Regime Switching and Technical Trading with Dynamic Bayesian Networks in High-Frequency Stock Markets

by

Aditya Tayal

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master in Mathematics
in
Statistics–Finance

Waterloo, Ontario, Canada, 2009

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Aditya Tayal

Abstract

Technical analysis has been thwarted in academic circles, due to the Efficient Market Hypothesis, which had significant empirical support early on. However recently, there is accumulating evidence that the markets are not as efficient and a new theory of price discovery, Heterogenous Market Hypothesis, is being proposed. As such, there is renewed interest and possibility in technical analysis, which identifies trends in price and volume based on aggregate repeatable human behavioural patterns.

In this thesis we propose a new approach for modeling and working with technical analysis in high-frequency markets: dynamic Bayesian networks (DBNs). DBNs are a statistical modeling and learning framework that have had successful applications in other domains such as speech recognition, bio-sequencing, visual interpretation. It provides a coherent probabilistic framework (in a Bayesian sense), that can be used for both learning technical rules and inferring the hidden state of the system. We design a DBN to learn price and volume patterns in TSE60 stock market and find that our model is able to successfully identify runs and reversal out-of-sample in a statistically significant way.

Acknowledgements

I would like to express my deep gratitude to my advisor and mentor Yuying Li for her support, patience and insightful comments at every stage of this thesis. She showed me how research can be a fun and rewarding process. Thank you for believing in me.

I would also like to thank my family and Adrienne. Their love, support and endless caring was always present in my everyday life.

To Adrienne.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Traditionally, there have been two main schools of thought in financial markets—technical analysis and fundamental analysis. Technical analysis is the study of trends in price and volume, while fundamental analysis concerns itself with economic factors and the projection of performance based on these factors. These two approaches need not be exclusive; indeed they can complement each other. Technical analysis tools can be used to draw significance to various economic trends and knowing economic trends can aid the technician in determining the potential significance of various technical signals and patterns [Murphy 99].

Despite the infiltration of technical analysis in industry practice, academic finance has been slow to accept it. In fact, among certain critics, technical analysis is viewed as a form of black magic. Indeed, in his influential book A Random Walk down Wall Street, [Malkiel 03] concludes that "[u]nder scientific scrutiny, technical analysis must share a pedestal with alchemy."

Much of the criticism of technical analysis has its roots in academic theory—specifically the efficient market hypothesis (EMH). EMH states that in a market populated with homogeneous, rational and fully informed agents, and in the absence of transaction costs, the market price will fully reflect all available information [Fama 70]. Thus the market's price is

always the correct one—any past trading information is already reflected in the price of the stock—and any attempt to predict price is useless.

EMH cannot be tested directly, since it assumes that the market price is actually the best estimate we have of the asset's intrinsic value. [Fama 70] offers a way to test the predictions of EMH on real-world markets, by identifying three sources of information, corresponding to three increasing degrees of informational efficiency that can be tested separately: in the weak form, prices are supposed to fully reflect all the information contained in historical information, so that no excess return can be achieved by following technical analysis strategies; in the semi-strong form, prices more generally reflect all sorts of information publicly available, so that prices quickly adjust to news and consequently even fundamental analysis is of little use in finding investment opportunities; finally, in the strong form, prices reflect all types of information, public or private, so that no one can use monopoly information to get excess profit. At the time, EMH was supported by a large body of empirical research [Samuelson 65, Fama 69, Jensen 67]. In particular, the weak form of EMH is consistent with a random walk model, such as brownian motion or more general Lévy processes.

However, recent work has questioned the validity of EMH [LeBaron 96]. Instead of assuming a homogenous market, in which all agents interpret news and react to it in the same way, a heterogenous market is proposed, in which agents act in different time horizons and in differing ways [Dacorogna 01]. Also, emerging discoveries in behavioural economics maintain that human psychology, not always rational, is intertwined with price processes [Kahneman 79]. Finally, random walk models are unable to explain properties of real world markets such as volatility clustering and correlations between waiting times of orders [Liu 99, Cont 01].

These results imply systematic patterns may exist in price action. [Lo 00] evaluated the effectiveness of chart patterns, and found that over a 31-year sample period several of them provided incremental information. Furthermore, stock prices are found to fluctuate far too much compared

with what could be expected from variations in the dividend process which are supposed to underlie the fundamental value [Shiller 81]. Also, studies have shown that volume can be a significant information source to price movement [Karpoff 87].

However, profiting using technical analysis is still open for debate. [Brock 92] show trading rules out-perform a buy-and-hold strategy on the Dow Jones Industrial Average Index. However, they do not include transaction costs in their analysis. [Bessembinder 98] replicate their work but include transaction costs and show that these subsume the profitability documented by [Brock 92]. A more comprehensive study of simple technical trading is surveyed in [Canegrati 08], with results that are mixed and dependent on the market and economy chosen.

While most studies have evaluated technical analysis at the daily frequency, technical analysis may be most useful at higher frequencies, when fundamentals are changing the least. The practice of day trading puts this to the test. Day traders engage in the buying and selling of securities many times during the course of a day based on short-term price volatility. They typically close out positions by the end of the trading day in order to avoid risk when the markets are closed.

Traditionally, the primary means of detecting trends and patterns has involved statistical methods such as clustering and regression analysis and more recently the Autoregressive Conditional Heteroscedastic (ARCH) model and its descendant, Generalized ARCH (GARCH) model. The mathematical models associated with these methods for financial forecasting, however, are linear and may fail to forecast the turning points because in many cases the data they model may be highly nonlinear. As a result, machine learning paradigms are becoming prevalent tools for analyzing markets since they inherently handle non-linear modeling. For example, [Austin 04] uses genetic algorithms to optimize a set of technical indicators for foreign exchange markets, [Nevmyvaka 06] addresses optimal execution strategies with a reinforcement learning algorithm, [Zhang 98] survey the use of artificial neural networks for financial forecasting, and

hidden Markov models have been used to estimate latent variables in models [Mamon 07].

In this thesis, we propose using Dynamic Bayesian Networks (DBNs) as a natural approach for financial forecasting and technical analysis. As far as we are aware it is the first study to investigate the use of DBNs in this capacity. DBNs are playing an increasingly important role in the design and analysis of machine learning algorithms. They provide a flexible and coherent probabilistic framework for modeling temporal data using the Bayesian network formalism—a marriage of probability theory and graph theory in which dependencies between variables are expressed graphically. Many of the classical multivariate probabilistic systems studied in statistics, systems engineering, information theory, pattern recognition and statistical mechanics are special cases of the general graphical model formalism—examples include mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models [Jordan 98]. Indeed, DBNs provide a powerful tool for dealing with uncertainty and complexity in a system that evolves over time [Murphy 02].

Furthermore, the DBN graphical model formalism provides a framework for the design of new systems—ideal for modeling high-frequency markets with embedded patterns. Fundamental to the idea of a graphical model is the notion of modularity: a complex system is built by combining simpler parts. Probability theory provides the glue whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data. They have been successfully employed in various other applications including speech recognition, gene sequencing, bio-informatics, visual object tracking, and medical diagnosis.

In this thesis, we design a DBN based on a hierarchical hidden Markov model to learn price and volume patterns in intraday data. We carefully design our features based on technical analysis priors of price and volume behaviour, and learn the model parameters using a historical window. The flexibility allowed by DBNs requires that one design the structure well—too complicated a structure results in over-fitting the learning data,

whereas too sparse a structure does not capture meaningful patterns.

Our model learns two distinct states in high-frequency data—one where volume and price behaviour indicates buying pressure (buying state) and another where it indicates selling pressure (selling state). Buying pressure is identified when price increases have accompanying volume increases and price decreases have accompanying volume decreases; selling pressure is identified when the reverse is true, that is when price increases have volume decreases and price decreases have volume increases. Distributions of price and volume are learned for the two unique states that maximize the likelihood of the observation sequence. We use intraday tick data from sixty stocks of the S&P/TSE 60 and find that during the buying state there is positive expectation in price and during the selling state there is negative expectation in price. We also investigate the predictive power of this model and obtain statistically significant evidence that high-frequency price and volume behaviour can identify intraday runs and reversals ex-ante. We conclude that dynamic Bayesian networks can provide a powerful approach for analyzing markets and is a promising technique upon which more complex models of market behaviour can be built.

1.1 Contributions

This thesis is an interdisciplinary work that involves aspects of machine learning, statistics and finance. The major contributions are:

1. Approach technical analysis using a coherent probabilistic framework (dynamic Bayesian networks) within a regime switching context.

2. Design a price and volume dynamic bayesian network for high-frequency stock markets based on technical analysis concepts. It is the first publicly available application of dynamic bayesian networks on high frequency stock data.

3. Verify significance of regimes in high-frequency markets (S&P500 TSE60) ex-ante; investigate profitability and properties of simple trading strategy based on model.

1.2 Thesis outline

Chapter 2 reviews basic financial theory about markets and introduces technical analysis; chapter 3 describes dynamic Bayesian network theory; chapter 4 motivates and proposes the price and volume model; finally, chapter 5 discusses and analyzes results of the model when applied to market tick data.

Chapter 2

Financial Theory

There have been two main approaches to financial markets, fundamental analysis and technical analysis. The former attempts to ascertain intrinsic value of financial assets, while the latter attempts to identify trends in them. The goal of both methods is to forecast or project performance of the asset.

In this chapter we review background theory relating to financial markets, providing context for both fundamental and technical analysis. We also review recent studies which have shed light on the complexities in the market. We begin by defining the Efficient Market Hypothesis and its implications. We then describe an alternate theory, Heterogenous Market Hypothesis, that explains some of the empirical findings; finally, we explain the basic idea of technical analysis and why it may have validity.

2.1 Efficient Market Hypothesis

Information is what drives investment decisions and trading. Security prices are a result of this process: as market participants engage in trading—buyers meet sellers—and new price levels are established. Whether this market price actually reflects the intrinsic value of the asset

is a difficult question. The Efficient Market Hypothesis (EMH) addresses this question by turning the problem upside down. [Fama 70] defines, "A market in which prices always fully reflect available information is called efficient." Therefore, in efficient markets, information should be reflected in prices with an accuracy that leaves no investor an incentive to search for more information or to trade. As a result, it assumes that the market price is actually the best estimator we have of the asset's intrinsic value.

An idealized "frictionless" market is defined as a market where the following hold [Schwartz 04],

- There are no taxes, no transaction costs, and no short-selling restrictions.
- All investors are fully informed ($\forall i, \phi_{i,t} \equiv \phi_t$) and, being fully informed, have the same (homogenous) expectations about what prices will be in the future ($\forall i, p_{i,t+1} \equiv E[p_{t+1}|\phi_t]$).
- Unlimited amounts can be borrowed or lent at a constant, risk-free rate.
- Markets are perfectly liquid.

where $\phi_{i,t}$ represents the information set known to investor i at time t, ϕ_t represents all information available at time t, p_{t+1} is the price of the asset at time $(t+1)$ and $p_{i,t+1}$ is investor i's expected price for time $(t+1)$ at time t.

In such a market, market efficiency is a result of two basic mechanisms: traders' rationality and arbitrage. Rational traders imply that traders base their demand function (their orders) on their expectations of an asset's fundamental value, centered around the fundamental value of the asset, $E[p_{t+1}|\phi_t]$. [Sharpe 98] defines arbitrage as "the simultaneous purchase and sale of the same, or essentially similar, security in two different markets for advantageously different prices". It is the means by which any new information quickly (instantaneously, in a frictionless

market) assimilates into the market causing the price to once again reflect all known information [Shleifer 97].

However, EMH cannot be tested directly since it requires an accurate model of asset pricing. Instead, examining whether or not traders can realize excess returns by trading on information becomes the test of market efficiency. The null hypothesis tests this in three increasing degrees of informational efficiency [Fama 70],

Weak form efficiency: Prices fully reflect the information implicit in the sequence of past prices. Thus excess returns cannot be realized by using trading rules based on past price movements, for instance by technical analysis or chartist methodologies. So past price changes cannot be used to improve predictions concerning the expected value of future price changes—consistent with the random walk hypothesis [Samuelson 65] (see Section 2.2).

Semi-strong form efficiency: Prices reflect all relevant information that is publicly available. In this situation, prices quickly adjust to new information available. The announcement of a piece of information is considered an event, and the studies are commonly referred to as event studies. For instance, [Fama 69] conducted a study of the effect of stock splits on share price. They found that prices adjust to news before the event occurred, and therefore profitable trading strategies cannot be developed in relation to an event after it has occurred. A number of other more recent event studies have substantiated the informational efficiency of the market in the semi-strong form of the hypothesis.

Strong form efficiency: Information that is known to any participant is reflected in market prices. Early identification of new information can provide a source of excess returns; for example, insiders who trade on the basis of privileged information can make substantial profits—violating strong form efficiency. However, the empirical

evidence shows that professional investment managers do not consistently realize superior portfolio returns. Mutual funds have been the most frequently studied of the institutions; [Jensen 67, Ippolito 89] show that they do not in general outperform the market.

More recently studies are showing empirical evidence against the EMH. Market prices were proven to exhibit excess volatility compared with the level we would expect from the movements in the underlying fundamentals [Shiller 81]. It was also shown that most price variations of the S&P500 stock index did not correspond to any news over more than 50 years of data [Cutler 89]. Finally, the 1987 stock market crash, the tech boom bubble and recently the credit crisis crash provide the most obvious evidence that prices do not simply reflect fundamental values.

As a result theoretical arguments against the EMH emerged. First, if all agents are rational and this is common knowledge, there should be no trade, since an agent will send an order only if they have private information not reflected in price, in which case other agents will refuse to trade [Milgrom 82]. This contrasts with the high level of intra-day activity witnessed in real-markets, suggesting that trades occur when participants have heterogenous beliefs. Moreover, as [Grossman 80] pointed out, if markets are efficient and prices actually reflect all available information, what is the incentive for rational traders and arbitrageurs to collect this information in the first place—and if they do not, what will ensure that the price reflects it? Finally, arbitrage was shown to be risky and consequently limited [Shleifer 97]. As such, there is no guarantee that the price will mean revert toward its intrinsic value once driven away by irrational traders. An alternative to the EMH is the heterogenous market hypothesis [Dacorogna 01], which addresses these issues. This is discussed in Section 2.4.

2.2 The Random Walk Hypothesis

When successive price changes have an expected value of 0 and are statistically independent and identically distributed, the security's price is said to follow a random walk. This is consistent with the weak form of market efficiency, as past prices cannot be used to improve prediction of future prices. Thus deviations from random walk provides us evidence of market inefficiencies.

Denoting $W_{t,\Delta t}$ as the log price increment (i.e., $W_{t,\Delta t} \equiv \log P_t - \log P_{t-\Delta t} = \log \frac{P_t}{P_{t-\Delta t}}$), the following conditions are necessary for the process to be a random walk [Daniel 06],

$$E[W_{t,\Delta t}] = 0$$

$$var(W_{t,\Delta t}) = \Delta t \sigma^2 < \infty$$

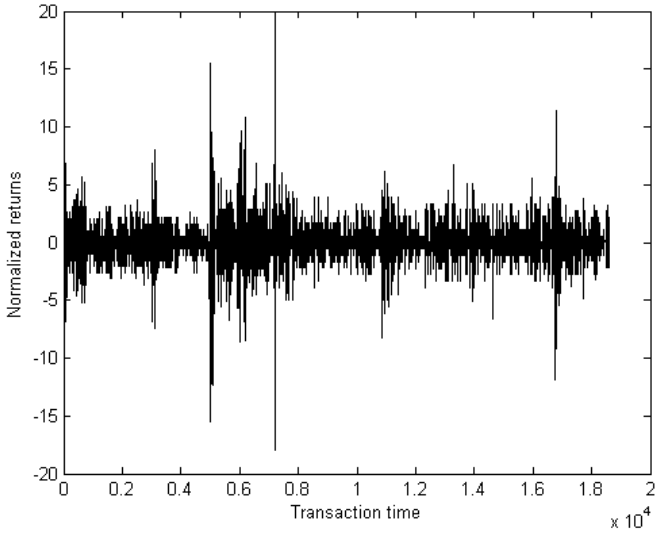
$$cov(W_{t,\Delta t}, W_{s,\Delta t}) = 0, \qquad \forall s \neq t$$

noting that a zero autocorrelation is only a necessary, but not sufficient condition for independence.

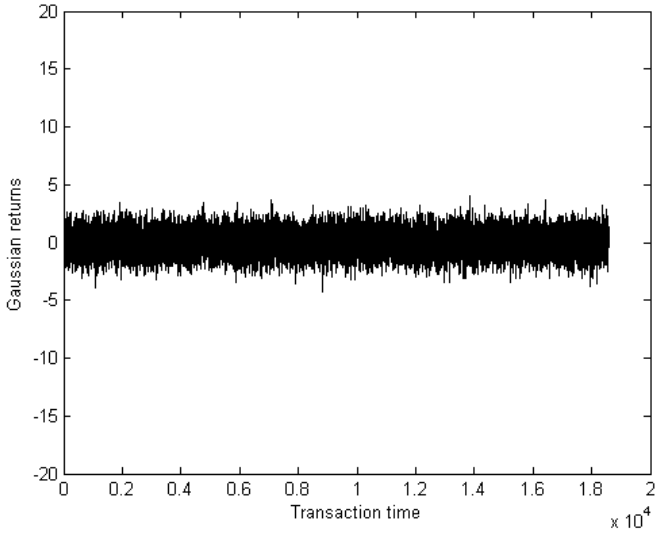
Different increment distributions result in different random walks. For instance, when $W_{t,\Delta t} \sim \mathcal{N}(0, \Delta t \sigma^2)$ (Gaussian white noise), we have geometric Brownian motion.

In fact, if price changes are independent and identically distributed, with finite second moment, then the Central Limit Theorem guarantees that the centered log-returns ($W_{t,\Delta t}$) will be normally distributed since returns are additive. Moreover, if we increase the time scale to large values ($\Delta t' = n\Delta t$), the distribution of $W_{t,\Delta t'} \equiv \log P_t - \log P_{t-\Delta t'}$ will remain normal, simply scaled by a factor $\sqrt{n}$, since $W_{t,\Delta t'} \sim \mathcal{N}(0, n\Delta t \sigma^2) = \sqrt{n}\mathcal{N}(0, \Delta t \sigma^2)$.

Figure 2.1 shows log returns for Thomson Corp., and for comparison purposes shows Gaussian increments. We can see that Gaussian price changes are a poor model for actual price changes. In particular, actual return distributions have fat tails (leptokurtic), exhibiting excess kurtosis



(a)



(b)

Figure 2.1: A comparison of a) transaction log returns for TSE:TOC over week of May 7, 2007 normalized to have mean 0, standard deviation 1, and b) Gaussian increments.

when compared with the normal distribution. Levy stable distributions are the only family of distributions stable by addition (i.e., linear combination of two independent copies of the variable has the same distribution) that can be obtained by summing independent identical random variables. This is known as the Generalized Central Limit Theorem, where normal limiting distribution becomes a special case. Levy stable distributions are bell shaped and can exhibit leptokurtosis. In particular, the characteristic exponent parameter, α, defines the tail's thickness and scaling behaviour (scales with exponent $1/\alpha$). When $0 < \alpha < 2$, the distributions are non-Gaussian and characterized by excess kurtosis with tails that decay as a power-law with exponent α. Consequently, their second moment is not defined and their first moment exists only when $\alpha > 1$ [Cont 03]. Thus the assumption of independent increments with infinite variance could then elegantly explain the excess kurtosis observed in empirical data while preserving the parsimony of the random walk hypothesis. However, recent studies using transaction data (for example [Liu 99]) report that the distribution of log returns exhibits a power-law behaviour for high-frequency with a tail index $\alpha \approx 3$, well outside of the Levy regime $0 < \alpha < 2$. This scaling breaks down when the sampling window increases for $\Delta t \approx 16$ days for individual stocks and $\Delta t \approx 4$ days for indices, after which a slow convergence to a Gaussian distribution is observed [Daniel 06].

2.3 Stylized facts

[Cont 01] conducted a comprehensive study of various financial assets and found that the return distributions all contained similar properties or stylized facts listed below. These facts further reveal the inability of the random walk approach to model real financial series.

Autocorrelations: Linear autocorrelations of asset returns are often insignificant, except for high-frequency small intraday time scale (<

20 minutes).

Fat tails: The unconditional distribution of returns has a tail index which is finite, higher than two and less than five. This excludes Levy stable laws with infinite variance and the normal distribution.

Distribution asymmetry: Distributions are negatively skewed, with a greater chance of larger drawdowns in prices but not equally large upward movements.

Time scaling: The shape of the distribution is not the same at different time scales. Particular in high-frequency (smaller time scales) the distribution vary the most, and as one increases the time scale distributions look more like a normal distribution.

Bursts: Returns at any time scale exhibit irregular bursts.

Volatility clustering: Volatility displays a significant positive autocorrelation, indicating that periods of high volatility are followed by high volatility and periods of lower volatility are followed by low volatility. (This can be modeled by using GARCH type models, where volatility σ is stochastic parameter that follows an autoregressive model.)

Conditional heavy tails: Even after correcting returns for volatility clustering using GARCH-type models, residual time series still exhibit fat tails.

Autocorrelation of absolute returns: Absolute returns can be another measure of volatility. Autocorrelation function of absolute returns decays slowly as a function of time lag, indicating long-range dependence.

Leverage effect: Volatility is negatively correlated with returns.

Volume/volatility correlation: Volume is correlated with volatility.

2.4 Heterogeneous Market Hypothesis

With more and more evidence accumulating against the EMH and given the stylized facts described above, a model of market dynamics called heterogeneous market hypothesis has surfaced [Dacorogna 01]. The heterogenous market hypothesis is in contrast to the assumption of a homogenous market where all participants interpret news and react to news in the same way.

Various participants in a market can have radically different time perspectives and motives for placing an order. A fund manager may be prepared to wait several days to execute a large order, whereas a day trader will want an extremely fast turnaround. Some participants trade because of their own analysis of information, others do so for liquidity reasons and some trade on the basis of technical analysis. All these flows are broken down into atomic transactions that meet in real time on the exchange. The different dealing frequencies clearly mean different reactions to the same news in the same market. The market is heterogenous with a "fractal" structure of the participants' time horizons as it consists of short-term, medium-term and long-term components. Each component has its own reaction time to news, related to its time horizon and dealing frequency.

Furthermore, participants may have divergent expectations, as they differ in their assessments of information. Information sets are vast, complex and challenging to understand. Different participants possess only a subset of the information that is publicly available, and some have private information. Also, to be useful, raw information has to be processed and analyzed, which may not be done in identical ways. And participants may also reassess their individual valuations based on what they come to know others are thinking. (Note, even though the assumption of homogenous expectations is unrealistic, models based on it, such as the standard capital asset pricing model, continue to provide insight into how the market determines prices for various assets according to their risk and return characteristics.)

This heterogeneity in time horizon and expectations can explain why volatility is positively correlated with market volume. In a homogenous market, the more participants that are present, the faster the price should converge to the intrinsic value on which all agents have rational expectation. In this case we would expect volatility to be negatively correlated with market presence (volume) and activity. In a heterogenous market, different market participants are likely to settle for different prices and decide to execute their transactions in different market situations—thus generating volatility.

2.5 Technical analysis

Technical analysis uses past price and volume information to forecast the direction of the market. Technical analysts, sometimes called "chartists", may employ models and trading rules based on price and volume transformations, such as the relative strength index, moving averages, regressions, inter-market and intra-market price correlations, cycles or, classically, through recognition of chart patterns. The basic principle of technical analysis is that market price reflects all relevant information. In fact technical purists even believe it is redundant to do fundamental analysis, since the price reflects this already (for example, prices adjust to news before the event occurs). On this point, technical analysts agree with one of the premises of EMH. The key basis for forecasting then is that price action tends to repeat itself because investors collectively tend toward patterned behaviour. Thus technicians' attempt to identify trends and conditions.

There is a considerable number of trading rules based on observations of past price and volume movements that have been developed. In general, the patterns are interpreted as shifts in demand and supply which can be identified by investigating market action in the form of price and volume movements. Thus, by understanding the emotions in the market and studying the market itself, as opposed to its fundamental components, technicians attempt to determine what direction, or trend, will continue

in the future.

The basic definition of price trend is the one put forward by Dow Theory [Murphy 99] in the early 1900s. Market price action can be represented as a sequence of *zig-zags*, defined as local extrema of a smoothed market price curve [Lo 00]. Zig-zags form as the market goes through periods of price discovery and consolidation in the direction of the overall trend. An uptrend is classified as a series of higher highs and higher lows; while a downtrend is one of lower lows and lower highs.

Volume is also considered a critical ingredient in technical analysis. It is used to confirm trends and chart patterns. Dow Theory describes how price and volume behaviour may be interpreted together [Ord 08]. Any price movement up or down with higher volume is seen as a signal that the price move is being supported and as such this represents the "true" market view. If many participants are active in a particular security, and the price moves significantly in one direction, Dow maintained that this was the direction in which the market anticipated continued movement. A move with weak volume indicates the market is merely consolidating. Furthermore, Dow Theory stipulates that this analysis can be done at all time scales. In this thesis, we design a dynamic Bayesian network that attempts to capture this behaviour in a high-frequency window. See Section 4.1 for a more thorough description of the technical analysis principle used.

Dow theory, and in general technical analysis, has not been received well by academics, although it is widely used among traders and financial professionals. One of the biggest challenges in assessing the validity of technical analysis is its highly subjective nature. Technical analysis has received much criticism from fundamentalists for this reason—as these patterns are attributed to be in "the eyes of the beholder". On the other hand, professional chartists profess that technical analysis is an art more than a science, requiring skills and judgement.

Literature has generally focused on evaluating simple technical trading rules such as filter rules and moving average rules that are fairly straight-

forward to define and implement. [Lo 00] moved this literature forward by evaluating more complicated trading strategies used by chartists that are hard to define and implement objectively. His work mitigates some of the stigma of technical analysis by proposing a systematic and automatic approach to technical pattern recognition. He evaluated the effectiveness of chart patterns, and found that over a 31-year sample period several of them provided incremental information.

Furthermore, emerging discoveries in behavioural economics claim that human psychology is not always rational, and in fact intertwined with the price process [Kahneman 79]. Using an innovative approach, [Kahneman 79], established human beings are subject to framing (decision depends on the way problem is presented), perform badly at estimating probabilities and are sensitive to relative wealth variation rather than absolute wealth level. Kahneman was awarded the Nobel Prize in Economics in 2002 for his work. His results imply emotion and sentiment may play a large part in price discovery, thus indirectly supporting technical analysis tenets.

Whether technical analysis actually works continues to be a matter of controversy, however. Recent studies have yielded mixed results. [Brock 92] show trading rules out-perform a buy-and-hold strategy on the Dow Jones Industrial Average Index. However, they do not include transaction costs in their analysis. [Bessembinder 98] replicate their work but include transaction costs and show that these subsume the profitability documented by [Brock 92]. [Canegrati 08] conducted the largest econometric study ever made to demonstrate the validity of technical analysis for companies listed on the FTSE. By analyzing more than 70 technical indicators, some of them almost unknown until then, the study demonstrated how market returns can be predicted, at least to a certain degree, by some technical indicators.

Most studies have evaluated technical analysis at the daily frequency. However, in practice technical analysis is used more frequently at higher frequencies for intraday trading, when fundamentals are changing the

least. In this thesis, we investigate the viability of price and volume patterns, while introducing a new and probabilistically consistent approach (dynamic Bayesian networks, which are discussed in the next chapter) for analyzing technical signals.

Chapter 3

Dynamic Bayesian Networks

Dynamic Bayesian networks (DBNs) can be used used to model stochastic processes that generate a sequence of observable quantities, or observations, as they evolve over time in a non-deterministic way. Stochastic processes occur in a large range of application areas, and there is set of common themes that allow them to be classified within a well developed taxonomy. The principal distinctions are

- Continuous time versus discrete time. Continuous time processes occur naturally in models of physical systems. Brownian motion can be considered as a simple continuous time DBN. Discrete time models can be used as approximations to continuous models and also occur naturally in many areas of economics, communications and computer science. Speech recognition and genome mapping are some examples of Discrete time models.

- Use of hidden state. Many time-series modeling techniques work exclusively with observable quantities. More complex techniques posit the existence of one or more hidden underlying states, whose interaction and values determines in some way the observed quantities. Hidden Markov models, Kalman filters are examples that use hidden states.

- Continuous variables versus discrete variables. When modeling systems, continuous variables are natural choice in some applications, whereas discrete variables are more appropriate in other domains. We can also obtain mixed continuous and discrete variables. In particular, it is common, for instance, in speech recognition, to have discrete state variables and continuous observations variables.

In this thesis we will be concerned with hidden-state, discrete-time and discrete-variable models.

The models use the concept of probability and Bayesian inference. We refer the reader to [Pearl 88] for a thorough discussion of the significance of probability, causality and process modeling.

Dynamic Bayesian networks, and more generally graphical probabilistic models, use a graph to describe a stochastic process. The graph contains a qualitative part, its topography, and a quantitative part, a set of conditional probability functions. The entire model can be thought of as "a compact and convenient way of representing a joint probability distribution over a finite set of variables" [Bengtsson 99].

The power of these models comes from the conditional independence assertions encoded in the topography. Conditional independence assertions allow for local inferences—so that calculations of joint probability distributions of conditionally independent subsets of variables can be performed separately, reducing complexity. Such conditionally independent subsets can be combined to form complex structures in a modular way. Use of conditional independence assertions result in sparse networks, and will itself create at least three important advantages compared to fully connected models:

- Sparse network structures have fewer computational and memory requirements,

- Sparse networks are less susceptible to noise in training data and less prone to overfitting (since there is less freedom in the form

of a restricted number of random variables, there is less risk that spurious regularity in data will be treated as significant),

- Resulting structure and parameters reveal useful knowledge about the underlying problem domain that was previously inconspicuous.

Graphical models are very versatile. They combine useful traits from graph theory and probability theory and offer an intuitive, visual representation of conditional independence, efficient algorithms for inference and strong representational power. Many important current models, such as mixture models, factor analysis, hidden Markov models (and variants), Kalman filtes and Ising models, can be expressed as particular instances of graphical models. Furthermore, specific algorithms for each of the models turn out to be just special cases of graphical model inference [Bilmes 00]. Indeed, the framework is flexible enough to subsume many existing techniques and is viewed as a unifying statistical framework, facilitating experimentation in new and complex ways. These properties make it an ideal tool for use in financial analysis. Financial data is rapt with noise, randomness and uncertainty. Thus this flexible statistical framework, that can learn dependencies and infer hidden states in an interpretable and efficient manner can become an invaluable approach for analyzing and manipulating financial data.

We shall first present Bayesian networks, describing their representation and usage. Then we look at dynamic Bayesian networks, which extends Bayesian networks by incorporating a time series component. We end by showing an example of how DBNs may be used for regime switching models in economic analysis.

3.1 Bayesian networks

A Bayesian network is a graphical model for representing conditional independencies between a set of random variables. They are constructed from directed and acyclic graphs. Nodes represent random variables—a

measured parameter, a hidden or latent variable or a hypothesis. The absence of edges imply conditional independencies. To each node or variable a conditional probability distribution is defined. Thus they encode the joint probability distribution of all the variables in a compact manner. Most of the theory for Bayesian networks is due to [Pearl 88].

The graph topology accounts for the qualitative part of the Bayesian network—i.e., which variables are conditioned on which. The quantitative part consists of defining the conditional probability functions or densities involved. For discrete ranges, the probability is typically stored in a conditional probability table. For continuous variables, Gaussian mixtures may be used. The directed edges of a Bayesian network provide an informal representation of causality, so that an edge goes from a cause to a consequence. This idea can be useful for constructing Bayesian nets by hand or for interpreting automatically derived ones. However, it is important to understand that this construct of causality is informal—while it is true that a graph corresponds to a particular joint probability distribution, the converse is not true. A given joint probability distribution may be factorized in different ways, giving rise to different graphs. For example, by Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

edges can be reversed and hence have inverted causal interpretations. Refer to [Pearl 88] for more detailed discussion of causality in Bayesian networks.

For example, consider the Bayesian network shown in Figure 3.1 (adopted from [Murphy 02]), where the four random variables, C, S, R, W are binary (i.e., have values in $\{0, 1\}$) and represent the events *C*loudy, *S*prinkler is on, *R*aining, and grass is *W*et respectively. We see that the event "grass is wet" ($W = 1$) has two possible causes: either the water sprinker is on ($S = 1$) or it is raining ($R = 1$). The strength of this relationship is shown in the conditional probability table (CPT) (refer to

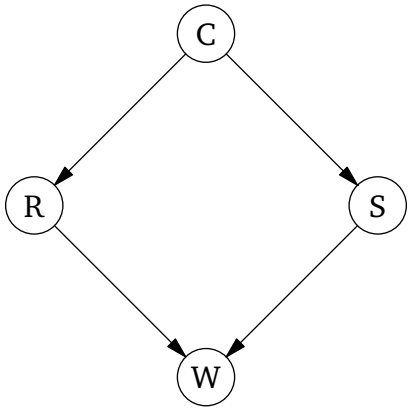


Figure 3.1: Simple Bayesian network with four random variables, C, S, R, W. (Example adopted from [Murphy 02])

	$P(C=0)$	$P(C=1)$
	0.5	0.5

C	$P(S=0)$	$P(S=1)$
0	0.5	0.5
1	0.9	0.1

C	$P(R=0)$	$P(R=1)$
0	0.8	0.2
1	0.2	0.8

S	R	$P(W=0)$	$P(W=1)$
0	0	1	0
1	0	0.1	0.9
0	1	0.1	0.9
1	1	0.01	0.99

Table 3.1: Conditional probability table for simple Bayesian network.

Table 3.1). For example, we see that $P(W = 1|S = 1, R = 0) = 0.9$, and hence, $P(W = 0|S = 1, R = 0) = 1 - 0.9 = 0.1$. Since the C node has no parents, its CPT specifies the prior probability that it is cloudy (in this case, 0.5).

Bayesian networks were motivated by the need of a flexible model with a rigorous probabilistic foundation, that allows top-down (semantic) and bottom-up (perceptual) evidences to be combined, permitting bi-directional inferences. They can be used for predictions, diagnosis and learning [Murphy 02]. We can use a Bayesian network to perform some inference tasks. The idea is that if we observe some evidences, that is, we know the values of some variables in the network, we could use those evidences to infer the values of other variables. Unknown variables are also known as hidden nodes and known value variables as observable nodes. Note that if all nodes are observed, there is no need to do inference, although we might still want to do learning.

Using the chain rule of probabilities, the joint probability distribution of the Bayesian net shown in Figure 3.1 can be expressed as,

$$P(C, S, R, W) = P(C)P(S|C)P(R|C, S)P(W|C, S, R)$$

However, this form does not consider the possible simplifications due to the assumed conditional independencies. If we do, we may from each factor exclude all conditional independent variables, arriving at a simpler joint distribution factorization,

$$P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|S, R)$$

More generally, consider a set of random variables denoted by $X = \{X_i\}$ associated with a set of nodes in a graph $G = (V, E)$, where X_i denotes the random variable associated with node $i, (i \in V)$. The definition of conditional independence in Bayesian networks states that a node is conditionally independent of its non-descendants given its parents [Pearl 88]. Thus, in general, the joint probability distribution associated with a given

graph can be factorized as follows,

$$\begin{aligned} P(X) &= P(X_1)P(X_2|X_1)\cdots P(X_n|X_1,...,X_{n-1}) \\ &= \prod_i P(X_i|X_{1:i-1}) \\ &= \prod_i P(X_i|\text{parents}(X_i)) \end{aligned}$$

where parents(X_i) is the set of parents of X_i in the graph. The function $P(X_i|\text{parents}(X_i))$ is called node i's conditional probability distribution (CPD). This can be an arbitrary distribution—for example multinomials encoded as conditional probability tables (CPT) can be used when the variables are discrete.

In addition to causal and diagnostic reasoning, Bayesian nets support the powerful notion of "explaining away". If a node is observed, then its parents become dependent, since they are rival causes for explaining the child's value. For the example in Figure 3.1, the two causes, S and R, compete to explain the observed data W. Hence, S and R become conditionally dependent given that their common child, W, is observed, even though they are marginally independent. For example, suppose the grass is wet, but that we also know that it is raining, then the posterior probability the sprinkler is on goes down: $P(S = 1|W = 1, R = 1) = 0.1945$.

In general, the conditional independence relationships encoded by a Bayesian net are described using the notion of d-separation [Neapolitan 03]. Two disjoints sets of nodes A and B are conditionally independent given set C, if C d-separates A and B—that is, if along every undirected path between a node in A and a node in B there is a node D such that: (1) D has converging arrows and neither D not its descendants are in C, or (2) D does not have converging arrow and D is in C. (Converging arrows implies the the node is a child of both the previous and following nodes in the path). Therefore one can infer many independence relations from visual inspection of the graph, without explicitly grinding through Bayes'

rule. For example, in Figure 3.1, C is conditionally independent from W given the set $C = \{S, R\}$, since both $S \in C$ and $R \in C$ are along the path between C and W and do not have converging arrows. However, C is not conditionally independent from W given R only.

3.2 Dynamic Bayesian networks

The Bayesian networks discussed so far all specify a certain point in time—they are static. They need to be extended in order to account for temporal processes such as financial time series (or, more generally, sequences of any kind, for instance speech). This is accomplished by a straightforward extension.

Dynamic Bayesian Networks (DBNs) are Bayesian networks which include directed edges pointing in the direction of time [Murphy 02, Bilmes 03]. A set of variables X_t denotes the system state at time t, where $X_t = \{X_t^1, X_t^2, ..., X_t^2\}$ and X_t^i denotes the X^i node of the underlying Bayesian network at time t. We only consider discrete-time stochastic processes so we will increase the value of t by one at each time step. (For a treatment on continuous-time DBNs we refer the reader to [Nodelman 07]). The structure and parameters are assumed to repeat for each time slice (i.e., the process is assumed to be stationary), so the conditional probabilities associated with $X_t^i, 1 \leq t \leq T$, are the same.

Formally, a dynamic Bayesian network is defined to be a pair (B_1, B_2), where B_1 is a Bayesian network which defines the prior $P(X_1)$ and B_2 is a two-slice temporal Bayesian network which defines $P(X_t|X_{t-1})$ by means of a directed acyclic graph as

$$P(X_t|X_{t-1}) = \prod_{i=1}^{N} P(X_t^i|\text{Parents}(X_t^i))$$

If the processes modeled are assumed to be Markovian (i.e., the future is conditionally independent of the past given the present) then dependency edges are only permitted between time frame t and $t+1$. In this

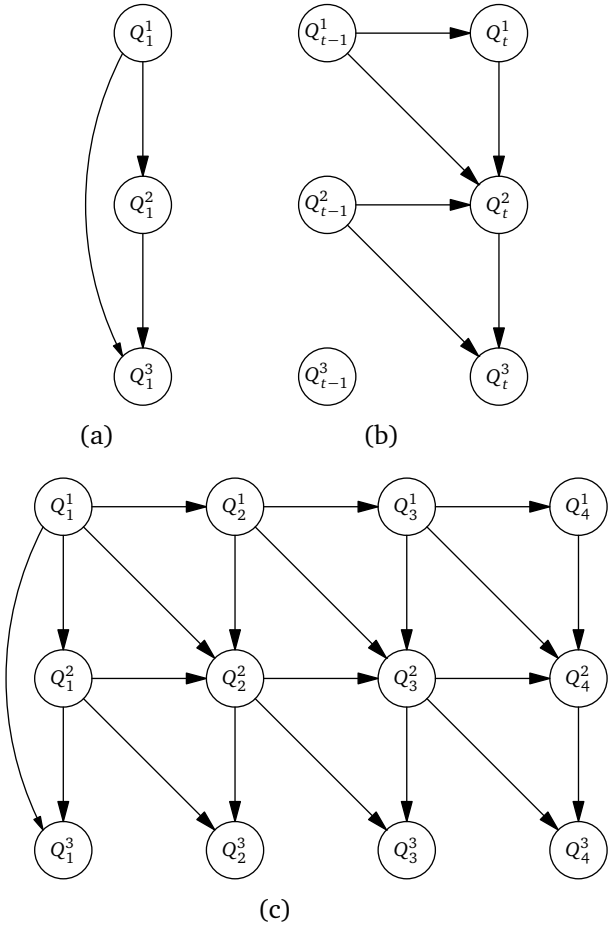


Figure 3.2: An example of a DBN representation and the unrolling mechanism (a) Initial network. (b) Transition network. (c) Unrolled DBN for four time slices.

case, it is enough to specify the the initial network (see Figure 3.2a) and the edges connecting two consecutive time slices (2TBN, see Figure 3.2b) and then repeat them as necessary (see Figure 3.2c for four time slices). Note, there is fundamentally no reason why we cannot allow arcs to skip across slices. Intuitively, directed arcs within a slice represent "instantaneous" causation [Murphy 02].

Conceptually, DBNs can be seen as "unrolling" a one-frame network for T time steps [Friedman 98] and adding time-dependencies, in effect creating a Bayesian network of size NxT. The resulting joint distribution is then given by,

$$P(X_1,...,X_T) = \prod_{t=1}^{T}\prod_{i=1}^{N} P(X_t^i|\text{Parents}(X_t^i))$$

We will introduce hidden Markov models (HMM) and hierarchical hidden Markov models (HHMMs) in the next sections and show that they are really just special cases of DBNs. In particular, we will later design our price and volume model as hierarchical hidden Markov model and then transform it to a DBN for learning and inference purposes.

3.2.1 Hidden Markov models

The basic idea of a hidden Markov model is that the observation sequence is generated by a system that can exist in one of a finite number of states. At each time-step, the system makes a transition from the state it is in to another state, and emits the observable quantity according to a state-specific probability distribution.

We will use S_t to denote the hidden state and Y_t to denote the observation. If there are K possible states, then $S_t \in \{1,...,K\}$. Y_t can be a discrete symbol, $Y_t \in \{1,...,L\}$, or a feature vector, $Y_t \in \mathbb{R}^L$.

The parameters of the model are the initial state distribution, $\pi(i) \equiv P(S_1 = i)$, the transition model, $A(i,j) \equiv P(S_t = j|S_{t-1} = i)$, and the observation model $P(Y_t|S_t)$.

The structure of matrix A is often depicted graphically, for example Figure 3.3, which depicts a left-to-right transition matrix. Note, the graph in Figure 3.3 should not be confused with the DBN graphs we discussed in the previous section. Here nodes represent states, in contrast to DBNs, where nodes represent random variables and can *take on* states.

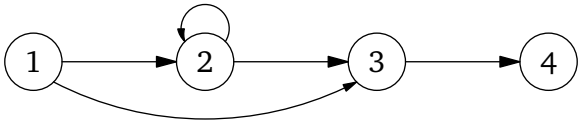


Figure 3.3: Simple left-to-right state transition diagram for a 4-state HMM. Nodes represent states and arrows represent allowable transitions (i.e., transitions with non-zero probabilities).

If the observation are discrete symbols, we can represent the observation model as a matrix, $B(i,k) \equiv P(Y_t = k|S_t = i)$. If the observations are vectors in $\mathbb{R}^L$, we can use for instance a Gaussian, $P(Y_t = y|S_t = i) = \mathcal{N}(y;\mu_i,\Sigma_i)$ where $\mathcal{N}(y;\mu,\Sigma)$ is the Gaussian density with mean μ and covariance Σ evaluated at y.

We can represent an HMM as a DBN as shown in Figure 3.4. The DBN represents the conditional independence assumptions, $S_{t+1} \perp S_{t-1}|S_t$ (Markov property) and $Y_t \perp Y_{t'}|S_t$, for $t' \neq t$.

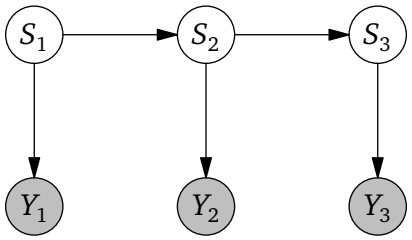


Figure 3.4: An HMM representation as an instance of a DBN, unrolled for three time slices.

The conditional probability distribution of each node given its parents are,

$$P(S_1 = i) = \pi(i)$$

$$P(S_t = j|S_{t-1} = i) = A(i,j)$$

$$P(Y_t = j|S_t = i) = B(i,j)$$

where π, A and B are as defined for the HMM and observations are discrete. The hidden Markov model consists of one hidden state; DBNs are more general in that they allow the hidden state to be specified by a set of random variables, $S_t^1, ..., S_t^N$, thus using a distributed representation of state which in itself can contain dependencies. Consequently, by representing HMMs as DBNs it becomes easy to create variations on the basic theme. For examples and discussion we refer the reader to [Murphy 02].

3.2.2 Hierarchical hidden Markov models

Hierarchical HMMs were introduced by [Fine 98] as extensions of HMMs. They are structured multi-level stochastic processes. They generalize HMMs by making each of the hidden states an autonomous probabilistic model on its own, that is, each state is an HHMM as well (i.e., recursive definition). An HHMM generates observation sequences by a recursive activation of one of the substates of a state (called abstract state), which in turn can activate one of its substates. This recursive activation continues until we reach a leaf state (called production state), which emits an observation according to a distribution specific to the state of the stack in the hierarchy. When the sub-HHMM is finished, control is returned to wherever it was called from. The calling context is stored using a depth-limited stack.

The observation sequence is denoted by $Y = y_1, y_2, ..., y_T$, where $y_i \in \{1, ..., L\}$ for discrete observations. A state of an HHMM is denoted by $q_i^d (d \in \{1, ..., D\})$ where i is the state index and d is the hierarchy index. The hierarchy index of the root is 0 and of the production states is D. We denote the number of substates of an abstract state q_i^d by $|q_i^d|$. In addi-

tion to its model structure, an HHMM is specified by the state transition probability between the internal states and the output distribution vector of production states. That is, for each internal state $q_i^d (d \in \{0, ..., D-1\})$, there is a state transition probability matrix denoted by $A^{q^d} = (a_{ij}^{q^d})$, where $a_{ij}^{q^d} = P(q_j^{d+1} | q_i^{d+1})$ is the probability of making a horizontal transition from the ith state to the jth state, both of which are substates of q^d. Similarly, $\Pi^{q^d} = \{\pi^{q^d}(q_i^{d+1})\} = \{P(q_i^{d+1} | q^d)\}$ is the initial distribution vector over the substates of q^d, which is the probability that state q^d will initially activate the state q_i^{d+1}. If q_i^{d+1} is an internal state, then $\pi^{q^d}(q_i^{d+1})$ may be interpreted as the probability of making a vertical transition—entering substate q_i^{d+1} from its parent state q^d. Each production state q^D is parameterized by its output probability vector $B^{q^D} = \{b^{q^D}(k)\}$, where $b^{q^D}(k) = P(Y_k = y_k | q^D)$ is the probability that the production state q^D will output the symbol $y_k \in \{1, ..., L\}$. The entire set of parameters is denotes by,

$$\lambda = \{\lambda^{q^d}\}_{d \in \{0,...,D\}} = \{\{A^{q^d}\}_{d \in \{0,...,D-1\}}, \{\{\Pi^{q^d}\}_{d \in \{0,...,D-1\}}, \{B^{q^D}\}\}$$

Refer to Figure 3.5 for an example HHMM. A string is generated by starting from the root state and choosing one of substates at random according to Π^{q^1}. Similarly, for each internal state q that is entered, one of q's substates is randomly chosen according to Π^q, until a production state q^D is reached at which point a single symbol is emitted according to distribution B^{q^D}. After completing the recursive string generation, the internal state that started the recursion chooses the next state in the same level according to the level's state transition matrix A^q. Each level (excluding root) has a final state q_{end}^d, which terminates the level transitions and returns control to the parent of the hierarchy.

[Murphy 02] shows how we can represent an HHMM as a DBN, using the structure shown in Figure 3.6. We assume production states are at the bottom of the hierarchy. The state of the HMM at level d and time t is represented by Q_t^d. The state of the whole HHMM is encoded by the

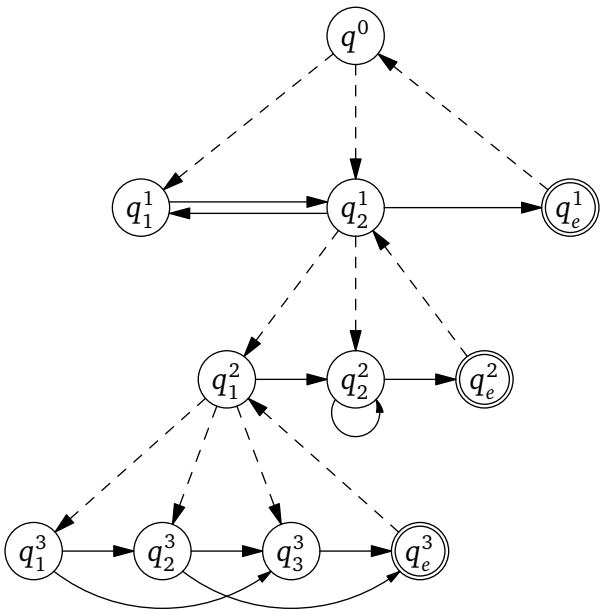


Figure 3.5: An illustration of an HHMM of four levels. Dashed and solid edges respectively denote vertical and horizontal transitions. Dashed edges upward denote (forced) returns from the end state of each level to the level's parent state. For simplicity, the production states are omitted from the figure.

vector $Q_t = \{Q_t^0, ..., Q_t^D\}$. The vector Q_t encodes the contents of the stack, that specifies the complete path to take from the root to leaf state in the state transition diagram.

F_t^d is an indicator variable that equals 1 if the HMM at level d and time t has finished, otherwise has value 0. Note if $F_t^d = 1$, then $F_t^{d'} = 1$ for all $d' > d$—that is, the number of F nodes that are off represents the effective height of the stack which represents the level of the hierarchy we are currently on.

The downward arcs between the Q variables represent the fact that a state activates a sub-state. The upward arcs between the F variables enforce the fact that a higher-level HMM can only change state when the lower level one is finished.

We define the CPDs of each node types below.

Bottom level ($d = D, t = 2 : T - 1$): Q^D follows a Markov chain, determined by which sub-HMM it is in (encoded by $Q_t^{0:D-1} \equiv k$). Instead of Q^D entering its end state, it turns on F^D to signal that higher level HMMs can now change state. Thus,

$$P(Q_t^D = j | Q_{t-1}^D = i, F_{t-1}^D = f, Q_t^{0:D-1} = k) = \begin{cases} \widetilde{A}_k^D(i,j), & \text{if } f = 0 \\ \pi_k^D(j), & \text{if } f = 1 \end{cases}$$

where $i, j \neq end$, where end represents the end-state for this HMM. Because Q_t^d does not take on the value "end" (there is no corresponding observation), the DBN and HHMM transition matrices are not identical. However we can obtain the DBN transition matrix, $\widetilde{A}_k^D$ from A_k^D by rescaling,

$$\widetilde{A}_k^D(i,j)(1 - A_k^D(i, end)) = A_k^D(i,j)$$

Similarly, π_k^D is the initial distribution for level D given context is in

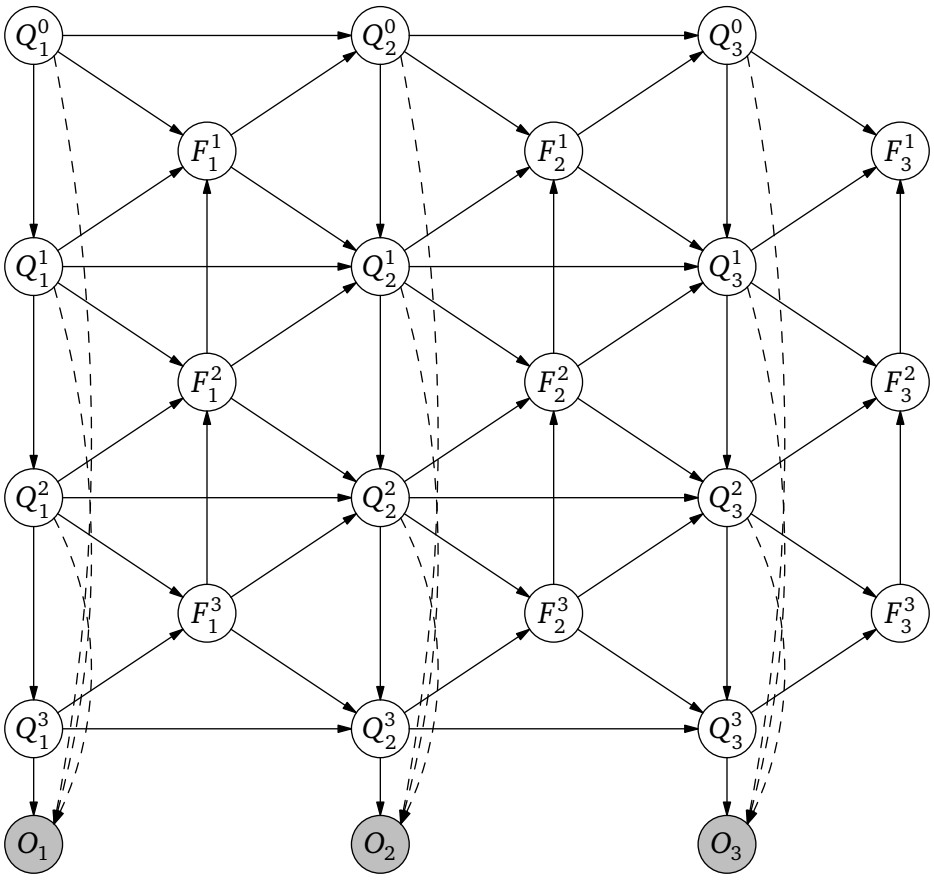


Figure 3.6: A 4-level HHMM represented as a DBN. Q_t^d is the state at time t, level d; $F_t^d = 1$ if the HMM at level d has finished (entered its exit state), otherwise $F_t^d = 0$. Shaded nodes are observed, clear nodes are hidden. The dotted arcs can be added to make the observation conditional on the hierarchical stack state.

state k,

$$P(F_t^D = 1 | Q_t^{0:D-1} = k, Q_t^D = i) = A_k^D(i, end)$$

Intermediate/Top levels ($d = 0 : D-1, t = 2 : T-1$): Similar to the bottom level, Q^d follows a Markov chain determined by $Q^{0:d-1}$ and F^d specifies whether we should use the transition matrix or the prior. The difference is that we now also get a signal from below, F^{d+1}, specifying whether the sub-model has finished or not. If it has, we can change state, otherwise we must remain in the same state. Thus,

$$P(Q_t^d = j | Q_{t-1}^d = i, F_{t-1}^{d+1} = b, F_{t-1}^d = f, Q_t^{0:d-1} = k) = \begin{cases} \delta(i, j), & \text{if } b = 0 \\ \widetilde{A}_k^d(i, j), & \text{if } b = 1, f = 0 \\ \pi_k^d(j), & \text{if } b = 1, f = 1 \end{cases}$$

We re-scale the transition matrix as before,

$$\widetilde{A}_k^d(i, j)(1 - A_k^d(i, end)) = A_k^d(i, j)$$

F^d should turn on only if Q^d is allowed to enter a final state, the probability of which depends on the context $Q^{1:d-1}$,

$$P(F_t^d = 1 | Q_t^d = i, Q_t^{0:d-1} = k, F_t^{d+1} = b) = \begin{cases} 0, & \text{if } b = 0 \\ A_k^d(i, end), & \text{if } b = 1 \end{cases}$$

Initial slice ($t = 1, d = 0 : D$): For the top level CPDs are, $P(Q_1^1 = j) = \pi^1(j)$ and for $d = 1, .., D$, we have $P(Q_1^d = j | Q_1^{0:d-1} = k) = \pi_k^d(j)$.

Final slice ($t = T, d = 0 : D$): To ensure that all sub-HMMs have reached their end states by the time we reach the end of sequence, we can clamp $F_T^d = 1$ for all d.

Observations: Observations can be conditioned on the entire stack, as

$P(O_t|Q_t)$. Alternatively, we can condition O_t only on some of its parents.

One of [Murphy 02]'s main contributions was showing the equivalence of an HHMM in DBN form as summarized above. In doing so, he was able to improve the inference algorithm proposed in the original HHMM paper by [Fine 98] from $O(T^3)$ time to $O(T)$ time, where T is the number of time slices (assuming per slice complexity as constant, see Section 3.3). Transforming HHMMs into DBNs allows us to leverage generic DBN inference and learning procedures, instead of deriving HHMM specific methods. Also, it becomes easier to vary the model as the DBN framework is more flexible and general.

3.3 Inference

We could use dynamic Bayesian networks to perform inference on hidden nodes, that is to calculate the posterior distribution of the node, given some evidence or values of observable variables. Before we receive evidence, the network represents our a priori belief about the system that it models; after we receive evidence, the network may be updated to denote our a posterior beliefs.

Probabilistic inferences in dynamic bayesian networks can be accomplished by "unrolling" the DBN for T time-slices and then applying a static Bayesian network inference algorithm.

The inference problem requires us to compute $P(X_Q|X_E = x_E)$, where X_Q is a set of query variables and X_E is a set of evidence variables. The most common exact inference (computing the probabilities exactly) methods are: variable elimination, which eliminates (by integration or summation) the non-observed non-query variables one by one by distributing the sum over the product; clique tree propagation, which caches the computation so that many variables can be queried at one time and new evidence can be propagated quickly; and recursive conditioning, which allows for a

space-time tradeoff and matches the efficiency of variable elimination when enough space is used. All of these methods have complexity that is exponential in the graph's treewidth [Robertson 84, Elidan 08]. In this thesis we describe the variable elimination algorithm, which is the basis of other exact inference algorithms. We refer the reader to [Murphy 02] for variants which improve upon the basic algorithm.

In variable elimination, the posterior probability of variables X_Q, given some evidence $X_E = x_e$ is computed using Bayes rule,

$$\begin{aligned} P(X_Q|X_E) &= \frac{P(X_Q,X_E)}{P(X_E)} \\ &= \frac{\sum_{h \not\in Q \cup E} P(X_H = h, X_Q, X_E)}{\sum_{h \not\in E} P(X_H = h, X_E)} \end{aligned}$$

Thus, inference boils down to marginalizing joint distributions. If variables can at most take on K states, then computing $\sum_h P(X)$ takes $O(K^N)$ time, where N is the total number of nodes. This is exponential in the number of nodes and becomes intractable for graphs of any significant size.

Since our Bayesian net represents a conditionally factored distribution, we can do better by taking advantage of conditional independence relations to marginalize efficiently. The joint distribution represented by a Bayesian network can be written in factored form,

$$P(X) = \prod_{i=1}^{N} P(X_i|\text{Parents}(X_i))$$

Thus,

$$P(X_Q,X_E) = \sum_{h \not\in Q \cup E} P(X_H = h, X_Q, X_E)$$

$$= \sum_{h \notin Q \cup E} \prod_{i=1}^{N} P(X_i | \text{Parents}(X_i), X_E)$$

This expression can be significantly simplified by summing out variables in an arbitrary elimination ordering such that, every time a variable $X \notin \{X_H, X_E\}$ is eliminated, only the factors containing X are multiplied and the resulting potential is marginalized over X. This process of ordering the factors (potentials) and the sum (variables) is the basis of variable elimination algorithms.

For example, referring to the Bayesian network in Figure 3.1, the joint probability distribution is,

$$P(C,S,R,W) = P(C)P(S|C)P(R|C)P(W|S,R)$$

So for instance,

$$\begin{aligned} P(W=w) &= \sum_c \sum_s \sum_r P(c,s,r,w) \\ &= \sum_c \sum_s \sum_r P(c)P(s|c)P(r|c)P(w|s,r) \\ &= \sum_c P(c) \sum_s P(s|c) \sum_r P(r|c)P(w|s,r) \end{aligned}$$

noting that as we perform the innermost sums, we create new terms, which need to be summed over in turn. Thus, the amount of work we perform when computing a marginal is bounded by the size of the largest term that we encounter. Any permutation of the variables to be eliminated could be used as elimination sequence. Therefore choosing a summation (elimination) ordering to minimize this is important for the efficiency of the algorithm. Since the problem of finding an optimal elimination ordering is NP-complete [Arnborg 87], several heuristic approaches have been proposed in order to achieve close to optimal elimination sequences [Zhang 99]. The most used method defined in the context of a greedy

algorithm, is to select the next variable to be eliminated, X, by minimizing the weight (state space size) of the new potential obtained during the process of eliminating X [Kjaerulff 90].

Note, that variable elimination takes $O(NK^M)$ time, where N is the number of nodes in the graph, K is the maximum number of states a node can take on, and M is the largest number of variables in a factor [Pearl 88]. If we want to compute $P(X_i|X_E)$ for all $i \notin E$, we could call variable elimination $O(N)$ times, once for each node, thus it would take $O(N^2K^M)$ time. As mentioned, clique tree propagation, caches the computation so that many variables can be queried at one time, thus providing a way to compute all N marginals in $O(NK^M)$ time [Murphy 02].

If we are interested in the most likely explanation of the set of query variables for the evidence (instead of the posterior distribution), the inference problem becomes,

$$x_Q^* = \arg\max_{x_Q} P(X_Q = x_q | X_E = x_E)$$

This is known as the Viterbi problem. We can solve this problem using the variable elimination algorithm, replacing sum-product with max-product as follows,

$$\begin{aligned} x_Q^* &= \arg\max_{x_Q} \frac{P(x_Q, x_E)}{P(x_E)} \\ &= \arg\max_{x_Q} P(x_Q, x_E) \\ &= \arg\max_{x_Q} \prod_{i=1}^{N} P(X_i|\text{Parents}(X_i), X_E) \end{aligned}$$

The difference is that Viterbi assigns to a node the probability of the single best assignment, while the posterior calculation assigns the sum of probabilities over all possibilities to a node.

3.3.1 Types of inference for DBNs

For a given dynamic Bayesian network, there are a variety of inference problems we might be interested in (see Figure 3.7 for a summary) [Murphy 02]. Let S_t represent the hidden nodes and Y_t represent the observable nodes,

Filtering Computing $P(S_t|Y_{1:t})$, i.e., monitoring (tracking) the state over time.

Prediction Computing $P(S_{t+h}|Y_{1:t})$ for some horizon $h > 0$ into the future.

Fixed-lag smoothing (Look-ahead) Computing $P(S_{t-l}|Y_{1:t})$, i.e., estimating what happened $l > 0$ steps in the past given all the evidence up to the present.

Fixed-interval smoothing (Look-ahead) Computing $P(S_t|Y_{1:T})$, i.e., estimating what happened in the past given the entire set of evidence. This is used for training as well.

Viterbi decoding Computing $\arg\max_{S_{1:t}} P(S_{1:t}|Y_{1:t})$, i.e., determining the most likely explanation of the observed data.

Look-ahead Viterbi Computing $\arg\max_{S_{1:t}} P(S_{1:t}|Y_{1:T})$, i.e., determining the most likely explanation for the entire set of observed data.

Classification Computing $P(Y_{1:t}) = \sum_{X_{1:t}} P(X_{1:t}, Y_{1:t})$, to compute the likelihood of a sequence under different models.

We can solve filtering, smoothing and prediction problems by applying evidence at appropriate times and then running the variable elimination inference algorithm described. If we are interested in Viterbi filtering or smoothing, we replace sum-product operator with max-product in the variable elimination algorithm. For a DBN with T time slices, this would take $O(T)$ time (where we have assumed per slice complexity as constant).

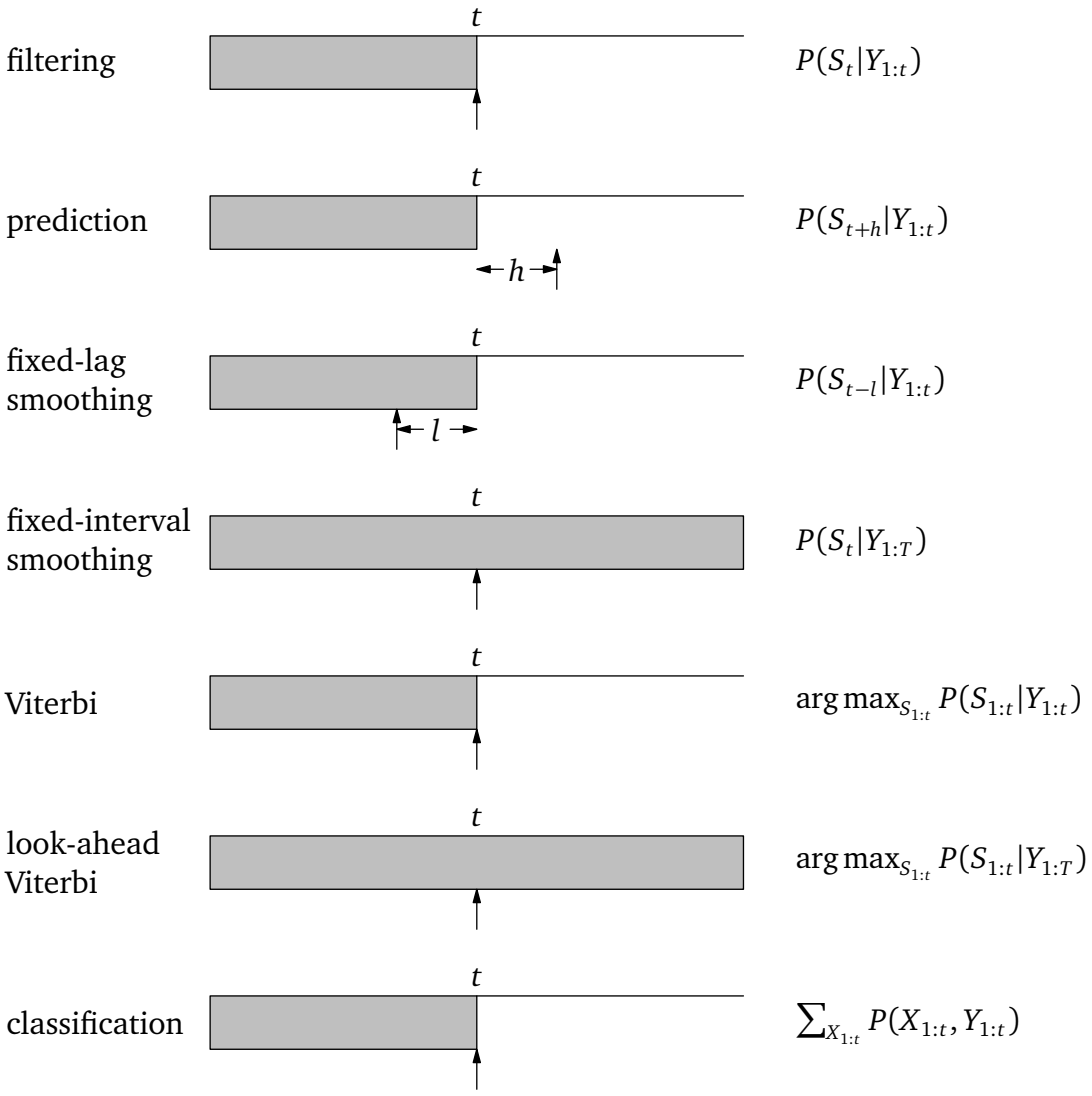


Figure 3.7: The main kinds of inference for DBNs. The shaded region is the interval for which we have data. The arrow represents the time step at which we want to perform inference. t is the current time, and T is the sequence length. h is a prediction horizon and l is a time lag. (Adopted from [Murphy 02])

In DBNs we often want to dynamically filter or smooth at each time step. This requires rebuilding the entire time history of the process for each time step, requiring $O(T^2)$ time.

However, for a DBN that is stationary and Markovian, we can do inference incrementally. Recall, that a stationary DBN implies that the node relationships within a time slice t and the transition function from time slice t to time slice $(t+1)$ do not depend on t; Markovian property implies that the time slice $(t+1)$ only depends on the time slice t and not on any previous time slices (i.e., the set of nodes in a time slice d-separates the past from the future).

Thus we can represent a DBN using a 2TBN and do inference just using this structure. In this case dynamic inference boils down to doing static inference on the 2TBN and then using generalized forward-backward operators to step through the DBN. (The forward-backward algorithm is used to do inference for an HMM. [Murphy 02] generalizes the forward and backward operators as "message" passing operators from which we can compute $P(X_t^i|Y_{1:T})$ and $P(X_t^i, \text{Parents}(X_t^i)|Y_{1:T})$ for any node X_t^i and its parents.) Two variants are described in the literature: 1) Frontier algorithm (originally presented in [Zweig 96]), which uses the full set of hidden nodes at the current time slice that d-separates the process into two segments, and 2) Interface algorithm [Murphy 02], which is slightly more efficient than the frontier algorithm since it uses only the out-going nodes between time-slices to d-separate the process. Both algorithms result in $O(T)$ time for inference.

3.4 Learning

When the structure of a dynamic Bayesian network is given, the learning task becomes one of estimating the model parameters. Generally, we are interested in finding the maximum likelihood estimates (MLEs) of the parameters of each node's conditional probability distribution—that is, the parameter values which maximize the likelihood of the evidence or

training data. If there are a small number of training cases compared to the number of parameters, we could use a prior to regularize the problem. In this case, we call the estimates maximum a posterior (MAP) estimates, and use Bayesian estimation, as opposed to maximum likelihood estimation. We refer the reader to [Murphy 02] for more information on this approach. We will primarily be concerned with maximum likelihood learning in this thesis.

If the network is fully observed—so that there is no hidden or unobserved nodes—the problem reduces to finding the MLE for a given sample. Training data can contain S sequences, assumed to be independent, each of which has the observed values for all n nodes per slice for each of T slices. For notational simplicity, we assume each sequence is of the same length. Thus we can imagine "unrolling" a two-slice DBN to produce a (static) Bayesian netowrk with T slices.

We assume the parameters values for all nodes are tied (i.e., constant) across time, so that for a time series of length T, we get one sample for each CPD in the initial slice, and $(T-1)$ data points for each of the other CPDs. If $S=1$, we cannot reliably estimate the parameters of the nodes in the first slice, so we usually assume these are fixed a priori. That leaves us with $N = S(T-1)$ samples for each of the remaining CPDs.

The joint probability (as discussed in Section 3.3) of all the nodes in the graph is,

$$P(X_1,...,X_m) = \prod_i P(X_i|\text{Parents}(X_i))$$

where $m = n(T-1)$ is the number of nodes in the unrolled network, excluding the first slice. The normalized log-likelihood of the training set $D = \{D_1,...,D_S\}$ is a sum of terms, one for each node,

$$L = \frac{1}{N}\log\prod_{l=1}^{S} P(D_l|G)$$

$$= \sum_{i=1}^{N}\sum_{l=1}^{S} \log P(X_i|\text{Parents}(X_i), D_l)$$

We see that the log-likelihood decomposes according to the structure of the graph. Thus we can maximize the contribution to the log-likelihood of each node independently and consequently estimate the parameters of each CPD given its local data.

If the CPD is in the exponential family, the parameters can be determined using its sufficient statistic. For instance, in the case of tabular CPD's (where the node has a multinomial distribution) we can define the parameters as,

$$\theta_{i,j,k} \equiv P(X_i = k|\text{Parents}(X_i) = j),$$

and the log-likelihood becomes,

$$\begin{aligned} L &= \sum_i \sum_l \log \prod_{j,k} \theta_{ijk}^{1_{ijkl}} \\ &= \sum_i \sum_l \sum_{j,k} 1_{ijkl} \log \theta_{ijk} \\ &= \sum_{i,j,k} N_{ijkl} \log \theta_{ijk} \end{aligned}$$

where $1_{ijkl} \equiv I(X_i = k, \text{Parents}(X_i) = j|D_l)$ is 1 if the event $(X_i = k, \text{Parents}(X_i) = j)$ occurs in case D_l, 0 otherwise. Thus, $N_{ijk} \equiv \sum_l I(X_i = k, \text{Parents}(X_i) = j|D_l)$ is the number of times the event $(X_i = k, \text{Parents}(X_i) = j)$ was seen in the training set. The resulting MLE is given by,

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_k N_{ijk}}$$

which can be verified by taking derivatives and using a Lagrange multiplier. This turns out to be simply the frequency of sample observations.

When the network contains hidden or unobserved nodes, the log-

likelihood cannot be decomposed into sum of local terms, one per node. Instead we obtain,

$$\begin{aligned} L &= \sum_l \log P(D_l) \\ &= \sum_l \log \sum_H P(H, D_l) \end{aligned}$$

where H is the set of hidden variables, and $\sum_H$ is the sum (or integral) over H required to obtain the marginal probability of the data. Thus the MLE is the argument that maximizes L. The obvious way to maximize likelihood is to do gradient ascent. However, a simpler and more straightforward algorithm is expectation maximization (EM). In fact, EM is implicitly a gradient method [Salakhutdinov 03, Salojärvi 05].

The basic idea behind EM is to apply Jensen's inequality to our likelihood function to get a lower bound on the the log-likelihood, and then to iteratively maximize this lower bound. Jensen's inequality says that, for any concave function f,

$$f\left(\sum_j \lambda_j y_j\right) \geq \sum_j f(y_j)$$

where $\sum_j \lambda_j = 1$. In other words, f of any weighted average is bigger than the average of the f's. Since the log function is concave, we can apply Jensen's to the likelihood function to get,

$$\begin{aligned} L &= \sum_l \log \sum_H P(H, D_l; \theta) \\ &= \sum_l \log \sum_H q(H) \frac{P(H, D_l; \theta)}{q(H)} \\ &\geq \sum_l \sum_H q(H) \log \frac{P(H, D_l; \theta)}{q(H)} \end{aligned}$$

$$= \sum_l \sum_H q(H) \log P(H, D_l; \theta) - \sum_l \sum_H q(H) \log q(H)$$

where q is any function s.t. $\sum_H q(H) = 1$ and $0 \leq q(H) \leq 1$.

Maximizing the lower bound with respect to q gives

$$q(H) = P(H|D_l; \theta)$$

This is called the E (expectation) step and makes the bound equality.

Maximizing the lower bound with respect to the free parameters θ is equivalent to maximizing the expected complete data log-likelihood,

$$L(\theta) = \sum_l \sum_H q(H) \log P(H, D_l; \theta)$$

so we obtain,

$$\theta' = \arg\max_\theta \sum_l \sum_H q(H) \log P(H, D_l; \theta)$$

This is called the M (maximization) step.

If we use $q(H) = P(H|D_l; \theta)$, and starting from some initial parameters θ_0, we get

$$\theta_{k+1} = \arg\max_\theta \sum_l \sum_H P(H|D_l; \theta_k) \log P(H, D_l; \theta)$$

[Dempster 77] proved that θ_{k+1} is guaranteed to ensure $P(D|\theta_{k+1}) \geq P(D|\theta_k)$, because using $q(H) = P(H|D_l; \theta)$ in the E step makes the lower bound touch the actual log likelihood curve, so raising the lower bound will also raises the actual log-likelihood curve.

Thus the EM method allows us to find a local maximum with an initial starting point θ_0. It is worth noting that generally the likelihood surface is heavily multi-modal—and so local search algorithms such as EM are prone to get stuck in local optima. A simple solution, which we will use in our experiments, is multiple restarts. An alternative, is to use deterministic annealing [Murphy 02]. This works by enforcing a certain level of entropy

(noise) in the system, which is gradually reduced. The idea is to multiply the entropy by a temperature term T; initially the temperature is high, which "smooths out" the energy surface, so it is easy to find the maximum. Then the temperature is gradually reduced to $T = 1$ corresponding to the original problem. Note, this is similar, but distinct from simulated annealing which also works my gradually reducing the free energy, but with random moves. We leave it to future work to consider this approach more fully.

3.5 Example: using DBNs for regime switching

In econometrics, a model with a fixed density distribution or single set of parameters may not be sufficient to account for structural changes in financial series. Time varying parameter models have been used to address this limitation. In particular, regime switching models, in which parameters move discretely between a fixed number of regimes have been used. In this section, we develop the Markov regime switching model presented by [Hamilton 89] and show the equivalent representation in a DBN framework.

Regime switching models have a rich history in financial econometrics, dating back to at least [Goldfeld 73], where a latent state variable controlling the regime follows a Markov chain. [Hamilton 89], extended regime switching models, allowing the parameters of an auto-regression to be controlled by the outcome of a discrete-state Markov process. Many authors have subsequently employed Markov switching to model regime changes in economic time series. Examples include investigations of business cycle asymmetry [Hamilton 89, Lam 90], heteroskedasticity in time series of asset prices [Schwert 94, Garcia 99], the effects of inflation on UK commercial property values [Barber 97], the effects of oil prices on U.S. GDP growth [Raymond 97], labor market recruitment [Storer 95], the dividend process [Driffill 98], government expenditure [Ruge-Murcia 95], and the level of merger and acquisition activity [Town 92].

A regime switching model increases flexibility of a static econometric model by allowing dynamic parameters. That is, each regime specifies a set of model parameters, and the regime switching model combines these parameter sets into one system. Depending on the most likely regime the system is in at any particular time, the corresponding set of parameters is applied.

Regime switching models are better able to fit economic data than their static counterparts—a natural consequence of introducing additional state parameters [Nelson 01]. Moreover, regime switches can be viewed as structural changes in the economy which can be associated with events such as financial crisis [Jeanne 00, Cerra 03, Hamilton 05], abrupt changes in government policy [Sims 06, Davig 04], or economic cycle transitions [Chauvet 05].

Consider how we may describe a structural change for a single variable y_t. Suppose that typical bahaviour of y_t follows a first-order autoregression:

$$y_t = c_1 + \phi_1 y_{t-1} + \epsilon_t \tag{3.1}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is Guassian noise.

Say this adequately describes the behaviour of y_t for $t = 1, 2, ..., t_0$ and that at $t = t_0$ there is a structural change in the economy that causes a significant change in the average level of the series. Furthermore we believe that this change also entails a different degree of dependence on the past. Thus for $t = t_0 + 1, t_0 + 2, ...$ we would like to model the data as:

$$y_t = c_2 + \phi_2 y_{t-1} + \epsilon_t \tag{3.2}$$

We can combine the piecewise models (3.1) and (3.2) in a larger encompassing model:

$$y_t = c_{s_t} + \phi_{s_t} y_{t-1} + \epsilon_t \tag{3.3}$$

where

$$s_t = \begin{cases} 1 & \text{(Regime 1)} \\ 2 & \text{(Regime 2)} \end{cases}$$

A complete specification would require a probabilistic model of what caused the change from $s_t = 1$ to $s_t = 2$. A simple specification is the realization of a two-state Markov chain with

$$\mathrm{P}(s_t = j|s_{t-1} = i, s_{t-2} = k, ..., y_{t-1}, y_{t-2}, ...) = \mathrm{P}(s_t = j|s_{t-1} = i) = p_{ij}$$

Assuming that we do not observe s_t directly, but infer its operation through the observations of y_t, the parameters necessary to fully describe this process are the variance of the noise σ^2, the auto-regression coefficients ϕ_1 and ϕ_2, the two levels c_1 and c_2, and the two state transition probabilities p_{11} and p_{22}, noting $p_{12} = 1 - p_{11}$ and $p_{21} = 1 - p_{22}$.

If we specify $p_{22} = 1$, then regime 2 is an absorbing regime and represents a permanent shift into state 2. The Markov formulation allows a more general possibility that $p_{22} < 1$, allowing a non-zero probability of switching back to state 1 once in state 2, i.e. $p_{21} = 1 - p_{22} > 0$. This is natural in business cycles or financial crisis situations where the structural change is rarely permanent.

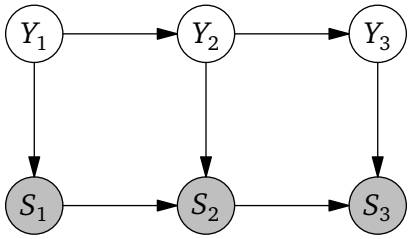


Figure 3.8: An auto-regressive HMM.

This Markov regime switching model can be represented as a discrete state, continuous observation DBN. Refer to Figure 3.8 for the structure of the DBN. $S_t = s_t$ is the discrete regime, which can take on values 1 or 2; and Y_t are observations. Note that Y_t is dependent on both the current latent regime S_t, and the previous observation Y_{t-1}. Thus this model differs from a standard HMM, by allowing Y_{t-1} to help predict Y_t. This is called an auto-regressive HMM [Murphy 02].

By specifying the conditional probability density for Y as,

$$P(Y_t = y_t | S_t = i, Y_{t-1} = y_{t-1}) = \mathcal{N}(y_t; c_i + \phi_i y_{t-1}, \sigma^2)$$

we have an equivalent DBN model. We can now apply the general inference and learning algorithms to find the parameters or infer the latent state. If we felt that the regime switch also affected the variance of the Gaussian noise, we can easily incorporate this behaviour by allowing σ to be dependent on i as well.

The unifying perspective of DBNs brings out connections between models that had previously been considered quite different, and we refer the reader to [Murphy 02] for a more detailed "laundry list" of examples of such models. [Pesaran 06] for instance, uses a hierarchical hidden Markov model for modeling time series subject to multiple structural breaks. If we start thinking in terms of graphical networks we can introduce far more flexibility in the models—all able to leverage the same paradigm for inference and estimation.

While regime switching models have been mostly used for low frequency economic series modeling to account for structural changes, they can also be a powerful tool for technical analysis in high frequency stock markets to account for behavioural changes of market participants. In the next chapter we develop such a regime switching model based on dynamic Bayesian networks and technical analysis principles.

Chapter 4

Price and Volume Model

In this chapter we design a dynamic Bayesian network for modeling runs and reversals in high frequency stock markets within a regime switching framework. There are two notable distinctions from past regime switching models which are unique to our endeavour:

- Regimes generally correspond to longer term horizons. This work models regimes or states that persist for short intraday periods which manifest as runs or reversals in the price process.

- Regimes are used to classify particular economic environments based on fundamentals. We extend this concept to permit regimes to be identified by technical analysis paradigms, recognizing that a technical chartist is essentially working under a regime switching model in which buy and sell signals are used to mark a regime switch.

- We use non-synchronous tick observations, rather than synchronous data observed at sampled time intervals.

We first review price and volume relationships in technical analysis, explain our feature extraction approach and finally describe the model.

4.1 Price and volume relationships

[Murphy 99] provides a definition of technical analysis: "[the] study of market action through the use of charts for the purpose of forecasting future price trends." The term market action includes two principal sources of information—price and volume. The assumption technicians use is that all of the fundamental information and current market opinions are already reflected in the current price and when viewed in conjunction with past prices often reveals recurring price and volume patterns that provide information to potential future price movement. The patterns are interpreted as shifts in demand and supply which can be identified by the study of market action. They are generally horizon invariant, with the claim that similar patterns should exist at each frequency horizon, whether it is intraday, daily or long term [Murphy 99].

Most patterns are based on *zig-zags*, which are defined by a sequence of local extrema, $\{E_k\}$, of the price process at the points where price changes direction. Here $E_k = (t_k, p_k)$ is a coordinate, where t_k is the time and p_k is the price at the extrema. The price path may be smoothed first, with kernels of varying bandwidth, to obtain different horizons or granularities of zig-zags [Lo 00]. By construction the series of extrema contains alternating minima and maxima—that is, if the kth extremum is a maximum, then the $(k+1)$th extremum is a minimum and vice versa. The minima form *support* levels, and the maxima form *resistance* levels, as these identify the points at which demand and supply levels cross [Murphy 99]. A zig-zag *leg* is defined as the vector from one extrema to the next, $\mathbf{l}_k = \overrightarrow{E_{k-1}E_k}$, so zig-zags may also be defined in terms of a sequence of legs, $\{\mathbf{l}_k\}$.

[Ord 08] describes how price and volume behaviour may be interpreted. Volume is supposed to push price. If volume is increasing on the upward legs and contracting on the downward legs then a bullish trend should continue. The explanation is that when prices go up, sellers are more willing to meet buyers at those higher levels and volume increases.

When prices go down, sellers are not as interested since they expect to sell at higher prices, consequently volume decreases. This activity reflects a bullish undertone. On the other hand, if volume increases on the downward legs and decreases on the upward legs, then the market has a bearish sentiment, since in order to attract buyers, prices have to go lower. When prices head higher, buyers are not interested, believing they will be able to obtain lower prices, thus creating a bearish undertone. Refer to Figure 4.1 for a visual example. [Ord 08] suggests using average volume over a zig-zag leg to measure the force in a leg, instead of total volume, since it reflects the buying or selling pressure in a normalized way.

As is common in technical analysis, the definition of 'increasing' or 'decreasing' average volume and the bandwidth used to obtain zig-zags can be subjective. We will return to making them precise for our use in section 4.3 as the feature extraction process is described.

4.2 Market microstructure

An electronic stock exchange, such as the Toronto Stock Exchange (TSE) uses a continuous double auction mechanism. It consists of an order book, in which limit orders are sorted by price (and subsequently time and volume) and stored in two stacks, the bid side for buy orders and the ask side for sell orders. When a new buy (respectively sell) limit order reaches the book, it either triggers a trade if its limit price is higher than the best offer (respectively lower than the best bid), or it is stored in the book at the appropriate level based on the price, timing and volume. Other types of orders may also be submitted, such as market buy and sell orders, which are executed immediately by consuming the top of the ask and bid stacks, respectively.

Each transaction that results in a trade can be represented as logical unit called a *tick*. This forms the most granular level of the price process, and it includes the time, price and volume of the transaction. Participant information such as buyer and seller names may also be available. A

natural consequence of this process is that tick data is asynchronous, since the arrival of trades are not uniform in time.

The process of day trading—entering and exiting trades within a short span of minutes or even seconds—attempts to profit from short term price volatility. [Schwartz 04] divides short term volatility into two components: *fundamental volatility* and *technical volatility*. Fundamental volatility characterizes price adjustments that are attributable to news concerning fundamental values. Technical volatility is process driven and characterizes price changes that are attributable to market friction caused by the order book mechanism. Technical volatility accentuates volatility and manifests as swings—runs and reversals over intraday intervals—in response to the arrival of buy and sell orders in the market. As a result, it is generally viewed as the source of trading cost to portfolio managers, in the form of spreads, execution costs and market impact. But on the flip side, it compensates dealers and limit order traders for the risks they take in settling prices for other players [Schwartz 04].

Technical analysis can be thought of as an approach to inferring where a stock's price is relative to an unobserved consensus equilibrium value. As such, day traders that use technical analysis claim they can benefit from accentuated volatility by exploiting intraday runs and reversals. Moreover, some technologically sophisticated hedge funds have discovered that they can earn the spread rather than pay it by timing technical volatility. [Schwartz 04]

One of the sources of technical volatility is the bid-ask spread. Transaction prices bounce between the bid and the ask, with staggered arrivals of market sell orders that execute at the bid, and arrivals of market buy orders that execute at the ask (see Figure 4.2). This behaviour is known as the *bid-ask bounce* and can be viewed as bouncing between tick support and resistance levels [Schwartz 04].

For our analysis we shall construct our zig-zags using bid-ask support and resistance points in attempt to capture technical volatility trends due to spreads. We leave the analysis of other sources of technical volatility to

future research. Using [Ord 08]'s volume and price analysis approach, we associate increasing or decreasing volume indicators to the zig-zag legs in order to assess whether there is buying or selling pressure. Then, using a temporal probabilistic model, i.e. DBNs, we infer what short term trend is likely to form.

4.3 Feature extraction

Tick series can be defined as a sequence of triples, $\{y_k\}$, $y_k = (t_k, p_k, v_k)$, where $t_k \leq t_{k+1}$ is the time stamp in seconds, p_k is the trade price, and v_k is the trade volume. The sequence is ordered by the occurrence of trades and forms the direct market price process. Note, there can be more than one trade within a second.

Using the tick series $\{y_k\}$, we derive zig-zags that capture the bid-ask bounce, to obtain a new series $\{z_n\}$, $z_n = (i_n, j_n, e_n, \phi_n)$, where e_n are local extrema prices, and i_n, j_n are indices to $\{y_k\}$, with $i_n \leq j_n$, representing the starting and ending point of the extrema. More precisely, $e_n = p_k$ for all k where $i_n \leq k \leq j_n$ and $p_{i_n-1} < e_n < p_{j_n+1}$ (for local maxima) or $p_{i_n-1} > e_k > p_{j_n+1}$ (for local minima). Note when $i_n < j_n$ a zig-zag extrema consists of a consecutive sequence of ticks which form a plateau or valley in the price process (see Figure 4.3). ϕ_n measures the average volume per second during the zig-zag leg ending at e_n. That is,

$$\phi_n = \frac{1}{t_{j_n} - t_{i_{n-1}} + 1} \sum_{k=i_{n-1}}^{j_n} v_k$$

where we have normalized the volume by $t_{j_n} - t_{i_{n-1}} + 1 = \Delta t_n + 1$, adding 1 to avoid division by zero in situations where the entire zig-zag leg occurs within the same second. Also, note that calculation of average volume over a leg is inclusive of end-point extrema volume, consistent with [Ord 08]'s methodology. Figure 4.4 shows the distribution of the number of ticks in a zig-zag leg for GoldCorp Inc (TSE:G) over the month of May, 2007.

In practice we do not observe a realization of a zig-zag point as soon as

it is completed. Instead there is a one tick lag between the leg completion and the time of detection. This will become important in Section 5 when we analyze the predictability of the model. In particular, the realization of the nth zig-zag point z_n, is made after observing the (j_n+1)th tick point y_{j_n+1}, that is one tick after it has completed. We use this tick point as the reference when analyzing predictability and therefore do not use any forward information which may cause a "look-ahead" bias.

Discrete features are defined based on the zig-zag series $\{z_n\}$. For each zig-zag point, z_n, there is a corresponding feature set $O_n = (f_n^0, f_n^1, f_n^2)$ which are used to form a new series $\{O_n\}$. f_n^0 represents the direction of the zig-zag leg, f_n^1 indicates whether there is a trend, and f_n^2 indicates whether average volume increased or decreased. These are defined more precisely below:

$$f_n^0 = \begin{cases} +1, & \text{if } e_n \text{ is a local maximum (zig-zag leg was positive)} \\ -1, & \text{if } e_n \text{ is a local minimum (zig-zag leg was negative)} \end{cases}$$

$$f_n^1 = \begin{cases} +1, & \text{if } e_{n-4} < e_{n-2} < e_n \text{ and } e_{n-3} < e_{n-1} \text{ (up-trend)} \\ -1, & \text{if } e_{n-4} > e_{n-2} > e_n \text{ and } e_{n-3} > e_{n-1} \text{ (down-trend)} \\ 0, & \text{otherwise (no trend)} \end{cases}$$

Before we define f_n^2 we first define some intermediary variables:

$$\theta_n^1 = \frac{\phi_n}{\phi_{n-1}}, \quad \theta_n^2 = \frac{\phi_n}{\phi_{n-2}}, \quad \theta_n^3 = \frac{\phi_{n-1}}{\phi_{n-2}}$$

These represent average volume ratios associated with the current zig-zag leg and its predecessors. We discretize each of the above ratios $(j = 1, 2, 3)$ to obtain:

$$\widetilde{\theta}_n^j = \begin{cases} +1, & \text{if } \theta_n^j - 1 > \alpha \\ -1, & \text{if } 1 - \theta_n^j > \alpha \\ 0, & \text{if } |\theta_n^j - 1| \leq \alpha \end{cases}$$

for some α level, which specifies what percentage change is necessary to identify significant increases or decreases in the average volumes of the zig-zag legs. If $\alpha = 0$, all changes in average volume are identified, and if $\alpha > 0$ then small increases or decreases in average volume are not recognized. By experimentation, we found $\alpha = 0.25$ detects changes in volume in high-frequency data appropriately. So for instance if the current leg's average volume had increased by more than 25% of the previous leg's, then $\widetilde{\theta}_n^1 = +1$, but if it had decreased by more than 25% instead, $\widetilde{\theta}_n^1 = -1$; had the change been less than 25%, $\widetilde{\theta}_n^1 = 0$.

Using $\widetilde{\theta}_n^1, \widetilde{\theta}_n^2, \widetilde{\theta}_n^3$ directly in our feature space would result in $3^3 = 27$ possible permutations, and when combined with features f_n^0 and f_n^1, there is a total of $2 \cdot 3 \cdot 27 = 162$ possibilities. To reduce the feature space size and simultaneously capture essential aspects we use the following grouping to define f_n^2,

$$f_n^2 = \begin{cases} +1, & \text{if } \widetilde{\theta}_n^1 = 1, \widetilde{\theta}_n^2 > -1, \widetilde{\theta}_n^3 < 1 \text{ (volume strenghtens)} \\ -1, & \text{if } \widetilde{\theta}_n^1 = -1, \widetilde{\theta}_n^2 < 1, \widetilde{\theta}_n^3 > -1 \text{ (volume weakens)} \\ 0, & \text{otherwise (volume is indeterminant)} \end{cases}$$

Thus f_n^2 characterizes whether volume is strengthening or weakening in the direction of the corresponding zig-zag leg.

Up Legs		Down Legs	
Symbol	Vector (O_n)	Symbol	Vector (O_n)
U_1	$(1,1,1)$	D_1	$(-1,1,-1)$
U_2	$(1,-1,1)$	D_2	$(-1,-1,-1)$
U_3	$(1,1,0)$	D_3	$(-1,1,0)$
U_4	$(1,0,1)$	D_4	$(-1,0,-1)$
U_5	$(1,0,0)$	D_5	$(-1,0,0)$
U_6	$(1,0,-1)$	D_6	$(-1,0,1)$
U_7	$(1,-10)$	D_7	$(-1,-1,0)$
U_8	$(1,1,-1)$	D_8	$(-1,1,1)$
U_9	$(1,-1,-1)$	D_9	$(-1,-1,1)$

Table 4.1: Enumeration of observation feature space. Ranges from bullish observation at the top to bearish observations at the bottom.

The final observation feature space consists of $2 \cdot 3 \cdot 3 = 18$ possibilities, nine for positive direction legs, and nine for negative direction legs. These are enumerated in Table 4.1 from bullish observations to bearish observations based on [Ord 08]'s price and volume prescriptions. Here, the terms bullish and bearish are referring to price behaviour over a high-frequency window that may last just minutes or even seconds. Bullish observations have strengthening volume ($f_k^2 = 1$) along positive zig-zag legs ($f_k^0 = 1$) and weakening volume ($f_k^2 = -1$) along negative zig-zag legs ($f_k^0 = -1$), with a trend ($f_k^1 = \pm 1$) emphasizing more significance. Conversely, bearish observations have strengthening volume ($f_k^2 = 1$) along negative zig-zag legs ($f_k^0 = -1$) and weakening volume ($f_k^2 = -1$) along positive zig-zag legs ($f_k^0 = 1$), again with a trend ($f_k^1 = \pm 1$) emphasizing more significance. Figure 4.5 shows an example of the unconditional distribution of the observations for GoldCorp Inc. (TSE:G) over the month of May, 2007.

4.4 Model specification

We specify the model first as a hierarchical hidden Markov model (HHMM), which we shall transform into dynamic Bayesian form for learning and

inference purposes. The HHMM is a well formalized tool suitable to model complex patterns in long temporal sequences. Figure 4.6 shows the hierarchical hidden Markov model (HHMM) we propose to use.

There is one root node, q^0, and two top level states, q_1^1 and q_2^1. These two states represent distinct modes the market can be in, in particular specifying whether the asset is in a run or a reversal. We do not explicitly designate a priori which state is associated to a run or a reversal, rather we allow the model to learn two different states (noting the symmetry) and subsequently label its meaning based on the in sample behaviour (Refer to Section 4.5).

Each of the top level states activates its own probabilistic model, which is a simple HMM. As discussed in Section 3.2.2, the semantics of a hidden Markov model requires that internal nodes (i.e. q^0, q_1^1 and q_2^1) undergo a vertical transition first; subsequently, after completing a depth first traversal of the tree, control returns to the activating node and then a horizontal transition is applied. Here, the state q_1^1 activates the internal state q_1^2; subsequently, it transitions horizontally, so that it is always alternating between states q_2^2 and q_1^2. These are production states which emit observations, X, with a distribution over possible feature vectors. Refer to Table 4.1 for an enumeration. In particular, q_1^2 emits negative zig-zag legs, i.e. $\{X|q_1^2\} = \{D_1, ..., D_9\}$ and q_2^2 emits positive zig-zag legs, i.e. $\{X|q_2^2\} = \{U_1, ..., U_9\}$. While in the top level state q_1^1, there is a non-zero probability of entering the termination state only from q_1^2, at which point control is returned back to the top level and a horizontal transition to q_2^1 is effected (note, there are no loop-backs). The q_2^1 state is symmetrical to the q_1^1. It activates on q_3^2 and subsequently alternates between q_3^2 and q_4^2, emitting positive zig-zag legs and negative zig-zag legs respectively, so $\{X|q_3^2\} = \{U_1, ..., U_9\}$ and $\{X|q_4^2\} = \{D_1, ..., D_9\}$. The q_2^1 state terminates only from q_3^2, at which point the top level state transitions back to q_1^1, and the process continues. The restriction on the activation and termination nodes is to enforce that all possible observation sequences are well behaved—alternating between positive and negative zig-zag legs—even as the top

level state undergoes transitions.

In addition to the model structure, to complete the definition of the HHMM, we need to specify the state transition probabilities between the states and the output distribution vector of the production states. For the root node and each top-level state q^k ($k = \{0, 1\}$ is the hierarchy index), there is a state transition probability matrix denoted by $A^{q^k} = (a_{ij}^{q^k})$ where $a_{ij}^{q^k} = P(q_j^{k+1} | q_i^{k+1}, q^k)$ is the probability of making a horizontal transition from the ith state to the jth while in the internal state q^k.

Transitions at layer 1 can be specified as,

$$A^{q^0} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

where the (i, j)th element corresponds to the probability of transitioning from q_i^1 to q_j^1 in Layer 1.

For the top level states,

$$A^{q_1^1} = \begin{pmatrix} 0 & p_1 & 0 & 0 & 1-p_1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$A^{q_2^1} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_2 & 1-p_2 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In both $A^{q_1^1}$ and $A^{q_2^1}$ the (i, j)th element corresponds to transitioning from q_i^2 to q_j^2 in Layer 2, noting that $j = 5$ represents the termination node for

both q_1^1 and q_2^1.

Similarly, $\Pi^{q^k} = \{\pi^{q^k}(i)\} = \{P(q_i^{k+1}|q^k)\}$ is the distribution vector over the substates of q^k, which is the probability that state q^k will initially activate the state q_i^{k+1}. For the root node, we assign an initial probability of starting in a particular top level state,

$$\Pi^{q^0} = \begin{pmatrix} p_i & (1-p_i) \end{pmatrix}^T$$

For the top-level states the activation is deterministic, thus,

$$\Pi^{q_1^1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}^T$$

and

$$\Pi^{q_2^1} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \end{pmatrix}^T$$

so q_1^1 activates q_1^2 and q_2^1 activates q_3^2.

Each production state, q^2, is parameterized by its output probability vector $B^{q^2} = \{b^{q^2}(k)\}$, where $b^{q^2}(k) = P(x_k|q^2)$ is the probability that the production state q^2 will output the symbol $x_k \in \Omega$. The entire set of parameters is denoted by

$$\lambda = \{\lambda^{q^d}\}_{d\in\{0,1,2\}} = \{\{A^{q^d}\}_{d\in\{0,1\}}, \{\Pi^{q^d}\}_{d\in\{0,1\}}, \{B^{q^2}\}\}$$

To summarize, the top level states prescribe a unique distribution over price and volume observations. The production states visited gives rise to a sequence of observation symbols according to these distributions. Furthermore, each top-level state consists of both positive and negative zig-zags in a symmetrical way. The symmetry of the two states, does not inherently assume that q_1^1 is bullish and q_2^1 is bearish or vice-versa. In fact, the feature vectors are price change agnostic, that is, for any realization of the observations there is equal number of positive and negative legs—identified with $+1$ and -1 for feature element f_k^0—irrespective of the

magnitude of the price change. The model focusses on capturing how volume interacts with price to identify whether there is buying or selling pressure.

The duration of each top level state, $q_i^1, i \in \{1,2\}$ has a geometric distribution with success probability parameter of p_i. Specifically, the probability we remain in state q_i^1 for exactly $(2d-1)$ steps is $P_i(d) = (1-p_i)p_i^{d-1}$ where $1-p_i$ is the probability of transitioning to the end state. We leave it to future work to incorporate and evaluate more general duration distribution models.

We can represent this HHMM as a DBN [Murphy 02] shown in Figure 4.4. Each node in a time slice represents a random variable and the model unfolds over discrete time steps. Note the time steps need not be uniformly spaced. In our case, our observations are realized at irregular time intervals as zig-zags are formed. In standard econometric treatises, we usually consider synchronous time steps. We would have to assign a business time scale transformation or sample at synchronous times. In comparison, by using a DBN framework we inherently are able to address non synchronous observations without any added complexity or loss of samples.

In the DBN model,

$$Q_t^1 = \begin{cases} 1, & \text{when top-level state is } q_1^1 \\ 2, & \text{when top-level state is } q_2^1 \end{cases}$$

$$Q_t^2 = \begin{cases} 1, & \text{when state } q_1^2 \text{is active} \\ 2, & \text{when state } q_2^2 \text{is active} \\ 3, & \text{when state } q_3^2 \text{is active} \\ 4, & \text{when state } q_4^2 \text{is active} \end{cases}$$

$$F_t = \begin{cases} 0, & \text{indicates we continue in the same } Q_t^1 \text{ state} \\ 1, & \text{indicates the } Q_t^1 \text{ state must transition} \end{cases}$$

$$O_t = j, \quad j \in \begin{cases} \{U_1, ..., U_9\}, & \text{if } Q_t^2 = q_1^2 \text{ or } Q_t^2 = q_3^2 \\ \{D_1, ..., D_9\}, & \text{if } Q_t^2 = q_2^2 \text{ or } Q_t^2 = q_4^2 \end{cases}$$

We now define the conditional probability distributions (CPDs) of each of the node types below, which will complete the definition of the model. We consider the top (Q_t^1) and bottom (Q_t^2) layers of the hierarchy separately (since they have different local topology), as well as the first, middle and last time slices.

Layer 1:

$$P(Q_t^1 = q_j^1 | Q_{t-1}^1 = q_i^1, F_{t-1} = f) = \begin{cases} \delta(i,j), & \text{if } f = 0 \\ A^{q^0}(i,j), & \text{if } f = 1 \end{cases}$$

where $\delta(i,j)$ implies we must stay in the same state.

Layer 2:

$$P(F_t = f | Q_t^2 = q_i^2, Q_t^1 = q_k^1) = A^{q_k^1}(i, end)$$

$$P(Q_t^2 = j | Q_{t-1}^2 = q_i^2, Q_t^1 = q_k^1, F_{t-1} = f) = \begin{cases} \tilde{A}^{q_k^1}(i,j), & \text{if } f = 0 \\ \Pi^{q_k^1}(j), & \text{if } f = 1 \end{cases}$$

where

$$\tilde{A}^{q_k^1}(i,j) = \frac{A^{q_k^1}(i,j)}{1 - A^{q_k^1}(i, end)}$$

is used for the transition matrix since Q_t^2 never actually enters an end state. Instead, the F_t node is used to signal the termination of the level.

Output level:

$$P(O_t = x_j | Q_t^2 = q_i^2) = b^{q_i^2}(j)$$

Note that for the output distributions we only need to condition on layer 2 nodes as the event specified by random variable Q_t^2 uniquely maps to an event specified by Q_t^1 of layer 1. In particular, when $Q_t^2 = q_1^2$ or $Q_t^2 = q_2^2$, then $Q_t^1 = q_1^1$ and when $Q_t^2 = q_3^2$ or $Q_t^2 = q_4^2$, then $Q_t^1 = q_2^1$.

The current model assumes only two regimes representing a 'buy state' and 'sell state'. The most natural addition would be a third state, neutral, in which no position is advocated. We could also consider a 5 hidden state model, where we distinguish a strong buy from a weak buy and a strong sell from a weak sell. By reducing the number of states (and hence parameters) we limit ourselves from over-fitting in sample (and therefore are more robust out of sample). Whether two regimes suffice is not clear. However, with even three regimes, we found higher likelihoods do not necessarily increase out-of-sample accuracy. This is partly because we are not labeling our hidden nodes and hence our conditional distributions are not 'unique' enough for the given sample amount (i.e. we would need more samples do learn three or more unlabeled hidden states). With more samples, faster implementation and better learning heuristics (such as deterministic annealing) we can revisit more complicated models and experiment further. Also another possibility is semi-supervised learning, where we label at least some of the hidden states—but we would need to do this very carefully since as discussed in Section 4.5, it is not always clear what state corresponds to a particular observation. Finally, there are algorithms to learn the structure of the Bayesian network, which we could apply to see how many hidden states best fit the data. We leave this to future work.

4.5 Learning and inference

We use the EM algorithm to learn the two different top-level states based on the high-frequency data observations alone. The top-level states are not labeled in the learning phase. They can be considered high-frequency regimes, where the latent top-level variable (q_i^1) specifies the active regime

and each regime determines a unique observation distribution over the price and volume feature space.

Our goal is to distinguish between bullish trends (run) and bearish trends (reversals) in the high-frequency window. However, there are many ways to label runs and reversals. Often trends are identified at different frequencies by using retracement levels. A retracement level is the percentage change from a minimum or maximum that must be reached to change the direction of the current trend. If we use different retracements levels we can observe any particular observation may be classified as either part of a run or part of a reversal. Refer to Figure 4.8, in which a stock series has been shown with 5% and 3% retracement levels. We can see that depending on how we choose to measure runs or reversals, we obtain different classifications for the same point in time. Thus, to alleviate the possibility of inappropriately labeling the regime, we allow the EM algorithm to best explain the observation sequence based on switching between two hidden regimes. Thus an unsupervised learning methodology is ideal to protect from inappropriate labeling.

Moreover since these two top-level states are symmetrical in structural semantics (refer to Section 4.4), there is no built-in bias associated with the regime. A regime switch simply implies structural changes in the price and volume observations of the series, signalling when buying and selling pressure has changed. Once the model has been learned we must analyze the results to investigate what the regimes represent, if anything useful (see Section 5.2).

Since the likelihood surface in EM is riddled with local optima, we start with several different initial parameters and pick the results that provide highest likelihood in search of the global optimum.

Once we have learned a model based on samples we may use it in the future for inference. Viterbi inference at each time step is used to determine the top level state based on observation up to the current time t. Thus,

$$\hat{Q}_t^1 = \arg\max_{Q_{1:t}^1} P(Q_{1:t}^1 | O_{1:t})$$

This is in contrast to using the Filter inference in which

$$\hat{Q}_t^1 = \max_{Q_t^1} P(Q_t^1 | O_{1:t})$$

is calculated, since in Viterbi the sequence of all possible states are examined to see which one is most probable given the data up to time t.

Figure 4.1: Bema Gold Corp. chart showing that in a bullish trend, volume increases as price increases, and volume decreases as price declines. (Adopted from [Ord 08])

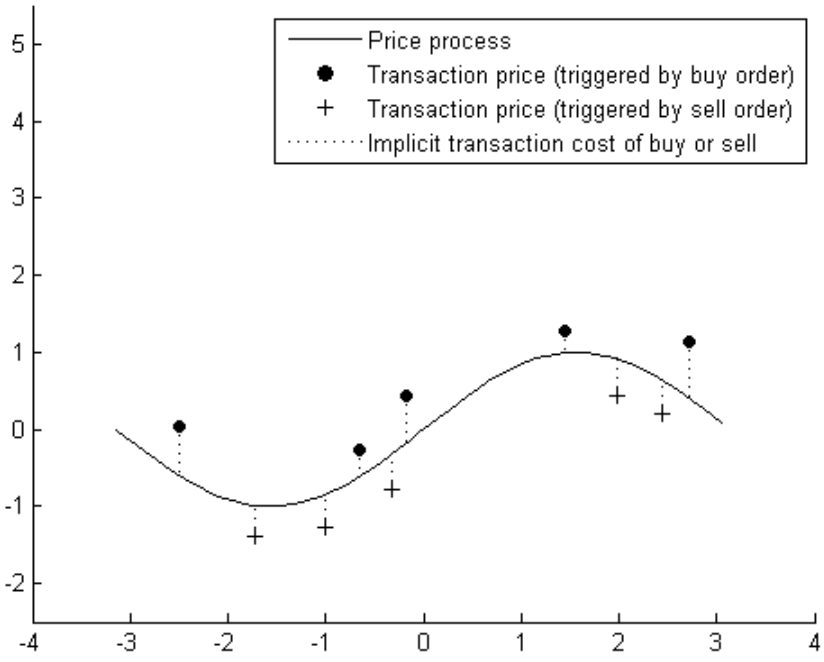


Figure 4.2: Evolution of the transaction price and the bid-ask bounce.

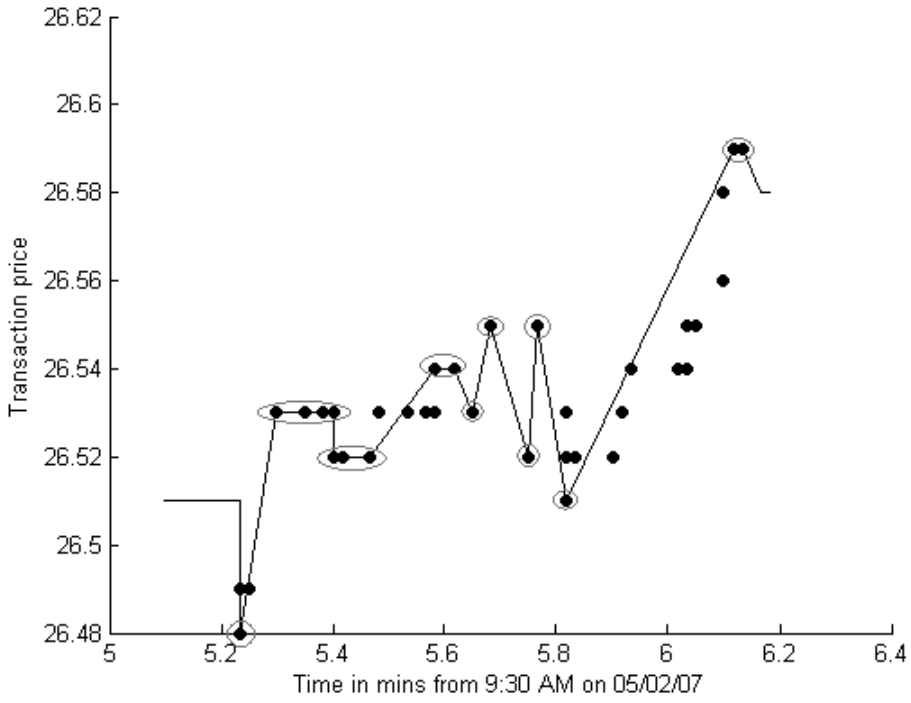


Figure 4.3: Sample tick level zig-zags extracted from transaction price for Goldcorp Inc (TSE:G). Red circles indicate local extrema points (or plateaus) which form tick level support and resistance levels.

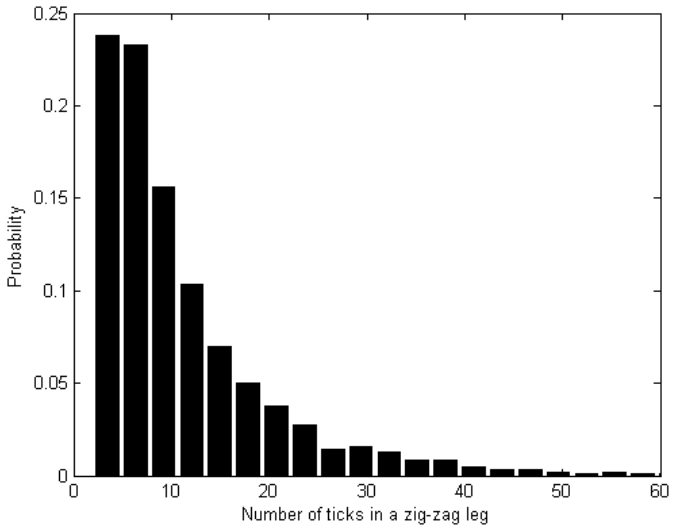


Figure 4.4: Distribution of length of zig-zag leg in number of ticks for GoldCorp Inc (TSE:G) for May 2007.

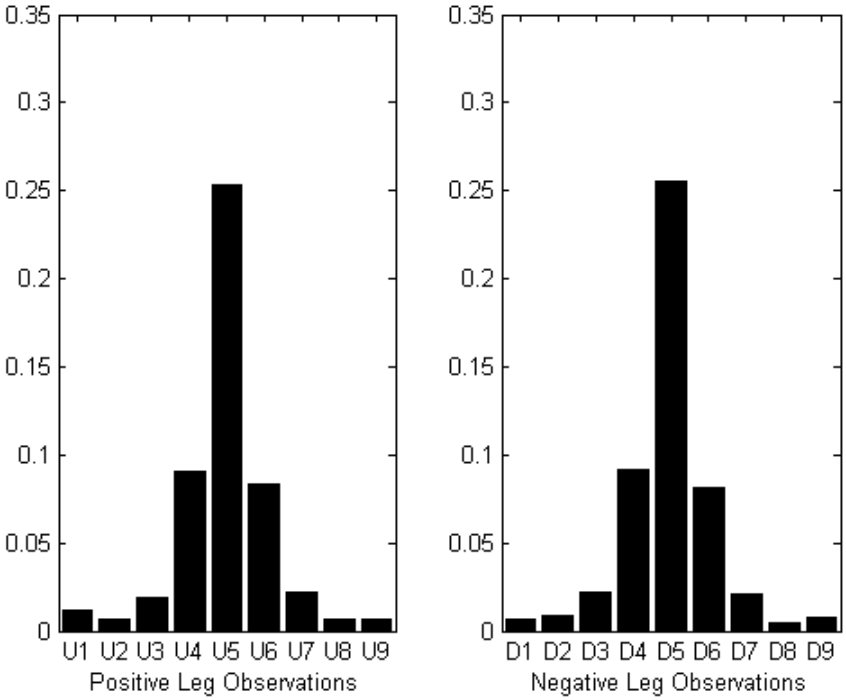


Figure 4.5: Unconditional distribution of observations for GoldCorp Inc. (TSE:G) over the month of May, 2007.

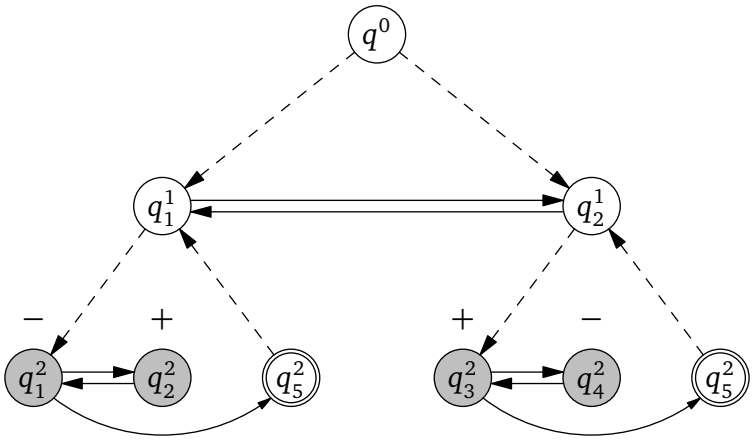


Figure 4.6: Hierachical hidden Markov model for price and volume analysis. q_1^1 and q_2^1 are top level states representing runs or reversals. q_1^2 and q_4^2 represent negative zig-zag legs, while q_2^2 and q_3^2 represent positive zig-zag legs. These are production nodes, filled in gray, that emit an observation symbol according to some distribution. Transitions enforce the alternating sequence of positive and negative legs. q_5^2 is the termination nodes (note, there are two of them) at which point control is returned back to the parent node in layer 1.

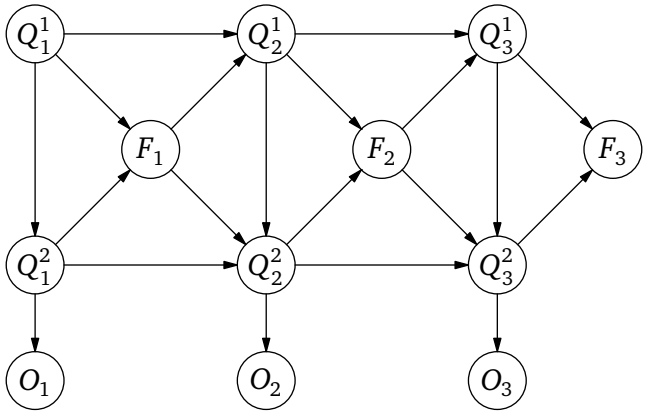


Figure 4.7: First three time slices of equivalent DBN for price and volume analysis.

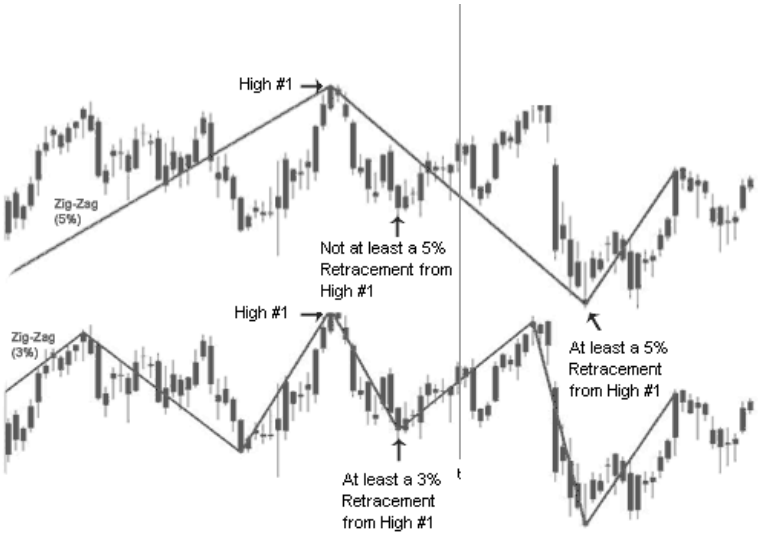


Figure 4.8: Example of how zig-zags extracted with different retracement levels can allow the same point in time to be labeled differently. For instance, along the gray vertical line, using 5% retracement we would classify the observation point as belonging to a downtrend, while using 3% retracement we would classify the observation as belonging to an uptrend.

Chapter 5

Computational Results

In this section we learn and evaluate the forecasting ability of the model we developed. We first consider a simulated example to show how the model can encode price and volume patterns and how the EM algorithm can effectively learn back the patterns out of sample. Subsequently, we use historical Toronto Stock Exchange data to learn the model on a rolling window basis and evaluate the results statistically.

5.1 Simulated example

We illustrate how the DBN structure can model price and volume technical patterns with a hypothetical set of parameters. Define top-level state q_1^1 as a bullish state and q_2^1 as a bearish state. Refer to Figure 4.6. In the bullish state the observations favour strengthening volume on the upward zig-zag legs, and weakening volume on the downward zig-zag legs; in the bearish state the observations tend towards weakening volume on the upward legs and strengthening volume on the downward legs. This behaviour can be expressed by specifying the conditional observation distributions, $b^{q_i^2}(j)$, for $i \in \{1,2,3,4\}$ and $j \in \{U_1,...,U_9,D_1,...,D_9\}$, as described below.

Let $\Theta(i;\mu,\sigma) = \Phi(i+0.5;\mu,\sigma) - \Phi(i-0.5,\mu,\sigma)$, where $\Phi(x;\mu,\sigma)$ is the cumulative normal with mean μ and standard deviation σ. This effectively

discretizes the normal probability distribution.

For the bullish state, q_1^1, define conditional observation distributions as:

$$b^{q_1^2}(U_k) = \Theta(k; \mu_1, \sigma_1), \quad b^{q_1^2}(D_k) = 0$$

$$b^{q_2^2}(D_k) = \Theta(k; \mu_1, \sigma_1), \quad b^{q_2^2}(U_k) = 0$$

For the bearish state, q_2^1, define conditional observation distributions as:

$$b^{q_3^2}(U_k) = \Theta(k; \mu_2, \sigma_2), \quad b^{q_3^2}(D_k) = 0$$

$$b^{q_4^2}(D_k) = \Theta(k; \mu_2, \sigma_2), \quad b^{q_4^2}(U_k) = 0$$

where $k = 1, ..., 9$. By choosing $\mu_1 = 3, \mu_2 = 7, \sigma_1 = \sigma_2 = 2.5$, we obtain conditional observation distributions as shown in Figure 5.1. As discussed in Section 4.3, U_1 to U_9 are ranked from bullish upward legs to bearish upward legs and D_1 to D_9 are ranked from bullish downward legs to bearish downward legs. Thus we have a higher likelihood for observations that have supporting volume on positive zig-zag legs and weakening volume on negative zig-zag legs conditioned on being in the bullish state (q_1^1). Conversely, we have a higher probability of supporting volume on negative zig-zag legs and weakening volume on positive zig-zag legs conditioned on the bearish state (q_2^1). In addition we set the probability of remaining in the top level state as $p_1 = p_2 = 0.8$, encoding the expected duration of the trends with a geometric distribution. Figure 5.2 shows the unconditional distribution of observations based on simulating the DBN for 1000 time-steps. This is a mixture of the conditional observation distributions. Figure 5.3 shows the duration distribution of the top-level state of the simulated data, where duration is the number of zig-zags before a top-level state switch.

We divide the 1000 samples into two groups of 500. We use the first 500 samples to learn the parameters of the model using expectation maximization (discussed in Section 3.4) assuming all we know is the observation sequence. We do not assume we can assign labels to the

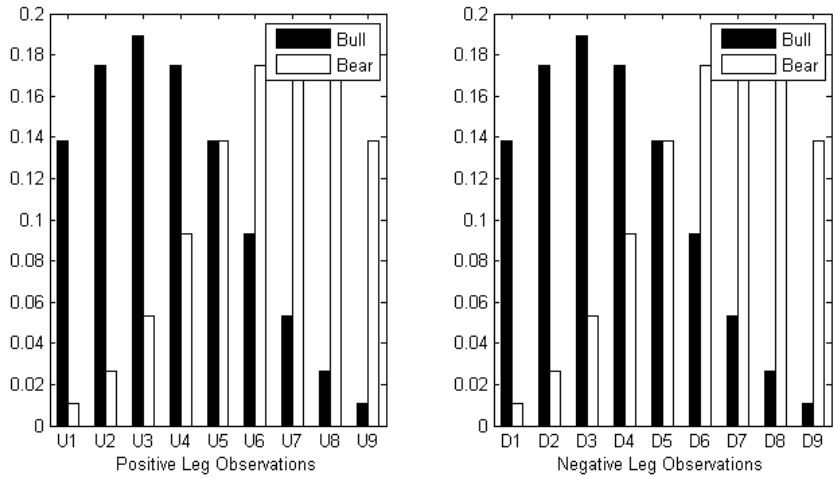


Figure 5.1: Conditional distribution of observations for simulation parameters $\mu_1 = 3, \mu_2 = 7, \sigma_1 = \sigma_2 = 2.5$.

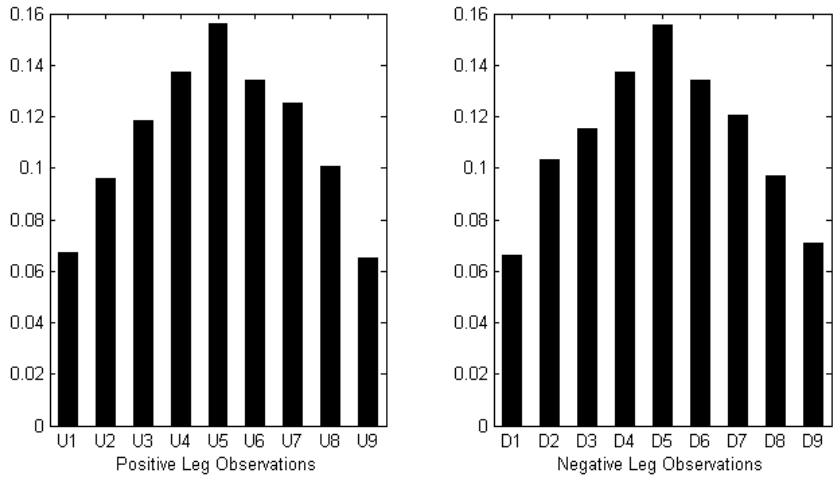


Figure 5.2: Unconditional distribution of observations based on simulation of DBN for 1000 time steps with parameters $\mu_1 = 3, \mu_2 = 7, \sigma_1 = \sigma_2 = 2.5$.

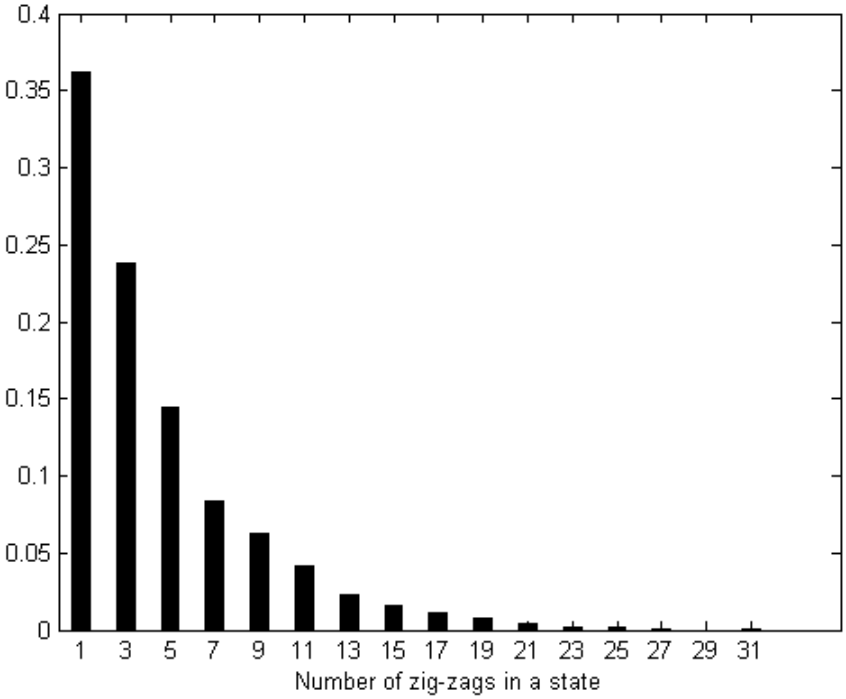


Figure 5.3: Duration distribution of the top-level state of the simulated data.

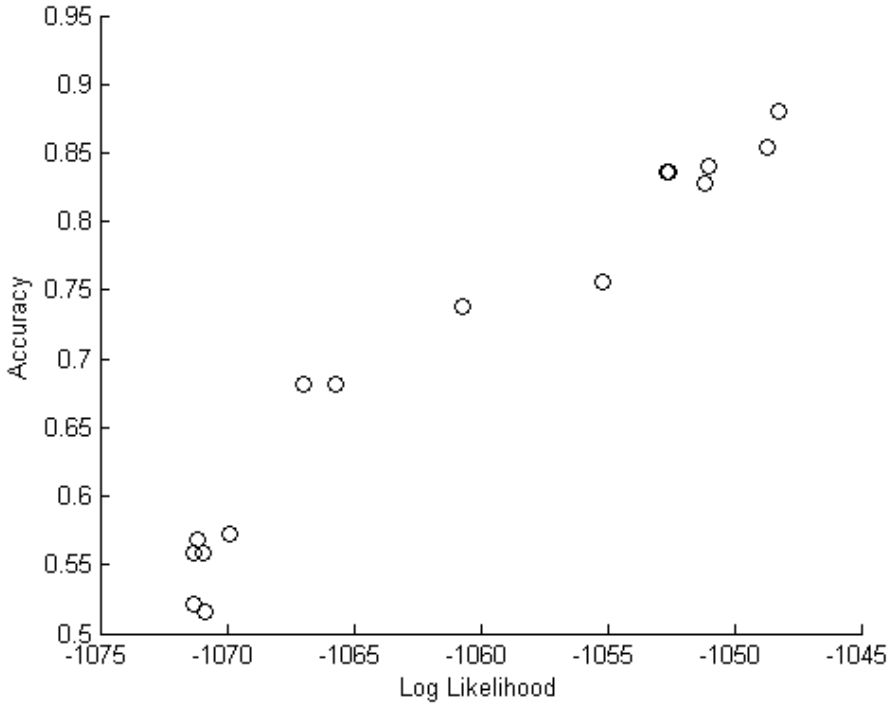


Figure 5.4: Model learned likelihood versus percentage accuracy.

top-level states for reasons explained in section 4.5.

Once we have learned the parameters, we use them on the second untouched set of 500 samples. The learned parameters are used to infer the top-level states of the second group of 500 out-of-sample observations with the Viterbi algorithm. Thus,

$$\hat{Q}_t^1 = \arg\max_{Q_{1:t}^1} P(Q_{1:t}^1 | O_{1:t})$$

for $t = 501, ..., 1000$.

We start with twenty different initial parameters chosen randomly. For each attempt, we optimized the likelihood and calculated the out-of-sample percentage accuracy. Figure 5.4 shows a scatter plot of likelihood versus percentage accuracy. For two possible top-level states, simple guessing yields 50% accuracy in expectation. However, the learned model obtains up to 90% accuracy. We see that obtaining a higher likelihood in general provides a higher accuracy rate—consequently, likelihood maximization is able to effectively learn back the latent model parameters out of sample.

5.2 TSE60 experiment and analysis

Using historical high-frequency Toronto Stock Exchange data, we analyze 60 stocks of the TSE60 for May 2007. The data consists of 22 business days excluding holidays and weekends. In addition, a manual data cleansing process revealed three days of unusable data due to significant errors, which leaves us with 19 days, labeled as $D_1, ..., D_{19}$. Transactions are identified in the data to form the raw tick series. In aggregate, over the 60 stocks this consisted of 3,449,363 ticks and after applying the feature extraction process, 646,692 number of zig-zag observations. At a daily frequency this would account for 13,688 years of observation data and on average 228 years of data per stock. So although 19 days may not seem a lot, at a high-frequency scale, it provides for ample observation points. We

Table 5.1: Quartile groupings by average daily volume in the month of April 2007.

Largest Quartile		2nd Quartile		3rd Quartile		Smallest Quartile	
Ticker	Vol (mil)	Ticker	Vol (mil)	Ticker	Vol (mil)	Ticker	Vol (mil)
SJR.B	22.60	MFC	2.10	YLO.UN	1.24	IMO	0.59
TOC	10.50	CCO	1.99	TRP	1.18	CP	0.56
BCE	9.40	RY	1.98	T	1.17	SC	0.56
BBD.B	5.06	ABX	1.96	BAM.A	1.03	NA	0.54
MG.A	4.70	PCA	1.91	TA	1.02	SNC	0.52
G	4.11	YRI	1.85	MDS	0.91	BVF	0.50
SXR	3.95	CNR	1.78	SLF	0.91	FM	0.45
TLM	3.78	BNS	1.73	HSE	0.85	L	0.45
LUN	3.53	RCI.B	1.54	POT	0.83	NCX	0.44
CTC.A	3.40	RIM	1.48	AEM	0.79	THI	0.40
K	3.32	COS.UN	1.43	AGU	0.77	GIL	0.38
SU	2.92	BMO	1.36	ENB	0.72	FTS	0.33
TCK.B	2.70	TD	1.30	CM	0.65	ERF.UN	0.32
ECA	2.23	CNQ	1.29	AER.UN	0.65	IMN	0.29
NXY	2.19	NT	1.27	PWT.UN	0.62	WN	0.13

process the raw series to obtain zig-zag features as described in Section 4.3.

We divide the 60 stocks into four groups (quartiles) based on the average volume of trades in the previous month, labeling them $G_1, ..., G_4$, where G_1 consists of 25% of the stocks that had the most volume, and G_4 consists of 25% of the lowest volume stocks. Refer to Table 5.1 for a listing of each group. The reason for doing this is to be able to investigate the cross-sectional performance of the model. (In our analysis we also considered dividing the stocks into ten group or by decile, results were similar. To highlight the important characteristics we are showing quartile results.) We would expect that higher volume stocks would be more likely to have price and volume patterns embedded, since price movements in these highly liquid stocks require more volume synchronicity; consequently a volume move would carry more significance.

Figures 5.5, 5.6, 5.7 and 5.8 show the unconditional distribution of the observations for each quartile. We can observe that all four quartiles have remarkably similar unconditional distributions—each with relatively normal skew (-0.01) and a kurtosis lower than that of a Gaussian distribution

(1.3).

We use the price and volume dynamic Bayesian network described in Section 4.4 to model the features extracted. Model parameters are learned based on a rolling five day window using the EM algorithm as described in Section 3.4, thus maximizing the posterior probability of the parameters λ (see Section 4.4) given the observations O,

$$\lambda^* = \arg\max_{\lambda} \log P(O;\lambda) = \arg\max_{\lambda} \log \Sigma_H P(O,H;\lambda)$$

where $H \in q_1^1, q_2^1$ represents the hidden state variables. We attempted EM trials using five initial starts and chose the best set of parameters λ^* that maximized the joint likelihood. Since the EM algorithm can be computationally expensive on large sets of data we limited number of EM trials to five—a future course of study can investigate more trials and/or combine other searching techniques such as genetic algorithms in attempt to find the global maximum likelihood.

The above step provides us with a learned set of parameters for the DBN model over a five day historic window. Recall, in the corresponding HHMM model the two top level states (q_1^1, q_2^1) are symmetric and we do not assign semantic meaning to them prior to the learning phase. Once we've learned the parameters, we asses the trade performance of these states over the five day historic in-sample period. Noting that the expectation step in the EM algorithm has already marked each hidden state with the most likely state value, we can evaluate the in-sample expected trade return of each state as follows,

$$E(R_{q_1^1}) = \frac{1}{N_{q_1^1}} \sum_{k=1}^{N_{q_1^1}} \frac{p_{q_1^1}^{f_k} - p_{q_1^1}^{i_k}}{p_{q_1^1}^{i_k}} \tag{5.1}$$

$$E(R_{q_2^1}) = \frac{1}{N_{q_2^1}} \sum_{k=1}^{N_{q_2^1}} \frac{p_{q_2^1}^{f_k} - p_{q_2^1}^{i_k}}{p_{q_2^1}^{i_k}} \tag{5.2}$$

where $p_{q_1^1}^{i_k}$ ($p_{q_2^1}^{i_k}$) is the initial price at the beginning of the kth continuous

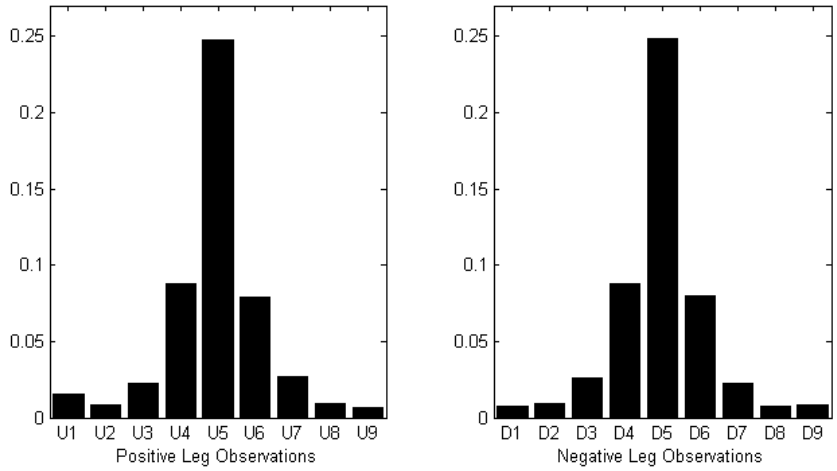


Figure 5.5: Unconditional distribution of observations for G_1.

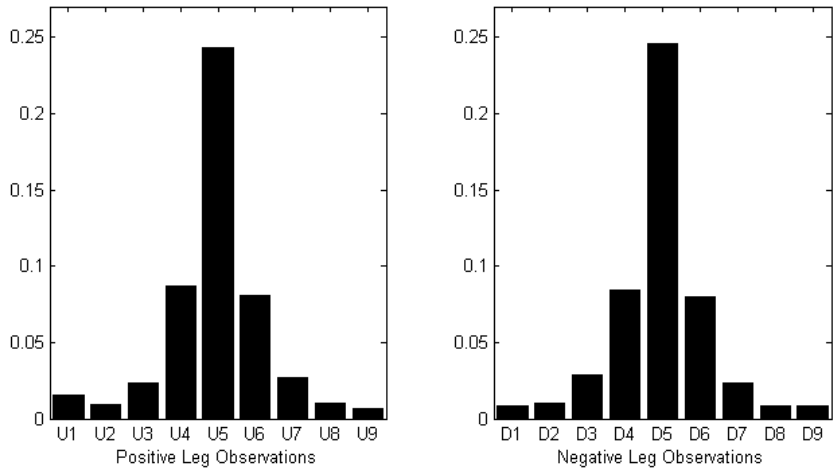


Figure 5.6: Unconditional distribution of observations for G_2.

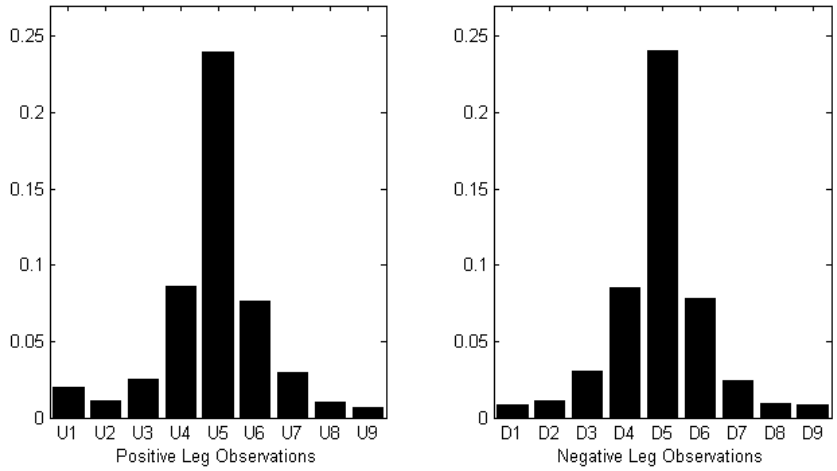


Figure 5.7: Unconditional distribution of observations for G_3.

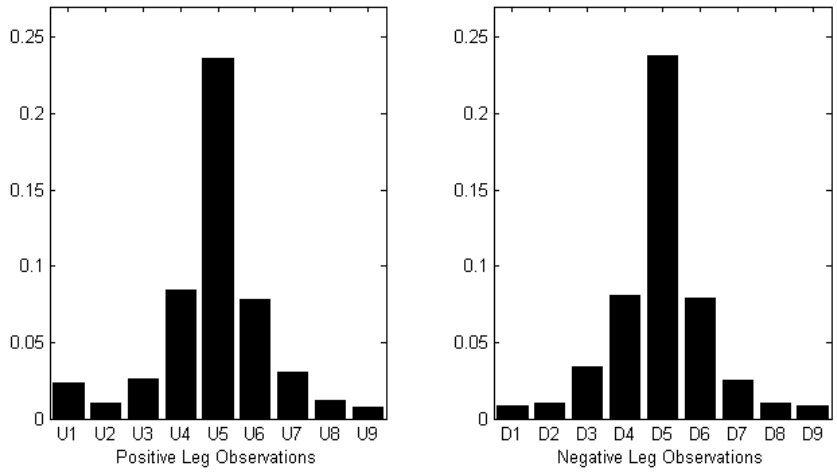


Figure 5.8: Unconditional distribution of observations for G_4.

block of q_1^1 (q_2^1) states, $p_{q_1^1}^{f_k}$ ($p_{q_2^1}^{f_k}$) is the final price before the top-level state switch, and $N_{q_1^1}$ ($N_{q_2^1}$) is the number of samples of state switches. We can now assign meaning to each state: if $E(R_{q_1^1}) > E(R_{q_2^1})$ we designate state q_1^1 as a run (bullish) and q_2^1 as a reversal (bearish), otherwise we designate q_2^1 as a run (bullish) and q_1^1 as a reversal (bearish).

Figures 5.9, 5.10, 5.11 and 5.12 show the in-sample conditional (conditioned on being in the bullish or bearish top-level state) distribution of feature observations learned for each quartile. The learned conditional distributions are similar across quartiles with the same characteristics showing up in each.

Recall Table 4.1 lists the feature vectors, which we ranked a priori based on volume and price technical analysis concepts so that U_1,...,U_4 (for the positive legs) and D_1, ..., D_4 (for the negative legs) are bullish observations whereas U_6,...,U_9 (for the positive legs) and D_6, ..., D_9 (for the negative legs) are bearish observations. We can see that there is a strong tilt towards the bullish observations in the bullish state, and a strong tilt towards the bearish observations in the bearish state. This behaviour is learned from the unconditional distribution directly using EM.

This validates that the intraday price process can be split into two regimes, with unique distributions over the feature vectors. Each regime is distinguished from the unconditional distribution with a tilt towards price and volume characteristics so that the bullish state is more likely to have observations where price increases are supported by volume increases and price decreases are not supported by volume decreases, whereas the bearish state is tilted toward observations where price decreases are supported by volume increases and price increases are not supported by volume decreases.

The exception is U_2 and U_8 for up legs and D_2 and D_8 for down legs. The bullish state is more likely to have observation feature U_8 and D_8 than U_2 and D_2, which counters our a priori ranking shown in Table 4.1. U_2 (equiv. D_2) refers to feature observation (1, -1, 1) (equiv. (-1, -1, -1)), and

U_8 (equiv. D_8) refers to feature observation (1, 1, -1) (equiv. (-1, 1, 1)). Recalling f_k^0 is the direction of the leg, f_k^1 is a price momentum indicator and f_k^2 is the volume indicator, we see that in these cases momentum is more significant than the price and volume formation. Using DBNs we are thus able to learn characteristics which can be used for technical analysis accounting for exceptions in a coherent probabilistic framework.

Now that we have a fully specified model, we can use this model for inference purposes on the sixth out-of-sample day. Recalling Section 3.3 and Figure 3.7, we have several options for inference: filtering, Viterbi, prediction, fixed-lag smoothing, offline fixed-interval smoothing, and offline fixed-interval Viterbi smoothing. We investigate two of these inference types, offline fixed-interval Viterbi smoothing and Viterbi.

Offline fixed-interval Viterbi smoothing infers the hidden state at time t as,

$$\hat{\hat{Q}}_t^1 = \arg\max_{Q_{1:t}^1} P(Q_{1:t}^1 | O_{1:T})$$

where $Q_{1:t}^1$ represent the sequence of hidden states up to present time t, and $O_{1:T}$ represents all the evidence for the sixth out-of-sample day (i.e. T corresponds to the final DBN slice or zig-zag of the sixth day). We are inferring the state at some earlier time t using information known for the remainder of the day, and thus permitting a look-ahead bias. The information obtained about the inferred state using this approach cannot be traded upon, however, it provides us with an upper bound benchmark of our model. Consequently, it illustrates whether the designed model was capable of learning meaningful patterns out-of-sample. Refer to Figure 5.13 for an example out-of-sample day where the offline fixed-interval viterbi was used for inference.

Viterbi inference estimates the hidden state at time t as,

$$\hat{Q}_t^1 = \arg\max_{Q_{1:t}^1} P(Q_{1:t}^1 | O_{1:t})$$

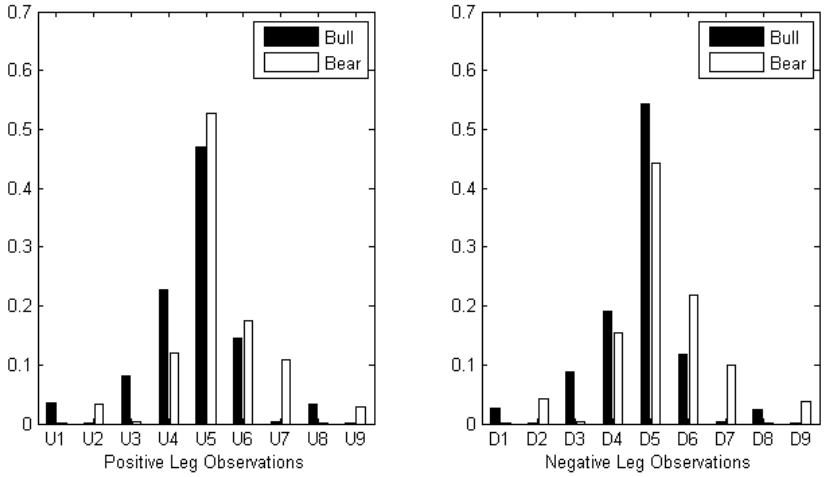


Figure 5.9: Conditional distribution of observations for G_1.

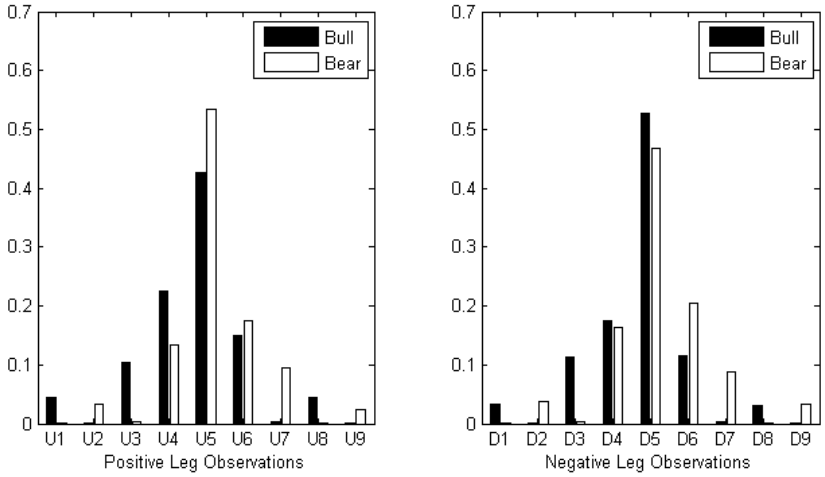


Figure 5.10: Conditional distribution of observations for G_2.

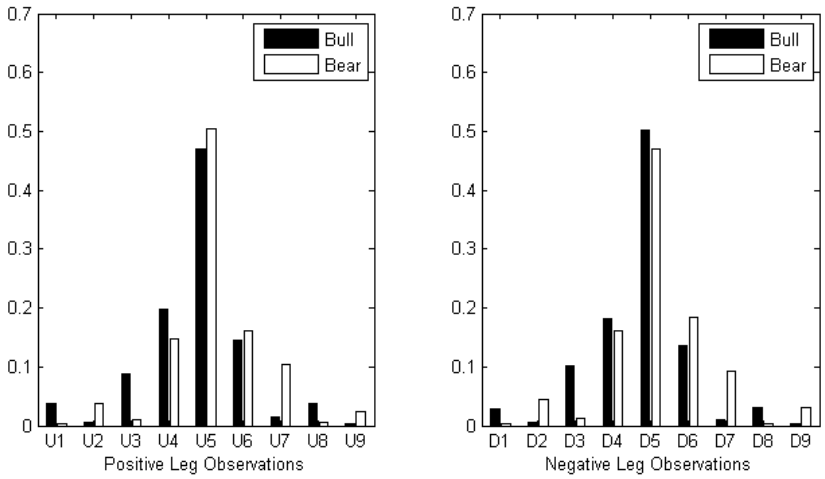


Figure 5.11: Conditional distribution of observations for G_3.

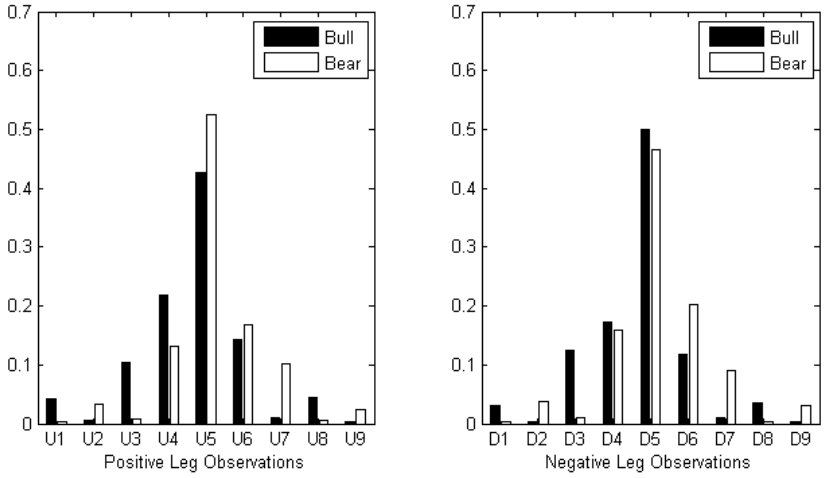


Figure 5.12: Conditional distribution of observations for G_4.

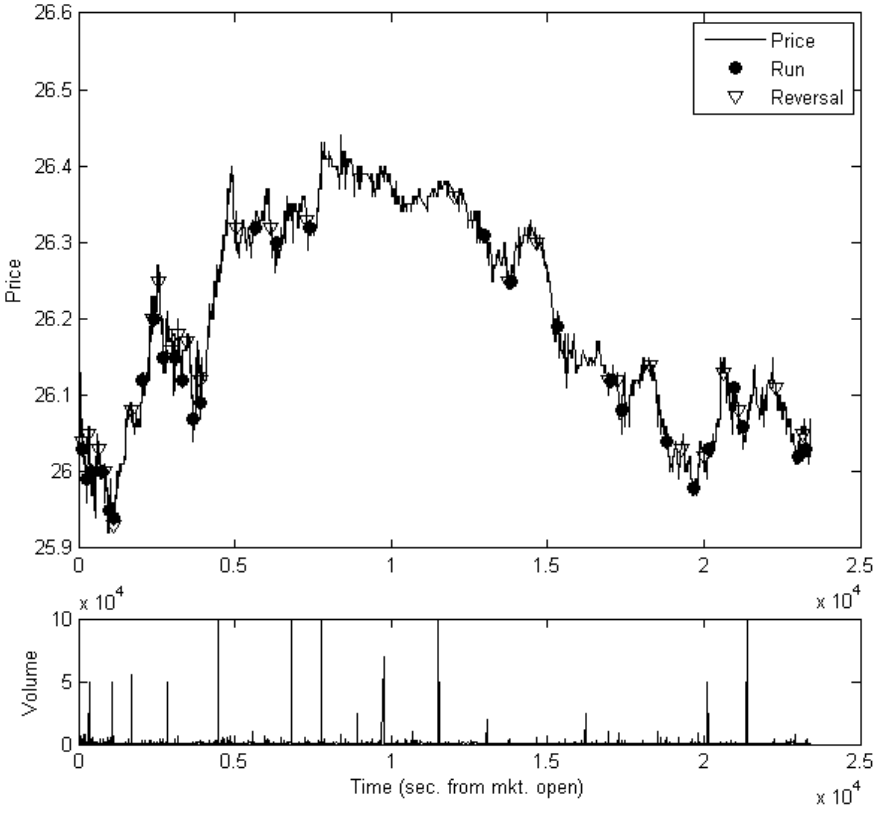


Figure 5.13: Example of out-of-sample offline fixed-interval look-ahead viterbi inference. Sample day showing results for GoldCorp Inc. (TSE:G) on May 11, 2007. Filled circles are the start of the bullish state and upside-down triangles are the start of the bearish state.

where $Q^1_{1:t}$ represent the sequence of hidden states up to present time t, and $O_{1:t}$ represents the observational evidence up to present time t for the sixth out-of-sample day. In this case we do not incur any look-ahead bias—we are only using information known up to the present time to infer the hidden state at the present time. Thus, the inferred states can be traded upon. Refer to Figure 5.14 for an example sixth day where viterbi was used for inference.

We shall investigate the trade return distribution based on the inference out-of-sample. The kth trade return is given by

$$R_k = \frac{p^{f_k} - p^{i_k}}{p^{i_k}}$$

where p^{i_k} is the initial price at the beginning of a top-level state switch, p^{f_k} is the final price before the top-level state switches again. As mentioned in Section 4.3, we do not observe a realization of a zig-zag point (and hence the corresponding observation) as soon as it is completed—rather there is a one tick lag between the leg completion and the time of detection. We assume we trade at the next tick, after the zig-zag leg is completed and we are able to detect the zig-zag leg, ensuring we do not have any look-ahead bias when evaluating the return.

5.2.1 Goodness-Of-Fit tests

A natural first step in the analysis of the model is to gauge the information content of the top-level (q^1_1, q^1_2) learned states. We propose to do this by investigating the unconditional empirical distribution of trade returns with the corresponding conditional empirical distribution, conditioned on the top-level state, for each quartile. If the model has learned two distinct states, conditioning on them should alter the empirical distribution of trade returns, otherwise if the model has simply over-fit the observation features out-of-sample then the conditional and unconditional distributions of trade returns should be close. Although this is a weaker test of the effectiveness of the model—informativeness does not guarantee a prof-

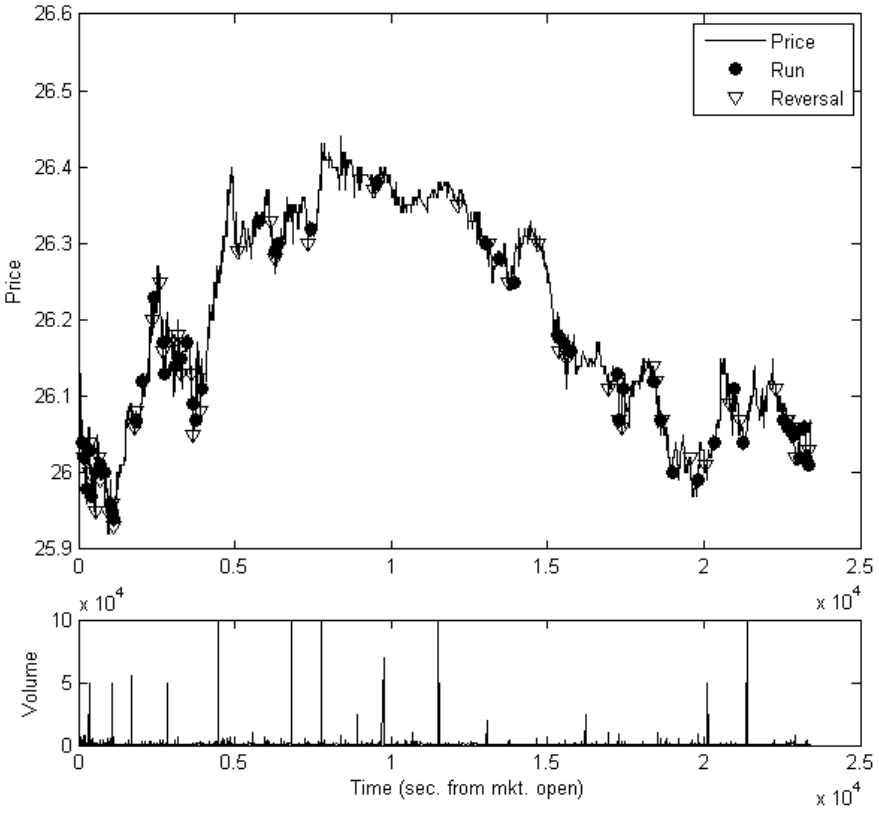


Figure 5.14: Example of out-of-sample viterbi inference. Sample day showing results for GoldCorp Inc. (TSE:G) on May 11, 2007. Filled circles are the start of the bullish state and upside-down triangles are the start of the bearish state.

itable trading strategy—it is nevertheless a more fundamental assessment of whether the model has even learned anything at all.

Table 5.2 and Figures 5.15, 5.16, 5.17 and 5.18 show the conditional and unconditional trade return distributions for each quartile. We see the trade return distributions are quite unlike normal distributions with significant skew and kurtosis. In particular, the bullish state trade return distribution is positively skewed with significant kurtosis, while the bearish state trade return distribution is negatively skewed also with significant kurtosis. This implies most of the action is occurring at the tails. Looking at the graphs we see that the bullish distribution lies above the bearish distribution in the right tail and below in the left tail. This is the case for all the four quartiles, with results being stronger for the top quartile and diminishing as we go to lower volume quartiles.

Chi-Square test

The chi-square Goodness-of-fit test [Snedecor 89] is used to test if a sample of data came from a population with a specific distribution. We can use this tool to test the informativeness of the two top-level states by checking if the conditional trade return distribution is statistically equivalent to the unconditional trade distribution. If conditioning on the regime provides no incremental information, the conditional trade returns should be similar to those of the unconditional returns.

The test requires that the data first be grouped. The actual number of observations in each group is compared to the expected number of observations and the test statistic is calculated as a function of this difference. The number of groups and how group membership is defined affects the statistical power of the test (i.e. how sensitive it is to detecting departures from the null hypothesis). The power of the test is also affected by the sample size and shape of the null and underlying (true) distributions. In general, power is maximized by choosing endpoints such that group membership is equiprobable (i.e. the probabilities associated with an observation falling into a given group are divided as evenly as possible

Table 5.2: Summary characteristics of the conditional trade return distributions for each quartile.

	Avg. State	Bullish State				Bearish State			
	Time (min)	Mean	Std.	Skewness	Kurtosis	Mean	Std.	Skewness	Kurtosis
Largest Quartile	6.25	0.010%	0.228%	2.73	35.07	-0.017%	0.205%	-2.00	24.31
2nd Quartile	5.49	0.003%	0.162%	2.62	31.54	-0.004%	0.158%	-1.64	14.89
3rd Quartile	9.36	0.003%	0.188%	0.71	13.98	-0.005%	0.182%	-0.59	14.45
Smallest Quartile	11.84	0.006%	0.238%	2.66	26.97	-0.001%	0.232%	-0.82	16.07
All stocks	7.47	0.005%	0.201%	2.48	32.29	-0.008%	0.190%	-1.41	20.24

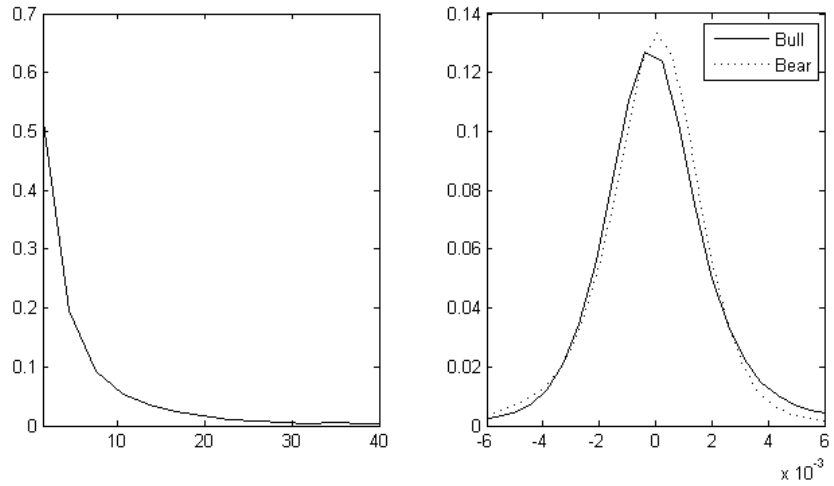


Figure 5.15: Left: Distribution of the length of a state; right: conditional distribution of trade returns for G_1.

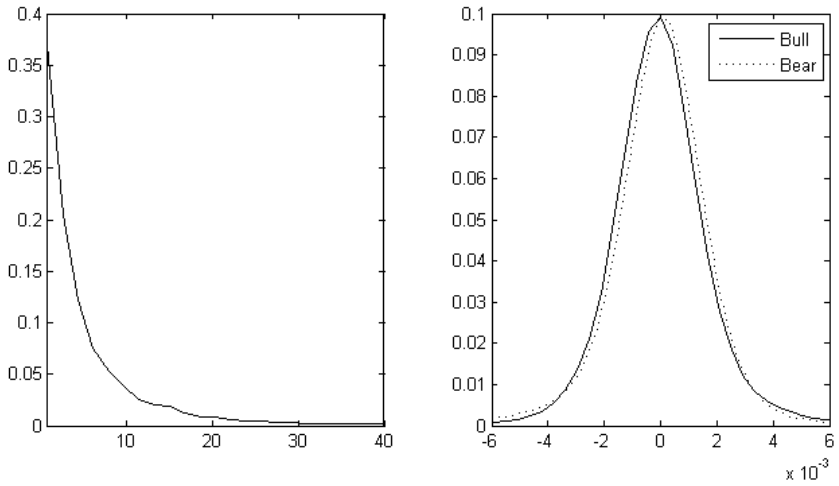


Figure 5.16: Left: Distribution of the length of a state; right: conditional distribution of trade returns for G_2.

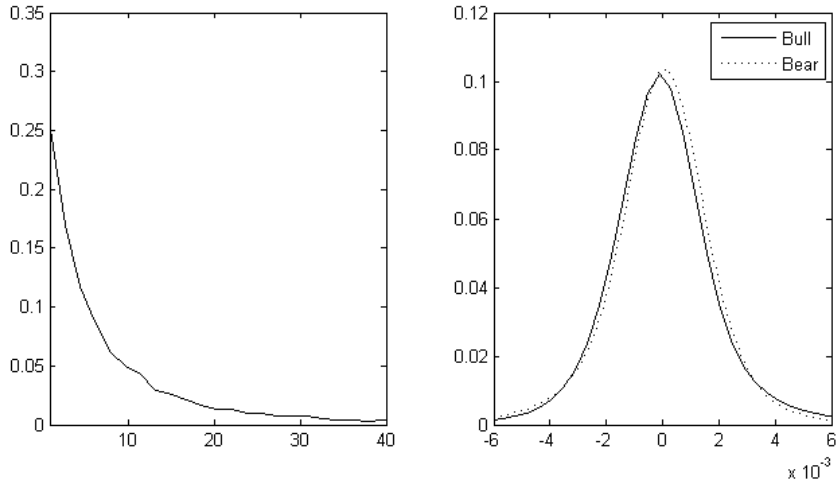


Figure 5.17: Left: Distribution of the length of a state; right: conditional distribution of trade returns for G_3.

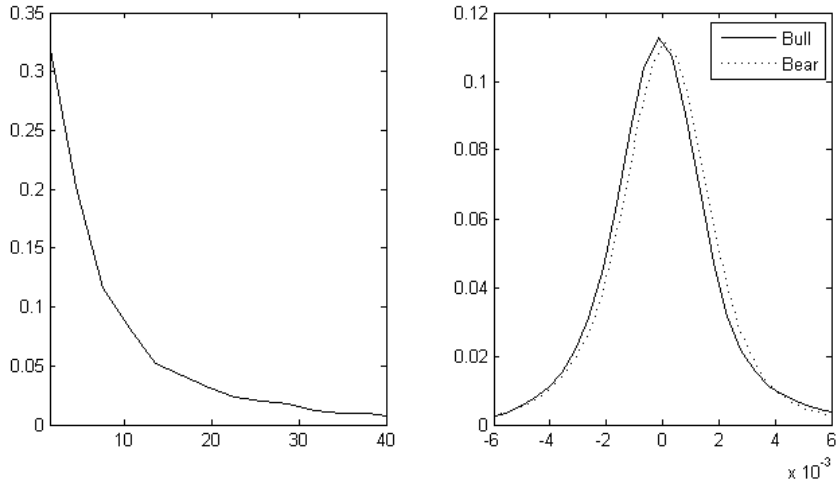


Figure 5.18: Left: Distribution of the length of a state; right: conditional distribution of trade returns for G_4.

across the intervals) [Sheskin 00].

Thus we could use groups based on quantiles of the conditional returns with the unconditional returns. In particular, we compute the deciles of unconditional trade returns, thus grouping the data into 10 bins. We tabulate the relative frequency $\hat{\delta}_j$ of conditional trade returns falling into decile j of the unconditional returns, $j = 1, ..., 10$,

$$\hat{\delta}_j \equiv \frac{\text{number of conditional returns in decile j}}{\text{total number of conditional returns}}$$

Assuming that the trade returns are independent and identically distributed, the chi-square test defined is,

H_0:	Conditional and unconditional trade distributions are identical.
H_a:	Conditional distribution is not the same as the unconditional distribution.
Q Statistic:	$Q \equiv \sum_{j=1}^{10} \frac{(n_j - 0.10n)^2}{0.10n} \sim \mathcal{X}_9^2$
Asymptotic:	$\sqrt{n}(\hat{\delta}_j - 0.10) \sim N(0, 0.10(1-0.10))$

where n_j is the number of observations that fall in decile j and n is the total number of observations. If conditioning on the regime provides no information, the expected percentage falling in each decile is 10% with variance decreasing in the order of n^{-1}. Also, note that the sampling distributions are derived under the assumption that returns are IID, which is not reasonable for financial data. We normalize the trade returns, by subtracting its mean and dividing by its standard deviation in attempt to address this issue. However, this does not eliminate the dependence or heterogeneity in the data observations. (Note, similar assumption is made in [Lo 00] in their statistical tests.) We hope to extend analysis to more general non-IID case in future work.

Kolmogorov-Smirnov test

Another comparison of the conditional and unconditional distributions of returns is provided by the non-parametric Kolmogorov-Smirnov test ([Chakravarti 67, Sheskin 00]). The Kolmogorov-Smirnov test can be used to decide if a sample comes from a population with a specific distribution. In the two sample version, it can be used to compare whether two samples came from the same distribution. In this case, the Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the two samples.

Let $\{Z_{1n}\}_{n=1}^{N_1}$ and $\{Z_{2n}\}_{n=1}^{N_2}$ be two samples that are each independent and identically distributed with cumulative distributions functions $F_1(z)$ and $F_2(z)$, respectively. The empirical cumulative distribution function, $\hat{F}_i(z)$ of each sample is given by,

$$\hat{F}_i(z) = \frac{1}{N_i} \sum_{k=1}^{N_i} I_{Z_i \leq z}, \qquad i = 1, 2$$

where $I_{Z_i \leq z}$ is the indicator function, equal to 1 if $Z_i \leq z$ otherwise 0. The test is defined as,

H_0:	$F_1(z) = F_2(z)$, the samples are drawn from the same distribution.
H_a:	$F_1(z) \neq F_2(z)$, samples have different distributions.
Statistic:	$\gamma_{N_1,N_2} \equiv \left(\frac{N_1 N_2}{N_1 + N_2} \right)^{\frac{1}{2}} \sup_{-\infty < z < \infty} \lvert \hat{F}_1(z) - \hat{F}_2(z) \rvert$
Asymptotic:	$\lim_{\min(N_1,N_2) \to \infty} \mathbb{P}(\gamma_{N_1,N_2} \leq x) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 x^2), \quad x > 0$

An approximate α-level test of the null hypothesis can be performed by computing the statistic and rejecting the null if it exceeds the upper α percentile for the null distribution given by the asymptotic distribution. Thus we can calculate p-values with respect to the asymptotic distribution.

This test assumes the sample trade returns are independent and identically distributed—we normalize the trade returns once again, keeping in mind that this does not eliminate dependence or heterogeneity of the samples. (Similar assumption is made in [Lo 00]).

What these goodness-of-fit tests tell us

Chi-square test results are summarized in Tables 5.3 (in-sample), 5.4 (out-of-sample with lookahead bias) and 5.5 (out-of-sample without look ahead bias) for each quartile. For each state, the percentage of conditional trade returns that falls within each of the 10 unconditional return deciles is tabulated. If conditioning on the state provides no information, the expected percentage falling in each decile is 10%. Asymptotic z-statistics for this null hypothesis are reported in parenthesis, and the $\mathcal{X}^2$ goodness-of-fitness test statistic Q is reported in the last column with the p-value in parenthesis below the statistic. We see that the relative frequency of the conditional returns are significantly different from those of the unconditional returns for both the bullish and bearish state and across all the quartiles. In-sample results have extreme z-scores, indicating that the learned states define two clearly distinct return distributions. This persists in the out-of-sample look-ahead Viterbi, and to a lesser extent in the out-of-sample without look-ahead Viterbi. In all case and across all quartiles the p-value is 0.00% (at two points of accuracy) showing that conditioning on the state alters the trade distribution and hence contains information.

This result is further supported with the Kolmogorov-Smirnov test results, summarized in Tables 5.6 (in-sample), 5.7 (out-of-sample with lookahead bias) and 5.8 (out-of-sample without look ahead bias). The p-values are with respect to the asymptotic distribution of the Kolmogorov-Smirnov test statistic for the equality of conditional and unconditional trade return distribution. In-sample and look-ahead Viterbi results have p-value of 0.00% (at two decimal points of accuracy) across all quartiles. In Viterbi without look ahead the statistical significance declines particularly

for Quartile 3 at 13.43%. The other quartiles still show a statistically significant deviation from unconditional (at the 5% level).

5.2.2 Regime return characteristics

The goodness-of-fit tests provided us with information about whether there is information content in the two states. In this section we will test whether the two states correspond to runs (high frequency bullish periods) and reversals (high frequency bearish periods). We propose to do this by testing the means of the trade returns in each state.

Bullish versus bearish

First we would like to verify whether the bullish regime has a higher mean than the bearish regime. As discussed in Section 5.2 we designate a top-level state to be bullish or bearish based on the in-sample trade return performance after we learn the maximum likelihood model parameters via EM. Using inference we can subsequently generate out-of-sample trade returns for the bullish regime and the bearish regime, which we designate as $\{X_{1n}\}_{n=1}^{N_1}$ and $\{X_{2n}\}_{n=1}^{N_2}$, respectively.

We shall use the t-test for two independent samples [Sheskin 00]. In conducting the test, the two sample means (denoted by $\overline{X}_1$ and $\overline{X}_2$) are used to estimate the values of the means of the populations (μ_1 and μ_2) from which the samples are derived. If the result of the t-test for two independent samples is significant, it indicates that there is a high likelihood that the samples represent population with different mean values. The test is useful when the underlying population variances are unknown, and therefore must be estimated by computing the unbiased sample standard deviation ($\overline{S}$),

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$$

The test is defined as,

Table 5.3: Goodness-of-fit diagnostics for the in-sample conditional state trade return distribution.

		In sample										
		Decile										Q
		1	2	3	4	5	6	7	8	9	10	(p-value)
Largest Quartile	Run	4.05	7.58	9.21	10.57	10.53	8.62	8.67	10.67	13.01	17.08	2088.18
		(-27.86)	(-11.33)	(-3.71)	(2.69)	(2.50)	(-6.46)	(-6.23)	(3.14)	(14.10)	(33.15)	(0.00%)
	Reversal	15.93	12.42	10.77	9.44	9.19	11.65	11.32	9.34	6.99	2.94	2105.33
		(27.81)	(11.36)	(3.63)	(-2.63)	(-3.79)	(7.73)	(6.19)	(-3.08)	(-14.13)	(-33.09)	(0.00%)
2nd Quartile	Run	3.19	7.57	9.82	10.82	10.86	9.41	9.38	10.26	12.20	16.48	2385.06
		(-34.83)	(-12.41)	(-0.91)	(4.17)	(4.41)	(-3.02)	(-3.15)	(1.34)	(11.27)	(33.13)	(0.00%)
	Reversal	16.81	12.43	10.17	9.18	8.94	10.81	10.61	9.75	7.79	3.52	2401.88
		(34.83)	(12.42)	(0.88)	(-4.20)	(-5.44)	(4.12)	(3.10)	(-1.29)	(-11.29)	(-33.13)	(0.00%)
3rd Quartile	Run	4.59	8.54	9.59	10.64	10.93	8.97	9.13	9.79	11.94	15.88	985.99
		(-20.97)	(-5.64)	(-1.57)	(2.48)	(3.60)	(-3.98)	(-3.38)	(-0.82)	(7.53)	(22.76)	(0.00%)
	Reversal	15.42	11.45	10.41	9.36	9.07	11.04	10.87	10.20	8.07	4.13	985.43
		(20.98)	(5.63)	(1.58)	(-2.49)	(-3.61)	(4.02)	(3.36)	(0.78)	(-7.49)	(-22.75)	(0.00%)
Smallest Quartile	Run	4.34	7.71	10.11	11.73	10.55	9.34	9.28	10.19	11.18	15.59	741.57
		(-18.85)	(-7.65)	(0.36)	(5.76)	(1.82)	(-2.21)	(-2.41)	(0.62)	(3.92)	(18.63)	(0.00%)
	Reversal	15.65	12.31	9.89	8.27	9.44	10.68	10.72	9.81	8.82	4.40	741.20
		(18.81)	(7.69)	(-0.35)	(-5.76)	(-1.85)	(2.25)	(2.39)	(-0.62)	(-3.92)	(-18.64)	(0.00%)
All stocks	Run	3.89	7.79	9.67	10.74	10.85	8.99	9.16	10.25	12.26	16.39	6097.87
		(-52.64)	(-19.02)	(-2.82)	(6.39)	(7.28)	(-8.66)	(-7.24)	(2.19)	(19.50)	(55.03)	(0.00%)
	Reversal	16.11	12.21	10.33	9.24	9.09	11.08	10.84	9.74	7.74	3.61	6116.28
		(52.62)	(19.04)	(2.82)	(-6.51)	(-7.80)	(9.30)	(7.24)	(-2.20)	(-19.48)	(-55.04)	(0.00%)

Table 5.4: Goodness-of-fit diagnostics for the out-of-sample look ahead viterbi conditional state trade return distribution.

		Out-of-sample Look-ahead Viterbi										
		Decile										Q
		1	2	3	4	5	6	7	8	9	10	(p-value)
Largest Quartile	Run	4.02	7.33	10.35	10.72	10.56	7.90	8.14	10.33	13.46	17.19	442.83
		(-12.35)	(-5.51)	(0.73)	(1.49)	(1.16)	(-4.33)	(-3.84)	(0.68)	(7.14)	(14.84)	(0.00%)
	Reversal	15.97	12.67	9.60	9.31	9.44	12.07	11.89	9.65	6.56	2.84	441.72
		(12.34)	(5.52)	(-0.83)	(-1.42)	(-1.15)	(4.28)	(3.90)	(-0.72)	(-7.12)	(-14.81)	(0.00%)
2nd Quartile	Run	3.26	7.67	10.73	9.98	11.49	9.05	8.79	10.24	12.42	16.39	475.66
		(-15.29)	(-5.30)	(1.66)	(-0.05)	(3.37)	(-2.16)	(-2.75)	(0.53)	(5.48)	(14.49)	(0.00%)
	Reversal	16.76	12.34	9.27	10.00	8.40	11.07	11.22	9.77	7.58	3.59	479.55
		(15.31)	(5.31)	(-1.66)	(0.01)	(-3.62)	(2.41)	(2.76)	(-0.53)	(-5.48)	(-14.51)	(0.00%)
3rd Quartile	Run	5.15	7.90	10.69	12.29	10.61	8.47	9.12	9.35	11.68	14.73	164.37
		(-8.27)	(-3.58)	(1.17)	(3.91)	(1.04)	(-2.60)	(-1.50)	(-1.11)	(2.87)	(8.08)	(0.00%)
	Reversal	14.86	12.07	9.32	7.72	8.67	12.23	10.89	10.66	8.29	5.27	174.92
		(8.29)	(3.54)	(-1.15)	(-3.89)	(-2.26)	(3.80)	(1.52)	(1.13)	(-2.91)	(-8.06)	(0.00%)
Smallest Quartile	Run	4.74	8.74	10.88	11.08	11.28	9.39	8.89	9.84	11.13	14.03	104.14
		(-7.84)	(-1.88)	(1.32)	(1.62)	(1.91)	(-0.92)	(-1.66)	(-0.25)	(1.69)	(6.01)	(0.00%)
	Reversal	15.29	11.28	9.07	8.92	8.17	11.13	11.18	10.13	8.87	5.96	110.12
		(7.87)	(1.90)	(-1.38)	(-1.60)	(-2.72)	(1.68)	(1.75)	(0.19)	(-1.68)	(-6.01)	(0.00%)
All stocks	Run	4.08	7.76	10.67	10.95	10.87	8.66	8.67	9.99	12.42	15.93	1135.74
		(-22.58)	(-8.56)	(2.57)	(3.62)	(3.33)	(-5.12)	(-5.06)	(-0.02)	(9.21)	(22.62)	(0.00%)
	Reversal	15.92	12.25	9.33	9.05	8.90	11.57	11.33	10.01	7.58	4.06	1151.51
		(22.58)	(8.56)	(-2.55)	(-3.63)	(-4.21)	(5.99)	(5.06)	(0.05)	(-9.22)	(-22.63)	(0.00%)

Table 5.5: Goodness-of-fit diagnostics for the out-of-sample viterbi conditional state trade return distribution.

		Out-of-sample Viterbi										
		Decile										Q
		1	2	3	4	5	6	7	8	9	10	(p-value)
Largest Quartile	Run	9.59	10.48	11.61	10.55	10.08	9.21	8.92	8.67	9.29	11.59	58.94
		(-1.05)	(1.24)	(4.14)	(1.42)	(0.21)	(-2.04)	(-2.77)	(-3.42)	(-1.82)	(4.10)	(0.00%)
	Reversal	10.41	9.51	8.40	9.44	9.89	10.81	11.08	11.32	10.71	8.42	58.96
		(1.06)	(-1.27)	(-4.12)	(-1.44)	(-0.28)	(2.10)	(2.79)	(3.39)	(1.84)	(-4.07)	(0.00%)
2nd Quartile	Run	8.90	11.87	11.63	10.14	9.70	10.76	9.61	8.77	7.98	10.63	99.25
		(-3.07)	(5.22)	(4.55)	(0.40)	(-0.84)	(2.11)	(-1.08)	(-3.43)	(-5.62)	(1.75)	(0.00%)
	Reversal	11.11	8.11	8.37	9.86	8.90	10.62	10.41	11.21	12.04	9.37	106.30
		(3.08)	(-5.26)	(-4.54)	(-0.39)	(-3.06)	(1.73)	(1.13)	(3.36)	(5.68)	(-1.74)	(0.00%)
3rd Quartile	Run	9.44	11.34	11.27	10.30	9.32	9.70	9.85	9.01	8.51	11.27	36.32
		(-1.17)	(2.81)	(2.65)	(0.64)	(-1.43)	(-0.64)	(-0.32)	(-2.07)	(-3.13)	(2.65)	(0.00%)
	Reversal	10.54	8.67	8.72	9.68	10.67	10.32	10.16	11.00	11.48	8.75	36.14
		(1.14)	(-2.79)	(-2.68)	(-0.66)	(1.41)	(0.66)	(0.34)	(2.10)	(3.10)	(-2.63)	(0.00%)
Smallest Quartile	Run	9.40	11.12	11.63	10.51	10.80	9.75	8.73	9.05	8.76	10.26	29.36
		(-1.12)	(2.09)	(3.04)	(0.96)	(1.49)	(-0.47)	(-2.37)	(-1.78)	(-2.31)	(0.48)	(0.06%)
	Reversal	10.60	8.88	8.37	9.49	9.20	10.25	11.27	10.95	11.24	9.74	29.44
		(1.13)	(-2.09)	(-3.04)	(-0.96)	(-1.50)	(0.47)	(2.38)	(1.78)	(2.32)	(-0.48)	(0.05%)
All stocks	Run	9.33	11.21	11.58	10.32	10.17	9.72	9.34	8.70	8.65	10.98	191.10
		(-3.18)	(5.70)	(7.46)	(1.53)	(0.80)	(-1.32)	(-3.13)	(-6.13)	(-6.36)	(4.64)	(0.00%)
	Reversal	10.68	8.79	8.40	9.69	9.65	10.47	10.67	11.30	11.35	9.02	196.50
		(3.19)	(-5.69)	(-7.53)	(-1.48)	(-1.67)	(2.20)	(3.14)	(6.11)	(6.37)	(-4.63)	(0.00%)

Table 5.6: Kolmogorov-Smirnov test of the equality of the in-sample conditional and unconditional trade return distribution.

	In sample			
	Bullish State		Bearish State	
	γ	p-value	γ	p-value
Largest Quartile	8.15	0.00%	8.16	0.00%
2nd Quartile	7.73	0.00%	7.76	0.00%
3rd Quartile	5.34	0.00%	5.36	0.00%
Smallest Quartile	4.83	0.00%	4.83	0.00%
All stocks	12.84	0.00%	12.87	0.00%

Table 5.7: Kolmogorov-Smirnov test of the equality of the out-of-sample look ahead viterbi conditional and unconditional trade return distribution.

	Out-of-sample Look-ahead Viterbi			
	Bullish State		Bearish State	
	γ	p-value	γ	p-value
Largest Quartile	5.49	0.00%	5.43	0.00%
2nd Quartile	4.80	0.00%	4.82	0.00%
3rd Quartile	2.80	0.00%	2.82	0.00%
Smallest Quartile	2.33	0.00%	2.35	0.00%
All stocks	7.64	0.00%	7.64	0.00%

Table 5.8: Kolmogorov-Smirnov test of the equality of the out-of-sample viterbi conditional and unconditional trade return distribution.

	Out-of-sample Viterbi			
	Bullish State		Bearish State	
	γ	p-value	γ	p-value
Largest Quartile	1.39	4.25%	1.38	4.40%
2nd Quartile	2.07	0.04%	2.08	0.04%
3rd Quartile	1.18	12.58%	1.16	13.43%
Smallest Quartile	1.74	0.46%	1.72	0.54%
All stocks	2.97	0.00%	2.96	0.00%

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$

t Statistic:
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{(N_1-1)S_1^2 + (N_2-1)S_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$$
with $(N_1 + N_2 - 2)$ degrees of freedom

We would like to point out that the t-test for two samples is based on the following assumptions:

1. Each sample has been randomly selected from the population it represents (IID samples from each population distribution);

2. The distribution of data in the underlying population from which each of the samples is derived is normal;

3. The third assumption, which is referred to as the homogeneity of variance assumption, states that the variance of the underlying population represented by Sample 1 is equal to the variance of the underlying population represented by Sample 2 (i.e., $\sigma_1 = \sigma_2$).

If any of these assumptions are violated, the reliability of the t-test statistic may be compromised. An alternative is to consider the analogous nonparametric test—which will have relatively fewer or less rigorous assumptions. However, numerous empirical sampling studies have demonstrated that under most conditions a parametric test like the t test for two independent samples is reasonably robust. That is, it provides information about the underlying sampling distribution, in spite of the fact that one or more of the test's assumptions have been violated. In addition, parametric tests, such as the t test for two independent samples, are more powerful than their nonparametric analogs. We risk adjust the trade returns, by dividing by its standard deviation in attempt to address these assumptions. We shall leave it to future work to consider a more general analysis.

Runs and reversals

Finally, we would also like to test if the bullish regime has a positive mean while the bearish regime has a negative mean. This tests whether our learned states captures runs and reversals out-of-sample. We can use the single sample t-test to compare each regime's mean trade return (μ_1 and μ_2 for bullish and bearish, respectively) from 0 [Sheskin 00]. In the one tail test, if the result of the single-sample t test yields a significant positive (negative) value, we can conclude there is a high likelihood the sample is derived from a population with a positive (negative) mean. The test is used when the underlying population standard deviation (σ) is unknown, and therefore must be estimated by computing the unbiased sample standard deviation (S).

The test for the bullish regime is,

$$\begin{aligned} &H_0: && \mu_1 \leq 0 \\ &H_a: && \mu_1 > 0 \\ &\text{t Statistic:} && t = \frac{\overline{X}_1}{\frac{S_1}{\sqrt{N_1}}} \text{ with } (N_1 - 1) \text{ degrees of freedom} \end{aligned}$$

The test for the bearish regime is,

$$\begin{aligned} &H_0: && \mu_2 \geq 0 \\ &H_a: && \mu_2 < 0 \\ &\text{t Statistic:} && t = \frac{\overline{X}_2}{\frac{S_2}{\sqrt{N_2}}} \text{ with } (N_2 - 1) \text{ degrees of freedom} \end{aligned}$$

If the absolute value of the t-statistic is less than α-level of the t-distribution we can reject the null hypothesis. Alternatively, we can calculate the p-value as the probability of obtaining a t value equal to or more extreme than that obtained from the sample data when H_0 is true.

The following two assumptions apply to the single-sample t test:

1. The sample has been randomly selected from the population it represents (IID samples from population distribution);

2. The distribution of data in the underlying population the sample represents is normal.

We risk adjust the trade returns, by dividing by its standard deviation in attempt to address these assumptions. We leave it to future work to explore more general analysis which do not rely on these assumptions.

What the regime mean tells us

Results for regime mean tests are summarized in Tables 5.9 (in-sample), 5.10 (out-of-sample with lookahead bias) and 5.11 (out-of-sample without look ahead bias) for each quartile. P-values can be calculated as the probability of obtaining a test statistic value equal to or more extreme than that obtained from the sample data when the null hypothesis is true. We find that for in-sample and out-of-sample with look-ahead bias, the results are statistically significant at 0.00% (accurate to two decimal points). This is the case for all three tests, namely 1) bullish mean is greater than bearish mean, 2) bullish mean is positive, 3) bearish mean is negative. For out-of-sample Viterbi (without look-ahead bias) we see weaker results, but statistically significant for the top three quartiles at the 5% level. For the 4th quartile, while the positive bullish mean test is significant at 4.76%, the negative bearish mean is much weaker at 36.97%.

We can conclude that the price and volume DBN model is indeed able to learn two different states that represent intraday bullish and bearish properties—with results stronger for larger volume stocks.

Table 5.9: Regime mean t-tests for in-sample conditional state trade return distribution.

	In sample									
			$H_0 : \overline{X}_1 \leq 0$				$H_0 : \overline{X}_2 \geq 0$		$H_0 : \overline{X}_1 \leq \overline{X}_2$	
	$\overline{X}_1$	N_1	t-stat	p-value	$\overline{X}_2$	N_2	t-stat	p-value	t-stat	p-value
Largest Quartile	0.3001	19,745	41.96	0.00%	-0.3579	19,769	-50.22	0.00%	65.17	0.00%
2nd Quartile	0.3333	23,541	51.04	0.00%	-0.3099	23,524	-47.41	0.00%	69.61	0.00%
3rd Quartile	0.2701	13,497	31.31	0.00%	-0.2780	13,499	-32.16	0.00%	44.88	0.00%
Smallest Quartile	0.2852	9,993	28.35	0.00%	-0.2469	9,975	-24.59	0.00%	37.44	0.00%
All stocks	0.3035	66,776	78.14	0.00%	-0.3082	66,767	-79.37	0.00%	111.37	0.00%

Table 5.10: Regime mean t-tests for out-of-sample look-ahead Viterbi conditional state trade return distribution.

Out-of-sample Look-ahead Viterbi										
			$H_0 : \overline{X}_1 \leq 0$				$H_0 : \overline{X}_2 \geq 0$		$H_0 : \overline{X}_1 \leq \overline{X}_2$	
	$\overline{X}_1$	N_1	t-stat	p-value	$\overline{X}_2$	N_2	t-stat	p-value	t-stat	p-value
Largest Quartile	0.2970	3,834	18.30	0.00%	-0.3642	3,844	-22.54	0.00%	28.87	0.00%
2nd Quartile	0.3173	4,631	21.55	0.00%	-0.3083	4,618	-20.90	0.00%	30.01	0.00%
3rd Quartile	0.2224	2,620	11.37	0.00%	-0.2348	2,617	-11.98	0.00%	16.51	0.00%
Smallest Quartile	0.2247	2,003	9.97	0.00%	-0.2041	1,995	-9.09	0.00%	13.48	0.00%
All stocks	0.2782	13,088	31.69	0.00%	-0.2941	13,074	-33.49	0.00%	46.09	0.00%

Table 5.11: Regime mean t-tests for out-of-sample Viterbi conditional state trade return distribution.

	Out-of-sample Viterbi								
			$H_0 : \overline{X}_1 \leq 0$				$H_0 : \overline{X}_2 \geq 0$		$H_0 : \overline{X}_1 \leq \overline{X}_2$	
	$\overline{X}_1$	N_1	t-stat	p-value	$\overline{X}_2$	N_2	t-stat	p-value	t-stat	p-value
Largest Quartile	0.0437	5,952	3.36	0.04%	-0.0759	5,964	-5.85	0.00%	6.51	0.00%
2nd Quartile	0.0372	6,990	3.11	0.10%	-0.0274	6,977	-2.29	1.11%	3.81	0.01%
3rd Quartile	0.0333	3,950	2.09	1.84%	-0.0360	3,945	-2.26	1.20%	3.07	0.11%
Smallest Quartile	0.0298	3,139	1.67	4.76%	-0.0060	3,131	-0.33	36.97%	1.42	7.85%
All stocks	0.0372	20,031	5.26	0.00%	-0.0402	20,017	-5.68	0.00%	7.73	0.00%

5.2.3 Trading strategy and results

Now that we have statistical evidence that the model has indeed learned two states, one indicating a run and the other a reversal, we shall evaluate the trading profits of simple trading system based on the model. We shall position ourselves long one unit when top level state Q_t^1 switches to bullish (i.e. run) and short one unit when Q_t^1 switches to bearish (i.e. reversal). Again, we ensure no look-ahead bias is present by placing the trade one tick after the observation is complete, since this is the point at which the zig-zag leg is identifiable, at which point we can extract the observation feature.

Table 5.12 shows the results obtained for the four quartiles. The buy and hold results are compared to the look-ahead Viterbi and Viterbi without look-ahead strategy. We note that the developed strategy based on the DBN model has very low correlation to the buy and hold strategy, indicating that it can perform in both bullish and bearish periods. Look-ahead Viterbi has exceptional performance. Though this strategy cannot be traded upon, it is indicative of the significance of the learned price and volume patterns out-of-sample. Without look-ahead bias the performance drops, but it is still significant compared to buy and hold. In general, high volume stocks perform better than lower volume stocks. This was seen with the statistical tests as well, with stronger significant results for the larger quartiles.

Performance without the look-ahead bias drops mainly due to instability of the regime state, which results in a higher number of trades (see Figure 5.13 versus 5.14 for a visual example). In comparison, the look ahead bias has smoothed switching since we estimate the hidden state using information from the past, present and the future. It remains an interesting course of investigation to see how we can enhance the DBN inference framework to minimize instable regime switches.

Figures 5.19, 5.20, 5.21 and 5.22 show the accrued profit and loss (P&L) of $1 invested in the strategy using Viterbi inference (without look-

Table 5.12: Summary results of trade performance using DBN price and volume model.

	Vol ($ Mil)	Buy & Hold		Look-ahead Viterbi			Viterbi		
	April	Ann. Return	Return	Ann. Return	Corr to B&H	Trades/Day	Ann. Return	Corr to B&H	Trades/Day
	April	Return	Return	Corr to B&H	Trades/Day	Return	Corr to B&H	Trades/Day	
Largest Quartile	5.63	-64%	758%	0.06	36.56	189%	-0.02	56.74	
2nd Quartile	1.66	-13%	649%	0.02	44.42	43%	-0.05	66.96	
3rd Quartile	0.89	0%	302%	-0.01	24.94	36%	0.00	37.60	
Smallest Quartile	0.43	11%	296%	-0.14	19.04	25%	-0.11	29.86	
All stocks	2.15	-16%	501%	-0.02	31.24	73%	-0.04	47.79	

ahead bias) compared to the buy and hold approach for each quartile. Comparing to the simple buy and hold strategy, we can clearly see that our model has the capacity to capture significant profit from technical intraday volatility.

When we place the trade after the one-tick lag, we assume we can fulfil the trade at that price—that is, we are not accounting for trading costs. Although this may not be reasonable for the general public, it may be ok for a market maker that is generally expected to earn the spread (not taking into account large block trades which would necessarily span several transactions and have significant market impact if traded all at once). In fact, as discussed in Section 4.2, the model uses bid-ask resistance points to capture technical volatility arising due to spreads. Thus we have been able to generate a predictive model of ultra-high frequency moves—useful for a market maker or possibly as an input for an optimal execution engine. Also, passive public traders who use limit orders could benefit from this model as well, and an interesting extension would be to combine order book information to provide precise limit order trade signals. Another interesting future course of study could attempt to learn from features which characterize other sources of technical intraday volatility apart from bid-ask spreads such as market impact, price discovery and momentum effects.

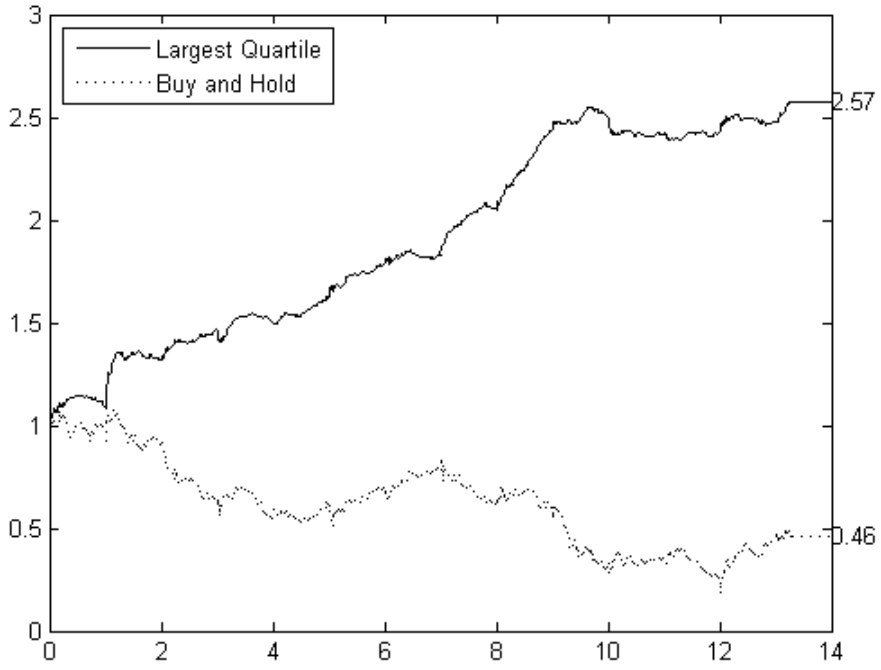


Figure 5.19: Value of \$1 invested in G_1 stocks.

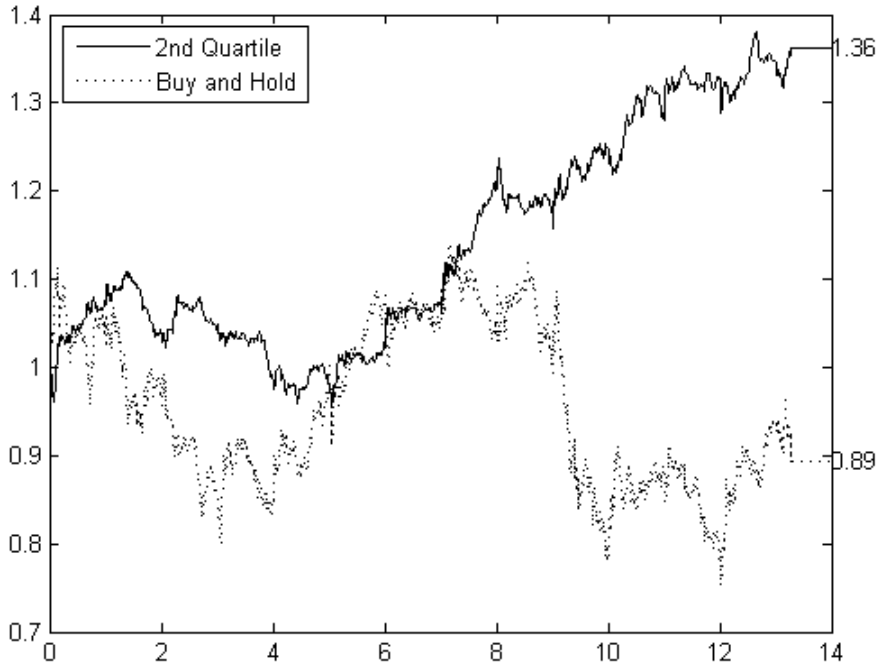


Figure 5.20: Value of \$1 invested in G_2 stocks.

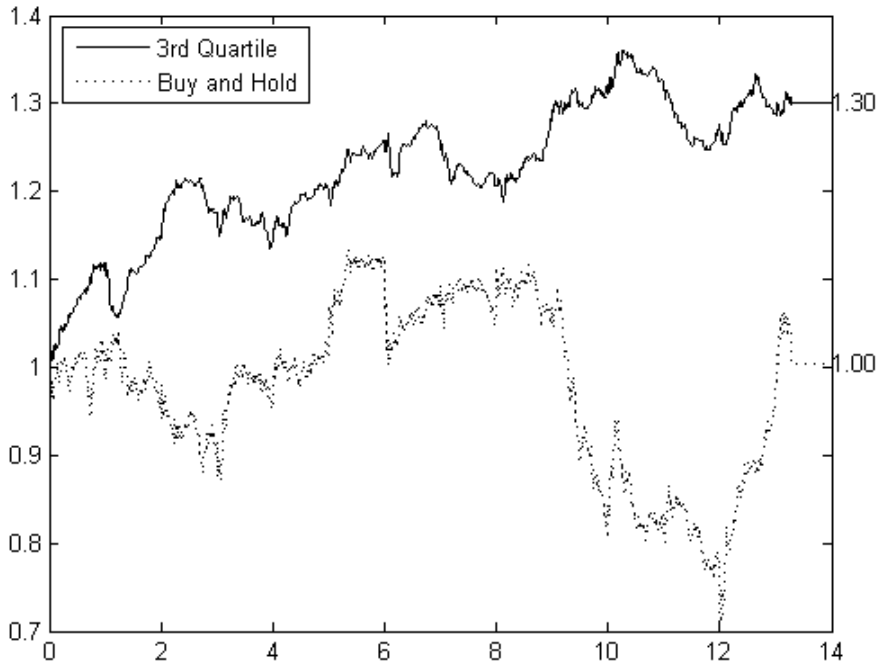


Figure 5.21: Value of \$1 invested in G_3 stocks.

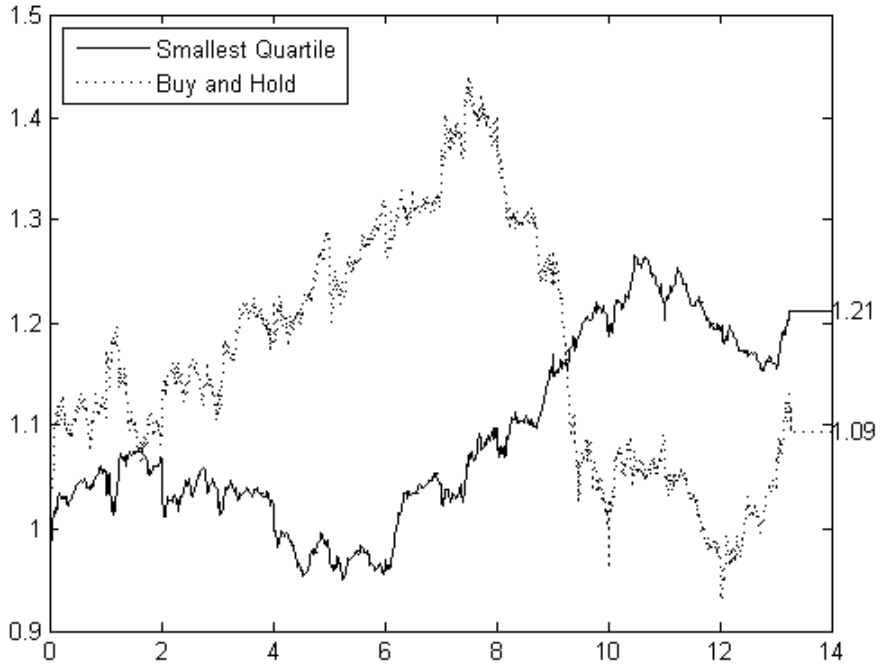


Figure 5.22: Value of \$1 invested in G_4 stocks.

Chapter 6

Conclusion

This chapter concludes this thesis. We begin by summarizing the key points of the work, what guided us in this direction and what we learned from our model and experiments. Finally, we review our contributions and suggest future avenues of investigation.

6.1 Summary of key ideas

In this thesis, we have proposed a new approach for modeling and working with technical analysis. The idea of technical analysis has been thwarted from the beginning in academic circles. This is mostly due to the Efficient Market Hypothesis, which had significant empirical support early on. The idea of the efficiency of markets is that price changes are due to fundamental value changes and that these changes cannot be anticipated a priori. In its weakest form, it implies that the market follows a random walk, and therefore past price information cannot be used to forecast future price. However, recent research has started to question the Efficient Market Hypothesis—both on a theoretical basis and with empirical results. In particular, the return distribution of prices seem to exhibit specific characteristics, such as volatility clustering and excessive leptokurtosis, that is difficult to explain with a random walk model.

Technical analysis also advocates that price (and volume) information reflects all known information, but in addition, it believes that human behaviour tends to repeat itself, forming trends and patterns in market prices. Thus by diligently studying price and volume behaviour chartists identify market sentiment and attempt to forecast price direction. However, technical analysis remains mostly an "art" with much room left for subjective interpretation.

This thesis promotes a new toolset to work with technical analysis: dynamic Bayesian networks (DBNs). DBNs are a statistical modeling and learning framework that have had successful applications in speech recognition, bio-sequencing, visual interpretation, and other areas. It subsumes several popular paradigms including mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models. Technical analysis may well be the next frontier for such methods. By providing a coherent probabilistic framework (in a Bayesian sense), it can be used for both learning technical rules and inferring the hidden state of the system.

6.2 Results and contributions

We present a model based on dynamic Bayesian networks to learn price and volume patterns in high-frequency markets. It is the first study to apply DBNs to financial data, and one of the few that investigate technical analysis in high-frequency markets.

We carefully define high-frequency features based on zig-zags that characterize price moves with corresponding volume moves in a trending or non-trending environment. This feature set encodes price and volume technical analysis tenets. We design a hierarchical hidden Markov model (HHMM) that distinguishes between runs and reversals by learning distinct distributions over the feature space. The HHMM is designed to prevent overfitting, however complex enough to learn significant patterns in the feature space. We subsequently transform the HHMM into a DBN for efficient learning and inferences purposes. One of the key differences

from other DBN applications, such as speech recognition, is that we do not label our hidden states in the training phase. In financial series, it is not clear whether a particular point is in a run or reversal since it depends on the time scale of concern. As such, training financial DBNs poses additional challenges.

We investigate TSE60 stocks and found that we are able to learn two regimes that successfully captured runs and reversals out-of-sample. We present statistical tests, verifying that two regimes captured unique return trade distributions both in- and out-of-sample. Moreover, we showed that the bullish regime resulted is a positive trade mean, while the bearish regime resulted in a negative trade mean in a statistically significant way out-of-sample. In general, we found higher volume stocks lend themselves more favourably to technical analysis, indicating that price and volume behaviour is more persistent in larger aggregates. Finally we illustrate the results of a trading system based on the model, which yielded substantial positive results compared to the buy and hold strategy. We have not considered the impact of trading costs—since our window frequency was at the transaction level, we expect the model to be more useful for market makers, optimal execution kernels and limit order traders.

Our results validate the presence of technical predictability; we conclude that there is information content available in price and volume data in transaction data that can be used to predict intraday trends. Moreover, our methods suggest that dynamic Bayesian networks can be used to improve upon traditional technical analysis approaches.

6.3 Applications and future work

This work just scratches the surface of using DBNs for financial analysis. Below we list a few possible way we can extend our work,

Parameter learning: Improve learning for the global maximum likelihood parameters, for example by considering deterministic anneal-

ing or incorporating a genetic evolutionary approach.

Structure learning: Consider learning the structure (topography) of the DBN [Friedman 98], instead of designing it manually.

Chain graphs: DBN semantics encode conditional probabilities in the form of directed arcs; we can also consider including undirected arcs representing correlation (known as chain graphs).

Order flow: Incorporate order flow data in the feature space. [Fama 70] shows that order flow data significantly improves technical analysis signals.

Time scales: Consider alternate time scales (i.e., hours, days, etc.), by using features that correspond to that time-scale. In addition, consider modeling a multi-resolution DBN, as is done for language and speech recognition. (Refer to [Filali 06] for discussion on multi-dynamic Bayesian networks).

Cross-sectional analysis: Markets exhibit a high degree of interdependency; consider linking individual asset models to form a large integrated multi asset DBN.

Continuous features: Use continuous features instead of just discrete, allowing for more informative features.

Fundamental value: Consider fundamental data in the feature space. In particular, consider designing a DBN that captures equilibrium dynamics and consequently identifies mispricings from equilibrium.

Dynamic Bayesian networks, and more generally, graphical probabilistic networks is an encompassing and general framework to work within. We hope that our considerations illustrate that they can be powerful tools for market analysis, and hope to explore more possibilities in future research.

Bibliography

The numbers at the end of each entry list pages where the reference was cited. In the electronic version, they are clickable links to the pages.

[Arnborg 87] Stefan Arnborg, Derek G. Corneil & Andrzej Proskurowski. *Complexity of Finding Embeddings in a k-Tree*. SIAM Journal on Algebraic and Discrete Methods, vol. 8, no. 2, pages 277–284, 1987. 40

[Austin 04] Mark Austin, Graham Bates, Michael Dempster & Stacy Williams. *Adaptive systems for foreign exchange trading*. Eclectic, vol. 18, pages 21–26, 2004. 3

[Barber 97] Colin Barber, Donald Robertson & Andrew Scott. *Property and Inflation: The Hedging Characteristics of U.K. Commercial Property, 1967-1994*. The Journal of Real Estate Finance and Economics, vol. 15, no. 1, pages 59–76, July 1997. 49

[Bengtsson 99] Henrik Bengtsson. Bayesian networks - a self-contained introduction with implementation remarks. Master's thesis, Mathematical Statistics, Lund Institute of Technology, September 1999. 22

[Bessembinder 98] Hendrik Bessembinder & Kalok Chan. *Market Efficiency and the Returns to Technical Analysis*. Financial Management, vol. 27, no. 2, 1998. 3, 18

[Bilmes 00] Jeff Bilmes. *Dynamic Bayesian Multinets*. In UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, pages 38–45. Morgan Kaufmann Publishers Inc., 2000. 23

[Bilmes 03] Jeffrey A. Bilmes. *Graphical models and automatic speech recognition*. In Mathematical Foundations of Speech and Language Processing. Springer-Verlag, 2003. 28

[Brock 92] William Brock, Josef Lakonishok & Blake LeBaron. *Simple Technical Trading Rules and the Stochastic Properties of Stock Returns*. Journal of Finance, vol. 47, no. 5, pages 1731–64, December 1992. 3, 18

[Canegrati 08] Emanuele Canegrati. *A Non-Random Walk down Canary Wharf*. Rapport technique, University Library of Munich, Germany, August 2008. 3, 18

[Cerra 03] Valerie Cerra & Sweta Chaman Saxena. *Did Output Recover from the Asian Crisis?* IMF Working Papers 03/48, International Monetary Fund, April 2003. 50

[Chakravarti 67] I. M. Chakravarti, R. G. Laha & J. Roy. Handbook of methods of applied statistics, volume I. John Wiley and Sons, 1967. 99

[Chauvet 05] Marcelle Chauvet & James D. Hamilton. *Dating Business Cycle Turning Points*. NBER Working Papers 11422, National Bureau of Economic Research, Inc, June 2005. 50

[Cont 01] Rama Cont. *Empirical properties of asset returns: stylized facts and statistical issues*. Quantitative Finance, vol. 1, pages 223–236, 2001. 2, 13

[Cont 03] Rama Cont & Peter Tankov. Financial modelling with jump processes. Chapman and Hall, 2003. 13

[Cutler 89] David M. Cutler, James M. Poterba & Lawrence H. Summers. *What Moves Stock Prices?* Journal of Portfolio Management, vol. 15, pages 4–12, 1989. 10

[Dacorogna 01] Michel M. Dacorogna, Ramazan Gençay, Ulrich A. Müller, Richard B. Olsen & Olivier V. Pictet. An introduction to high-frequency finance. Academic Press, 2001. 2, 10, 15

[Daniel 06] Gilles Daniel. *Asynchronous Simulations of a Limit Order Book*. PhD thesis, University of Manchester, 2006. 11, 13

[Davig 04] Troy Davig. *Regime-switching debt and taxation*. Journal of Monetary Economics, vol. 51, no. 4, pages 837–859, May 2004. 50

[Dempster 77] A. P. Dempster, N. M. Laird & D. B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pages 1–38, 1977. 48

[Driffill 98] John Driffill & Martin Sola. *Intrinsic bubbles and regime-switching*. Journal of Monetary Economics, vol. 42, no. 2, pages 357–373, July 1998. 49

[Elidan 08] Gal Elidan & Stephen Gould. *Learning Bounded Treewidth Bayesian Networks.* In Daphne Koller, Dale Schuurmans, Yoshua Bengio & Léon Bottou, editeurs, NIPS, pages 417–424. MIT Press, 2008. 39

[Fama 69] Eugene F. Fama, Lawrence Fisher, Michael C. Jensen & Richard Roll. *The Adjustment of Stock Prices to New Information*. International Economic Review, vol. 10, pages 1–21, 1969. 2, 9

[Fama 70] Eugene F Fama. *Efficient Capital Markets: A Review of Theory and Empirical Work*. Journal of Finance, vol. 25, no. 2, pages 383–417, May 1970. 1, 2, 8, 9, 122

[Filali 06] Karim Filali & Jeff Bilmes. *Multi-dynamic Bayesian Networks*. In nips, December 2006. 122

[Fine 98] Shai Fine & Yoram Singer. *The Hierarchical Hidden Markov Model: Analysis and Applications*. In Machine Learning, pages 41–62, 1998. 32, 38

[Friedman 98] N. Friedman, K. Murphy & S. Russell. *Learning the Structure of Dynamic Probabilistic Networks*. In UAI, 1998. 30, 122

[Garcia 99] Marcio G P Garcia & Pierre Perron. *Unit Roots in the Presence of Abrupt Governmental Interventions with an Application to Brazilian Data*. Journal of Applied Econometrics, vol. 14, no. 1, pages 27–56, Jan.-Feb. 1999. 49

[Goldfeld 73] Stephen M. Goldfeld & Richard E. Quandt. *A Markov model for switching regressions*. Journal of Econometrics, vol. 1, no. 1, pages 3–15, March 1973. 49

[Grossman 80] Sanford J Grossman & Joseph E Stiglitz. *On the Impossibility of Informationally Efficient Markets*. American Economic Review, vol. 70, no. 3, pages 393–408, June 1980. 10

[Hamilton 89] James D Hamilton. *A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle*. Econometrica, vol. 57, no. 2, pages 357–84, March 1989. 49

[Hamilton 05] James D. Hamilton. *What's Real About the Business Cycle?* NBER Working Papers 11161, National Bureau of Economic Research, Inc, February 2005. 50

[Ippolito 89] Richard A Ippolito. *Efficiency with Costly Information: A Study of Mutual Fund Performance, 1965-1984*. The Quarterly Journal of Economics, vol. 104, no. 1, pages 1–23, February 1989. 10

[Jeanne 00] Olivier Jeanne & Paul Masson. *Currency crises, sunspots and Markov-switching regimes*. Journal of International Economics, vol. 50, no. 2, pages 327–350, April 2000. 50

[Jensen 67] Michael C. Jensen. *The Performance of Mutual Funds in the Period 1945-1964*. Journal of Finance, vol. 23, no. 2, pages 389–416, 1967. 2, 10

[Jordan 98] Michael I. Jordan. Learning in graphical models (adaptive computation and machine learning). The MIT Press, 1998. 4

[Kahneman 79] Daniel Kahneman & Amos Tversky. *Prospect Theory: An Analysis of Decision under Risk*. Econometrica, vol. 47, no. 2, pages 263–91, March 1979. 2, 18

[Karpoff 87] Jonathan M. Karpoff. *The Relation between Price Changes and Trading Volume: A Survey*. Journal of Financial and Quantitative Analysis, vol. 22, no. 01, pages 109–126, March 1987. 3

[Kjaerulff 90] Uffe Kjaerulff. *Triangulation of graphs : algorithms giving small total state space*. Rapport technique, Department of Mathematics and Computer Science, March 1990. 41

[Lam 90] Pok-sang Lam. *The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series*. Journal of Monetary Economics, vol. 26, no. 3, pages 409–432, December 1990. 49

[LeBaron 96] Blake LeBaron. *Technical Trading Rule Profitability and Foreign Exchange Intervention*. Rapport technique 5505, National Bureau of Economic Research, Inc, March 1996. 2

[Liu 99] Yanhui Liu, Parameswaran Gopikrishnan, Cizeau, Meyer, Peng & Eugene H. Stanley. *Statistical properties of the volatility of price fluctuations*. Physical Review E, vol. 60, no. 2, 1999. 2, 13

[Lo 00] Andrew W. Lo, Harry Mamaysky & Jiang Wang. *Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation*. Journal of Finance, vol. 55, no. 4, pages 1705–1770, 08 2000. 2, 17, 18, 54, 98, 100

[Malkiel 03] Burton Malkiel. A random walk down wall street. W.W. Norton and Co., 2003. 1

[Mamon 07] Rogemar S. Mamon & Robert J. Elliott. Hidden markov models in finance. Springer, 2007. 4

[Milgrom 82] Paul Milgrom & Nancy Stokey. *Information, trade and common knowledge*. Journal of Economic Theory, vol. 26, no. 1, pages 17–27, February 1982. 10

[Murphy 99] John Murphy. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications. Prentice Hall Press, 1999. 1, 17, 54

[Murphy 02] K. Murphy. *Dynamic Bayesian Networks: Representationm, Inference, and Learning*. PhD thesis, University of California, Berkeley, 2002. xv, xvi, 4, 24, 25, 26, 28, 30, 32, 33, 38, 39, 41, 42, 43, 44, 45, 48, 51, 52, 64

[Neapolitan 03] Richard E. Neapolitan. Learning bayesian networks. Prentice Hall, 2003. 27

[Nelson 01] Charles R Nelson, Jeremy Piger & Eric Zivot. *Markov Regime Switching and Unit-Root Tests*. Journal of Business & Economic Statistics, vol. 19, no. 4, pages 404–15, October 2001. 50

[Nevmyvaka 06] Yuriy Nevmyvaka, Yi Feng & Michael Kearns. *Reinforcement learning for optimized trade execution*. In ICML '06: Proceedings of the 23rd international conference on Machine learning, pages 673–680. ACM Press, 2006. 3

[Nodelman 07] Uri D. Nodelman. *Continuous Time Bayesian Networks*. PhD thesis, Stanford University, 2007. 28

[Ord 08] Tim Ord. The secret science of price and volume. Wiley, 2008. xvi, 17, 54, 55, 57, 60, 69

[Pearl 88] Judea Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann, 1988. 22, 24, 26, 41

[Pesaran 06] Hashem Pesaran, Davide Pettenuzzo & Allan Timmermann. *Forecasting Time Series Subject to Multiple Structural Breaks*. Review of Economic Studies, vol. 73, no. 4, pages 1057–1084, October 2006. 52

[Raymond 97] Jennie E Raymond & Robert W Rich. *Oil and the Macroeconomy: A Markov State-Switching Approach*. Journal of Money, Credit and Banking, vol. 29, no. 2, pages 193–213, May 1997. 49

[Robertson 84] Neil Robertson & Paul D. Seymour. *Graph minors. III. Planar tree-width*. J. Comb. Theory, Ser. B, vol. 36, no. 1, pages 49–64, 1984. 39

[Ruge-Murcia 95] Francisco J Ruge-Murcia. *Credibility and Changes in Policy Regime*. Journal of Political Economy, vol. 103, no. 1, pages 176–208, February 1995. 49

[Salakhutdinov 03] Ruslan Salakhutdinov, Sam Roweis & Zoubin Ghahramani. *Optimization with em and expectation-conjugate-gradient*. In Proceedings, Intl. Conf. on Machine Learning (ICML, pages 672–679, 2003. 47

[Salojärvi 05] Jarkko Salojärvi, Kai Puolamäki & Samuel Kaski. *Expectation maximization algorithms for conditional likelihoods*. In ICML '05: Proceedings of the 22nd international conference on Machine learning, pages 752–759. ACM, 2005. 47

[Samuelson 65] Paul Samuelson. *Proof that properly anticipated prices fluctuate randomly*. Industrial Management Review, vol. 6, page 41Ű49, 1965. 2, 9

[Schwartz 04] Robert Schwartz. Equity markets in action: The fundamentals of liquidity, market structure & trading. Wiley, 2004. 8, 56

[Schwert 94] G.W. Schwert. *Mark-up Pricing in Mergers and Acquisitions*. Papers 95-01, Rochester, Business - Financial Research and Policy Studies, 1994. 49

[Sharpe 98] William Sharpe. Investments (6th edition). Prentice Hall, 1998. 8

[Sheskin 00] David J. Sheskin. Handbook of parametric and nonparametric statistical procedures. Chapman and Hall, 2 edition, 2000. 98, 99, 101, 107

[Shiller 81] Robert J Shiller. *Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?* American Economic Review, vol. 71, no. 3, pages 421–36, June 1981. 3, 10

[Shleifer 97] Andrei Shleifer & Robert W Vishny. *The Limits of Arbitrage*. Journal of Finance, vol. 52, no. 1, pages 35–55, March 1997. 9, 10

[Sims 06] Christopher A. Sims & Tao Zha. *Were There Regime Switches in U.S. Monetary Policy?* American Economic Review, vol. 96, no. 1, pages 54–81, March 2006. 50

[Snedecor 89] George W. Snedecor & William G. Cochran. Statistical methods. Iowa State University Press, 8 edition, 1989. 94

[Storer 95] Paul Storer & Marc A. Van Audenrode. *Unemployment Insurance Take-Up Rates in Canada: Facts, Determinants, and Implications*. Canadian Journal of Economics, vol. 28, no. 4a, pages 822–35, November 1995. 49

[Town 92] R J Town. *Merger Waves and the Structure of Merger and Acquisition Time-Series*. Journal of Applied Econo-

metrics, vol. 7, no. S, pages S83–100, Suppl. De 1992. 49

[Zhang 98] Guoqiang Zhang, B. Eddy Patuwo & Michael Y. Hu. *Forecasting with artificial neural networks: The state of the art*. International Journal of Forecasting, vol. 14, no. 1, pages 35–62, March 1998. 3

[Zhang 99] Nevin Zhang & David Poole. *On the role of context-specific independence in probabilistic reasoning*. In Proc. IJCAI, 1999. 40

[Zweig 96] G. Zweig. A forward-backward algorithm for inference in bayesian networks and an empirical comparison with hmms. Master's thesis, Department of Computer Science, U.C. Berkeley, 1996. 44