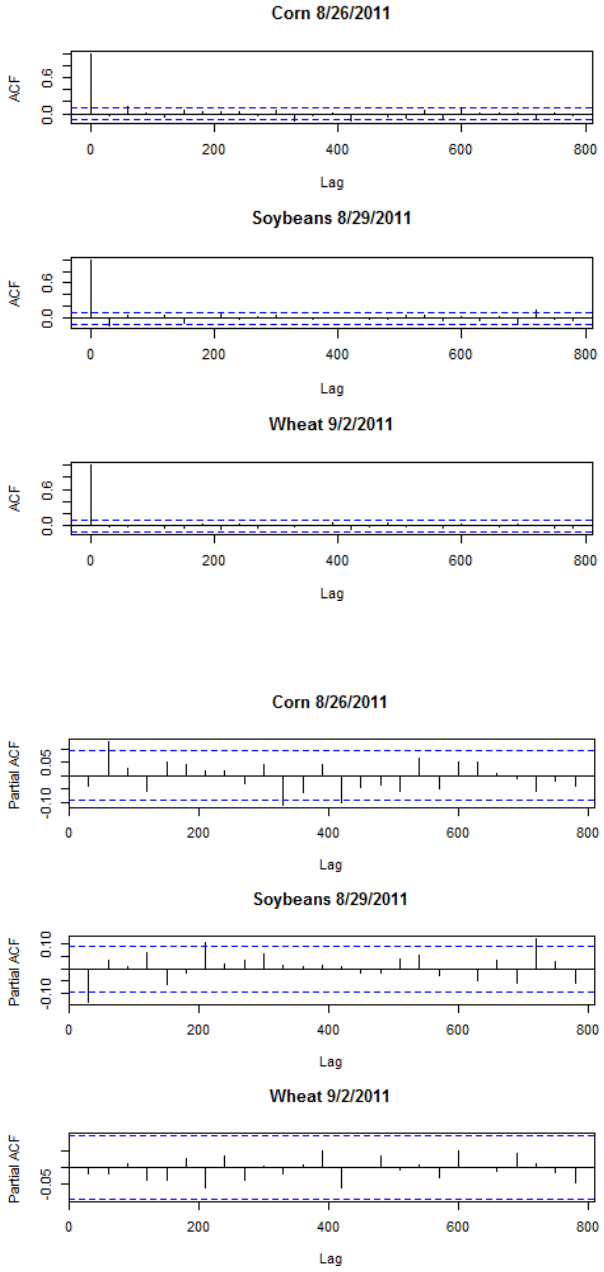
Another tool used to assess appropriate lag lengths for fitting time-series models is the partial autocorrelation function (PACF) which measures autocorrelations at a given lag removing linear dependence among observations over that lag length.

Figure 2.18: PACFs for returns, 1-second aggregation



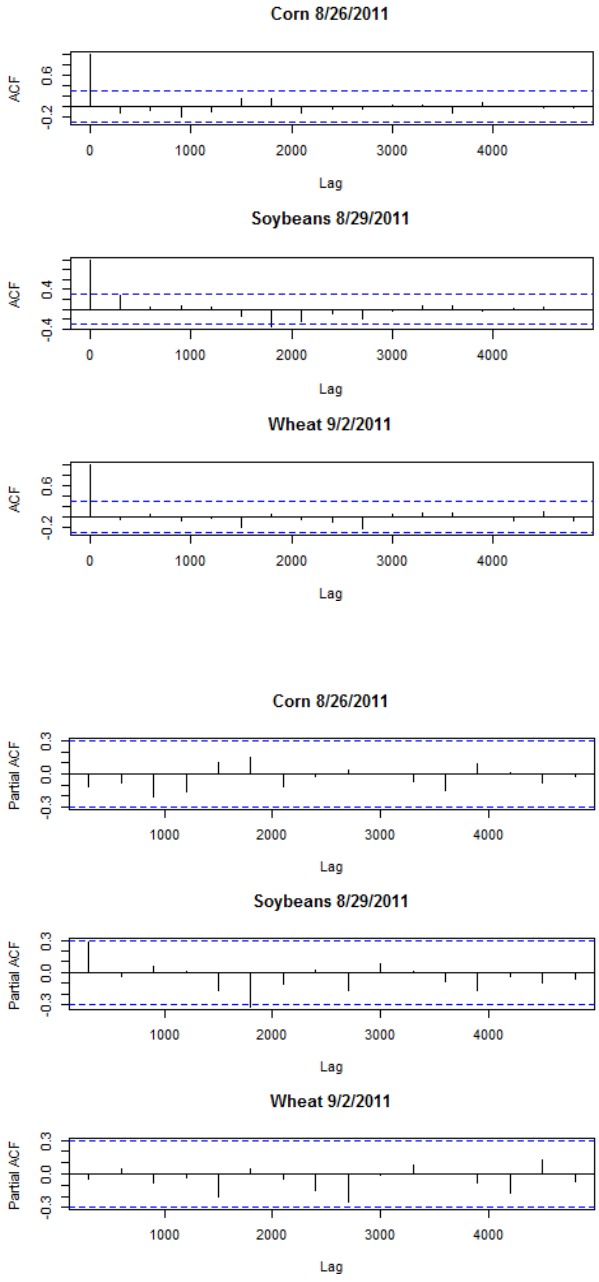
The PACF of returns at the 1-second frequency is displayed in Figure 2.18. These suggest returns are weakly negatively auto-correlated at a 1-period lag, with progressively weaker correlation at higher order lags. Moving to lower-frequency series doesn't increase the apparent autocorrelation in returns. If anything, in fact, the autocorrelation gets weaker.

Figure 2.19 ACFs and PACFs for returns, 30-second aggregation



The ACFs and PACFs for the three commodities at the 30-second sampling frequency are displayed in Figure 2.19.

Figure 2.20 ACFs and PACFs for returns, 5-minute aggregation



The ACFs and PACFS for the three commodities at the five-minute sampling frequency are displayed in Figure 2.20.

Taken together, these plots illustrate another side the sampling frequency trade-off in addition to the one that applies to estimating the distribution of returns. Sampling at

lower frequencies has benefits in terms of filtering out microstructure noise and identifying jump events, but it appears to obscure the autocorrelation relationships in the price data. This structure is important if one wants to study the volatility of prices using standard the standard time series models like those used in financial econometrics.

Chapter III

Methods

In this chapter I give a description of the methods of this study. Following a brief review of other approaches taken in the high-frequency finance literature, I detail the theory underlying the proposed wavelet-based analysis.

3.1 Review of nonparametric volatility analysis

To make up for the shortcomings of other time-series methods in capturing some of the unique properties of high-frequency data, econometricians have developed new models. One set of models incorporates a nonparametric approach to modelling intraday volatility at high frequencies. The analysis in Chapter 4 is a variant of this approach, so reviewing it will provide a good introduction.

Because of market microstructure complications, using (informationally) efficient-market theory to guide the choice of models may not yet be useful. For this reason, researchers have turned to non-parametric (data-driven or descriptive) approaches to study volatility at high frequencies.

This large and growing literature (see Ait-Sahalia and Jacod (2012) for a recent review) continues to produce new approaches. I survey a small subset of these models in this section. I focus on a method (realized-volatility estimators) of which many other approaches are generalizations or modifications.

3.1.1 *Defining volatility*

It is common practice in the high-frequency volatility literature to begin with an assumption that the (true) log-price follows a jump-diffusion process, as stated in (4).

$$X_t = \int_0^t \mu_s \, ds + \int_0^t \sigma_s ds + \sum_{l=1}^{N_t} J_l \qquad (4)$$

In (4) the first two terms represent the standard drift and diffusion parts of a Brownian motion. The third term represents the jumps and is often treated as a poisson counting

process. It should be noted that the integrals in the above equation are stochastic (Ito) integrals (the process is an Ito semi-martingale).

When an asset price is treated as a standard Brownian motion (that is, without jumps), the parameter σ is taken to represent volatility. But for my purposes, I will need a definition that can incorporate jumps. More generally, an asset's volatility is the variation over time of its period-by-period returns. Anderson et. al. (2010) provide a comprehensive overview, treating volatility as a realization of the quadratic variation (stochastic) process, as written in (5).

$$[X,X]_t = X_t^2 - 2\int_0^t X_s ds = \lim_{||\mathcal{P}||\to 0} \sum_{t=1}^{n} (X_t - X_{t-1})^2 \quad (5)$$

Here $\mathcal{P}$ is a partition of the time interval [0, t]. In the jump-diffusion model in (6), quadratic variation corresponds to the diffusion and jump terms,

$$[X,X]_t = \int_0^t \sigma_s^2 ds + \sum_{l=1}^{N_t} J_l^2 \quad (6)$$

The first term in (6) is the *integrated volatility* and the second term is the *jump variation*. Both are components of volatility under my definition. The actual volatility of an asset—which Anderson *et al.*, call the *notional volatility*—is taken to be the quadratic variation of its returns. *Expected volatility* is the expectation of the notional volatility at some time given the notional volatility up to that point.

However, one cannot observe the notional volatility because to do so would require prices in continuous time. Moreover, due to market microstructure noise, the observed log-price p_t is taken to be an imperfect realization of the underlying true log-price process: $p_t = x_t + \epsilon_t$, where the error term ϵ_t is taken as a measure of noise. In practice, realized volatility must be estimated in the presence of noise, which requires estimating the noise parameters as well.

3.1.2 *Realized-Volatility (RV)*

The general idea underlying RV estimators is to measure volatility on a cumulative basis as the sum of squared returns. To begin, I re-visit the definition of returns, which I defined above as one-period returns and put it into a general h-period form, $r_{t,h} = p_t - p_{t-h}$. I define realized volatility for a given period in (7).

$$RV_{t+hn}(h) = \sum_{i=1}^{n} r_{t+ih,h}{}^2 \quad (7)$$

Here n is the number of observations corresponding to the sampling frequency over the time period of interest. For example, if measuring daily realized volatility for returns at the 5-minute frequency, n = 45 in that case (the exchange operates for 3 hours, 45 minutes a day).

RV is a simple measure: it is the second (uncentered) moment of h-period returns over the interval of interest. And some variant of it has been around for a long time as a measure of historical volatility (see Poterba and Summers, 1986, or Schwert, 1989). However, it has some appealing properties which make it useful in practice.

Anderson *et al.* (2003) show that as sampling intervals approach zero and the time horizon approaches infinity, RV estimators converge to the notional (actual) volatility. Furthermore, RV provides a consistent estimate of expected volatility. This makes it potentially useful in a forecasting context.

The simple RV has some drawbacks. The results mentioned above apply to RV as an estimator sampling from a continuous time diffusion process in which the (log) price is perfectly observed. However, its estimates of volatility are not robust to jumps in the process. Furthermore, due to market microstructure noise, the actual data available is observed imperfectly. In practice, RV estimates increase rapidly as sampling frequencies increase, suggesting that microstructure noise is significant. The methods I adopt are an attempt to overcome these shortcomings.

3.1.3 *Generalizations of RV estimators*

One generalization that has been suggested to the basic RV estimate is *realized bi-power variation* (RPV), due to Barndorff-Nielson and Shepherd (2002). In this formulation $RPV_t = \sum_{t=2}^{n} |r_t| * |r_{t-1}|$. Rather than take the sum of squared returns, RPV sums the products of consecutive (absolute) returns. This simple modification produces a volatility estimator that is robust to jumps.

Another approach is to average over different time scales. Zhang, Mykland and Ait-Sahalia (2005) introduce *Two-Scales RV* (TSRV) where $TSRV = RV(h) - (N/h)^{-1}RV(1)$. This measure is an average RV over two time scales, in which (a

weighted multiple of) the estimate corresponding to the highest sampling frequency is removed. As estimates at the highest frequencies are polluted by noise, this measure is robust to microstructure noise.

A further generalization is *Multiple-Scale RV* due to Zhang (2006). This is a generalization of the TSRV where more than two subsampling frequencies are used in the estimate. There are numerous other volatility measures with different attributes as relate to consistency, asymptotics, efficiency and robustness to noise and jumps.

3.2 Method of Analysis

With the literature on high-frequency volatility estimation as a guide, I can approach the problem of analyzing volatility in the data set. I seek a method that provides useful volatility estimates in the presence of market microstructure noise and jumps. Once I have such a method, I can apply it to my dataset and use the results for practical applications.

My approach is inspired by that of Fan and Wang (2007), who make use of wavelets to disentangle jumps in the time series of prices. Wavelets also allow us to decompose the time series at a number of different time scales or resolutions, which gives us another angle on the frequency-selection trade-offs discussed earlier. Wavelets have seen limited application in econometrics [see Ramsey (2002) for a survey; Yogo (2008) provides a very concise introduction with an application to business cycle analysis]. The mathematical theory underlying wavelets is sophisticated and I intend to give only a conceptual introduction. The next section describes the basic ideas behind wavelets as they apply to my study.

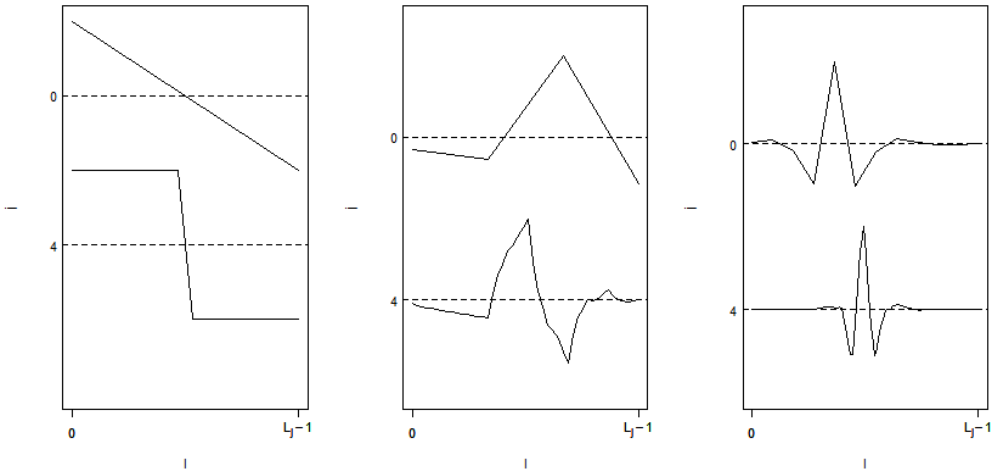
3.2.1 Wavelets

I review here some of the basic concepts of wavelets as they apply to my analysis. Though they are based on some old ideas in mathematics, the tools for practical wavelet analysis have only developed since the 1980s with the development of the Continuous Wavelet Transform and especially since the 1990s after the introduction of the Discrete Wavelet Transform and efficient methods for computing it.

Wavelets have found many applications— bioinformatics, signal processing and data compression (the JPEG image format uses wavelet based compression, for example), geophysics and seismology, and in many areas of statistics including timer series analysis. The literature on wavelets is now quite large and my treatment will be by necessity brief. A few useful references on wavelets include Percival and Walden (2000); who give a thorough treatment in the time-series context; Hubbard (1998) is an introduction for popular audiences that also gives a historical narrative of the development of wavelets; and Fugal (2009) gives a readable conceptual introduction. Najimi (2012) is a recent treatment that is compact and elegant.

3.2.1.1 An intuitive description. The basic idea underlying wavelet analysis is simple: Given a signal (a time-series in this case), take weighted averages over intervals of the signal and compare them across time periods and over different time scales. The weighting (or filtering) is done with a class of functions known as wavelets. This allows us to decompose the sample variance of a time series into pieces associated with different scales.

Figure 3.1: Some wavelets



I begin by reviewing several important properties of wavelet functions. Wavelet literally means “little wave.” Unlike a wave function, which is periodic and is defined over the real line, wavelets only take on values over a particular interval.

Three standard wavelets are plotted in Figure 3.1. The first panel illustrates a Haar wavelet, named for the mathematician who introduced it in 1910. The second panel illustrates a Daubeches (4) or D4 wavelet, a useful class developed by Ingrid Daubeches, a pioneer of the Discrete Wavelet Transform. The third panel illustrates a Coiflet (6) wavelet, which has proven popular for financial applications.

In a wavelet transform, the integral (average) of the product of the signal with a series of stretched and time-shifted wavelet functions is computed. The result is a series of coefficients that indicate sharp changes in the signal over a given scale. These coefficients can be used to detect jumps in the context of this study. The coefficients can also be used to reconstruct a high-resolution signal using much less information than that contained in the signal itself, which is why wavelets have been so useful for data compression applications.

An instructive analogy for the wavelet transform is found in Fourier analysis. In the Fourier Transform, a signal is decomposed into a sum of sinusoids (waves) of different frequency with which it correlates. But this treats a signal as stationary, which may not be a good approximation in applications like ours. The dynamic or windowed Fourier transform is a smoothing approach that applies a Fourier transform over a fixed time interval, or, window recursively along the length of the signal, like a moving average.

The Discrete Wavelet Transform (DWT) is similar to the dynamic Fourier transform, except instead of using sinusoids as a basis it uses wavelets. More importantly, in the DWT the width of the window changes and the signal is analyzed multiple times. Begin by applying the wavelet filter at the finest scale, weighting the average of each pair of observations in the signal, generating a number of coefficients equal to half the number of observations. Then the filter is stretched to twice its length (and half its amplitude) and apply it to the signal averaging each two-observation pair. The process continues iteratively until a filter over the length of the signal is run. The set of all the coefficients generated by this process is known as wavelet coefficients. Finally, take a coefficient that represents the sample mean of the whole signal, known as a scaling coefficient. The results of a DWT simply encode the information in the signal. But they

contain information that can be used to identify jumps in a signal path at many different scales, if one knows how to interpret the results.

3.2.1.2 *Wavelet basis functions*. Here I describe two important functions that constitute the basis of analysis by wavelets. The notation in this section follows Percival and Walden (2000, Ch. 4).

A (real-valued) *wavelet filter* of width L, $\{h_l: l = 0, \ldots, L-1\}$, where L is an even integer and $h_l \equiv 0$ for $l < 0$ and $l \geq L$, is a function that satisfies the following three properties:

$$\sum_{l=0}^{L-1} h_l = 0$$

$$\sum_{l=0}^{L-1} {h_l}^2 = 1$$

$$\sum_{l=0}^{L-1} h_l \, h_{l+2n} = \sum_{l=-\infty}^{\infty} h_l \, h_{l+2n} = 0$$

Here n is a nonzero integer. In words, the first property states that wavelet filter integrates to zero, the second property states that its square integrates to unity (is normalized), and the third property states that it is orthogonal to even shifts. The second and third properties are known as *orthonormality*. This is an important feature of wavelets that I return to in the next section.

Given the wavelet filter, h_l, I define the *scaling filter*, $g_l \equiv (-1)^{l+1} h_{L-1-l}$, which satisfies the following four properties:

$$\sum_{l=0}^{L-1} g_l = \sqrt{2}$$

$$\sum_{l=0}^{L-1} {g_l}^2 = 1$$

$$\sum_{l=-\infty}^{\infty} g_l \, g_{l+2n} = 0$$

$$\sum_{l=-\infty}^{\infty} g_l \, h_{l+2n'} = 0$$

Here n, n' are nonzero integers. The wavelet and scaling filters are sometimes referred to as the "mother wavelet" and "father wavelet" respectively. This is because

they can be used to generate a series of coefficients that can be used to analyze (or synthesize) a signal, as demonstrated in the next section.

3.2.1.3 *The Discrete Wavelet Transform (DWT)*. Begin with a time series $\{X_t: t = 0, \ldots, N-1\}$ of length $N=2^J$. Represent the column vector of X_t by **X**. The DWT gives us $N=2^J$ coefficients $\{Wn : n = 0, \ldots, N-1\}$ and represent by **W** the column vector of DWT coefficients. Then,

$$\mathbf{W} = \mathcal{W}\mathbf{X} \tag{8}$$

Here, $\mathcal{W}$ is an N*N matrix in $\mathbb{R}$ and $\mathcal{W}^T\mathcal{W} = \mathbb{I}_N$. In signal processing terms, the above equation represents an analysis of signal **X**. The DWT matrix $\mathcal{W}$ is an *orthonormal transform*, such that

$$\mathbf{X} = \mathcal{W}^T\mathbf{W} \tag{9}$$

And

$$\| \mathbf{W} \|^2 = \| \mathbf{X} \|^2 \tag{10}$$

Here $\| \mathbf{X} \|^2$ is the squared norm of X: $\| \mathbf{X} \|^2 = < \mathbf{X}, \mathbf{X} > = \mathbf{X}^T\mathbf{X} = \sum_{t=0}^{N-1} {X_t}^2$, which is also called the *energy* of $\{X_t\}$. The first condition represents the synthesis of **X**, in signal processing terms. The second condition is called the energy-preserving condition. The Wn coefficients are associated with scale λ, where

$$\bar{X}_t(\lambda) = \frac{1}{\lambda}\sum_{t=1}^{\lambda-1} X_{t-1} \tag{11}$$

For scale $\tau_j = 2^{j-1}$, j=1, … , J, ($J = \log_2 N$). There are $N/(2\tau_j)$ coefficients associated with scale τ_j. Thus there are N/2 coefficients at scale $\tau_1 = 1$, N/4 coefficients at scale $\tau_2 = 2$, and so on, with one coefficient, W_{n-2}, at scale $\tau_J = 2^{J-1} = N/2$. There is also another coefficient W_{N-1} at scale N, which is just proportional to the sample average of **X**. The first N-1 coefficients are the *wavelet coefficients* and the W_{N-1} is the *scaling coefficient*.

The DWT can be computed quickly using a method called the Pyramid Algorithm, which is faster than the well-known Fast Fourier Transform (see Percival and Walden, 2000, 4.4-4.6, for details).

3.2.1.4 *The Maximal Overlap Discrete Wavelet Transform (MODWT)* One shortcoming of the DWT is its requirement that the number of observations in a time series, N, be a power of 2. Technically, in my analysis I will make use of a variant called the maximal overlap (or undecimated) DWT, which does not have this requirement. The MODWT is not strictly an orthonormal transform, but has similar properties to the DWT such that it can be used to analyze the sample variance of a time series for components at different scales. This decomposition is the topic of the next section.

3.2.1.5 *Multi-resolution (time-frequency) Analysis (MRA).* Begin with a partition of **W**, the vector of DWT coefficients, into J+1 sub-vectors: $\mathbf{W}_j$ (j = 1, …, J) and $\mathbf{V}_J$, such that **W**j contains all DWT coefficients for scale τ_j and $\mathbf{V}_J$ contains the scaling coefficient W_{N-1} so that,

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_j \\ \vdots \\ \boldsymbol{V}_J \end{bmatrix} \tag{12}$$

Likewise, one can partition the DWT matrix $\mathcal{W}$ into J+1 submatrices $\mathcal{W}_j$ and $\mathcal{V}_J$, where $\mathcal{W}_j$ is the $N/2^j$*N submatrix for j = 1, …, J and $\mathcal{V}_J$ is a length-N row vector whose values are all equal to 1/N.

Let $\mathcal{D}_j \equiv \mathcal{W}_j^T \mathbf{W}_j$, j=1, …, J, which is an N-dimensional column vector whose elements represent changes in **X** at scale τ_j, which is called the j-th level wavelet *detail*. Also, let $\mathcal{S}_J \equiv \mathcal{V}_J^T \mathbf{V}_J$, which is an N-dimensional column vector whose elements are all equal to the sample mean. Now one can write the **X** vector as,

$$\mathbf{X} = \sum_{j=1}^{J} \mathcal{D}_j + \mathcal{S}_J \tag{13}$$

which expresses the signal **X** as the sum of a constant (sample mean) vector $\mathcal{S}_J$ and J wavelet detail vectors. Each detail vector contains a time series that expresses changes in **X** and different time scales. Equation (13) is a *Multi-Resolution Analysis* (MRA).

As previously mentioned, the results of a DWT, in the form of DWT coefficients or MRA, can be used to identify discontinuities or other jumps in a time series. This will prove useful for my analysis.

3.2.2 Wavelet-based decomposition of jumps and volatility

I use wavelet methods to identify and filter jumps in high-frequency time-series and to estimate volatility. The methods may be better understood through their application.

3.2.2.1 Identifying jump locations. The DWT coefficients summarize the variation in a time series over different time scales and at different locations in the series. They contain information on jumps, like those demonstrated in the high-frequency data earlier in this chapter. At a higher resolution (lower scale), a jump is represented by a large value for a coefficient, where most coefficients in an otherwise smooth series would be close to zero.

To identify jump locations I set a threshold for the wavelet coefficients and find the locations of those points where the corresponding $|W_n|$ exceeds the threshold. I denote the estimated jump locations, $\hat{t}_l$.

Furthermore, because of the multi-scale nature of the DWT, it provides a useful lens through which to examine the tradeoff between aggregating tick-by-tick data at higher or lower frequencies. Using the highest available frequency in a wavelet analysis, the coefficients at larger scales inform us about variation at lower-frequency aggregations, but only approximately because in the DWT and especially the MODWT the analysis is on a subset of all the data points. The DWT can identify jumps, which is of interest for several purposes.

3.2.2.2 *Analyzing jumps*. The identified, jumps are useful for two purposes. First, returning to the jump diffusion model above, they provide data that can be used to estimate the parameters of the jump term(s) corresponding to jump variation. Ait-Sahalia and Jacod (2012) review this type of analysis. Recall from the jump-diffusion model above that jump variation corresponds to

$$\psi = \sum_{l=1}^{N_t} J_l^2 \qquad (14)$$

Following Fan and Wang (2007), I can estimate this variation by taking the estimated jump locations $\hat{t}_l$, choosing a small neighborhood $\hat{t}_l \pm \delta_n$ around that location

and averaging over it, representing by $\bar{p}_{\hat{t}_l+}$ and $\bar{p}_{\hat{t}_l-}$ the averages over $[\hat{t}_l, \hat{t}_l + \delta_n]$ and $[\hat{t}_l, \hat{t}_l + \delta_n]$, respectively. Then $\hat{J}_l = \bar{p}_{\hat{t}_l+} - \bar{p}_{\hat{t}_l-}$ is an estimate of the jump size. The estimator for jump variation is written as,

$$\hat{\psi} = \sum_{l=1}^{\hat{q}} (\bar{p}_{\hat{t}_l+} - \bar{p}_{\hat{t}_l-})^2 = \sum_{l=1}^{\hat{q}} \hat{J}_l^2 \quad (15)$$

Second, in addition to jump variation, the wavelet estimated-jumps can provide summary statistics on frequency and autocorrelation. Furthermore, one may ask whether there are other variables that correlate with jumps, such as changes in the bid-ask spread or jumps in other assets.

3.2.2.3 *Integrated-volatility without jumps*. Another benefit to identifying jumps is that they can be removed from the dataset, with the remaining time series used to estimate integrated volatility. To remove jumps, I begin by estimating the counting process at time t by $\widehat{N}_t = \sum_{l=1}^{\hat{q}} 1, (\hat{t}_l \leq t)$, and jumps at time t by $\hat{\psi}_t = \sum_{l=1}^{\widehat{N}_t} (\bar{p}_{\hat{t}_l+} - \bar{p}_{\hat{t}_l-}) = \sum_{\hat{t}_l \leq t} \hat{J}_l$. The jump-adjusted data is written,

$$p_t^* = p_t - \hat{\psi}_t = p_t - \sum_{\hat{t}_l \leq t} \hat{J}_l \quad (16)$$

I can obtain estimates of integrated volatility by computing nonparametric volatility estimators on p*. Since there is a plethora of volatility estimators that are robust to microstructure noise but not to jumps (e.g., TSRV) this provides a useful means to treat the issues of noise and jumps separately. This is the primary contribution of Fan and Wang (2007), who demonstrate that the estimators derived by this means have desirable statistical properties, and it is the primary template followed in my analysis.

3.2.2.4 *Wavelet-based volatility estimation*. In addition to using existing volatility estimators on the jump-censored data, Fan and Wang also introduce their own wavelet-based volatility estimator. To begin, I denote the observed jump adjusted returns as $r^*_t = p^*_t - p^*_{t-1}$. Let $W_{n,j}$ represent the DWT coefficients of r^*_t for scale j, which are taken up to some scale $j_i < J = \log_2 N$. The corresponding wavelet RV is written,

$$WRV = 1/N \sum_{l=1}^{2^{j_i}} \sum_{j=1}^{log_2 N - j_i} \sum_{n=1}^{2^j} W_n \quad (17)$$

Then the wavelet estimator of integrated volatility can be written,

$$\widehat{\Theta}_w = WRV - 1/2^{j_i} \sum_{t=1}^{N} {r^*}_t^2 \quad (18)$$

As already mentioned, statistical methods are often better illustrated through a practical application. In the analysis of the next chapter I will better illustrate these methods.

Chapter IV
Empirical Analysis

4.1 Introduction

Following the statistical procedures detailed in Chapter Three, I can detect jump events in high-frequency financial time series and use these to estimate the [parameters of the] jump [components of the] process generating /describing the data. I can also remove jump events from the time series and fit realized volatility measures for the remaining data or, further, fit series containing jumps to other estimates for the process that are robust to discontinuities (in particular, the wavelet realized volatility estimator outlined in sec. 3.2.2.4).

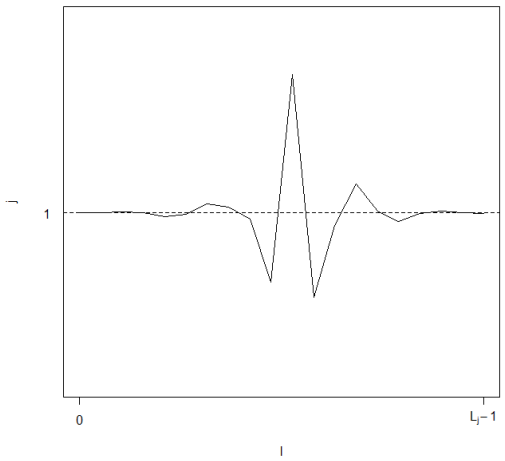
In this chapter I carry out such an analysis, in the order described above. But first, a section will use graphics to illustrate how multi-resolution analysis can be used to detect jumps, using series for particular commodities on illustrative days, as in Chapter Two. Then follow the results of a jump-filtration exercise. The next section is an exploratory analysis of the jump events, examining such questions as how different subsampling (time-aggregation) schemes change the quantity and character of jump events identified. The final two sections contain the results of analyses using various estimators of realized volatility, first with the jumps removed and then for estimators that can accommodate jumps.

4.2 Graphical multi-resolution analysis

Recall (from sec. 3.2.1) that the wavelet filter is an orthonormal function with zero area that can be stretched, shifted and correlated—or "convolved"—with a signal to analyze it and reveal patterns or changes over varying scales. The output is a set of coefficients known as a discrete wavelet transform, or DWT.

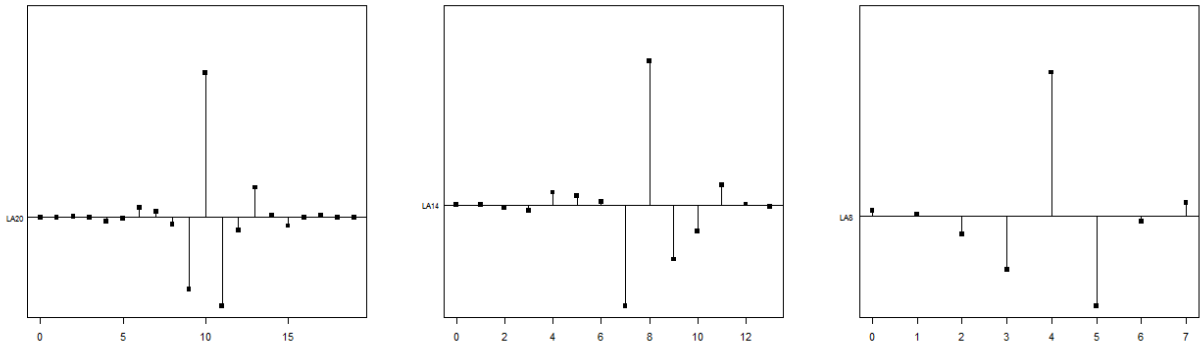
In this section I will illustrate the DWT through an example from my data set. It will be useful to begin by visualizing a wavelet filter.

Figure 4.1: The S(20) wavelet filter



In Figure 4.1, I display a wavelet filter known as a the "Daubechies Least Asymmetric" wavelet, or Symlet, of length 20, which I abbreviate S(20). Though the wavelet basis functions are continuous, the filters used in practice take the form of a set of coefficients "sampled" from the function at regular intervals. The number of points in a given filter is the aforementioned "length."

Figure 4.2: Symlet filters of different lengths



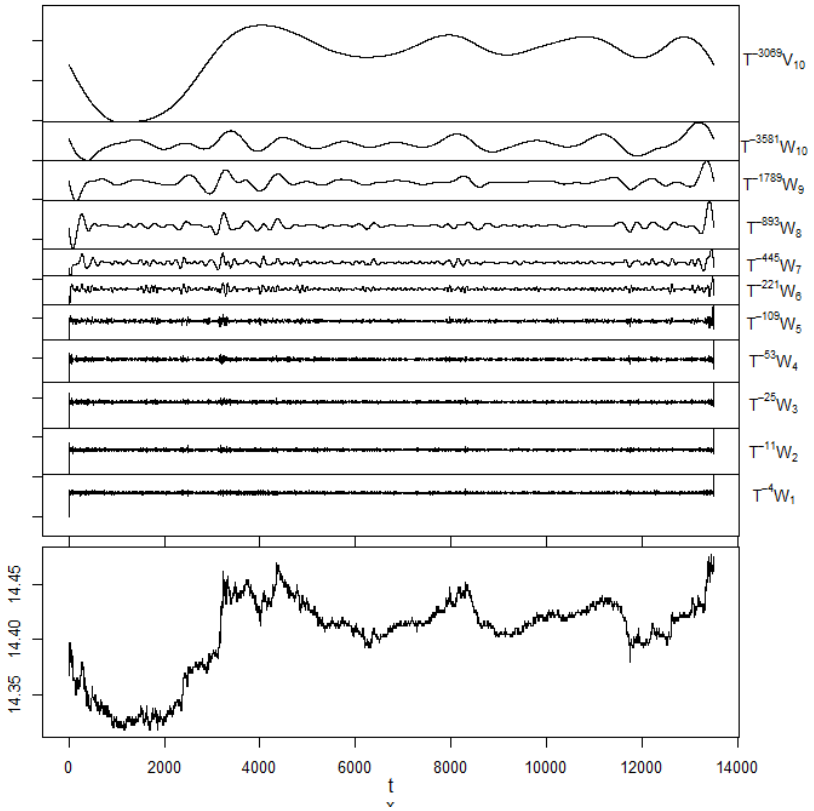
The S(20) filter can be represented as above, as a continuous function, but a more illustrative display is to plot the discrete filter coefficients. The S(20), S(14) and S(8) wavelet filters are plotted in that fashion in Figure 4.2. The use of the term "length" to

distinguish filters can be misleading. Wavelet filters used in the DWT span a variety of time intervals, and the length of the filter choice doesn't change that. Length in this sense refers only to the number of points along the wavelet function used to determine the filter coefficients.

The Multi-Resolution Analysis (MRA) I demonstrate in the remainder of this section makes use of the S(8) filter. This is due to the use of this filter for the illustrations in Fan and Wang (2007), on which much of my methods are based.

A MRA decomposes a time series into a sum of a constant vector, with a series of J vectors that correspond with its variations at different scales, or levels. This can be illustrated in a set of plots.

Figure 4.3: MODWT plot of soybean prices, 8/29/11, 1-second frequency

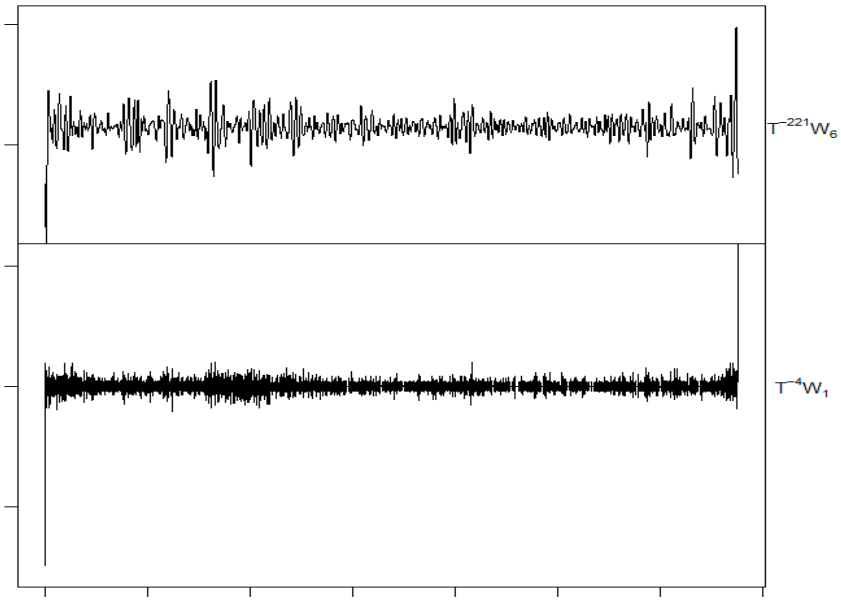


I depict the original time series for the price of soybeans on 8/29/2011 in lower most plot in Figure 4.3 (the same example that was used to illustrate time aggregation in sec. 2.2.2). The plots above it display the DWT coefficients for successive levels of

decomposition, corresponding to variation in the series over different scales. The very top plot corresponds to the scaling coefficients.

The jump in price between 10:15 a.m. and 10:30 a.m. is a good illustration. The lowest level wavelet coefficients hardly respond to the jump, but it is evident in the higher levels.

Figure 4.4: MODWT levels 1 and 6 coefficients for soybean prices, 8/29/11, 1-second frequency



The coefficients for only levels 1 and 6 are plotted in Figure 4.4. Most of the coefficients in either case are close to zero, as they are when variation over a given (time) scale is low. The level 1 coefficients, corresponding to the second-by-second time scale are uniformly low, consistent with minor variation over that scale. The jump in near the end of the first hour of the day is evident, but only barely.

The level 6 coefficients, corresponding to variation at roughly the minute-by-minute frequency, show a more notable response to that jump. It is also evident in some of the higher levels.

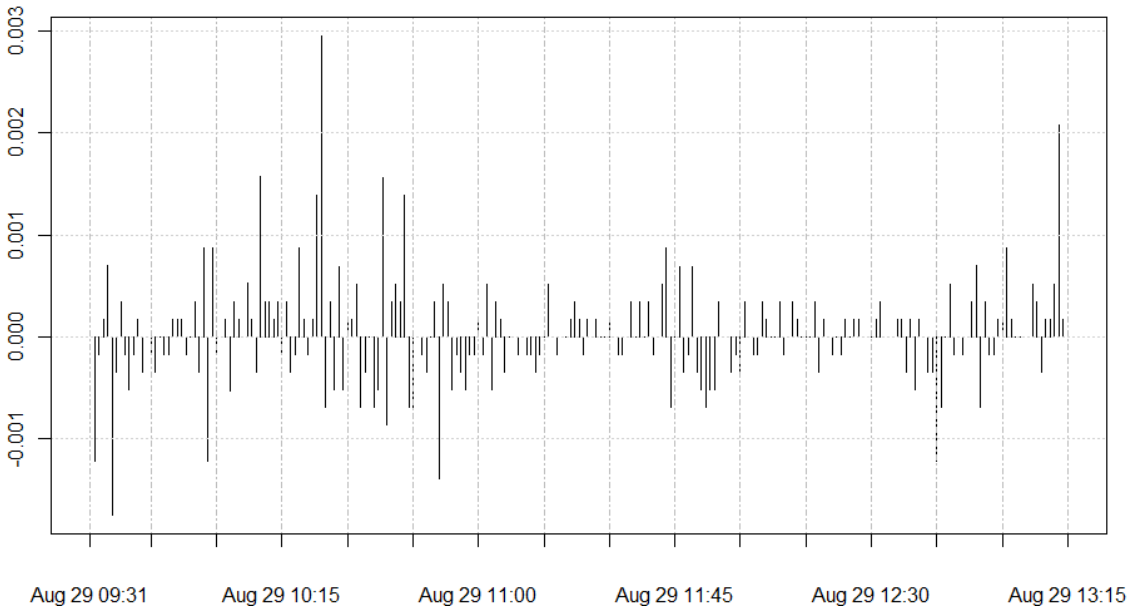
By decomposing a time series in this manner, wavelet methods allow detection of jump events that add noise to volatility estimates or may be of interest in and of themselves. This isn't of major significance when dealing with one time series over the

course of a day; one could simply look at the chart and pick out the jumps visually (though filtering with wavelets is much more precise!). But it will be much more useful in analyzing data at the high-frequency data for several commodities over longer periods, as discussed in the next section.

4.3 Jump detection

Using the same example data series, I can demonstrate the detection of intraday jumps in commodity prices by wavelet methods. The graphical MRA above suggested that the minute-by-minute frequency for the soybean prices might contain jumps. Returns for that frequency are displayed in Figure 4.5 below.

Figure 4.5: Soybean returns, 8/29/11, 1-minute frequency



The next step is to identify a threshold value for the wavelet coefficients such that values greater than the threshold indicate a jump in the returns. A standard value used in practice is referred to as the *universal threshold* by Percival and Walden (2000, pp. 400-402) given by:

$$\delta^u = \sqrt{[2\sigma_\epsilon^2 \ln(N)]} \qquad (19)$$

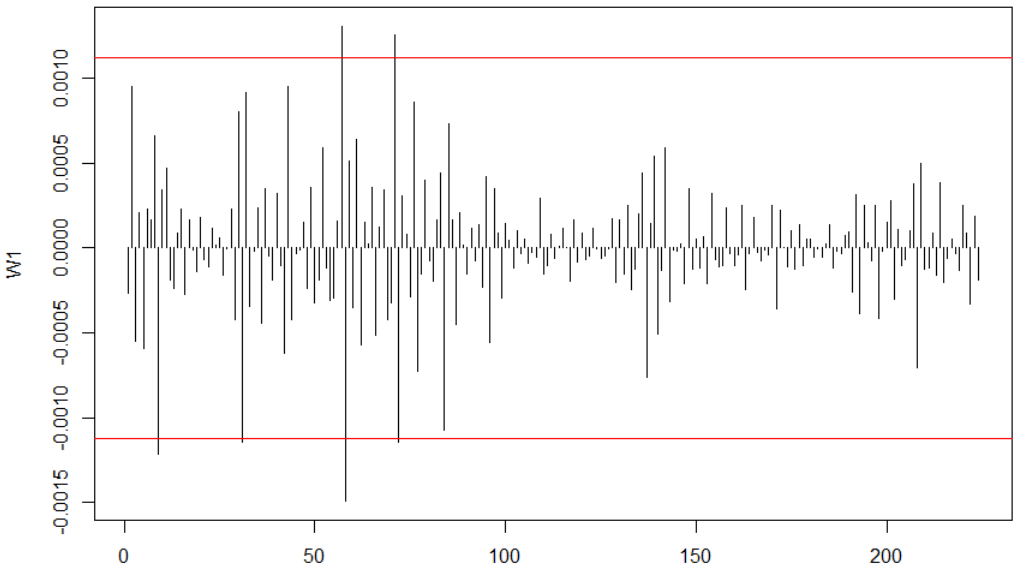
Where σ_ϵ^2 is the variance of the noise in the process $p_t = x_t + \epsilon_t$. Since σ_ϵ^2 is unknown, I must estimate it. Percival and Walden (2000, p. 429) recommend the *median absolute deviation* estimator:

$$\tilde{\sigma}_{mad} = \sqrt{2} * median(|W_{1,0}|, |W_{1,1}|, \dots, |W_{1,N-1}|)/0.6745 \quad (20)$$

Where $W_{1,k}, k = 0, \dots, N-1$, are the level-1 MODWT coefficients (recall that in contrast to the DWT, the MODWT yields N coefficients at each level).

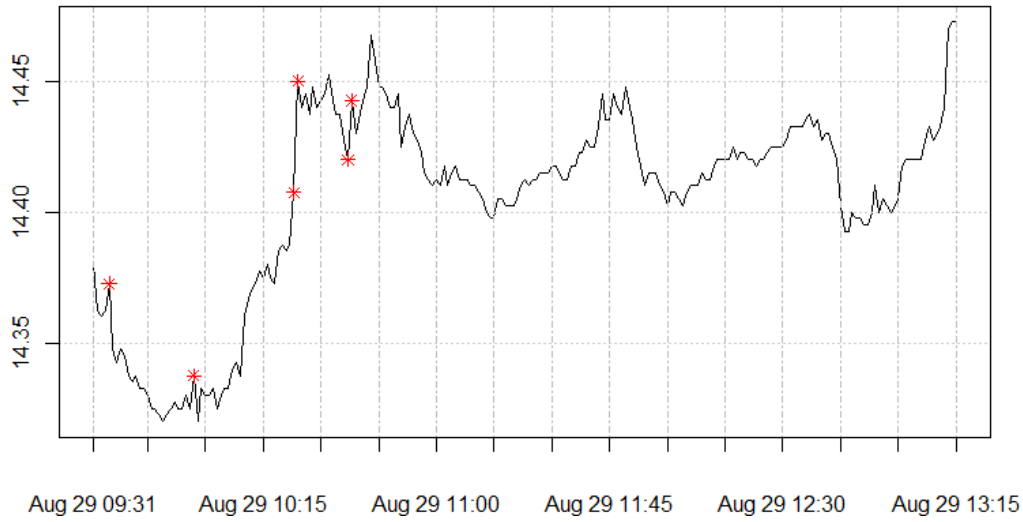
Applying a MODWT on the returns produces a set of wavelet coefficients that can be used to calculate these values. For one-minute returns on 8/29/11, the median absolute MODWT coefficient is 0.000756105057, giving a universal threshold value of $\delta^u = 0.001120986$. The level-1 wavelet coefficients are plotted in Figure 4.6, along with universal threshold values as red bars.

Figure 4.6: Level-1 MODWT coefficients with thresholds (red)



Coefficient values greater than the threshold correspond to large (absolute) minute-by-minute returns and can be mapped to jumps in the price series. Soybean prices at the 1-minute frequency are displayed in Figure 4.7, with indicators (in red) for jumps detected by the foregoing procedure.

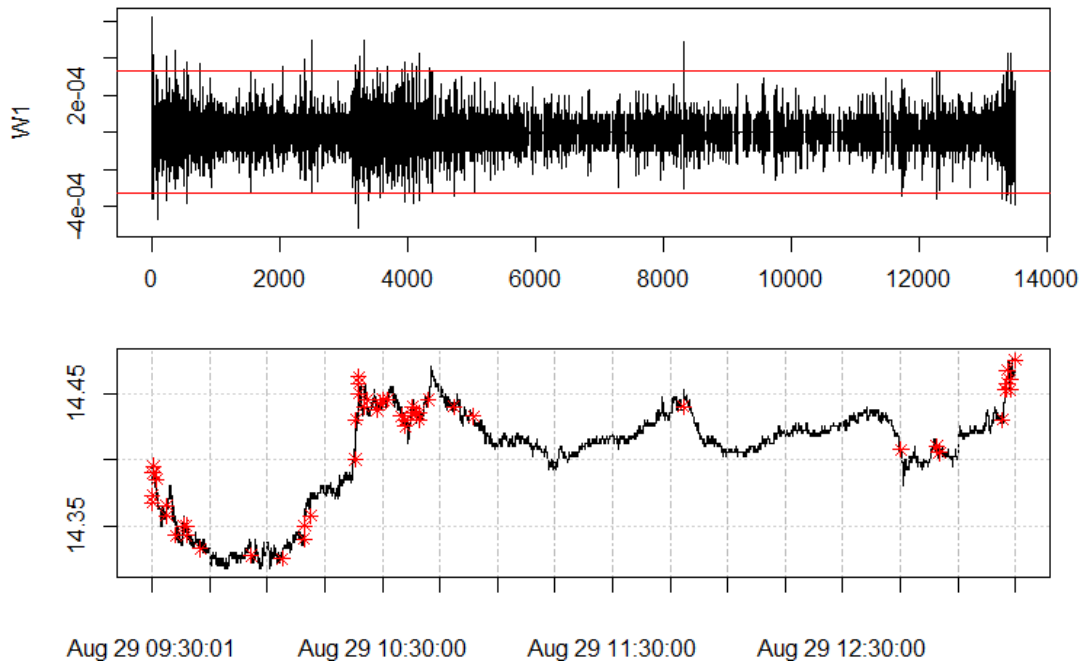
Figure 4.7: Soybean prices, 8/29/2011, 1-minute aggregation, with jump indicators (red)



The locations in the time series correspond to events that appear upon casual observation to be jumps, but don't include all the movements that appear to be jump events. This illustrates that this procedure has some usefulness in comparison to a less formal, exploratory approach (beyond the savings in time).

The choice of a one-minute frequency for detecting jumps in this example was driven by the multiresolution analysis of prices in the previous section. For comparison's sake, I can perform the same jump detection exercise on a higher frequency time scale to compare results. At a one-second aggregation, the level-one wavelet coefficients on the returns series produce a universal threshold value of $\delta^u = 0.0003294727$. In Figure 4.8, I plot the level-one wavelet coefficients for that frequency along with thresholds and the price series with identified jumps indicated as above.

Figure 4.8: Level-1 MODWT coefficients with thresholds (red) and Soybean prices, 8/29/2011, 1-second aggregation, with jump indicators (red)



The comparison is instructive. At a higher frequency, this procedure detects a greater number of jumps. This is to be expected as the higher frequency aggregation is a larger data set and should contain more information on jumps (which in theory are discontinuous events in the continuous time process).

However, the threshold for jumps is smaller at this frequency, and I appear to be detecting "jump" events that are relatively small. This is another illustration of the time scale tradeoff. Without a more detailed analysis it is hard to say more about the appropriate balance to strike on that trade-off. That is the topic of the next section.

4.4 Jump analysis

While in theory a jump in the price process is an instantaneous change in the price at an instant in continuous time, pragmatism necessitates detecting them over some discrete time window. My method can be applied over a variety of time scales, so it is interesting to see how the results change at different frequencies.

The total number of jump events detected using my method over six trading days for each commodity is given in Table 4.1.

Table 4.1: Number of jumps detected at various sampling frequencies

	Jumps detected by commodity		
Frequency	Corn	Soybeans	Wheat
1-second	73	193	291
5-second	7	13	24
15-second	4	13	12
30-second	2	4	14
1-minute	6	9	6
2-minute	4	6	3
5-minute	2	1	2
15-minute	0	2	1

By way of comparison to other studies, Fan and Wang (2007, pp. 1355-1358), on whose approach mine is based, illustrate their jump estimator by applying it to seven months of data on Euro-Dollar and Yen-Dollar exchange rates at the one-minute frequency. The number of daily jumps they detect varies between zero and eight. For the six days in my data set, applying the wavelet jump detection at the one minute frequency identifies between zero and six jumps daily.

As illustrated in the example of the last section, applying the wavelet threshold jump test at a higher frequency detects a greater number of jump events than the number detected over a longer window because the magnitude of the threshold is smaller. A greater number of smaller jumps could induce more volatility over time than a lesser number of larger jump events, or it could induce less. What about the actual contribution of the jumps to volatility?

Recall the definition of jump variation (JV) from chapter two as the summation of squared jump events over time (equation 14). After detecting jump events, estimating the variation in the price process induced by them is straight forward (equation 15). Estimates of JV over the six days are given in Table 4.2. These estimates are associated with risk, as I evaluate uopn in the discussion of Table 4.3 below.

While the highest frequency estimates of JV are greater than some of the lower frequency estimates, the relationship between the time scale and the magnitude of JV is not straightforward; JV estimates increase for some lower frequencies. The question of the optimal sampling frequency for estimating integrated volatility has been well studied; for an early discussion, see Ait-Sahalia, *et al*., (2005). However, less has been said about the proper sampling frequency for identifying jumps.

Table 4.2: Jump variation at various sampling frequencies

	Jump variation by commodity		
Frequency	Corn	Soybeans	Wheat
1-second	6.4559	4.9191	12.2358
5-second	2.9326	2.4575	7.1254
15-second	2.7854	2.8503	7.7747
30-second	3.8687	3.1304	9.0246
1-minute	5.8672	4.2612	6.6584
2-minute	5.4937	4.4493	6.9348
5-minute	6.6131	1.5801	4.9677
15-minute	0.0000	2.0527	1.4608

Note: Value is annualized, in percentage terms from cumulative jump variation over the six trading days

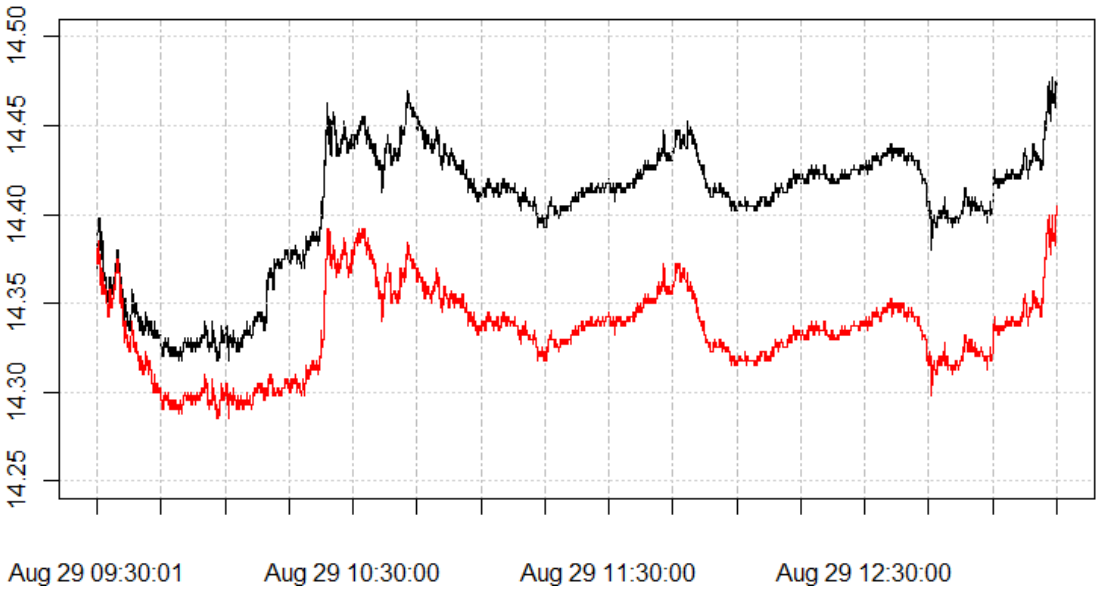
Since jump events happen on a small timescale, intuition dictates that using a higher frequency in testing for them is sensible. This intuition is supported by the recent findings of Christensen, Oomen and Podolskij (2014), who argue that JV estimates using a different method than that of my paper (to be discussed in the next section) are often overstated when using lower frequencies because at those frequencies bursts of volatility could be interpreted as jumps.

While jumps are of interest on their own, they can also improve my estimates of volatility. This approach can detect jumps at one frequency and generate estimates of integrated volatility at another. Doing so can also offer some insight into the questions about identifying jumps discussed above, as shown in the next section.

4.5 Jump-censored realized volatility

As discussed in section 3.2.2.3, after identifying jumps one can remove them in order to produce a simulated jump-free price series. An example is depicted in Figure 4.9, continuing with the soybean data in earlier illustrations. The observed prices at a 1-second frequency are plotted in black, while jump-censored prices are in red.

Figure 4.9: Soybean prices, 8/29/2011, 1-second aggregation, with jump-censored prices (red)



While the jump-free prices remain quite volatile from second to second, their total volatility over the day is obviously less. In this particular example, simple RV over the day is approximately 1.3×10^{-4}, while jump variation was about 1.4×10^{-5}. In other words, jump events accounted for 10.5 percent of observed intraday volatility on a second by second basis.

As this example illustrates, one informative interpretation of the identified jumps is to consider them in terms of the share they contribute to total volatility (quadratic variation). The starting point is to examine simple RV as a measure of observed volatility. Though imperfect, it provides a useful benchmark to quantify the importance of jumps and microstructure noise.

The RV for each commodity over the six trading days is displayed for various frequencies in Table 4.3. The value is annualized, in percentage terms. For example, a

value of 25, as approximately estimated from the returns for wheat at a 30 second frequency, indicates that is the volatility observed in our sample period held for a year, one would observe the price of wheat to vary within a band of about 25 percent of the expected price over that period. (Note that this represents the intraday RV from open to close. Close-to-close volatility includes activity that occurs outside of trading hours. This important additional source of volatility is discussed in Chapter 5.)

Table 4.3: Realized Volatility at various sampling frequencies

	Realized volatility by commodity		
Frequency	Corn	Soybeans	Wheat
1-second	30.2730	18.0117	31.7298
5-second	24.3249	15.2727	28.0803
15-second	21.8033	14.0601	26.1569
30-second	20.6801	13.8361	25.0856
1-minute	20.5099	13.6162	24.1902
2-minute	20.1556	13.6317	23.6297
5-minute	19.2236	12.5900	21.5147
15-minute	19.9663	12.5552	19.9750

Note: Value is annualized, in percentage terms from cumulative realized volatility over the six trading days

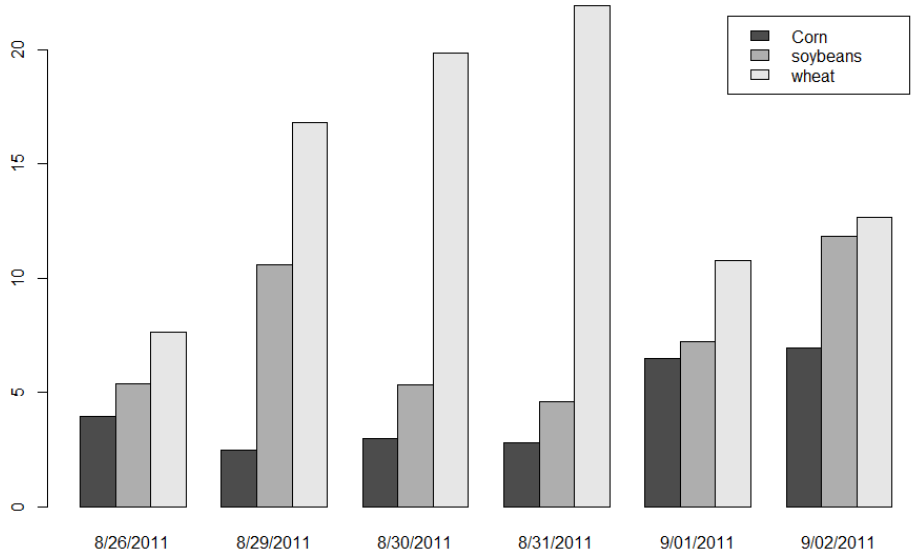
Table 4.4: Jump Variation as a share of intraday volatility

	Jump variation share by commodity		
Frequency	Corn	Soybeans	Wheat
1-second	4.5478	7.4588	14.8706
5-second	1.4535	2.5891	6.4390
15-second	1.6320	4.1095	8.8348
30-second	3.4997	5.1188	12.9421
1-minute	8.1834	9.7939	7.5764
2-minute	7.4292	10.6533	8.6130
5-minute	11.8342	1.5752	5.3314
15-minute	0.0000	2.6729	0.5348

Note: Percent

As indicated in Table 4.4, the share of volatility contributed by jumps can be substantial, but varies by commodity and by frequency. Moreover, the importance of jumps varies from day to day and commodity to commodity. The daily contribution of jumps to total volatility is displayed in Figure 4.10, for each of the three commodities over the six day window of my study, with both jump detection and RV estimates at the 1-second frequency.

Figure 4.10: Jump variation as a share of intraday volatility (percent, 1-second aggregation)



A comparison to other published results is again instructive. Ait-Sahalia and Jacod (2012, pp. 1040-1041), using a somewhat different approach than I employ, estimate the contribution of jumps to quadratic variation to individual stocks and the Dow Jones Industrial Average, over a variety of frequencies and jump thresholds. Though they find that jump variation increases somewhat with sampling frequency, their estimates of its contribution to total volatility are roughly 25 percent for individual stocks and 5 percent to 15 percent for the index, results not markedly different from mine.

Returning to the issue raised at the end of the last section—whether a higher frequency or lower frequency of time aggregation is appropriate for identifying jumps—it

is instructive to compare my method to another common method for quantifying jumps introduced by Anderson, Bollerslev and Diebold (2007, hereafter ABD). The ABD method compares RV to another estimator of integrated volatility: the Realized Bi-Power Variation (BPV) of Barndorff-Nelson and Sheppard (2002), discussed in section 3.1.3. Because BPV is robust to jumps and RV is not, taking the difference of the two is an indirect measure of jump variation; RV - BPV = JV. ABD also suggest this metric as the basis of a significance test for jumps.

In Table 4.5 I display BPV estimates for the same period, frequencies and commodities as the previous tables. It is readily apparent that these estimates are smaller than the RV estimates for the same data. The ABD estimate of JV implied by these is given in Table 4.6.

Table 4.5 Realized Bi-Power Variation at various sampling frequencies

	Bi-Power Variation by Commodity		
Frequency	Corn	Soybeans	Wheat
1-second	19.7112	11.7804	19.5532
5-second	21.1148	13.3533	23.6910
15-second	20.3495	13.1902	24.1603
30-second	19.6163	12.8448	23.4498
1-minute	19.8765	12.8249	22.3295
2-minute	18.8432	12.5040	22.1518
5-minute	18.4424	12.3048	19.5417
15-minute	18.7197	11.8310	17.5874

Note: Value is annualized, in percentage terms from cumulative bi-power variation over the six trading days

Table 4.6 Jump Variation as a percentage of realized volatility

	Jump variation share by commodity		
Frequency	Corn	Soybeans	Wheat
1-second	57.6048	57.2233	62.0246
5-second	24.6517	23.5565	28.8192
15-second	12.8909	11.9910	14.6843
30-second	10.0237	13.8158	12.6166
1-minute	6.0816	11.2852	14.7925
2-minute	12.5985	15.8616	12.1178
5-minute	7.9629	4.4792	17.5004
15-minute	12.0967	11.2037	22.4775

The estimates of JV implied by this method are much larger than those that come from the wavelet jump detection method. One advantage my method has over the ABD measure is that it estimates jump locations as well as size, whereas the ABD measure only quantifies JV over a given time period. But what can be said about the substantial difference in the estimates?

One reason the ABD estimates may be larger than the wavelet-based estimates is that while BPV is robust to jumps, it is not robust to microstructure noise (see chapter 1). It is apparent from Table 4.6 that the JV estimates derived by this method tend to decrease as the frequency of time aggregation gets lower. This is also true of RV and BPV estimates, because at higher frequencies, microstructure noise effects become larger.

But in principle jumps are jumps—instantaneous or very high frequency events. The jump variation of the price process is independent of the sampling frequency an econometrician uses to measure it. So JV estimates should not vary systematically with the sampling frequency. While the wavelet-derived JV estimates are not constant across frequencies, they also do not have an obvious correlation with sampling frequency.

This suggests that a better approach would be to estimate jump variation at a higher frequency, and use an estimator of integrated volatility on the jump-censored

prices that is robust to microstructure noise. The Two-Timescales Realized Volatility (TSRV) estimator due to Zhang, *et al*. (2005) is useful for this purpose.

Recall from section 3.1.3 that the TSRV estimates volatility for a "fast" time scale and a "slow" scale, over a moving window, and averages the two scales. The last scale admits information from high-frequency price fluctuations, while the slow scale smoothes out microstructure noise. This provides two dimensions on which to vary sampling frequency, but in practice, typically the highest frequency available is used for the fastest scale. TSRV estimates of volatility are displayed in Table 4.5 for different values for the slow scale, using the 1-second frequency for the fast scale.

Table 4.7: Two-Timescales Realized Volatility at various sampling frequencies

	Two Timescales RV by commodity		
Frequency	Corn	Soybeans	Wheat
30-second	20.7072	13.3946	24.3326
1-minute	20.5255	13.3108	23.3575
2-minute	20.6388	13.3093	22.5829
5-minute	19.8392	13.0554	21.2882
10-minute	19.6315	12.9667	19.8585
15-minute	19.9553	12.7167	19.6019

The values for TSRV estimates fall between those for RV and BPV estimates, generally closer to BPV estimates. Interestingly, they don't increase dramatically as sampling frequency gets higher. This is consistent both with their being robust to microstructure noise (as that noise increases at higher frequencies). It is also consistent with the jump detection and measurement results which suggests JV is lower for the data than what is implied by the ABD measure; since TSRV is not jump-robust, estimates to be significantly greater than BPV for those frequencies at which the ABD measure implies substantial jump variation.

Finally, using the procedure for detecting jumps and censoring them from the price series, a jump-adjusted-TSRV (JTSRV) measure can produce estimates of integrated volatility that are robust to microstructure noise. These JTSRV estimates for

the six trading days are reported in Table 4.6, using a 1-second frequency for the slow scale and the jump detection.

Table 4.8: Jump-adjusted Two-Timescales Realized Volatility at various sampling frequencies

	Jump TSRV by commodity		
Frequency	Corn	Soybeans	Wheat
30-second	19.5723	13.0694	23.3975
1-minute	19.2357	13.0066	22.2065
2-minute	19.2210	12.9164	21.4208
5-minute	18.2021	12.3323	20.0547
10-minute	17.5357	12.1251	18.7420
15-minute	17.0875	11.8672	18.3870

These estimates, while smaller than the unadjusted TSRV estimates, are not substantially so. The difference is consistent with my estimates of jump variation. Comparing the RV estimates to the TSRV and JTSRV estimates is instructive. At higher frequencies, microstructure noise dominates as a source of realized volatility.

As a practical matter, estimates of RV often use lower frequency subsamples data in which microstructure noise is less prevalent. Doing so not only discards information on high-frequency variation not induced by market microstructure, it also potentially ignores jump events that by definition occur at very high frequencies. Using tests for jumps or measurements of variation at lower frequencies can potentially misinterpret short bursts of volatility as jumps, while applying them at higher frequencies can lead estimates of integrated volatility polluted by noise.

The multiscale approach outlined above addresses these challenges. By identifying jumps at a high frequency, I can measure their contribution to total volatility and remove them from the returns series to estimate volatility at multiple scales. I can thereby empirically decompose quadratic variation into integrated volatility and jump variation.

As an academic matter, this decomposition is of interest for what it tells us about the underlying price process. But of what practical use is it? That is the topic of my final chapter.

Chapter V

Applications

5.1 Introduction

In this final chapter, I examine some applications of the methods developed in this paper. In a sense, the approach to quantifying volatility outlined in chapter two and the results presented in chapter three are their own application. Volatility estimation is of interest in itself. Since it is a key component of risk, volatility estimates can be of use in portfolio and value-at-risk analysis. However, for the time horizons over which most investors are concerned about risk are far greater than those examined here.

However, volatility is of great interest in the pricing and valuation of options and other derivatives. It is a centerpiece of the well-known Black-Scholes option pricing formula, though it cannot be directly observed. However methods have been derived recently for estimating the spot volatility component of that model from high frequency data, and my approach can be used to improve those estimates. Given that these financial instruments have their origins in the markets for agricultural commodities, it is curious that most of the research on them has focused on other assets.

Another useful application of volatility estimates like ours is in forecasting volatility. Future volatility is of interest to market participants, the time series of estimates available with high-frequency data, can be used to derive forecasts of future volatility. I examine two such models, one that forecast intraday volatility, and one that can forecast the important component of volatility due to trading after market close.

Spot volatility estimation and volatility forecasting are straightforward applications of my results. But there are other uses that could potentially be of interest as well. I discuss a few of these briefly before the concluding section.

5.2 Estimating Spot Volatility

Spot volatility is a measure of volatility per unit time over a given time frame (also referred to a instantaneous volatility by Anderson, *et al*. (2010)). It is of interest for a variety of reasons. For example, it has been well documented that volatility changes over

the course of the day and the week according to a rough periodic pattern. These patterns are of interest to traders and risk managers.

Recall from equation 4 that log-prices are assumed to follow a semi-martingale with volatility parameter σ_s and from equation 6 that quadratic variation is defined as $[X,X]_t = \int_0^t \sigma_s^2 ds + \sum_{l=1}^{N_t} J_l^2$. Then spot volatility can be defined in equation 21:

$$\sigma_t^2 = \frac{d[X,X]_t}{dt} \tag{21}$$

As with true prices, X, this quantity is unobservable, but can be estimated. Parametric approaches are similar in spirit to time-series models with seasonality; fit an intraday (constant) curve for the periodic component of volatility, and then measure movements about it. GARCH models with seasonality can also be used.

A typical nonparametric approach is a "piecewise constant volatility" method, in which the trading day is divided into blocks and realized volatility estimates are calculated over those blocks. Recent papers on this topic include Fan and Wang (2008) and Kristensen (2010), who focus on kernel estimators. But I focus on the recent contribution of Zu and Boswijk (2014), who base their estimator on TSRV. This is a natural complement to my approach.

Zu and Boswijk estimate spot volatility using TSRV estimates in a moving filter over a bandwidth h, where h is some interval of the trading day longer than the "slow" TSRV timescale k. They call this the Two-Timescales Realized Spot Volatility (TSRSV) estimator, defined as:

$$\hat{\sigma}_t^2 = \frac{TSRV_t - TSRV_{t-h}}{h} \tag{22}$$

In equation 22, time t is normalized to (0, 1), so h is a fraction of the length of the trading day. Zu and Boswijk also present a "smoothing" version of the TSRSV estimator in which the volatility estimators are calculated over h/2 lead and lagged returns.

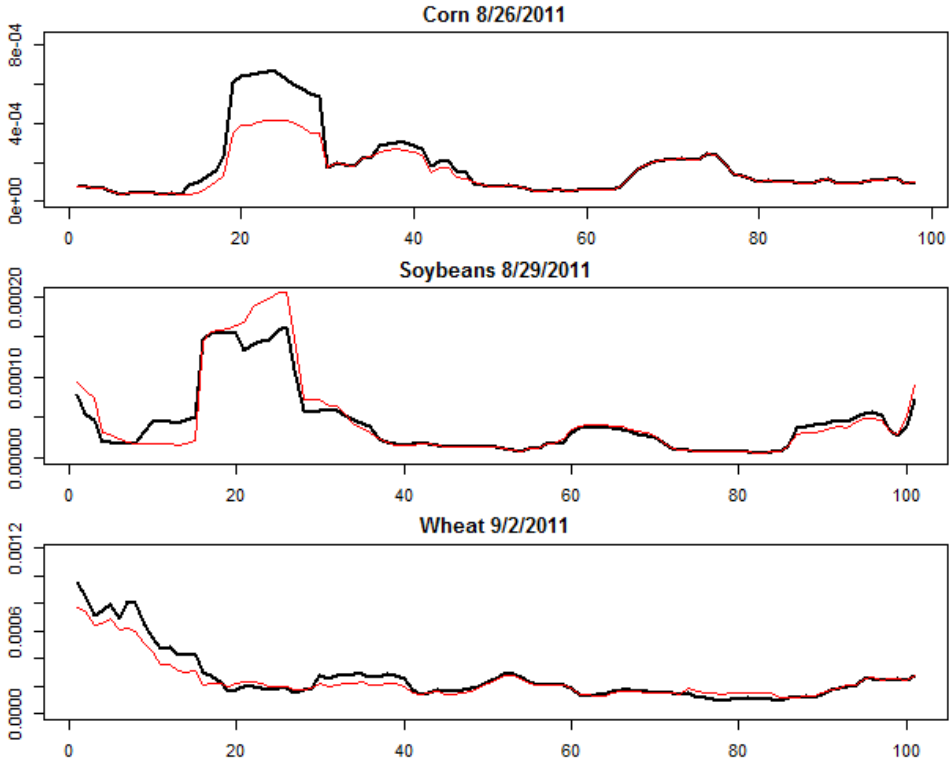
One application of nonparametric spot volatility estimates of this sort is the estimation of intraday changes in volatility. For example, using minute-by-minute returns and a bandwidth of 30 minutes, calculate RV over the first half-hour of the day, then calculate it for the half-hour beginning with the second minute of the day, and so on, producing spot volatility estimates for every minute of the trading day. One advantage of this sort of

procedure is that it allows us to measure time varying volatility on a day-by-day basis; I don't need many days of data to fit a periodic function representing it.

Given that the TSRSV estimator builds off of TSRV, I can apply the jump filtration technique and calculate it on jump-free returns, as in section 4.5. A primary reason doing so is of interest is because spot volatility estimates are intended to reflect the continuous component of volatility—integrated volatility—at a point in time. Jumps are discontinuous idiosyncratic events and the market participants for whom short term volatility matters likely aren't interested in them.

In Figure 5.1, I display an example. Spot volatility estimates are given for the three trading days and commodities used for examples in chapter two. The values in these plots were calculated using two-minutes as the "slow" scale and one second for the "fast" scale. The bandwidth, h, is equal to 1/9 of the trading day or 25 minutes (this was the value chosen by Zu and Boswijk from an optimal bandwidth selection analysis). The TSRSV spot volatility estimates are shown in black, while the jump-TSRSV estimates are in red.

Figure 5.1: Spot volatility estimates derived from TSRV (black) and JTSRV (red)



Though the spot volatilities plotted in Figure 5.1 appear somewhat different, they display common pattern of heightened volatility at four periods of the day: the opening of trading, before and after lunch and at the close of the day. Lunch-related volatility in the data appear less pronounced than for other exchanges, as trading in these commodities ends at 1:15 p.m.

These estimates can be of use in value at risk analysis and hedging, as discussed in the penultimate section of the chapter.

5.3 Forecasting Volatility

In addition to developing instantaneous spot volatility estimates, high frequency data like those for commodities in this study can be used to predict volatility going forward using time series methods. To the extent that they can improve on forecasts of volatility over those generated from day-by-day returns, these forecasts could be valuable to investors and market makers on commodities exchanges (Fleming, *et al*., 2003). Yet again, a procedure like ours that can decompose observed price variation into a jump variation component and an integrated volatility component offers further refinements in existing methods.

For these purposes, there is a distinction between the intraday volatility from open to close on exchanges, and the daily volatility from close to close. The latter includes a return from activity outside exchange hours that causes prices to change after hours. While this is an important source of volatility for market participants, it is less straightforward to model than intraday volatility.

I take these issues up in the two sections that follow. The first discusses a simple univariate class of models for forecasting intraday variation. The second section describes a more complicated set of models that can also forecast close-to-close volatility.

5.3.1 Intraday Volatility

A straightforward way to forecast realized volatility is to fit an autoregressive model that expresses RV on a given day as a function of RV the previous day, or as a function of lagged values of several previous days. The simple AR model can account for the

observation that high-volatility trading days tend to be followed by high-volatility days. But what about horizons longer than a day?

An extension proposed by Corsi (2009), called the Heterogeneous Autoregression model for Realized Volatility (HAR-RV), includes terms for the previous week (5 trading days), month (22 trading days), etc. The expectation of volatility for day t+1 on day t is given by the linear model:

$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \epsilon_{t+1} \quad (23)$$

where $RV_{t,\,t+h}$ is the average daily variation over h days: $1/h(RV_{t+1} + RV_{t+2} + \ldots + RV_t)$. Note that I use RV to represent the measure of variation in the notation, but I can use other measures, such as BPV or TSRV.

Anderson, Bolerslev and Diebold (2007), modified the HAR-RV model to incorporate jumps. They noted that, in addition to the effect jumps have on RV estimates, the jump component of variation was less persistent than the integrated volatility component. For those reasons, explicitly accounting for jumps in the model could improve forecast results. They incorporate a component for jump variation at day t in the HAR-RV model, similarly to an error term, referring to this version as HAR-RV-J:

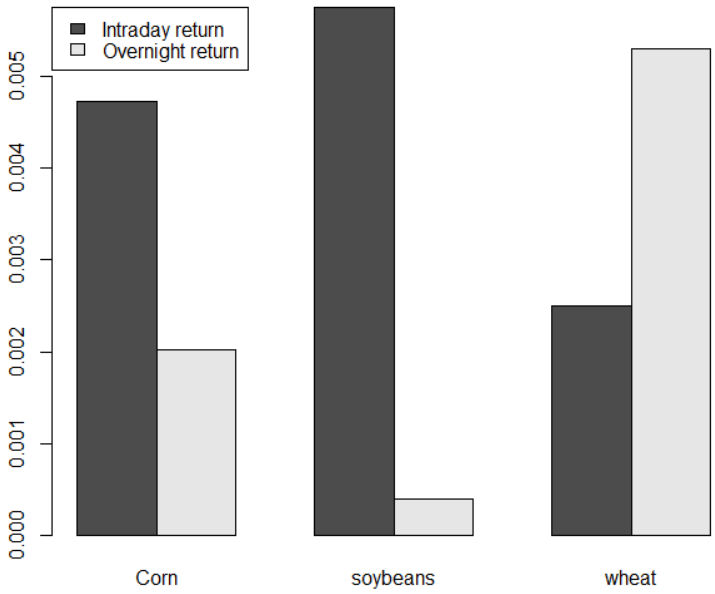
$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + J_t + \epsilon_{t+1} \quad (24)$$

Anderson, *et al*. suggested estimating the jump variation component using the ABD statistic described in section 3.4. However, as I note there, that measure may overstate jump variation. This is where my method can be applied. My estimates of jump variation can simply be plugged into the HAR-RV-J model to generate volatility forecasts.

5.3.2 Close-to-Close Volatility

While HAR models can be applied to forecasting intraday realized variation, they aren't informative about an important source of short term volatility: the overnight returns from afterhours price changes. As illustrated in Figure 5.2, which displays average absolute intraday and overnight returns for each of the three commodities in the data set over the six example days the overnight return is substantial.

Figure 5.2: Intraday vs. overnight returns (absolute) for three commodities



In order to account for these effects, Shephard and Shephard (2010) introduced a two equation model called the High-frEquency bAsed Volatility (HEAVY). Represent by $\mathcal{F}_{t-1}^{HF}$ the past series of daily returns, $r_1, r_2, \ldots, r_T$, and daily realized measure of volatility $RM_1, RM_2, \ldots RM_T$, based on the high frequency data set. The HEAVY model is the system:

$$var(r_t \mid \mathcal{F}_{t-1}^{HF}) = h_t = \omega + \alpha RM_{t-1} + \beta h_{t-1}, \qquad \omega, \alpha \geq 0, \beta \in [0,1) \qquad (25)$$

$$E(RM_t \mid \mathcal{F}_{t-1}^{HF}) = \mu_t = \omega_R + \alpha_R RM_{t-1} + \beta_R \mu_{t-1}, \omega_R, \alpha_R, \beta_R \geq 0, \alpha_R + \beta_R \in [0,1) \quad (26)$$

In contrast to the application of my jump detection method to the HAR model, in which I used the wavelet-derived measures of jump variation to parameterize the jump component, in the HEAVY model the jump-censored volatility measures can be used to fit the model.

5.4 Other Applications

The applications presented in this chapter are just two examples of uses that high frequency volatility and jump estimates could be put to. The following are examples of other applications to which these methods might lend themselves.

Hedging. The parameter estimates can be used to forecast returns and volatilities. The estimates are derived from spot prices on underlying commodities for which there are also exchange traded futures and options. Given the forecasts of returns and volatilities, these derivative prices can be used to construct a portfolio of commodities and derivatives that optimally hedges against risk while maximizing returns.

Option Valuation. The parameters describing volatility are fundamental in standard option pricing models such as Black-Scholes. Therefore, with the estimates of volatility derived in this study, one can derive the implied efficient or "true" price of options. The estimates can also be inputs into non-parametric option pricing models such as that of Ait-Sahalia and Duarte (2003). Volatility estimates that are robust to microstructure noise and that account for jumps would improve the estimates of option prices, and would be a value in options trading strategies that make use of opton pricing models.

Moreover, the jump estimates I derive are useful in applied option pricing. Since it has been long understood that a continuous Brownian motion assumed in Black-Scholes might not adequately capture asset prices that jump discretely, option pricing models based on a jump-diffusion process have been developed. Tsay (2010, sec. 6.9) gives an example of such a model, from which prices of call and put options can be derived using parameters describing the expected number of jump events during the time to expiry of the option, the average size of the jump event, and it's standard error. Each of these parameters can be derives empirically using the methods outlined in this paper.

High-Frequency Arbitrage. With the estimates of the efficient option price derived using my methods it is possible, in principal, to exploit short-term discrepancies between them and the market price to reap arbitrage profits. This violates the core no-arbitrage assumption at the heart of asset pricing theory. However, one rationale for this

assumption is that arbitrage opportunities that arise are bid away by trading, thus won't exist in equilibrium. In practice, arbitrage is big business. The intellectual appeal of high-frequency data is that it allows us to observe this equilibration process, but this may also open the door to arbitrage strategies. Whether these opportunities can actually be exploited in practice is another matter.

Manipulation Detection. Finally, I return to the brave new world of electronic trading covered in the first chapter. With this analysis it it possible to conduct an exercise in "forensic finance," that could be of interest to exchange operators and regulators. Using the jump detection technique, one can search for patterns in market activity in an attempt to detect strategic (possibly illegal) manipulation of market prices. The question here is whether jumps in the order book forecast jumps in the price of commodities, by how long, and in what way. I can only provide preliminary answers to these questions, but they hold great promise if electronic trading is to be made to work well.

Chapter VI
Summary and Conclusions

6.1 Summary

The recent availability of large datasets of tick-by-tick price and transaction information from exchanges presents an exciting new frontier for financial econometrics. Among other things, they allow us to observe the process of equilibration in real time. As I illustrate in the discussion in Chapter 2, these data create novel challenges for analysts. Methods based on the traditional assumption that prices reflect available information aren't adequate when prices are observed at a high enough frequency to reflect the arrival of information.

While the research literature that has developed so far has focused on data from equity and foreign exchange markets, the methods derived in these studies can be applied to data from commodities markets as well. This was the first objective of my study mentioned in section 1.4: *analyze volatility in high frequency commodity exchange data*.

In Chapter 3, I reviewed some of the many methods in from studies of high-frequency data on realized volatility, noting that while simple realized volatility measures can be used to estimate volatility from high-frequency data, they are biased by microstructure noise and by the presence of jumps in the price process. That relates to my second research objective: *identify jumps in the data* to disentangle the jump component, *analyze volatility separately from jumps*, and *quantify jumps and noise* as well.

In Chapter 4, I presented a selection of such methods that make it possible to identify "jump" events that may correspond to the sudden arrival of new information. By identifying such events, it is possible to disentangle the portion of observed volatility that corresponds to jumps from that portion that is associated with ongoing (integrated) volatility in the sense asset pricing models usually focus on.

My final research objective was to *measure instantaneous volatility and forecast future volatility* using my estimates. As I demonstrate in Chapter 5, these methods and the

estimates derived from them are pertinent to existing applications. Estimates of spot volatility and forecasts of future volatility are improved by using better estimates. These applications are of particular interest to participants in commodities markets, whether they are market makers, investors, or the actual holders of physical commodities who are exposed to these risks.

6.2 Conclusions

I began this thesis by emphasizing a basic trade-off to the analyst of a high-frequency financial time series: making use of all the information is such large datasets while accepting the noise that comes along with it, versus reducing noise by subsampling at lower frequencies but discarding potentially informative (and valuable) information.

What can be said about the how the approach followed in this paper resolves that trade-off? As I illustrate in Chapter 4, the wavelet-based method of filtering jumps can produce estimates of volatility that decompose it into separate components that measure the volatility contributed by jump events separately from the integrated volatility. The results indicate that this approach can lead to quite different estimates of realized volatility compared to measures that are not robust to jumps.

The results also differ substantially by that time scale is chosen. The methods developed in this study can retain the information contained in higher frequency prices. By applying estimators of realized volatility that are robust to microstructure noise, I can make use of this information without discarding it. In Chapter 5, I illustrate that these methods and the estimates derived from can make a difference in applications. Measures of spot volatility in particular can be improved by using better estimates.

6.3 Future directions for research

High-frequency financial econometrics is a young field, so there are many interesting questions analysts will tackle in the coming years. Here, I focus on a few that are direct consequences of my work in this thesis.

First, in my analysis I focused only on six days of trading from late 2011, but much more data has become available since. An obvious extension of this study would be

to apply my methodology to a much larger set of commodities and number of trading days. Are the estimates of volatility here consistent for a longer time horizon? A longer period would greatly improve the volatility forecasts discussed in section 5.3.

Another extension of my analysis that a longer span of data could shed light on is the question of statistical significance. My approach focuses, as much of the literature on jump detection does, on detecting jumps based on a given threshold. However, another approach, as in Lee and Mykland (2008), focuses on statistically testing for the presence of jumps over a given time scale. That is, over some period, is a high level of volatility that is attributed to jump events actually significantly greater than what might otherwise be observed as continuous volatility in the absence of jumps? While my approach has the advantage of detecting and quantifying individual jumps, a significance test could sort out jumps that were detected spuriously. This is particularly relevant at very high frequencies where the observed jump variation can be greater.

Another relevant question from an economist's perspective is how valuable these estimates of volatility are. High-frequency data are expensive, but are the worth it. As mentioned in Chapter 5, studies like those of Flemming et. al. (2003) have attempted to quantify the value to portfolio managers and others of volatility forecasts generated from high-frequency data. Further, section 5.4 discusses the potential value of realized volatility estimates like mine in option pricing. How valuable these estimates are to commodity market participants and whether they justify the price of the data is an open question.

Finally, a large portion of this thesis is dedicated to identifying and analyzing jump events in the price process. These are treated as random shocks, but should reflect the arrival of new information or other events. With the availability of much more high-frequency commodity market data, a more systematic analysis of jump events is possible—what, if any, "news" do they reflect, are they correlated with previous market events or price fluctuations, etc.

This last returns to what is most fundamentally interesting about high-frequency financial data. They can open up a window the process by which these markets arrive at equilibrium. They can illuminate discussions of the transition to electronic exchanges and

high-frequency trading on which so much coverage has focused. And, perhaps, they can ultimately inform policy.

References

Ait-Sahalia, Yacine and Jefferson Duarte. 2003. “Nonparametric option pricing under shape restrictions.” Journal of Econometrics 116 (1-2): 9-47.

Ait-Sahalia, Yacine and Jean Jacod. 2012. “Analyzing the spectrum of asset returns: Jump and volatility components in high-frequency data.” Journal of Economic Literature 50 (4): 1007-1050.

Ait-Sahalia, Yacine and Jailin Yu. 2008. “High frequency market microstructure noise estimates and liquidity measures.” NBER Working Paper No. 13825.

Ait-Sahalia, Yacine, Per Mykland and Lan Zhang. 2005. “How often to sample a continuous-time process in the presence of market microstructure noise.” Review of Financial Studies 18 (2): 352-416.

Anderson, Torsten, Tim Bolerslev and Francis X. Diebold. 2007. “Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility.” Review of Economics and Statistics 89 (4): 701-720.

Anderson, Torsten, Tim Bolerslev, Francis X. Diebold and Paul Labys. 2003. “Modeling and forecasting realized volatility.” Econometrica 71 (2): 579-625.

Anderson, Torsten, Tim Bolerslev, Francis X. Diebold and Paul Labys. 2010. “Parametric and nonparametric volatility measurement.” In Handbook of Financial Econometrics: Volume 1: Tools and Techniques, edited by Yacine Aıt -Sahalia and Lars Peter Hansen, 67-137. Amsterdam and Boston: Elsevier, North -Holland.

Barndorff-Nielsen, Ole E. and Neil Shephard. 2002. “Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models.” Journal of the Royal Statistical Society, Series B, 64: 253-280.

Barndorff-NielsonBarndorff-Nielsen, Ole E., and Neil Shephard. 2004. “Power and Bipower Variation with Stochastic Volatility and Jumps.” Journal of Financial Econometrics 2 (1): 1–37.

Budish, Eric, Peter Cramton and John Shim. 2015. “The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response.” *Quarterly Journal of Economics* 130: 1547-1621.

Campbell, John Y., Andrew W. Lo and A. Craig MacKinlay. 1997. The Econometrics of Financial Markets. Princeton, N.J.: Princeton University Press.

Corsi, Fulvio. 2009. “A simple approximate long-memory model of realized-volatility.” Journal of Financial Econometrics 7: 174–196.

Dodd, Randall. 2010. “Opaque trades.” Finance and Development. MF..arch, 2010.

Engle, Robert and John Russell. 2010. “Analysis of high frequency data.” In Handbook of Financial Econometrics: Volume 1: Tools and Techniques, edited by Yacine Aıt-Sahalia and Lars Peter Hansen, 67-137. Amsterdam and Boston: Elsevier, North-Holland.

Fan, Jianquing and Yazhen Wang. 2007. “Multi-scale jump and volatility analysis.” Journal of the American Statistical Association 102 (480): 1349-1362.

Fan, Jianquing and Yazhen Wang. 2008. “Spot volatility estimation for high-frequency data.” Statistics and its Interface 1: 279-288.

Flemming, Jeff, Chris Kirby and Barbara Ostidek. 2003. “The economic value of volatility timing using ‘realized’ volatility.” Journal of Financial Economics. 67: 473-509.

Fugal, Lee. 2009. Conceptual Wavelets in Digital Signal Processing. San Diego: Space and Signals Technical Publishing.

Hasbrouck, Joel. 2007. Empirical market microstructure. New York: Oxford University Press.

Hubbard, Barbara. 1998. The world according to wavelets. Natick, Mass.: AK Peters.

Johnson, Barry. 2010. Algorithmic trading and DMA: An introduction to direct access trading strategies. London: 4Myeloma Press.

Lee, Suzanne and Per Mykland. 2008. “Jumps in financial markets: A new nonparametric test and jump dynamics.” Review of Financial Studies 21 (6): 2553-63.

Lowenstein, Roger. 2012. “A speed limit for the stock market.” New York Times. Oct. 1, 2012.

Mandlebrot, Benoit. 1963. “The Variation of certain speculative prices.” Journal of Business 36: 394-419.

Poterba, James and Lawrence Summers. 1986. “The persistence of volatility and stock market fluctuations.” American Economic Review 76: 1124-1141.

Najimi, Amir-Homayoon. 2012. Wavelets: A Concise Guide. Baltimore: Johns Hopkins University Press.

Percival, Donald B. and Andrew T. Walden. 2000. Wavelet Methods for Time Series Analysis. New York: Cambridge University Press.

Ramsey, James B. 2002. “Wavelets in economics and finance: Past and future.” Studies in Nonlinear Dynamics & Econometrics 6 (3)

Schwert, G. W. 1989. “Why does stock market volatility change over time?” Journal of Finance, 44, pp. 1115–1153.

Shephard, Neil and Kevin Shephard. 2010. “Realising the future: Forecasting with high -frequency-based volatility (HEAVY) models.” Journal of Applied Econometrics 25:197 -231.

Shumway, Robert H. and David S. Stoffer. 2011. *Time Series Analysis and Its Applications*. New York: Springer-Verlag.

Tsay, Ruey S. 2010. Analysis of Financial Time Series. Hoboken, N.J.: John Wiley and Sons.

Wasserman, Larry. 2006. All of Nonparametric Statistics. New York: Springer-Verlag.

Yogo, Motohiro. 2008. “Measuring business cycles: A wavelet analysis of economic time series.” Economics Letters 100 (2): 208-212.

Zhang, Lan; Per Mykland and Yacine Ait-Sahalia. 2005. “A tale of two time scales: Determining integrated volatility with noisy high-frequency data.” Journal of the American Statistical Association 100 (472): 1394-1411.

Zu, Yang and Peter Boswijk. 2014. “Estimating spot volatility with high-frequency financial data.” Journal of Econometrics. 181 : 117-135.