Volatility and correlation: Modeling and forecasting using Support Vector Machines

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für Wirtschaftswissenschaften der
Universität Fridericiana zu Karlsruhe (TH)

genehmigte

DISSERTATION

von

Dipl.-Wi.-Ing. Amir Safari

Tag der mündlichen Prüfung: 14.7.2008
Referent: Prof. Dr. Detlef Seese
Korreferent: Prof. Dr. Svetlozar T. Rachev

Karlsruhe 2008

Acknowledgment

The assistance of many people has effectively helped to develop the present dissertation. I wish to convey my deepest gratitude to my supervisor, Prof. Dr. Detlef Seese, Institute AIFB, University of Karlsruhe, for providing so much of the supports, thoughtful feedbacks, suggestions, comments and collaborative spirit that was necessary in completing this project. His supports have not been solely limited to research. My gratitude also goes to Prof. Dr. Svetlozar T. Rachev, Institute SMW, University of Karlsruhe, and Department of Statistics and Applied Probability, University of California, at Santa Barbara, thesis's committee member, for his valuable suggestions and guidance. I would like to thank Prof. Dr. Wolfgang Gaul, Institute ETU, University of Karlsruhe, thesis's committee member. This is my great pleasure and proud to have agreement of Prof. Rachev and Prof. Gaul to be as thesis's board member and to have a joint published work with Prof. Rachev. I would like also to thank Prof. Dr. Clemens Puppe, the chair of my doctoral examination. I am also grateful to everybody at the Institute AIFB, University of Karlsruhe, in particular to Joachim Melcher, Tobias Dietrich, Andreas Mitschele and Markus Kress for their kind scientific and technical helps. I really appreciate comments and discussions of Dr. Cornelia Schön, Institute WIOR, University of Karlsruhe, Dr. Stephan Chalup, School of Electrical Engineering and Computer Science, University of Newcastle, Dr. Wei Sun, Institute SMW, University of Karlsruhe, and Erika Wiebe, Institute ETU, University of Karlsruhe in different subjects.

This dissertation was financially supported by Ministry of Science, Research and Technology of Iran in a scholarship framework. I am grateful for this support of the Ministry.

Summary

Realized Volatility (RV) is an estimate of daily volatility from a sample aggregation of squared or absolute values of intraday high frequency returns. Realized correlation is also conditionally constructed based on the volatility. The model-free RV is time-varying, observable and free of the distributional and model assumptions. Sampling as often as possible would theoretically produce consistent estimates of the true variance in the limit. However, the most important practical challenge to the realized volatility theory, beside the lack of continuous-time recorded prices, is market microstructure noise. The noise at high frequency levels does not allow realized volatility estimator to converge into its integrated volatility and it causes a considerable bias and inefficiency. To overcome the problem of noise, some approaches have been introduced. In addition, realized volatility literature usually assumes a Gaussian noise. Meanwhile, the noise in real world financial markets does not follow a Gaussian process. The present dissertation discusses volatility and correlation estimators and introduces new volatility and correlation estimators which converge faster and are consistent under Gaussian microstructure noise. They show lower error under non-Gaussian noise relative to Gaussian noise. Empirically, the new estimators reveal better some stylized facts of financial markets in analogous to their counterparts. Importantly, the proposed correlations exhibit negative asymmetry or heavy tail in dependence structure of stock market comovements consistent with other approaches such as Archimedean copulas.

It has been well documented that financial time series have in common some regularities. Nonstationarity and noise of stock market returns are among those characters. Considerable insight into the volatility dynamics is gained by looking at the data at several different scales. An efficient way of representing a time series with such complex dynamics is given by multiscale analysis. With the help of wavelet functions, the multiscale analysis is able to decompose a time series into several scales while preserving the time dimension. According to the multiscale theory, a function is described by means of a low resolution function plus a series of details from low to high resolution. Exploiting the multiscale analysis, a successful strategy can be designed to improve upon a function estimation performance. The proposed strategy consists of a combination of the multiscale analysis and Support Vector Regression (SVR) as an estimation tool to fit a nonparametric volatility model, namely CHARN model. In fact, the combination is expected to yield higher performance of learning the structures behind each separated scale. According to the strategy, the combination is designed so that the original time series of returns is decomposed into several scales or resolutions, each scaled time series is approximated separately by a SVR machine and then the fitted functions on different scales are summed up to reach a general function approximation for the original time series. The experiments show that the SVR machine outperforms other function approximation algorithms such as some kind of neural network. More important, the experiments show that the strategy yields promising results. The multiscale strategy yields better performance of estimation than the usual single scale strategy.

Contents

List of Figures

Chapter 1

Introduction

Volatility of and correlation between return series in financial markets are studied. In fact, an important topic in financial econometrics that has received significant attention in the finance literature is the modeling of second moment of asset returns. Importance of volatility measuring, modeling and forecasting for finance is quite clear. Perhaps volatility is one of the most important measures for determining behavior of a financial market and of any time series. The interest in stock market volatility has grown significantly, specially in recent years, with the general observation that stock markets around the world are becoming increasingly integrated and more volatile. This interest in stock market volatility has extended beyond the experience of developed markets, and has now focused on emerging markets. Even how the volatilities in financial markets are correlated is an important subject of interest for many researches. Application of financial market volatility, in turn, is indispensable in most financial problems including asset and derivative pricing, asset allocation, risk management, and hedging. Volatility and correlation forecasts are, in fact, fundamental statistical parameters for many financial models. As volatility is not a directly observable variable, large research areas have emerged that attempt to best address this problem. By far the most popular approach is to obtain volatility estimates using the statistical models that have been proposed in the ARCH and Stochastic Volatility literature. Another method of extracting information about volatility is to formulate and apply economic models that link the information contained in options to the volatility of the underlying asset. All these approaches have in common that the resulting volatility measures are only valid under the specific assumptions of the models used and it is generally uncertain which or whether any of these specifications provide a good description of actual volatility. Most of the models fail to explain stylized facts. True enough, they are only models, and as such perhaps only means for successful data fitting, but they are missing on something crucial.

A model-free measure of volatility is the sample variance of returns. Using daily data, for instance, it may be freely estimated using returns spanning over any number of days and, as such, one can construct a time series of model-free variance estimates. When one chooses the observation frequency of this series, an important trade-off has to be made, however. When the variances are calculated using a large number of observations (e.g. the returns over an entire year), many interesting properties of volatility tend to disappear (the volatility clustering and leverage effect, for instance). On the other hand, if only very few observations are used, the measures are subject to great error. At the extreme, only one return observation is used for each daily variance estimate.

The approach taken is to calculate the daily volatility from the sample sum of squared or of absolute values of intraday returns, the realized volatility (RV). Implied volatility model for asset pricing is based on the assumption of constant volatility, but it is now widely accepted that volatility is time varying. Specifically, the high frequency record of some stock indices or individual equities over a period is used to obtain a time series of daily realized volatilities. These are free of the assumptions necessary when the statistical or economic approaches are employed and we have an (almost) continuous record of returns for each day.

In theory, realized volatility computed from the highest possible frequency data should provide both consistent and efficient estimator for integrated volatility. However, the most important challenge to the realized volatility theory, beside the lack of continuous-time recorded prices, is market microstructure noise. The noise at high frequency levels does not allow realized volatility estimator to converge into its integrated volatility.

To overcome the problem, sparse sampling or applying lower frequencies have been recommended in literature to reduce the market microstructure effects. Obviously this recommendation is in contrast to the realized volatility theory under which the realized volatility estimator converges as the frequency continuously increases. Therefore, other approaches have been investigated in literature. The approaches include a kernel-based correction, a moving average filter, an autoregressive filter and a subsampling and averaging approach. It has been assessed to what extend correction for microstructure noise improves forecasting future volatility. The subsampling and averaging method has been documented to have the best performance among the noise corrector approaches. The subsampling and averaging approach constitutes the class of estimators that best predicts volatility. The present dissertation applies the subsampling approach on a wider class of the realized volatility, namely realized power volatility and introduces some new realized volatility and correlation estimators. The wider class is analogously more robust against large values, since it is constructed upon absolute transformation of return time series.

Simulation experiments suggest in general that the proposed estimators are consistent and contain comparatively faster convergence under Gaussian microstructure noise assumptions. In addition, all estimators indicate lower error under non-Gaussian noise compared to Gaussian microstructure noise. Empirical experiments imply that the proposed estimators are better able to display some dynamic behaviors and some stylized facts. While the realized correlation estimator possesses almost a normal distribution in comovement structure between stock markets, the proposed correlation is negatively skewed. We shall specifically see how the problem of noise can be solved by different estimators and how the estimators empirically behave.

Multiscale volatility estimation by Support Vector Regression (SVR) is also studied. It has been well documented that financial time series such as return series share common characteristics. Non-stationary character of stock market returns, for example, has been tested and documented repeatedly. An efficient way of representing a time series such as returns with such complex dynamics is given by wavelet methodology. With the help of a wavelet basis, discrete wavelet transform is able to break a time series with respect to a time-scale while preserving the time dimension.

Time-scale specific information is important, if one accepts the view that stock market consists of heterogeneous agents or investors operating at different time-scales. According to hypothesis of a heterogeneous market, the stock market consists of multiple layers of investment horizons (time-scales) varying from an extremely short (minutes) to long (years). The small time-scales are commonly thought to be related to speculative activity and the bigger time-scales to investment activity. Therefore, time-scale is one of the most important aspects in which trading behaviors differ. Considerable insight into the volatility dynamics is gained by looking at the data at several different time-scales or frequencies. At small time-scales, in particular, the locality of wavelet analysis allows one to fully exploit high frequency data.

The methodology used here is based on a wavelet multiscaling or multiresolution analysis (MRA), which decomposes the return data into its low and high frequency components, in combination with the SVR. The multiresolution analysis is implemented by means of an algorithm called maximal overlap discrete wavelet transform (MODWT). The MODWT is particularly useful for analyzing and forecasting time series that exhibit nonstationary characteristics, since time-dependent events at various scales are properly localized by MODWT. According to multiresolution theory, a function is described by means of a low resolution function plus a series of details from low to high resolution. The multiresolution analysis, indeed, provides different levels of frequency of a process. Exploiting the multiresolution analysis, a strategy or scheme for estimation is proposed. According to the strategy, employing the multiresolution analysis, a signal or time series is decomposed into an arbitrary number of different series from a smooth to detailed levels preserving the time dimension. Each individual series as input data feeds a volatility CHARN model to be approximated by the SVR machine. The aggregated estimation for original return series is then obtained by adding up the individual estimations.

Support vector regression (SVR) machine is a state-of-the-art sort of learning algorithm. In general, estimation and approximation applications involve making inference from observations that are distorted or corrupted in some unknown manner, when the information that one wishes to extract is unknown to the observer. The simplest way to approximate a function would be to take the mean of the observations. Choosing linear functions or more complicated bases of functions would be a more sophisticated approach, and the solution to obtain better results seems to enhance the complexity of the base. This is not true since one encounters the well-known effect of overfitting, which means that the complexity of the system of functions used is too high. For obtaining good approximations, one needs to take the complexity of the base of functions into account. There already exists a large set of approximation approaches, for instance splines and methods based on decomposition into orthogonal systems. All these methods suffer from shortcomings that are tried to be overcome by the support vector approach. Splines and decomposition approaches share the problem of exponential increase in the number of coefficients with the dimensionality of the problem. One solution is to use nonseparable expansions, e.g. neural networks, which allow tractable solutions of high dimensional problems. Their architecture has to be defined a priori or modified by some heuristics during training, which cannot assure that the optimal structure of the network is found for a particular problem. Moreover, the possibilities for controlling the complexity of the function base are rather limited, and the training algorithm can get stuck in local minima. Only for the asymptotic case and for the case of known prior probabilities optimal selection criteria have been obtained. In contrast, support vector machines (SVMs) possess a number of advantages. Their architecture does not have to be determined beforehand, and input data of any arbitrary dimension can be treated with only a linear cost in the number of input dimensions. Moreover, the training has a unique solution, and the modeling functions may be chosen within a rich function base having to satisfy only some conditions from functional analysis. Capacity is controlled efficiently by implementing a learning bias that involves a regularization term.

SVMs combine several results from statistical learning theory, optimization theory, and machine learning, and employ kernels as one of their most important ingredients.

The SVMs have been proved to pose excellent performance in many applications. Regression function approximation by support vector regression machine where the data is corrupted by noise, nonstationarity and locality will be particularly considered in this dissertation. The noise problem, for example, causes overfitting problem and in turn poor generalization. We shall demonstrate how SVMs can be applied in combination of multiresolution analysis to the specific problem of financial time series prediction.

Implementing the multiresolution strategy is expected to improve accuracy and precision performances of learning structures or approximating volatility model. In general, experiments on real data suggest that the multiscale estimation strategy yields better performance of estimation relative to a single resolution strategy. In particular, we shall indicate in detail how the multiresolution analysis can be used to help improving approximation power or learning patterns.

The present dissertation includes 2 main parts in addition to a general introduction as well as basic conclusions. Part I focuses on the realized volatility and correlation, and part 2 is devoted to multiscale modeling and forecasting volatility. Each part is provided by its own specific introduction. The specific introductions to the parts typically include literature review, problem description, motivation, objective, contribution, and structure of the corresponding part.

The current chapter 1, covering an introduction, provided a general overview on the subjects and issues of the chapters. Chapter 2 addresses the definition of realized volatility and correlation. Moreover, assumptions under which realized volatility and correlation estimators converge are explained. In chapter 3, the estimators are simulated under various assumptions and results are demonstrated. In addition, their distributional and dynamic behaviors are empirically experienced. Chapter 4 explains multiresolution analysis (MRA) by maximal overlap discrete wavelet transform (MODWT). Support vector regression machine, as an application of a theory called statistical learning theory, is illustrated in chapter 5. To nonparametrically approximate volatility, a model called conditional heteroskedastic autoregressive nonlinear (CHARN) is invoked to be approximated in chapter 6. Chapter 7 presents the results of volatility function estimation by SVR under different strategies. All fundamental conclusions, discussions, and open questions are gathered in chapter 8.

Part I

Realized volatility and correlation

Introduction to the part: High frequency finance has started to rapidly grow as a new field of finance after being availability of high frequency financial data. Exploiting the high frequency data, volatility and correlation can be measured more accurately in a model-free approach. The realized volatility, in a new approach to volatility, is claimed to be consistent under general nonparametric conditions. In other words, this type of measures provides more precise ex-post observations of the actual volatility compared to the traditional sample variances based on daily or coarser frequency data. The main idea is to aggregate intra-daily squared returns to construct realized volatility. In fact, sampling as often as possible would, in theory, produce exact estimates of the true variance in the limit. But in practice, the realized volatility and correlation estimators suffer from the market microstructure noise. The noise results in biased and imprecise estimators. This suggests that the estimators do not converge for high frequency levels, where the noise especially exists. To overcome the problem of noise, some approaches have been introduced. In addition, the realized volatility literature usually assumes a Gaussian microstructure noise. However, the noise in the real world financial markets does not follow the Gaussian process. The present part discusses volatility and correlation estimators and introduces new volatility and correlation estimators which converge faster and are consistent under Gaussian microstructure noises.

Literature review As it has been already discussed, the presence of market microstructure noise in high frequency financial data complicates the estimation of financial volatility and correlation making the approach unreliable. There is a considerable bias of estimation at the higher frequency due to intervention of the noise. While the realized volatility approach suggests sampling at the highest possible frequency to attain the highest precision, the market microstructure frictions exist at the highest levels of frequency. That is, this problem is most serious in high frequency data since the volatility of true price usually shrinks with the time interval, while the volatility of noise components such as the bid-ask spread usually does not. For this reason, sparse sampling or lower frequencies have been recommended to reduce the market microstructure contamination. But this is in contrast to the realized volatility theory. Optimal sampling schemes have been investigated by Bandi and Russell [Ban05b]. Bandi and Russell [Ban05b] argue that while it is theoretically necessary to sum squared returns that are computed over very small intervals to better identify the underlying volatility over a period, the summing of numerous contaminated return data entails substantial accumulation of noise. The resulting effect is the determination of a bias-variance trade-off. They quantify the trade-off in the presence of a realistic microstructure model of price determination and provide clear and easily implementable directions for optimally sampling high frequency data for the purpose of volatility estimation. The optimal sampling problem can be written as the minimization of the conditional MSE expansion of the realized volatility estimator. Specifically, they deem the (easy to implement) 15-minute sampling interval to be a valid (albeit conservative) choice of frequency. Such choice can be improved upon (i.e., lowered) in the case of very liquid stocks. Aït-Sahalia et al. [Ait05] in contrast conclude that even with optimal sampling, using say 5-min returns when transactions are recorded every second, a vast amount of data is discarded, in contradiction to basic statistical principles. They demonstrate that modeling the noise and using all the data is a better solution, even if one misspecifies the noise distribution. So the answer is: sample as often as possible. Therefore, researchers investigate some methods to cope with the problem at the highest available frequency in presence of the noise. A kernel-based correction proposed by Zhou [Zho96]; a moving average filter introduced by Maheu and McCurdy [Mah02]; an autoregressive filter introduced by Bollen and Inder [Bol02]; and a subsampling and averaging approach introduced by Zhang et al. [Zha05]. Ghysels and Sinko [Ghy07] assess to what extend correction for microstructure noise improves forecasting future volatility using the MIxed DAta Sampling (MIDAS) framework. The subsampling and averaging procedure has been experimentally documented by Ghysels and Sinko [Ghy07] to predict volatility the best among microstructure noise correctors. Their empirical results suggest that for 30 Dow Jones stocks data, within the class of quadratic variation measures, the subsampling and averaging approach constitutes the class of estimators that best predicts volatility.

Problem description The problem of market microstructure noise is well dealt with by the subsampling procedure of Zhang et al. [Zha05] where their proposed realized volatility is constructed upon squared intra-daily returns. However, the construction of realized volatility upon squared returns is only one of the alternatives for realized volatility and is indeed a specific case of what Barndorff-Nielsen and Shephard [Bar03] generally introduce. In fact, Barndorff-Nielsen and Shephard [Bar03] extend the realized squared volatility to realized power variation which covers realized squared as well as realized absolute volatility. Meanwhile, the special case, i.e., realized absolute volatility models have been reported to produce better volatility forecasts than models based on squared returns. But a specific problem we face to in this part is that the realized absolute volatility and the wider case of realized power volatility suffer from the microstructure noise. The realized squared correlation faces to the same problem of noise.

Motivation All in all, much of discussions in this part is motivated by the need for forecasting volatility and correlation of financial asset return series. In particular, inspired by superiority of the subsampling method to cope with the noise problem, advocating Zhang et al. [Zha05], the subsampling method is applied on the wider class of realized variation, namely the realized power volatility and especially on its specific case, that is, the realized absolute volatility.

Objective Applying the subsampling method to construct new realized volatility and correlation estimators, we aim to improve upon the convergence of realized power variation at higher frequencies under the presence of noise. Consistent and unbiased estimators for true volatility and correlation are desired. Furthermore, it is desired the estimators include dynamic behaviors and stylized facts as many as and as strong as possible.

Contribution The part contributes by new realized volatility and correlation estimators. The estimators are expected to converge faster and to be consistent under the market microstructure noise. The estimators are more robust against the large values. In addition, some new types of the microstructure noise are introduced and simulated. Empirically, the new estimators reveal better some dynamic behaviors and stylized facts. In terms of distributional characteristic, the new realized correlation estimators exhibit negative asymmetry or heavy tail.

Structure of the part The current part contains two chapters. The first chapter involves theories and the second to simulation and empirical experiments. In two first sections of chapter 2, importance of volatility and correlation modeling and forecasting for different areas of applied finance are mentioned. In section 2.3, it is discussed that whether volatility is a measure of risk as sometimes it is supposed. Section 2.4 addresses to different alternatives to realized volatility estimator. In this section, our idea for measuring volatility is formulated. Consequently, realized correlation estimators constructed based on realized volatility are explained in the next section. The Epps effect which yields a considerable bias when applying non-synchronous trading hours to estimate covariation is discussed in this section. We will see how to efficiently solve the problem of Epps effect. Before simulating and evaluating the realized volatility and correlation estimators, some assumptions about the price, return and noise processes have to clarified. These are explained in section 2.6. Moreover, section 2.7 explains some self-similar noise processes to be exploited in simulation studies. Under normality and non-normality assumptions of the noise, the realized volatility and correlation estimators are simulated to observe their convergence and error behaviors in the next chapter, sections 3.1 and 3.2. Then in section 3.3, distributional and dynamic behaviors of the estimators utilizing real data are empirically experimented. Observing some dynamic behavior of the estimators, we are intrigued to put into discussion the old issue of predictability of the financial markets. Very briefly, some discussions of predictability are provided in section 3.4. In section 3.5, an association of volatility and correlation is studied. The last section 3.6 temporarily covers some related conclusions and discussions about important issues. Further investigations are also discussed in this section.

Chapter 2

Realized volatility and correlation estimators

2.1 Importance of volatility modeling and forecasting

Return volatility is at the center of many theories within financial economics, be it asset and derivatives pricing or risk management, so it is hardly surprising that great effort has been made to determine reliable, if not optimal, procedures for forecasting future volatility. Likewise, the practical import of volatility for financial performance has spurred product innovation, leading to a rapid increase in organized trading of financial derivatives written directly on volatility variables, such as variance swaps and futures and options written on volatility indices as well as an over-the-counter market in variance and volatility swaps on individual assets. In short, the financial industry views volatility as a distinct asset class endowed with separate risk factors and novel opportunities for both strategic trading and hedging. Obviously, the latter developments also have generated a surge in the demand for practical volatility forecast procedures.

Volatility forecasting is an important task in financial markets, and it has held the attention of academics and practitioners over the last two decades. An extensive research reflects the importance of volatility in investment, security valuation, risk management, and monetary policy making. Volatility is the most important variable in the pricing of derivative securities, whose trading volume has quadrupled in recent years. To price an option, we need to know the volatility of the underlying asset from now until the option expires. In fact, the market convention is to list option prices in terms of volatility units. Nowadays, one can buy derivatives that are written on volatility itself, in which case the definition and measurement of volatility will be clearly specified in the derivative contracts. In these new contracts, volatility now becomes the underlying asset. So a volatility forecast and a second prediction on the volatility of volatility over the defined period is needed to price such derivative contracts.

“Financial risk management has taken a central role since the first Basle Accord was established in 1996” [Poo03]. This effectively makes volatility forecasting a compulsory and practically risk management exercise for many financial institutions around the world. Banks and trading houses have to set aside reserve capital of at least three times that of value-at-risk (VaR), which is defined as the minimum expected loss with a 1-percent confidence level for a given time horizon (usually one or ten days). Sometimes, a 5-percent critical value is used. Such VaR estimates are readily available given volatility forecast, mean estimate, and a normal distribution assumption for the changes in total asset value. When the normal distribution assumption is disputed, which is very often the case, volatility is still needed in the simulation process used to produce the VaR figures.

Poon and Granger [Poo03] state that “financial market volatility can have a wide repercussion on the economy as a whole. The incidents caused by the terrorists’ attack on September 11, 2001, and the recent financial reporting scandals in the United States have caused great turmoil in financial markets on several continents and a negative impact on the world economy. This is clear evidence of the important link between financial market uncertainty and public confidence. For this reason, policy makers often rely on market estimates of volatility as a barometer for the vulnerability of financial markets and the economy. In the United States, the Federal Reserve explicitly takes into account the volatility of stocks, bonds, currencies, and commodities in establishing its monetary policy. The Bank of England is also known to make frequent references to market sentiment and option implied densities of key financial variables in its monetary policy meetings” [Poo03].

2.2 Importance of correlation modeling and forecasting

Asset returns cross correlations is pivotal to many prominent financial problems such as asset allocation, risk management and option pricing.

The covariance of financial asset returns is of central importance in the theory of asset prices, and is a recurring theme throughout finance. Finding good empirical ex-post estimates of covariance is a key step to understand it better. For this purpose, there is an opportunity to draw on recent advances in the study of ex-post realized variances.

Correlations are critical inputs for many of the common tasks of financial management. Hedges require estimates of the correlation between the returns of the assets in the hedge. If the correlations and volatilities are changing, then the hedge ratio should be adjusted to account for the most recent information. Similarly, structured products such as rainbow options that are designed with more than one underlying asset have price that are sensitive to the correlation between the underlying returns. A forecast of future correlations and volatilities is the basis of any pricing formula. Asset allocation and risk assessment also rely on correlations. However, in this case a large number of correlations is often required. Construction of an optimal portfolio with a set of constraints requires a forecast of the covariance matrix of the returns. The quest for reliable estimates of correlations between financial variables has been the motivation for countless researches [Eng02].

2.3 Volatility is a measure of risk?

Bollerslev and Zhou [Bol07] show that the difference between model-free implied and realized variances, which they term the variance risk premium, provides remarkable accurate and stable forecasts for the quarterly market return, with high (low) premia predicting high (low) future returns.

Poon and Granger [Poo03] argue that volatility is not the same as risk. "When it is interpreted as uncertainty, it becomes a key input to many investment decisions and portfolio creations. Investors and portfolio managers have certain levels of risk which they can bear". A good forecast of the volatility of asset prices over the investment holding period is a good starting point for assessing investment risk. Many investors and generations of finance students often have an incomplete appreciation of the differences between volatility, standard deviation, and risk. It is worth elucidating some of the conceptual issues here. In finance, volatility is often used to refer to standard deviation, σ, or variance, σ^2, computed from a set of observations. The sample standard deviation statistic $\hat{\sigma}$ is a distribution free parameter representing the second moment characteristic of the sample. Only when σ is attached to a standard distribution, such as a normal or a t distribution, can the required probability density and cumulative probability density be derived analytically. Indeed, σ can be calculated from any irregular shape distribution, in which case the probability density will have to be derived empirically. In the continuous time setting, σ is a scale parameter that multiplies or reduces the size of the fluctuations generated by the standard wiener process. Depending on the dynamic of the underlying stochastic process and whether or not the parameters are time varying, very different shapes of returns distributions may result. So it is meaningless to use σ as a risk measure unless it is attached to a distribution or a pricing dynamic. When σ is used to measure uncertainty, the users usually have in mind, perhaps implicitly, a normal distribution for the returns distribution.

Standard deviation, σ, is the correct dispersion measure for the normal distribution and some other distributions, but not all. Other measures that have been suggested and found useful include the mean absolute return and the interquantile range. However, the link between volatility and risk is tenuous; in particular, risk is more often associated with small or negative returns, whereas most measures of dispersion make no such distinction. The Sharpe ratio, for example, defined as return in excess of risk free rate divided by standard deviation, is frequently used as an investment performance measure. It incorrectly penalizes occasional high returns. The idea of semi-variance, an early suggestion by Markowitz [Mar91], which only uses the squares of returns below the mean, has not been widely used, largely because it is not operationally easy to apply in portfolio construction.

Both concepts of volatility and risk are very broad. Hence, what are volatility and risk, depends on within which framework volatility and risk are identified. If there is a significant measurement error, volatility may not act as an appropriate measure of risk. Unconditional volatility measures average fluctuations around the unconditional mean. Conditional volatility measures time-series fluctuations around the conditional mean. A stochastic volatility measures time-series fluctuations around the stochastic mean. These volatility measures are obtained from the physical (or empirical) return distribution, the so-called P-measure of volatility. Alternatively an expected future volatility obtained from a set of options measures volatility, called Q-measure of volatility which is backed out from the risk-neutral distribution.

All of these measures of volatility can be treated as a measure of risk during normal functioning of the financial markets. However, during extremely volatile markets, volatility cannot be considered as an appropriate measure of risk even if the measurement error is zero. A downside risk measure such as Value at Risk, Expected Shortfall, Semi-variance, Tail Risk, and so on represents a measure of risk.

In summary, volatility interpreted as uncertainty, is one of the key variables in most models in modern finance. Risk is usually associated with small or negative returns (the so-called downside risk) whereas the most common measures of dispersion (e.g. standard deviation) make no such distinction. Furthermore, standard deviation is a useful risk measure only when it is attached to a distribution or a pricing dynamic.

2.4 Realized volatility estimators

In classical volatility literature, volatility is often calculated as the sample standard deviation. Figlewski [Fig97] notes that since the statistical properties of sample mean make it a very inaccurate estimate of the true mean, especially for small samples, taking deviations around zero instead of the sample mean typically increases volatility forecast accuracy. There are methods for estimating volatility that are designed to exploit or reduce the influence of extremes[1]. While the $\hat{\sigma}^2$ is an unbiased estimate of σ^2, the square root of $\hat{\sigma}^2$ is a biased estimate of σ due to Jensen inequality[2].

[1] For example, the Maximum likelihood method proposed by Ball and Torous [Bal84], the high-low method proposed by Parkinson [Par80] and Garman and Klass [Gar80].

[2] See Fleming [Fle98], and Cox and Rubinstein [Cox85] for explanation of how this bias can be corrected assuming a normal distribution for R_t. However, in most cases, the impact of this adjustment is small.

2.4.1 Realized Squared Volatility estimator

A voluminous literature has emerged for modeling the temporal dependencies in financial market volatility using ARCH and stochastic volatility models. Andersen and Bollerslev [And98] argue that while most of these studies have documented highly significant in-sample parameter estimates and pronounced intertemporal volatility persistence, traditional ex-post forecast evaluation criteria suggest that the models provide seemingly poor volatility forecasts. Contrary to this contention, Merton [Mer80] showed that the integrated volatility of a Brownian motion over a fixed interval can be approximated to an arbitrary precision using the sum of intraday squared returns, provided the data are available at a sufficiently high sampling frequency. In other words, the variance over a fixed interval can be estimated arbitrarily, although accurately, as the sum of squared realizations, provided the data are available at a sufficiently high sampling frequency. More formally, Andersen and Bollerslev [And98] show that volatility models produce strikingly accurate interdaily forecasts for the latent volatility factor that would be of interest in most financial applications. They discuss new methods for improved ex-post interdaily volatility measurements based on high frequency intradaily data. In fact, they demonstrate how high frequency intraday data may be used constructively in forming more accurate and meaningful ex-post interdaily volatility measurements.

The intuition behind the apparent poor predictive power of well-specified volatility models is straightforward. Let the return innovation be written as $r_t = \sigma_t.z_t$, where z_t, denotes an independent mean zero, unit variance stochastic process, while the latent volatility, σ_t, evolves in accordance with the particular model entertained. A common approach for judging the practical relevance of any model is to compare the implied predictions with the subsequent realizations. Unfortunately, volatility is not directly observed so this approach is not immediately applicable for volatility forecast evaluation. Still, if the model for σ_t^2 is correctly specified, then $E_{t-1}(r_t^2) = E_{t-1}(\sigma_t^2.z_t^2) = \sigma_t^2$, which appears to justify the use of the squared return innovation over the relevant horizon as a proxy for the ex-post volatility. However, while the squared innovation provides an unbiased estimate for the latent volatility factor, it may yield very noisy measurements due to the idiosyncratic error term, z_t^2. This component typically displays a large degree of observation-by-observation variation relative to σ_t^2, rendering the fraction of the squared return variation attributable to the volatility process low. Consequently, the poor predictive power of volatility models, when judged by standard forecast criteria using r_t^2 as a measure for ex-post volatility, is an inevitable consequence of the inherent noise in the return generating process.

The poor volatility forecast in ARCH models motivates a fundamentally different approach. Rather than seeking to perfect the forecast evaluation procedures-taking the noisy observations on volatility provided by fixed-horizon squared returns as given-it may prove fruitful to pursue alternative ex-post volatility measures. Specifically, building on the continuous-time stochastic volatility framework developed by Nelson [Nel90] and Drost and Werker [Dro96], Andersen and Bollerslev [And98] demonstrate how high frequency data allow for the construction of vastly improved ex-post volatility measurements via cumulative squared intraday returns.

Actually, applying the quadratic variation theory, Andersen and Bollerslev [And98] generalized the result of Merton [Mer80] to the class of special (finite mean) semimartingales, the so-called realized volatility (RV). This class encompasses processes used in standard arbitrage-free asset pricing applications, such as, Ito diffusions, jump processes, and mixed jump diffusions. In the standard arbitrage-free asset pricing framework, the log-price of a financial asset follows a continuous-time semi-martingale process with stochastic volatility and possibly jumps. The standard definition for an equally spaced returns series of the Realized Squared volatility $\widehat{RS}$ is

$$\widehat{RS} = \sum_{t_i}^{T} (Y_{t_{i+1}} - Y_{t_i})^2, \tag{2.1}$$

over a period t, with $0 = t_0 \leq t_1 \leq ...t_n = T$ and $i = 1, ..., n$ is ith intraday observation with an integer n. Here Y_{t_i} denotes a logarithmic price of an asset on day t at time i. According to the theory of quadratic variation, Andersen and Bollerslev [And98] suggest that as the observation frequency increases from a daily to an infinitesimal interval, this measure of volatility converges to genuine measurement of the latent volatility factor. In other words, the quantity $\widehat{RS}$ is thought to approximate the so-called Integrated Volatility (IV), i.e.,

$$IV(t) = \int_{t-1}^{t} \sigma^2(s) ds, \tag{2.2}$$

as $n \to \infty$.

In summary, the notion of realized volatility represents a model-free approach to (continuous-record) consistent estimation of the quadratic return variation under general assumptions based primarily upon arbitrage-free financial markets. As such it allows us to harness the information inherent in high frequency returns for assessment of lower frequency return volatility. It is thus the natural approach to measuring actual (ex-post) realized return variation over a given horizon. This perspective has now gained widespread acceptance in the literature, where alternative volatility forecast models are routinely assessed in terms of their ability to explain the distribution of subsequent realized volatility.

2.4.2 Realized Power Volatility estimator

Davidian and Carroll [Dav87] show absolute returns volatility specification is more robust against asymmetry and non-normality. There is some empirical evidence that deviations or absolute returns based models produce better volatility forecasts than models based on squared returns (Taylor [Tay86]; Ederinton and Guan [Ede00]; and McKenzie [Mak99]), but the majority of time series volatility models are squared returns models. Hence, Ding, Granger, and Engle [Din93] suggest measuring volatility directly from absolute returns. They propose realized Absolute Volatility, $\widehat{RA}$, as an aggregated intra-daily absolute returns as follows

$$\widehat{RA} = \sum_{t_i}^{T} \left| Y_{t_{i+1}} - Y_{t_i} \right| . \tag{2.3}$$

The estimator $\widehat{RA}$ is supposed to asymptotically converge to its true or integrated volatility, i.e., Integrated Absolute Volatility

$$IAV(t) = \int_{t-1}^{t} \sigma(s)ds, \tag{2.4}$$

as n increases. Barndorff-Nielsen and Shephard [Bar03a] have generalized the main idea of accumulative intradaily squared or absolute returns to a wider class called Realized Power variation of order r, that is, sums of absolute powers of increments of a process, $\widehat{RP}$,

$$\widehat{RP} = \sum_{t_i}^{T} \left| Y_{t_{i+1}} - Y_{t_i} \right|^r , \tag{2.5}$$

where $i = 1, ..., n$ is ith intraday observation with an integer n and r, the power or order, is a positive value. The quantity of Realized Power variation, $\widehat{RP}$ as a proxy, is supposed to approximate the daily increments of the power variation of the semimartingale that drives the underlying logarithmic price process, i.e., Integrated Power Volatility (IPV),

$$IPV(t) = \int_{t-1}^{t} \sigma^r(s)ds, \tag{2.6}$$

as $n \to \infty$ for a fixed t. The consistency result justifying this procedure is the convergence in probability of $\widehat{RP}$ to IPV as returns are computed over intervals that are increasingly small asymptotically or, equivalently, as $n \to \infty$ for a fixed t. Barndorff-Nielsen and Shephard [Bar03a] provided a limiting distribution theory for realized power variation. A special case of $\widehat{RP}$, where $r = 1$, is known as Realized Absolute (RA) volatility in (2.3). The estimator (2.1) is also a special case of $\widehat{RP}$ where $r = 2$.

In practice, it is infeasible the realized volatility estimators converge to their integrated volatility because of data limitations and a host of market microstructure features. In reality, there is a definite lower bound on the return horizon that can be used productively for computation of the realized volatility, both because we only observe discretely sampled returns and, more important, market microstructure frictions on intradaily level such as discreteness of the price grid, asymmetries in information, nonsynchronous trading effects, transaction costs, bid-ask spreads, lunch-time effects, and intraday periodic patterns such as U-shape volatility of trading volume over the day induce gross violations of the semimartingale property at the very highest return frequencies. This implies that we typically will be sampling returns at an intraday frequency that leaves a non-negligible error term in the estimate of integrated power volatility.

Several approaches have been introduced to correct the microstructure noise. A kernel-based correction introduced by Zhou [Zho96], an optimal sampling introduced by Bandi and Russell [Ban05b], a moving average filter introduced by Maheu and McCurdy [Mah02], an autoregressive filter introduced by Bollen and Inder [Bol02], and a subsampling and averaging approach introduced by Zhang et al. [Zha05]. It has been experimentally shown by Ghysels and Sinko [Ghy07] that the subsampling and averaging class of estimators predicts volatility the best among microstructure noise correctors.

2.4.3 Two-Scale Realized squared Volatility estimator

In order to deal with the market microstructure effects in approximating Integrated Volatility, $IV = \int_{t-1}^{t} \sigma^2(s)ds$, through the estimator $\widehat{RS}$, a well-accepted alternative approach called Two-Scale Realized Volatility (TSRV), based on a subsampling and averaging procedure has been proposed by Zhang et al. [Zha05]. Their device takes advantage of the rich sources of tick-by-tick data, and to a great extent corrects for the adverse effects of microstructure noise on volatility estimation.

The volatility estimator $\widehat{TSRV}$ combines the sum of squared estimators from two different time scales; $\widehat{RS}_{avg}$ from the returns on a slow time scale, whereas $\widehat{RS}_{all}$ is computed from the returns on a fast time scale using the latter as a means for bias-corrector of the measure. The $\widehat{RS}_{avg}$ estimator is constructed based on subsampling and averaging procedure. The $\widehat{TSRV}$ estimator approximates Integrated Volatility as unbiased and more precisely than $\widehat{RS}$ estimator under the microstructure frictions. It forms as

$$\widehat{TSRV} = \left(1 - \frac{\bar{n}}{n}\right)^{-1} \left(\widehat{RS}_{avg} - \frac{\bar{n}}{n}\widehat{RS}_{all}\right), \tag{2.7}$$

where the $\widehat{RS}_{all}$ estimator is the same as (2.1) and $\widehat{RS}_{avg} = \frac{1}{K}\sum_{t_{i+1},t_i \in g^{(k)}} (Y_{t_{i+1}} - Y_{t_i})^2$, when the K number of samples are regularly allocated to g subgrids. The estimator $\widehat{TSRV}$ averages the squared returns from sampling every data point, $\widehat{RS}_{all}$, and those from every Kth data point, $\widehat{RS}_{avg}$. Its asymptotic behavior derived by Zhang et al. [Zha05] when $n \to \infty$ and $\frac{n}{K} \to \infty$. It is a consistent and unbiased estimator for integrated volatility, IV, (2.2).

2.4.4 Two-Scale realized Power Volatility estimator

Motivated by the benefits of subsampling and averaging frequencies procedure in the Two-Scale squared Realized Volatility (TSRV), the general class of the realized power variation measure was extended to a Two-Scale realized Power Volatility (TSPV) measure [Saf07a], where the variation of the measure can be lessened by the averaging on samples and the bias can be vanished into zero by sampling on all data points.

To define the estimator, we start by defining the full grid of G arrival times, $G = \{t_0, ..., t_n\}$, partitioned into K nonoverlapping subgrids $g^{(k)}$ with $k = 1, ..., K$. The first subgrid starts from t_0 and takes every Kth arrival time, i.e., $g^{(1)} = (t_0, t_{0+K}, t_{0+2K}, ...,)$, the second subgrid starts from t_1 and takes every Kth arrival time, i.e., $g^{(2)} = (t_1, t_{1+K}, t_{1+2K}, ...,)$ and so on. Given the kth subgrid of arrival times, the corresponding realized variation estimator can be defined as $\widehat{RP}^{(k)} = \sum_{t_i, t_{i+1} \in g^{(k)}} \left|Y_{t_{i+1}} - Y_{t_i}\right|^r$, where t_i and t_{i+1} denote consecutive elements in $g^{(k)}$. Then the Two-Scale realized Power Volatility, $\widehat{TSPV}$ is estimated by

$$\widehat{TSPV} = \left(1 - \frac{\bar{n}}{n}\right)^{-1} \left(\widehat{RP}_{avg} - \frac{\bar{n}}{n}\widehat{RP}_{all}\right), \tag{2.8}$$

where $(1 - \frac{\bar{n}}{n})^{-1}$ is a small-sample adjustment and $\bar{n} = \frac{n-K+1}{K}$. The estimator $\widehat{TSPV}$ combines the realized power volatility estimators from two time scales. It combines the sum of power estimators from two different time scales; $\widehat{RP}_{avg}$ from the returns on a slow time scale, whereas $\widehat{RP}_{all}$ is computed from the returns on a fast time scale using the latter as a means for bias-corrector of the estimator. The $\widehat{RP}_{avg}$ estimator is constructed based on subsampling and averaging procedure. The estimator $\widehat{RP}_{all}$ is obtained via (2.5) and the estimator $\widehat{RP}_{avg}$ through

$$\widehat{RP}_{avg} = \frac{1}{K} \sum_{k=1}^{K} \sum_{t_i, t_{i+1} \in g^{(k)}} \left|Y_{t_{i+1}} - Y_{t_i}\right|^r \tag{2.9}$$

and, in a special case when the sampling points are regularly allocated, from

$$\widehat{RP}_{avg} = \frac{1}{K} \sum_{t_i, t_{i+1} \in g^{(k)}} \left|Y_{t_{i+1}} - Y_{t_i}\right|^r. \tag{2.10}$$

The averaging scale reduces the variance of the estimator while the all scale plays a bias-correcting role. The optimal number of subgrids K, where the bias induced by the noise is minimized, provided by Zhang et al. [Zha05] is expressed as

$$K = cn^{2/3}, \tag{2.11}$$

as $n \to \infty$ where c is estimated by $c = (\frac{16\sigma_\epsilon^4}{TE\eta^2})^{1/3}$ where $\eta^2 = \frac{3}{4}\int_0^t \sigma^4(s)ds$. The term σ_ϵ^4 is square of the variance of the noise, while $\int_0^t \sigma^4(s)ds$ is the integrated quarticity. The σ_ϵ^2 is estimated by $\widehat{\sigma_\epsilon^2} = \frac{1}{2n}\widehat{RP}$ and $\eta^2 = \frac{4}{3}(\widehat{RP})^2$ at some reasonable lower frequency, for example either every 15 or 20 minute [Bar06].

The estimator $\widehat{TSPV}$ is expected to consistently converge to its true estimator IPV even under the microstructure noise at every intadaily frequency, as $n \to \infty$ with $\frac{n}{K} \to \infty$ over a fixed interval of time t.

2.5 Realized correlation estimators

Exploiting the high frequency data has been advocated to improve the precision of asset volatility measurement and estimation. The so-call Realized Volatility approach was proposed to this end. As for the realized volatility approach, the idea of employing high frequency data in the computation of covariances and correlations between assets leads to the analogous concept of realized covariance (or covariation) and realized correlation. After introducing realized covariance and correlation by Andersen et al. [And01a] and Andersen et al. [And01b], several alternatives have been appeared.

2.5.1 Realized squared-based correlation estimator

Based on the realized variation theory, Andersen et al. [And01a] and Andersen et al. [And01b] have derived realized standard deviation, $\widehat{RS}_{std} = \widehat{RS}^{1/2}$; covariance, $\widehat{RCOV}_{xy} = \sum_{t_i}^{T}(Y_{t_{i+1}} - Y_{t_i})_x.(Y_{t_{i+1}} - Y_{t_i})_y$; and realized squared-based correlation, $\widehat{RSCOR}_{xy}$ in the form of

$$\widehat{RSCOR}_{xy} = \widehat{RCOV}_{xy}/(\widehat{RS}_{std,x}.\widehat{RS}_{std,y}), \tag{2.12}$$

where x and y are two assets or high frequency time series. Barndorff-Nielsen and Shephard [Bar04a] have provided an asymptotic distribution theory for these realized covariance and squared based correlation estimators allowing the returns to be a stochastic volatility semimartingale. The limit theory for the normalized estimation error for realized covariance, regression, and correlation of the returns of assets asymptotically results to $N(0, 1)$ as $n \to \infty$. This implies that their estimators converge in probability to the corresponding true covariance, regression, and correlation. The limit theory is robust as it does not require the empirical researcher to specify a model for the spot covolatility or the drift process. In this sense, it is semiparametric. They argue that an important theme in theoretical econometrics and statistics is that covariances are not very robust objects, as they are highly sensitive to large movements in asset prices. It may be desirable to construct economic theory and econometrics on more robust quantities such as mean absolute errors.

2.5.2 Realized absolute-based correlation estimator

Motivated by robustness of absolute transformation in analogous to square transformation and by availability of high frequency data, the concept of realized power-based volatility was extended to realized covariation in [Saf07a]. Thus, based on realized power variation, absolute-based realized power standard deviation, $\widehat{RP}_{std} = \widehat{RP}^{1/2}$ is derived according to the corresponding realized power variation. Realized covariance remains the same as $\widehat{RCOV}_{xy}$, and realized power-based correlation, $\widehat{RPCOR}_{xy}$ takes the form

$$\widehat{RPCOR}_{xy} = \widehat{RCOV}_{xy}/(\widehat{RP}_{std,x}.\widehat{RP}_{std,y}), \tag{2.13}$$

where all estimators are based on a fixed interval of time and where x and y are two assets or high frequency time series. Throughout of this part, we consider covariance and correlation estimators between only two assets. However, the estimators can be extended to the covariation between several assets. The corresponding Integrated Power-based Correlation is defined as

$$IRPCOR_{xy} = \frac{\int_{t-1}^{t} \Sigma_{xy}(s)ds}{\sqrt{\int_{t-1}^{t} \sigma_x^r(s)ds \int_{t-1}^{t} \sigma_y^r(s)ds}}, \tag{2.14}$$

where $\int_{t-1}^{t} \sigma_x^r(s)ds$ is integrated power volatility for asset x and so for y according to the previous notations, and $\int_{t-1}^{t} \Sigma_{xy}(s)ds$ is the true or Integrated Covariance.

2.5.3 Two-scale realized power-based correlation estimator

Due to the presence of microstructure noise or frictions in practice, the estimator $\widehat{RPCOR}_{xy}$, which is conditionally built on noisy realized covariance, would not consistently estimate $IRPCOR_{xy}$ and it would show a considerable bias and some slower convergence than it can when it is modified like the case of volatility by the same applied approach (subsampling). Hence, we further follow the subsampling method to construct a two-scale correlation estimator to be consistent and unbiased for $IRPCOR_{xy}$. We have $\widehat{TSPV}_{std} = \widehat{TSPV}^{1/2}$. The estimator Two-Scale Covariance $\widehat{TSCOV}_{xy}$ is proposed as follows

$$\widehat{TSCOV}_{xy} = \left(1 - \frac{\bar{n}}{n}\right)^{-1} \left(\widehat{RCOV}_{xy,avg} - \frac{\bar{n}}{n}\widehat{RCOV}_{xy,all}\right) \tag{2.15}$$

where $\widehat{RCOV}_{xy,all}$ is the same as $\widehat{RCOV}_{xy}$, built on the full grid. The estimator $\widehat{RCOV}_{xy,avg}$ is estimated by

$$\widehat{RCOV}_{xy,avg} = \frac{1}{K}\sum_{k=1}^{K} \sum_{t_i,t_{i+1}\in g^{(k)}} (Y_{t_{i+1}} - Y_{t_i})_x.(Y_{t_{i+1}} - Y_{t_i})_y. \tag{2.16}$$

Finally the Two-Scale Power-based Correlation, $TSPCOR_{xy}$, is estimated as

$$\widehat{TSPCOR}_{xy} = \widehat{TSCOV}_{xy}/(\widehat{TSPV}_{std,x}.\widehat{TSPV}_{std,y}), \tag{2.17}$$

where $\widehat{TSPCOR}_{xy}$ denotes the two-scale time-varying and instantaneous conditional correlation between the returns of two time series x and y. The $\widehat{TSPCOR}_{xy}$ estimator should converge asymptotically to the $IRPCOR_{xy}$ estimator under microstructure noise.

2.5.4 Synchronous covariance estimator

The standard way to compute the realized covariance is to first choose a time interval, construct an artificially regularly-spaced time series by means of some interpolation scheme and then take the contemporaneous sample covariance of those regularly-spaced returns. But simulations and empirical studies indicate that such covariance measure presents a bias toward zero which rapidly increases with the reduction of the time length of the fix interval chosen [Cor07]. As for the realized volatility, the presence of market microstructure can induce significant bias in standard realized covariance measure. However, the microstructure effects responsible for this bias are different. Corsi [Cor07] argues that “bid-ask bouncing, which is the major source of bias for the realized volatility, will just increase the variance of the covariance estimator but it will not induce any bias. On the contrary, the so called non-synchronous trading effect strongly affects the estimation of the realized covariance and correlation”. In fact, since the sampling from the underlying stochastic process is different for different assets, assuming that two time series are sampled simultaneously when, indeed, the sampling is non-synchronous gives rise to the non-synchronous trading effect. As a result, covariances and correlations measured with high frequency data will possess a bias toward zero which increases as the sampling frequency increases. This effect of a dramatic drop of the absolute value of correlations among stocks when increasing the sampling frequency was first reported by Epps [Epp79] and hence called the Epps effect. The absolute value of the correlation is biased toward zero. The effect implies that empirical correlations virtually disappear at high frequencies, while being far from zero at moderate intraday frequencies. Epps' findings have been replicated extensively in financial markets.

The Epps effect has been widely associated with non-synchronous trading, when fresh observations of transactions prices do not arise simultaneously across markets, but are separated by, e.g., a few seconds. See Scholes and Williams [Sch77]. If non-synchronous trading is the source of the Epps effect, there is a challenging consequence for realized covariation estimation. Indeed, Hayashi and Yoshida [Hay05] and Corsi [Cor07] develop an all-overlapping-returns estimator of covariation to do this, and Lunde and Voev [Lun07] and Zhang [Zha06] assess it when there is contamination or measurement error. However, evidence from Reno [Ren03] indicates that on equity and currency markets non-synchronous trading is not alone sufficient to explain Epps effects.

Let t_i and τ_j be the instants at which the prices x and y are being observed. Hayashi and Yoshida [Hay05] and Corsi [Cor07] proposed a covariance estimator

$$\widehat{RCOV}_{xy} = \sum_{t_i}^{T_n} \sum_{\tau_j}^{T_m} (Y_{x,t_i} - Y_{x,t_{i-1}}).(Y_{y,\tau_j} - Y_{y,\tau_{j-1}}). \tag{2.18}$$

$$I[\min(t_i, \tau_j) > \max(t_{i-1}, \tau_{j-1})],$$

where $I[.]$ is the indicator function which takes the value of one only when the observations of two returns instantaneously overlap. This estimator consistently estimates the covariance of non-synchronous processes.

2.6 Assumptions about processes

To evaluate performance of the realized estimators, some assumptions have to be imposed on price, return and noise processes. Let p denotes a price process. We observe logarithmic price Y =log p as

$$Y = Y^* + u, \tag{2.19}$$

where Y^* denotes the logarithmic equilibrium or efficient price of an asset, i.e., the price that would prevail in the absence of market microstructure frictions, and u denotes a microstructure contamination in the observed logarithmic price as induced by price discreteness and bid-ask bounce effects. We fix a certain time period t (a day, say) and assume availability of n high frequency prices over t. Given Eq. (2.19), we can readily define continuously-compounded returns over any intra-period interval of length $\frac{t}{n}$ and write

$$Y_{t_{i+1}} - Y_{t_i} = Y^*_{t_{i+1}} - Y^*_{t_i} + u_{t_{i+1}} - u_{t_i},$$

or

$$y_{t_i} = y^*_{t_i} + \epsilon_{t_i}, \tag{2.20}$$

where y_{t_i} is a return on day t at time i, and where $t = 1, ..., T$ and $i = 1, ..., n$. The following assumptions are imposed on the price process and market microstructure effects.

Price Process : The logarithmic price process, Y^*, is a continuous stochastic volatility semimartingale. Specifically,

1: The price process is decomposed as

$$Y^* = \alpha_t + m_t, \tag{2.21}$$

where α_t (with $\alpha_0 = 0$) is a continuous drift process of finite variation defined as $\int_0^t \phi(s)ds$ and m_t is a continuous local martingale defined as $\int_0^t \sigma(s)dWs$, with $\{W_t : t \geq 0\}$ denoting a standard Brownian motion.

2: The spot volatility process, σ_t, is cádlág and bounded away from zero.

3: The integrated variance process $\int_0^t \sigma^r(s)ds$ ($r = 2$ for integrated volatility) is bounded almost surely for all $t < \infty$.

Microstructure Noise : Considering the decomposition (2.19),

1: The microstructure frictions in the price process, $u^{'}_{t_i}$, have mean zero and are strictly stationary with joint density $f_n(.)$.

2: The variance of $\epsilon_{t_i} = u_{t_{i+1}} - u_{t_i}$ is $O(1)$ for all i and all n.

3: The $u^{'}_{t_i}$ are independent of the $Y^{*'}_{t_i}$ for all i and all n.

In agreement with asset pricing theory, the first assumption (price process) implies that the equilibrium return process evolves in time as a stochastic volatility martingale difference plus an adapted process of finite variation. The stochastic spot volatility can display jumps, diurnal effects, high-persistence (possibly of the long memory type), and nonstationarities. Furthermore, leverage effects (i.e., dependence between σ and the Brownian motion W) are allowed.

The second assumption permits general dependence features for the microstructure noise components in the recorded prices. The correlation structure of the microstructure noise contaminations can, for instance, capture first order negative autocorrelations in the recorded high frequency returns as determined by bid-ask bounce effects as well as higher order dependence in the market frictions as induced by clustering in order flows. In general, the characteristics of the noise returns ϵ's may depend on the sampling frequency.

While the equilibrium return process $y^*_{t_i}$ is modeled as being $O_p(\sqrt{\frac{t}{n}})$ over any intra-period time horizon of size $\frac{t}{n}$, the contaminations in the observed return process are $O_p(1)$. This result, which is a consequence of the assumptions of price and noise, implies that longer period returns are less contaminated by noise than shorter period returns. On the other hand, the size of the contaminations does not decrease in probability with the distance between subsequent time stamps. Provided sampling does not occur between high frequency price updates, the rounding of recorded prices to a grid (i.e., price discreteness) alone makes this feature of the set-up presented above empirically compelling.

Sometimes the dependence structure of the microstructure noise process can be simplified. Specifically, one can modify the assumption of noise as follows:

1: The microstructure frictions in the price process $u_{t_i}^{'}$ are i.i.d. mean zero.

2: The $u_{t_i}^{'}$ are independent of the $Y_{t_i}^{*'}$ for all i and all n.

If the microstructure noise contaminations in the price process, u_{t_i}, are i.i.d., then the noise returns, ϵ_{t_i}, display an MA(1) structure and are negatively correlated [Ban05a]. Importantly, the noise return moments do not depend on n, i.e., the number of observations over t or, equivalently, the sampling frequency. This is an important feature of the MA(1) model which has been exploited in recent works on volatility estimation. For example, Bandi and Russell [Ban05b] provide an alternative bias-correction in both the correlated noise case and in the MA(1) case. The subsampling and averaging methodology proposition of Zhang et al. [Zha05], indeed, consistently estimate integrated volatility in the presence of MA(1) microstructure noise. The MA(1) model, as typically justified by bid-ask bounce effects [Rol84], is known to be a realistic approximation in decentralized markets where traders arrive in a random fashion with idiosyncratic price setting behavior, the foreign exchange market being a valid example. It can also be a good approximation in the case of equities when considering transaction prices or even quotes posted on multiple exchanges.

While the abovementioned assumptions about noise like usual literature are basis for experiments in the present dissertation, the realized estimators will be examined also under some non-Gaussian noise processes. Next section explains some backgrounds for these non-Gaussian noise processes.

2.7 Self-similar noise processes

Self-similar processes are of great interest in modeling heavy-tailed and long-memory phenomena. Self-similar processes are invariant in distribution under suitable translations of time and scale. They are important in probability theory because of their connection to limit theorems. Lamperti [Lam62] uses the term semi-stable in order to underline that the role of self-similar processes among stochastic processes is analogous to the role of stable distributions among all distributions. A process $\{X(t)\}_{t\geq 0}$ is called self-similar [Lam62] if for some $H > 0$ and for every $a > 0$,

$$X(at) \overset{d}{=} a^H X(t),$$

where $\overset{d}{=}$ denotes equality of all finite-dimensional distributions of the processes on the left and right. The process $X(t)$ is also called H-self-similar process and the parameter H is called the self-similarity index or Hurst exponent. Weron et al. [Wer05] argue that if we interpret t as time and $X(t)$ as space then above equation tells us that every change of time scale $a > 0$ corresponds to a change of space scale a^H. The bigger H, the more dramatic is the change of the space coordinate. The equation, indeed, means a scale-invariance of the finite-dimensional distributions of $X(t)$. This property of a self-similar process does not imply the same for the sample paths. Therefore, pictures trying to explain self-similarity by some zooming in or out on one sample path, are, by definition, misleading. In contrast to the deterministic self-similarity, the self-similarity of stochastic processes does not mean that the same picture repeats itself exactly as we go closer. It is rather the general impression that remains the same.

Rachev et al. [Rac07] demonstrate that the normality as a distributional model for asset returns has been rejected conjecturing that financial return time series behave like non-Gaussian stable processes. The latter commonly are referred to as stable Paretian distributions or Levy stable distributions. In fact, Rachev et al. [Rac05a] explain that Stable Paretian is used to emphasize that the tails of the non-Gaussian stable density have Pareto power-type decay. Levy stable is used in recognition of the seminal work of Paul Levys' introduction and characterization of the class of non-Gaussian stable laws.

2.7.1 Fractional Gaussian noise

Sun, Rachev and Fabozzi [Sun06] discuss that "fractal processes (self-similar processes) are tightly connected with the analysis of long-range dependence. Many of the interesting self-similar processes have stationary increments". A process $\{X(t)\}_{t\geq 0}$ is said to have stationary increments if for any $b > 0$,

$$[X(t+b) - X(b)] \overset{d}{=} [X(t) - X(0)].$$

The fractional Brownian motion $\{B_H(t)\}_{t\geq 0}$ has the integral representation

$$B_H(t) = \int_{-\infty}^{\infty} [(t-u)_+^{H-1/2} - (-u)_+^{H-1/2}] dB(u), \tag{2.22}$$

where x_+=max(x,0) and $B(u)$ is a Brownian motion. It is H-self-similar stationary increments (H-sssi) and it is the only Gaussian process with such properties for $0 < H < 1$ [Sam94]. The classic Brownian motion $B(t)$, used by Einstein and Smoluchowski, is simply a special case of the fractional Brownian motion when $H = 1/2$.

In modeling of long-memory phenomena, the stationary increments of H-self-similar processes are of special interest since any H-self-similar process with stationary increments $\{X(t)\}_{t\in R}$ induces a stationary sequence $\{Y_j\}_{j\in Z}$, where $Y_j = X(j+1) - X(j)$ and $j = ..., -1, 0, 1, ...$. The sequence Y_j corresponding to the fractional Brownian motion is called fractional Gaussian noise [Mer03]. It is called a standard fractional Gaussian noise if varY_j=1 for every $j \in Z$. The fractional Gaussian noise has some remarkable properties. If H=1/2, then its autocovariance function $r(k) = R(0, k) = 0$ for $k \neq 0$ and hence it is the sequence of independent identically distributed (i.i.d.) Gaussian random variables. The situation is quite different when $H \neq 1/2$, namely the Y_j's are dependent and the time series has the autocovariance function.

2.7.2 Fractional stable noise

Rachev and Mittnik [Rac00] give a very detailed description on the stable Paretian models in finance. The stability property is highly desirable for asset returns. In the context of portfolio analysis and risk management, the linear combinations of different return series follow again a stable distribution. In fact, the Gaussian law shares this feature, but it is only one particular member of a huge class of distributions, which also allows for skewness and heavy tails.

Fractional Brownian motion can capture the effect of long-range dependence. But it has less power to capture heavy tailedness [Sun07]. The existence of abrupt discontinuities in financial data, combined with the empirical observation of sample excess kurtosis and unstable variance, confirms the stable Paretian hypothesis identified by Mandelbrot [Man83]. It is natural to introduce the stable Paretian distribution in self-similar processes in order to capture both long-range dependence and heavy tailedness. There are many different extensions of fractional Brownian motion to the stable distribution. The most commonly used extension of the fractional Brownian motion to the α-stable case is the linear fractional stable motion (also called the fractional Levy stable motion). Samorodnitsky and Taqqu [Sam94] define the process $\left\{Z_\alpha^H(t)\right\}_{t\in R}$ by the following integral representation

$$Z_\alpha^H(t) = \int_{-\infty}^{\infty} [(t-u)_+^{H-1/\alpha} - (-u)_+^{H-1/\alpha}] dZ_\alpha(u), \tag{2.23}$$

where $Z_\alpha(u)$ is a symmetric Levy α-stable motion. The integral is well defined for $0 < H < 1$ and $0 < \alpha \leq 2$ as a weighted average of the Levy stable motion $Z_\alpha(u)$ over the infinite past with the weight given by the above integral kernel denoted by $f_t(u)$.

The process $Z_\alpha^H(t)$ is the H-sssi. Assume that H-self-similarity follows from the above integral representation and the fact that the kernel $f_t(u)$ is d-self-similar with $d = H - 1/\alpha$, when the integrator $Z_\alpha(u)$ is $1/\alpha$-self-similar. This implies [Wer05] the following important relation

$$H = d + \frac{1}{\alpha}.$$

The process $Z_\alpha^H(t)$ is reduced to the fractional Brownian motion if one sets α=2. When H=$1/\alpha$, then the Levy α-stable motion is obtained which is an extension of the Brownian motion to the α-stable case. Contrary to the Gaussian case (α=2), the Levy α-stable motion ($0 < \alpha < 2$) is not the only $1/\alpha$-self-similar Levy α-stable process with stationary increments (this is true for $0 < \alpha < 1$ only). The increment process corresponding to the fractional Levy stable process is called a Fractional Stable Noise (FSN). By analogy to the case of α=2, fractional stable noise has the long-range dependence when $H > 1/\alpha$ and the negative dependence when $H < 1/\alpha$. If $H = 1/\alpha$, the increments of fractional Levy stable motion are i.i.d. symmetric α-stable variables. We note that there is no long-range dependence when $0 < \alpha \leq 1$ because H is constrained to lie in the interval (0, 1).

Some properties of these processes have been discussed in Maejima and Rachev [Mae87], Rachev and Mittnik [Rac00], Rachev and Samorodnitsky [Rac01], and Samorodinitsky and Taqqu [Sam94].

2.7.3 Simulation of the noise processes

A fast Fourier transform method for synthesizing approximate self-similar sample paths for Fractional Gaussian Noise has been presented by Paxson [Pax97]. The method is fast and appears to generate close approximations to true self-similar sample paths. A simulation procedure based on this method that overcomes some of the practical implementation issues has been prescribed by Bardet et al. [Bar03b]. Sun et al. [Sun07] explain procedure. "The procedure follows these steps:

1. Choose an even integer M. Define the vector of the Fourier frequencies $\Omega = (\theta_1, ..., \theta_{M/2})$, where $\theta_t = 2/M$ and compute the vector $F = f_H(\theta), ..., f_H(\theta_{M/2})$, where

$$f_H(\theta) = \frac{1}{\pi} sin(\pi H)\Gamma(2H+1)(1-cos\theta)\sum_{i\in N} \left|2\pi t+\theta\right|^{-2H-1},$$

and $fH(\theta)$ is the spectral density of fractional Gaussian noise.

2. Generate $M/2$ i.i.d. exponential (exp(1)) random variables $E_1, ..., E_{M/2}$ and $M/2$ i.i.d. uniform ($U[0,1]$) random variables $U_1, ..., U_{M/2}$.

3. Compute $Z_t = exp(2i\pi U_t)\sqrt{F_t E_t}$, for $t = 1, ..., M/2$.

4. From the M-vector: $\widetilde{Z} = (0, Z_1, ..., Z_{(M/2)-1}, Z_{M/2}, \bar{Z}_{(M/2)-1}, ..., \bar{Z}_1$.

5. Compute the inverse FFT of the complex Z to obtain the simulated sample path."

Using the Fast Fourier Transform (FFT) algorithm, Stoev and Taqqu [Sto04] provide an efficient method for simulation of a class of processes with symmetric α-stable (SαS) distributions, namely the linear fractional stable motion (LFSM) processes. The paths of the LFSM process are generated by using Riemann-sum approximations of its SαS stochastic integral representation. They introduce parameters $n, N \in \aleph$ and express the fractional stable noise $Y(t)$ as

$$Y_{n,N}(t) := \sum_{j=1}^{nN} \left(\left(\frac{j}{n}\right)_+^{H-(1/\alpha)} - \left(\frac{j}{n}-1\right)_+^{H-(1/\alpha)} \right) L_{\alpha,n}(nt-j), \tag{2.24}$$

Where $L_{\alpha,n}(t) := M_\alpha((j+1)/n) - M_\alpha(j/n)$, and $j \in \Re$. The parameter n is mesh size and the parameter M is the cut-off of the kernel function. The authors use the Fast Fourier Transformation (FFT) algorithm for approximating $Y_{n,N}(t)$. Consider the moving average process $Z(m), m \in \aleph$,

$$Z(m) := \sum_{j=1}^{nM} g_{H,n}(j) L\alpha(m-j), \tag{2.25}$$

where

$$g_{H,n}(j) := \left(\left(\frac{j}{n}\right)_{+}^{H-(1/\alpha)} - \left(\frac{j}{n} - 1\right)_{+}^{H-(1/\alpha)} \right) n^{-1/\alpha}, \tag{2.26}$$

and $L_\alpha(j)$ is the series of i.i.d. standard stable Paretian random variables. Since $L_{\alpha,n}(j) \overset{d}{=} n^{-1/\alpha} L_\alpha(j)$, where $j \in \Re$, then the latter equations (2.25) and (2.26) imply that $Y_{n,N}(t) \overset{d}{=} Z(nt)$, for $t = 1, ..., T$. Let $\tilde{L}_\alpha(j)$ be the $n(N+T)$-periodic with $\tilde{L}_\alpha(j) := L_\alpha(j)$, for $j = 1, ..., n(N+T)$ and let $\tilde{g}_{H,n}(j) := g_{H,n}(j)$, for $j = 1, ..., nN$, $\tilde{g}_{H,n}(j) := 0$, for $j = nN+1, ..., n(N+T)$. Then

$$\{Z(m)\}_{m=1}^{nT} \overset{d}{=} \left\{ \sum_{j=1}^{n(N+T)} \tilde{g}_{H,n}(j) \tilde{L}_\alpha(n-j) \right\}_{m=1}^{nT}, \tag{2.27}$$

because for all $m = 1, ..., nT$, the summation in equation (2.25) involves only $L_\alpha(j)$ with indices j in the range $-nN \leq nT-1$. Using a circular convolution of the two $n(N+T)$-periodic series $\tilde{g}_{H,n}$ and $\tilde{L}_\alpha$ computed by using their Discrete Fourier Transforms (DFT), the variables $Z(n)$, $m = 1, ..., nT$ (i.e., the fractional stable noise) can be generated.

2.8 Volatility and correlation modeling

2.8.1 Continuous-time volatility modeling

For evaluation of the volatility estimators, the GARCH approach of volatility modeling looks like a suitable framework. The GARCH(1,1) model has emerged as a work-horse for modeling volatility in financial markets, as it tends to provide a simple approximation to the main statistical features of the return series across a wide range of assets. For the simulation part of the present work, we advocate Andersen et al. [And98] and Andersen et al. [And99] and establish the diffusion foundation for analysis. Following Nelson [Nel90] and Drost and Werker [Dro96], the continuous-time diffusion limit of the GARCH(1,1) model is given by

$$dp_t = \sigma_t dW_{1,t}, \tag{2.28}$$

$$d\sigma_t^2 = \theta(\omega - \sigma_t^2)dt + (2\lambda\theta)^{1/2}\sigma_t^2 dW_{2,t}, \tag{2.29}$$

where $W_{1,t}$ and $W_{2,t}$ denote independent standard Brownian motions and where $\omega > 0$, $\theta > 0$ and $\lambda \in (0,1)$. According to Drost and Werker [Dro96] the discretely sampled returns from the continuous-time process defined by Eqs. (2.28) and (2.29), satisfy the weak GARCH(1,1) model

$$\sigma^2_{(n),t} = \psi_n + \alpha_n r^2_{(n),t-1/n} + \beta_n \sigma^2_{(n),t-1/n}, \tag{2.30}$$

with n observations per day t, where $\sigma^2_{(n),t} \equiv P_{(n),t-1/n}(r^2_{(n),t})$ denotes the best linear predictor of $r^2_{(n),t}$. The relationship between the discrete-time parameters ψ_n, α_n, and β_n and the continuous-time parameters ω, θ, and λ may be obtained in closed form, as outlined by Drost and Werker [Dro96]. Hence, in this weaker interpretation a GARCH(1,1) specification for any discrete frequency is compatible with the diffusion in Eqs. (2.28) and (2.29), and in this sense the setting provides a coherent framework for analysis of the model forecasts at different sampling intervals. Now, following Baillie and Bollerslev [Bai92] the h-period linear projection from the weak GARCH(1,1) model with returns that span $1/n$ day(s) is conveniently expressed as

$$\begin{aligned} P_{(n),t}(r^2_{(1/h),t+h}) &= P_{(n),t}\left(\left[\sum_{j=1,\ldots,nh} r_{(n),t+j/n}\right]^2\right) \\ &= \sum_{j=1,\ldots,nh} P_{(n),t}(r^2_{(1/h),t+j/n}) \\ &= \sum_{j=1,\ldots,nh} \left[\sigma^2_{(n)} + (\alpha_n+\beta_n)^j(\sigma^2_{(n),t} - \sigma^2_{(n)})\right] \end{aligned}$$

$$= nh\sigma^2_{(n)} + (\alpha_n - \beta_n)\left[1 - \alpha_n - \beta_n^{nh}\right] \times \left[1 - \alpha_n - \beta_n\right]^{-1}(\sigma^2_{(n),t} - \sigma^2_{(n)}), \tag{2.31}$$

where $\sigma^2_{(n)} \equiv \psi_n(1-\alpha_n-\beta_n)^{-1}$. Different realized volatility alternatives previously defined can be cast in the volatility term in the model formulated above.

2.8.2 Correlation modeling

Realized correlation is in essence a model-free estimator. Following Meddahi [Med02] and Barndorff-Nielsen and Shephard [Bar04b] one can estimate the difference between realized correlation and corresponding actual correlation estimators and then study and evaluate the consistency and unbiasedness of the realized correlation estimators. This difference indicates an error in estimation of integrated or true estimators. Meddahi [Med02] and Barndorff-Nielsen and Shephard [Bar04b] write actual correlation as

$$\frac{\int_{t-1}^{t} \Sigma_{xy}(s)ds}{\sqrt{\int_{t-1}^{t} \Sigma_{x}^{r}(s)ds \int_{t-1}^{t} \Sigma_{y}^{r}(s)ds}}, \tag{2.32}$$

where $\int_{t-1}^{t} \Sigma_{xy}(s)ds$ and $\int_{t-1}^{t} \Sigma_{x}^{r}(s)ds$ represent actual covariation between assets x and y and variation of order r for asset x respectively. If $r = 1$, then the above expression is actual correlation regarding to the absolute based volatility. If $r = 2$ which is equivalent to $\int_{t-1}^{t} \Sigma_{xx}(s)ds$ and $\int_{t-1}^{t} \Sigma_{yy}(s)ds$ for variations, then we have actual squared based correlation. Obviously both $\widehat{RPCOR}_{xy}$ and $\widehat{TSPCOR}_{xy}$ are estimating the same integrated correlation, i.e., $IRPCOR_{xy}$.

Chapter 3

Experiments on volatility and correlation estimators

3.1 Simulation experiments under normality

3.1.1 Simulation of volatility estimators

Having a suitable framework of volatility model prescribed in previous chapter, it is easy to evaluate consistency and unbiasedness of alternative estimators on finite samples. First the convergence of estimators is evaluated under normality assumption of the microstructure noise process. Then, the estimators are examined under more realistic non-Gaussian microstructure noise assumptions.

Simulation scheme : Advocated by Andersen et al. [And98] and Andersen et al. [And99], our theoretical assessment of the performance of the discrete-time GARCH(1,1) approximation in Eq. (2.31) for predicting the subsequent realized volatility models defined by the stochastic volatility diffusion in Eqs. (2.28) and (2.29) rely on numerical means. More specifically, sample-path realizations of the underlying stochastic volatility diffusion are obtained via simulation using an Euler scheme. The estimator TSAV of order 1 ($r = 1$) is compared with the estimators RA and TSRV. For theoretical evaluation of estimators, RMSE and Bias statistics are used. However, to accommodate the heteroskedastisity in forecast errors, following Andersen et al. [And99], we compute the corresponding heteroskedastisity adjusted statistics by

$$HRMSE = E[(1 - \text{estimator}/I(P)V)^2]^{1/2},$$

$$HBias = E[(1 - \text{estimator}/I(P)V)],$$

where I(P)V is the integrated (power) volatility for corresponding volatility estimator.

Advocated by Barndorff-Nielsen, Hansen, Lunde and Shephard [Bar04c], Bandi and Russell [2005b], Zhang et al. [Zha05] and Hansen and Lunde [2006] and recalling our assumptions about the price, return and market microstructure noise processes, we assume that the market microstructure noise, ϵ, follows a Gaussian process and is small. We assume a pure noise (i.e., noise is i.i.d and independent with the efficient price). Specifically, we set $(E\epsilon^2)^{1/2} = 0.01$, i.e., the standard deviation of the noise is 1% of the value of the variable of interest.

According to our daily real world data sample (will be described later in section 3.3) of NASDAQ from December 17, 2002 to January 31, 2007, we approximate the parameters of continuous-time GARCH(1,1) models (2.28) and (2.29) equal to θ=0.0173 (Std Error=0.0042, T stat.=4.15), ω=8.17e-007 (Std Error=2.9e-007, T stat.=2.74), and λ=0.974 (Std Error=0.0065, T stat.=150.84) by MLE parameter estimation. The GARCH parameters are fixed at the values obtained from maximum likelihood estimation based on real daily observations of NASDAQ for simulations. Random variables for simulations are generated by MATLAB. For generating data, we assume 250 working days a year as usual and generate data at different frequencies according to table 3.1. The simulations are based on 5 years of data samples and 7,000 sample paths (realizations). For two-scale based estimators we allow the sampling points to be regularly allocated. For all three alternative estimators, we assume equally distance sampling interval.

3.1.2 Results of volatility simulations

The results of Monte Carlo simulations of volatility estimators in terms of HRMSE and HBias are contained in Table 3.1 [Saf07a]. The table shows how the estimators converge to the integrated variation across frequencies when the sampling interval is going to diminish. Comparing the rows reveals asymptotic convergence in probability distribution. Moreover, it is clear from the table that how different estimators behave. A comparison between the columns of the table reveals convergence capability of the volatility estimators.

Table 3.1: Results of volatility simulation (displaying values*10,000)

Frequency at every	TSRV		RA		TSAV	
	HRMSE	HBias	HRMSE	HBias	HRMSE	HBias
60 min.	9.6922	0.4589	2.5602	0.4235	2.4511	0.3920
30 min.	9.0308	0.4127	2.5613	0.4243	2.0030	0.3917
15 min.	8.7350	0.4004	2.5674	0.4407	1.7182	0.3891
5 min.	7.8221	0.3612	2.7985	0.8939	1.3482	0.3197
1 min.	6.8790	0.2526	3.0134	1.4023	0.6201	0.1826
30 sec.	6.5025	0.2032	3.0141	1.7709	0.5618	0.1692
10 sec.	6.1408	0.1897	3.0375	2.2048	0.4803	0.1032
5 sec.	5.7112	0.1715	3.0392	2.6963	0.4403	0.0723

The table simply shows that the realized power volatility of order 1 (RA) is not an unbiased and consistent estimator of integrated power variation as the frequency increases. This finding is consistent with the literature around microstructure noise. Even the bias and variance of estimator is increasing across the frequencies caused by the market microstructure frictions. As a result, RA estimator is not a consistent estimator and according to the table, it obviously diverges. However, the two-scale estimators, as expected by the subsampling approach in line with Zhang et al. [Zha05], are consistent and unbiased estimator for the corresponding targets, i.e., Integrated Volatility (in our special case of order 2 for TSRV, i.e., r=2) and Integrated Power Volatility (in our special case of order 1 for RA and TSAV, i.e., r=1). A comparison between the two-scale estimators gives some informations. At each frequency the TSRV suffers from higher variation compared to the TSAV. In terms of bias the same condition holds. This implies that the rate of convergence and consistency differs between estimators, although both estimators gradually and eventually converge. Therefore, the two-scale absolute based estimator converges faster.

The difference between the two-scale estimators in convergence rate may be akin to the fact that absolute based estimators are inherently somewhat immune against large values in a relative sense. There is, indeed, empirical evidence that absolute returns based models produce better volatility forecasts than models based on squared returns. For example Taylor [Tay86], Ding, Granger, and Engle [Din93], McKenzie [Mak99], Ederinton and Guan [Ede00], Forsberg and Ghysels [For05], Andersen et al. [And06], and Ghysels et al. [Ghy06], show that a squared transformation of returns in squared based models of volatility in turn reinforces large values in return series and hence they appear in volatility series as larger values. Thus, TSRV seems theoretically not to be robust against large values, meanwhile construction of volatility based on realized power variation with absolute transformation is somewhat robust to rare values [Bar04d], in particular in case of $r = 1$ (or absolute based variation).

3.1.3 Simulation of correlation estimators

Simulation scheme : Considering different realized correlation estimators modeled in (2.12), (2.13), and (2.17), and actual correlation modeled in (2.32), the corresponding differences or errors can be studied. Specifically, we consider correlation estimators of (2.13) and (2.17) where $r = 1$, i.e., absolute-based correlations which are more common estimators. They are appeared in Table 3.2 with RACOR and TSACOR notations respectively. For two-scale based estimators, we allow the sampling points to be regularly allocated. Moreover, we consider again the assumptions about price, return and microstructure noise processes explained previously, and set the noise to be i.i.d. and equal to $(E\epsilon^2)^{1/2} = 0.01$. For evaluation, the HRMSE and HBias metrices will be utilized. The data are generated assuming 250 working days a year at different frequencies contained in table 3.2. The simulations are based on 5 years of data samples and 20,000 sample paths.

3.1.4 Results of correlation simulations

According to the results [Saf07a] contained in Table 3.2, the estimators have different convergence behavior at presence of the noise. The $\widehat{RSCOR}_{xy}$ and $\widehat{RACOR}_{xy}$ estimators not only do not converge but also explicitly diverge with increasing frequency where the noise intervene. Nevertheless, the two-scale estimator indicates consistency and unbiasedness due to its bias-corrector in addition to averaging procedure for reducing variation at higher frequencies. Essentially this estimator is included by a bias-corrector and averaging and therefore it converges as compared to other correlation estimators in terms of bias and variation. Obviously from the table it is seen that as the frequency increases for the $\widehat{TSACOR}$ correlation estimator, i.e., the number of intraday observations increases ($n \to \infty$), the HRMSE and HBias decrease.

Table 3.2: Results of correlation simulation (displaying values*10,000)

Frequency at every	RSCOR		RACOR		TSACOR	
	HRMSE	HBias	HRMSE	HBias	HRMSE	HBias
60 min.	10.1206	0.1164	3.2301	0.0673	3.0921	0.0662
30 min.	10.1259	0.1183	3.2523	0.0680	3.0905	0.0589
15 min.	10.1347	0.1197	3.2748	0.0687	3.0718	0.0540
5 min.	11.5602	0.1236	3.3407	0.0698	2.9743	0.0537
1 min.	12.9831	0.1405	3.3756	0.0786	2.7643	0.0449
30 sec.	13.3125	0.1427	3.3904	0.0820	2.7215	0.0416
15 sec.	13.8347	0.1436	3.4018	0.0835	2.6713	0.0401
5 sec.	13.8461	0.1477	3.4029	0.0849	2.5908	0.0393

While the realized volatility literature assumes that the market microstructure noise follows an i.i.d. process, Sun, Rachev, and Fabozzi [Sun07] conclude that an ARMA-GARCH model assuming a fractional stable noise outperforms other ARMA-GARCH models assuming independent and identically distributed (i.i.d.), stable, generalized Pareto, generalized extreme value and fractional Gaussian noises. They examine the model under different assumptions about noise and empirically compare them based on data of 27 German stocks included in DAX. As a result, non-Gaussian assumption about microstructure noise, and specifically the fractional stable noise seems more realistic.

3.2 Simulation experiments under non-normality

Motivated by the results of Sun, Rachev, and Fabozzi [Sun07], the realized volatility and correlation estimators under different assumptions about microstructure noise were simulated [Saf08b]. Particularly, the impact of different assumptions about the microstructure noise including i.i.d. or White noise, stable noise, fractional Gaussian noise, and fractional stable noise on accuracy and especially on the bias in the estimation are investigated and compared.

3.2.1 Simulation of volatility estimators

Simulation scheme : Now simulation and evaluation schemes of the realized volatility estimators $\widehat{TSRV}, \widehat{RA}$ and $\widehat{TSAV}$ are explained. The estimators are cast in the continuous-time volatility model (2.31) with different microstructure noise assumptions including the i.i.d noise, fractional Gaussian noise, stable noise, and fractional stable noise. For evaluation the bias and variance metrices of estimations are calculated. The Gaussian noise is set equal to 1% of the value of the variable of interest. Random variables for simulations are generated according to minute-by-minute frequency for 4 years assuming 252 working days a year. For generating the non-Gaussian noises, the described procedures in 2.7.3 are followed based on minute-by-minute frequency observations of CAC 40 and FTSE 100 explained in [Saf08b] subsection 5.1. The number of sample paths for all simulations is 15,000 realizations. Regarding to two-scale based estimators, we allow the sampling observations to be regularly allocated. For all three alternative estimators, we assume equally distance sampling interval. Three estimators including TSRV, RA, and TSAV (which the two latter estimators are the RP and TSPV estimators of the power 1) are compared. But what is more important here is the comparison of different microstructure noise assumptions. For evaluation of estimators, we use RMSE and Bias statistics.

3.2.2 Results of simulation for volatility

The results of Monte Carlo simulations in terms of RMSE and Bias of estimation are contained in Table 3.3. A horizontal comparison of different volatility estimators is an indication of different estimation power of the estimators. In general, the TSAV estimator yields less variation and bias than others at minute-by-minute simulation frequency. This is in line with the results of [Saf07a].

Table 3.3: Results of volatility simulations assuming different noise (simulated based on CAC data)

Assumptions	TSRV		RA		TSAV	
	RMSE	Bias	RMSE	Bias	RMSE	Bias
White noise	0.001612	1.686e-005	0.001125	1.749e-005	0.000853	1.383e-005
Fractional Gaussian noise	0.001487	1.675e-005	0.001060	1.688e-005	0.000820	1.347e-005
Stable noise	0.001489	1.676e-005	0.001072	1.692e-005	0.000821	1.349e-005
Fractional Stable noise	0.001306	1.502e-005	0.000885	1.644e-005	0.000819	1.318e-005

Table 3.4 provides the results of simulations using the simulated noise values based on 1 minute frequency real data of FTSE. The table yields the same results for CAC data. However, a vertical comparison of the values contained in the tables is more important purpose.

Both tables report that the GARCH(1,1) model assuming the fractional stable noise outperforms the other models assuming the White noise, Fractional Gaussian noise, and Stable noise. Among the models with different noise assumptions, the model assuming White noise has the worst results of fitting.

These results are consistent with those of obtained by Sun, Rachev, and Fabozzi [Sun07]. As it is emphasized in [Sun07], the results imply that the real microstructure noise is better characterized by the fractional stable noise. In fact, the i.i.d. noise, which is usually assumed in the realized volatility literature, is not the real case.

Table 3.4: Results of volatility simulations assuming different noise (simulated based on FTSE data)

Assumptions	TSRV		RA		TSAV	
	RMSE	Bias	RMSE	Bias	RMSE	Bias
White noise	0.002859	1.113e-004	0.002125	1.750e-004	0.001847	1.002e-005
Fractional Gaussian noise	0.002853	1.107e-004	0.002104	1.750e-004	0.001823	9.993e-006
Stable noise	0.002731	9.685e-005	0.001873	1.607e-004	0.001765	9.885e-006
Fractional Stable noise	0.002548	7.071e-005	0.001546	1.418e-004	0.001508	8.453e-006

3.2.3 Simulation of correlation estimators

Simulation scheme : For simulation of the correlation estimators we use again the same nonparametric scheme in the previously described subsection 2.8.2. To solve the problem of non-synchronous trading effect, in [Saf08b] we applied the scheme in model (2.18) for estimating the realized covariances. Regarding to the microstructure noise, the size of the White noise is set again equal to 1% of the generated data. Other types of the noise have been previously simulated and used for volatility estimators based on minute-by-minute real CAC and FTSE data. They will be exploited again here for correlation simulation. Other conditions for volatility simulations are held.

3.2.4 Results of simulation for correlation

Table 3.5 indicates that the White noise which is usually assumed when modeling of the realized volatility and correlation estimators, possesses the highest errors in terms of RMSE and Bias of estimation. Instead, models based on the Fractional Stable noise assumption have the best performance of estimation. This fact is true for the three correlation estimators.

Table 3.5: Results of correlation simulation

Assumptions	RSCOR$_{xy}$		RACOR$_{xy}$		TSACOR$_{xy}$	
	RMSE	Bias	RMSE	Bias	RMSE	Bias
White noise	0.003841	5.478e-005	0.003341	4.654e-004	0.002760	2.783e-005
Fractional Gaussian noise	0.003839	5.478e-005	0.003317	4.652e-004	0.002760	2.744e-005
Stable noise	0.003802	5.463e-005	0.003321	4.657e-004	0.002764	2.346e-005
Fractional Stable noise	0.003754	5.418e-005	0.003206	4.574e-004	0.002608	1.837e-005

The results suggest that the market microstructure noise includes some self-similarity and it does not follow a simple white process. In fact, the market microstructure noise possesses some long memory dependence structure as well as heavy tailedness.

3.3 Distributional and dynamic behaviors of estimators

It could be interesting to empirically study some more common and important distributional and dynamic behaviors and properties of the realized volatility and correlation estimators. In this section, such the behaviors are experimentally evaluated and compared for different estimators on real world data sets.

3.3.1 Data description

Volatility can be estimated arbitrarily well from an arbitrary short span of data, provided that returns are sampled sufficiently frequently. This suggests the use of high frequency data. Note that realized volatility at day t is based on information within day t as opposed to volatility from, e.g., GARCH models that depends on information up to day $t-1$. For this section, the empirical analysis is based on NASDAQ100 and DAX30 stock index data at every 5 minute frequency. Our sample indices cover longer than 4 years from December 17, 2002 to January 31, 2007 with 250 official business days a year. This period includes 1029 trading days totally with 76410 observations. Both indices encompass those equities with a high degree of liquidity and the related markets are very active markets. The variables of interest in our analysis are returns defined from time to time of aforementioned index values. We define return of an index by $Y_{t_{i+1}} - Y_{t_i} = log(Y_{t_{i+1}}) - log(Y_{t_i})$, which is the return from holding the index time t_i to time t_{i+1}, when Y_{t_i} is the observed index value.

Table 3.6: Basic statistics and test of return of indices

Statistic	NASDAQ100	DAX30
Minimum	-3.52e-02	-2.79e-02
Maximum	5.71e-02	2.43e-02
Mean	7.06e-06	5.44e-06
Median	0.00e+00	5.19e-06
Sum	5.37e-01	4.14e-01
Variance	1.87e-06	1.78e-06
Skewness	1.37e+00	-3.36e-01
Kurtosis	8.00e+01	3.62e+01
Jarque-Bera test	2.2e-16	2.2e-16

Table 3.6 describes some basic statistics of the time series. According to the table, positive mean and median returns explain an average positive return trend. In particular, excess kurtosis (peakedness) with skewness (asymmetry) shows obviously our time series depart from normality. In view of the fact that the kurtosis coefficient of distributions are much higher than 3 (coefficients equal to 80 for NASDAQ and to 36 for DAX) for a standard normal distribution, therefore distributions of our series are leptokurtic. Leptokurtosis is a sign of heavy tail in a distribution. This implies that there is a higher probability for extreme events in data than that is normally distributed. Negative coefficient of skewness for DAX (-0.336) series describes that the probability density function is negatively skewed and therefore that is asymmetric to the left side. However, this coefficient indicates an asymmetry to the right side for NASDAQ (1.37). At last, the Jarque-Bera test[1] for normality simply reveals that the investigated time series with P-value equal to 2.2e-16 do not form a normal distribution. Rachev et al. [Rac05b] write that "empirical evidence does not support the assumption that many important variables in finance follow a normal distribution".

3.3.2 Distributional behaviors

The realized volatility and correlation are observable and measurable in nature. Therefore, based on different construction of the volatility and correlation estimators, we can characterize their distributional behaviors with relying on conventional statistical procedures. Comparison of empirical distributions of different measures can be simply implemented.

Realized volatility estimators : Figure 3.1 depicts the time series of different realized volatility estimators. In fact, the figure unveils that volatility, constructed by realized measures, is time-varying. This is in contrast to the conventional approach which views the volatility as constant. Time-varying property of realized volatility suggests that volatility appears to change over time as a time series and hence may include some dynamic properties. Large values appeared in squared based volatility series are obvious. A comparison of realized volatilities with a traditional constant variance using Tables 3.6 and 3.7 detects that all realized measures tend to report volatility higher than a constant value. The variances of two indices in Table 3.6 are much smaller than the mean of realized volatilities in Table 3.7. However, the mean of TSRV volatility is smaller than that of others.

Andersen et al. [And01b] found that the distributions of realized daily variances are skewed to the right and are leptokurtic for exchange rate data. Consistent with this finding, Andersen et al. [And01a] find that the unconditional distributions of realized variances are highly right-skewed for stock exchange data. In line with these findings, our volatility series in Table 3.7 are rightward.

[1] The Jarque-Bera test of normality is the most widely used procedure for testing normality of economic time series returns. The algorithm provides a joint test of the null hypothesis of normality in that the sample skewness equals zero and the sample kurtosis equals three.

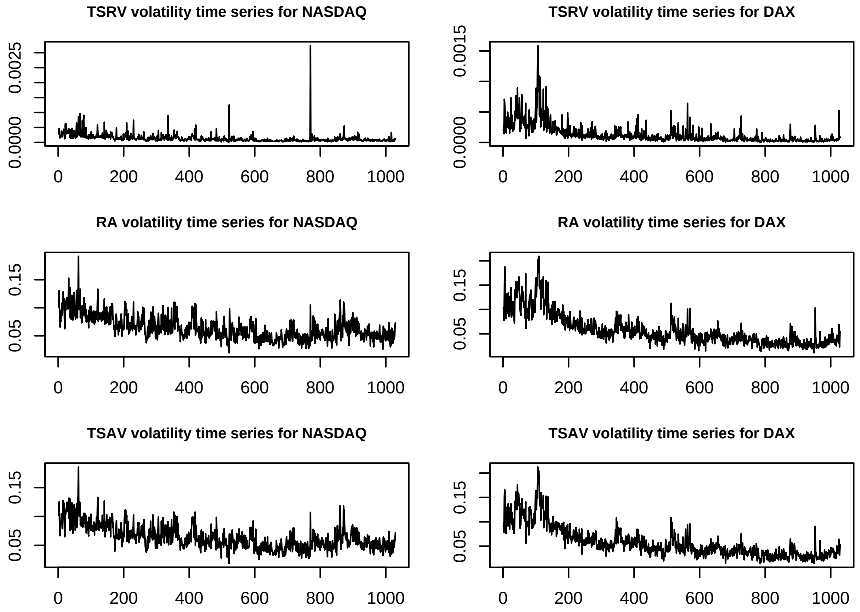


Figure 3.1: Time series of realized volatility measures constructed based on two-scale squared, absolute, and two-scale absolute transformations. Evidently volatility is viewed time-varying. Large values in squared based volatility are obvious.

Table 3.7: Basic statistics and tests of realized volatility measures

Statistic	NASDAQ100			DAX30		
	TSRV	RA	TSAV	TSRV	RA	TSAV
Mean	1.38e-04	6.39e-02	6.38e-02	1.32e-04	5.76e-02	5.76e-02
Median	1.00e-04	5.96e-02	6.00e-02	7.41e-05	4.86e-02	4.77e-02
Variance	2.37e-08	4.58e-04	4.42e-04	2.59e-08	1.06e-03	1.04e-03
Skewness	9.44e+00	9.62e-01	9.53e-01	3.16e+00	1.52e+00	1.52e+00
Kurtosis	1.64e+02	1.44e+00	1.28e+00	1.46e+01	2.44e+00	2.42e+00
Jarque-Bera	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16	2.2e-16
A.D-F of 20 lags	0.01	0.01	0.01	0.05	0.23	0.24
A.D-F of 30 lags	0.01	0.07	0.06	0.06	0.22	0.20

Four moments of realized volatility measures plus median are included in Table 3.7. The Skewness and kurtosis of the measures determine in more detail, none of the measures possess exactly a normal distribution. In terms of the Jarque-Bera test of normality reported in the table, none of the measures hold a normal distribution. With p-value 2.2e-16, normality for all measures is significantly rejected. However, a relative comparison may include informative facts. The Skewness coefficients of absolute based measures (RA and TSAV) are almost one-tenth of those of squared based measure (TSRV) in case of NASDAQ and a half in case of DAX, and closer to that of the normal distribution. All the coefficients are positive meaning that the distributions are skewed rightward. On the other hand, while the coefficients of kurtosis for absolute-based estimators are relatively close to 3 for a normal distribution, this coefficient is equal to 164 and 14.6 for NASDAQ and DAX for TSRV estimator. Higher kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations.

The main reason for difference among the distributions of volatility series may most likely be akin to different sensitivity to jumps or large values. Andersen et al. [And01a] argue that the squared returns approach, over the relevant return horizon, provides model-free unbiased estimates of the ex post realized volatility. Unfortunately, however, squared returns are also a very noisy volatility indicator and hence do not allow for reliable inference regarding the true underlying latent volatility. In fact, squared based volatility measures reinforce jumps and large values in return series to appear larger in realized volatility series as extreme values to shape a positive heavy tail. However, realized volatility constructed by absolute transformation when the power of transformation is around 1 ($r = 1$ in realized power volatility) seems relatively to be more monotonous. These arguments are also confirmed by figure 3.2. All distributions, especially distribution of TSRV are asymmetric to the right side. The shapes show long right tail. Presence of big jumps in squared based volatility is obviously evident in figure 3.2. As such, these jumps lead the time series of measure to form a longer right tail in distribution. The jumps or extreme values cause positive skewness coefficient (to the right side) in Table 3.7.

In general, all daily time series of estimators shape a kind of non-normal distribution meanwhile absolute based series seem nearer to normal. These findings are in agreement with that of Andersen et al. [And01a]. Of course, this phenomena was well documented as the fact of markets where the distribution of relative price changes is strongly nonGaussian: these distributions can be characterized by power law tails with an exponent close to 3 for rather liquid markets. Emerging markets have even more extreme tails, with an exponent that can be less than two - in which case the volatility is infinite [Bou02]. In spite of negative skewness in returns of DAX, implying existence of big negative jumps, figure 3.3 exhibits both tails of volatility distributions are positive. This is quite clear, because when constructing volatility measures, we restricted values to be positive by squared or absolute transformations. In essence, realized volatility is positive like the constant volatility. Meanwhile, differences in distribution of measures are obviously appeared. Plots depict that how much can the points match bisector line. The size of discrepancy from bisector represents deviation from normality. Obviously here, absolute based series have smaller jumps, and therefore closer distribution to the normal one.

Realized correlation estimators : The co-movement of world equity markets is often used as a barometer of economic globalization and financial inte-

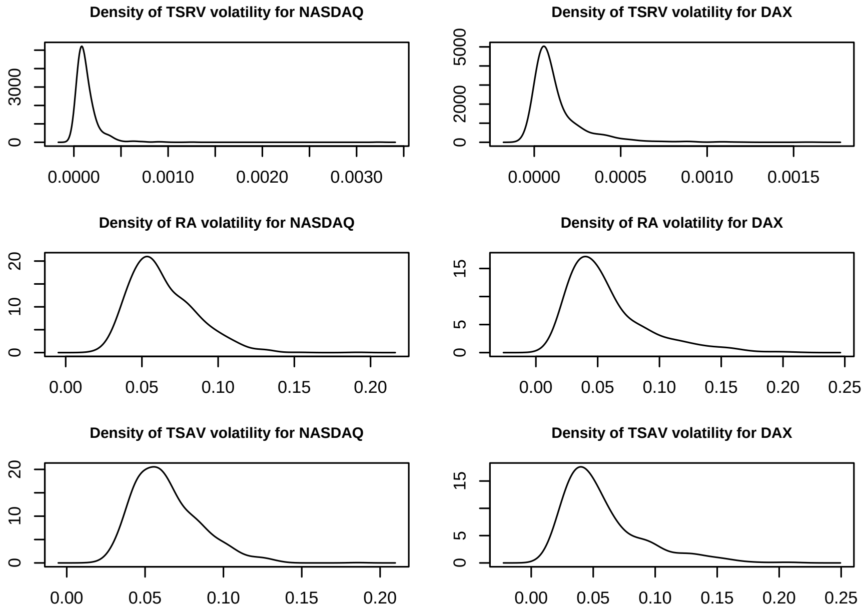


Figure 3.2: Kernel density distribution of different realized volatility series seem skewed rightward. However, the shapes are not the same. Asymmetry degree seems different among volatility series. Heavy tail in distributions is remarkable.

gration. Analyzing such co-movement is important for risk diversification of an international portfolio [Sun08]. The most commonly used measure to analyze comovements and cointegration among international equity markets is correlation analysis. Therefore, the realized correlation estimators are applied on 5 minute frequency stock indices to this end. Applying the correlation estimators (2.12, 2.13 and 2.17), our study is focused on correlation between the returns of the NASDAQ and DAX markets. A two-scale squared based correlation has not been yet appeared in literature. Thus, we proceed with the squared based correlation in (2.12). The models (2.13) and (2.17) are estimated with regard to $r = 1$.

In our analysis, both indices belong to very developed, active and liquid markets. A main difference of our correlation with that of traditional analysis includes time-variation, and hence probably nonlinearity and dynamics of dependence structure over time. In figure 3.4, some distributional properties of different realized correlation measures are graphically embodied. First row plots explicitly imply that realized correlation series, against classical formulation of correlation, are time-varying, what is a profound property of many financial phenomena, and that they may have some nonlinearities and dynamics. Their kernel density can be found in second row of the plots. As Andersen et al. [And01a] and Andersen et al. [And01b] reported, the distributions of realized squared correlation between stocks and between exchange rates are ap-

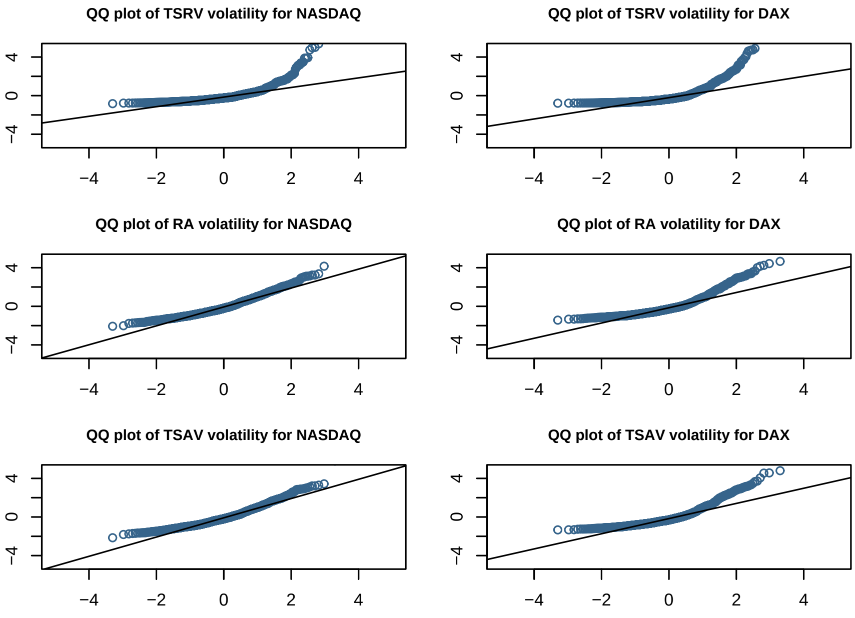


Figure 3.3: Quantile Quantile-normal plots compare the empirical volatility series to the theoretical distribution. The x and y axes of the plots represent theoretical quantiles and sample quantiles of the series respectively. More matching to the straight line means more approaching to normality.

proximately normal. In our experiment here, the distribution of $RSCOR_{xy}$ is normal too. However, we found that the distributions of $RACOR_{xy}$ and $TSACOR_{xy}$ look like non-normal in the shape.

Some basic distribution-related statistics of realized correlations are reported in Table 3.8. All the correlation estimators have negative mean. The $RSCOR_{xy}$ correlation shows stronger dependence between returns on average, while the estimators $RACOR_{xy}$ and $TSACOR_{xy}$ show a weaker degree of dependence between markets on average over our time period. These two latter correlations behave relatively more stable over the time, since they have much less variance than the $RSCOR_{xy}$ correlation has. Analogously, the $TSACOR_{xy}$ tends to be more stable than $RACOR_{xy}$ in terms of variance over the time. Comparing both mean and variance of different correlation estimators, we observe that $RSCOR_{xy}$ correlation shows a stronger (based on the mean value), and at the same time, more unstable (based on variance) dependence between markets. The $RSCOR_{xy}$ correlation is slightly skewed to the right, what is not consistent with a common sense. But $RACOR_{xy}$ and $TSACOR_{xy}$ estimators are negatively skewed so that the degree of skewness in $TSACOR_{xy}$ is much bigger than that of $RACOR_{xy}$ estimator. P-value of Jarque-Bera test for null normality is statistically significant at the 5 percent level for $RSCOR_{xy}$ correlation (0.464). Based on rather high skewness of absolute based correlation estimators, we find

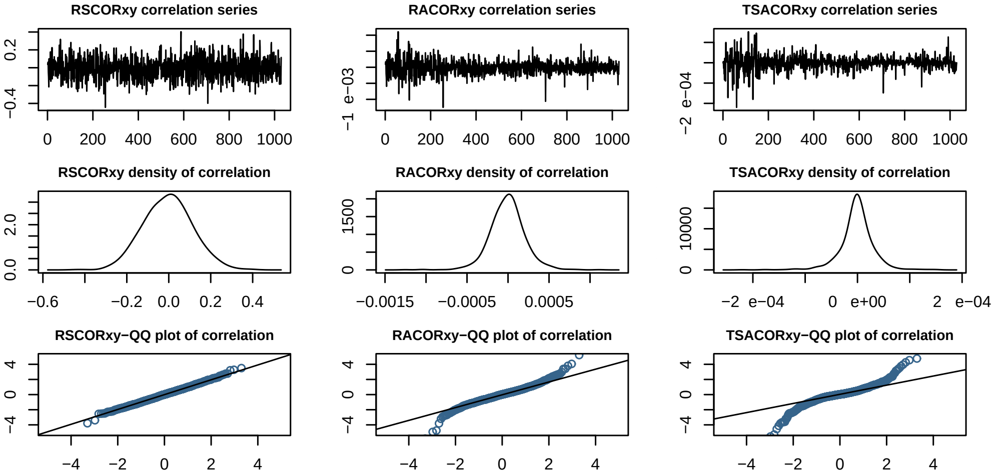


Figure 3.4: Distributional properties of realized correlations are graphically embodied. Evidently realized correlations, based on first row plots, fluctuate over the time. The squared based correlation possesses almost a symmetric density, while density of others are skewed. These findings are more informatively supplemented by QQ-normal plots. The x and y axes of the plots represent theoretical quantiles and sample quantiles of the series respectively. Longer left tail is documented in dependence structure of the absolute-based correlations.

that asymmetry is present in the conditional realized correlation distributions.

Table 3.8: Basic statistics and test of realized correlations

Statistic	RSCOR$_{xy}$	RACOR$_{xy}$	TSACOR$_{xy}$
Mean	-3.87e-03	-7.78e-06	-1.36e-06
Median	-2.78e-03	-4.53e-06	-6.04e-07
Variance	1.35e-02	4.56e-08	1.04e-09
Skewness	5.41e-02	-1.21e-01	-6.09e-01
Kurtosis	1.49e-01	3.32e+00	6.43e+00
Jarque-Bera test	0.464	2.2e-16	2.2e-16

In fact, when the relationship between the markets follows the $RSCOR_{xy}$ dependence structure, then based on our data, upside comoves are greater than downside ones. In contrast, negative asymmetry in $RACOR_{xy}$ and $TSACOR_{xy}$ correlations conveys that downside comoves are greater than upside comoves between markets. Asymmetry here has an important message: negative shocks in returns have greater impact than positive shocks in NASDAQ market on DAX market, if we assume DAX is affected by NASDAQ based on New York effect. According to behavioral finance theory, key observations made in behavioral finance literature include the lack of symmetry between decisions to acquire or keep resources, called colloquially the bird in the bush paradox, and the strong loss aversion or regret attached to any decision where some emotionally valued resources might be totally lost. In prospect theory, loss aversion refers to the tendency for people to strongly prefer avoiding losses than acquiring gains. “A bird in the hand is worth two in the bush”, reflects individual or social biases probably leading to negative asymmetry. These observations are confirmed and more informatively completed by QQ-normal plots in third row of the figure 3.4. Longer negative tail in multivariate absolute-based realized correlations can be documented in such a way that the extreme values are usually populated in the left tail of distributions.

Rachev et al. [Rac05b] discuss that “correlation is one particular measure of dependence among many. Another approach is to model dependency using copulas”. For continuous multivariate distribution functions, the univariate margins and the multivariate dependence structure can be separated, and the dependence structure can be represented by a copula [Emb03]. As Embrechts et al. [Emb03] argue, copulas provide a natural way to study and measure dependence between random variables. The popularity of linear correlation stems from the ease with which it can be calculated and it is a natural scalar measure of dependence in elliptical distributions (with well known members such as the multivariate normal and the multivariate t-distribution). However most random variables are not jointly elliptically distributed, and using linear correlation as a measure of dependence in such situations might prove very misleading. The copula-based Kendall’s rank correlation coefficient provides some advantages over the use of linear dependence in the elliptical distributions. On the other hand, the strength of non-linear dependence at joint extreme levels, which is a property of some copulas, may be informed by the tail dependence coefficients. Although the Student-t copula exhibits tail dependence against Gaussian copula which dose not, as an elliptical copula, Student-t copula show symmetric tails. Since elliptical distributions are radially symmetric, the coefficient of upper and lower tail dependence are equal. In financial applications, there is usually a stronger dependence between big losses (e.g. market crashes) than between big gains [Emb03] as it was shown in absolute-based realized correlation above. Clearly, such asymmetries cannot be modeled with elliptical copulas. In contrast to elliptical copulas, all commonly encountered Archimedean copulas (e.g., Frank, Gumbel, Clayton and mixtures) have closed form expressions. Their popularity also stems from the fact that they allow for a great variety of different dependence structures. Copulas differ not so much in the degree of association they provide, but rather in which part of the distributions the association is strongest. Through the choice of copula, a good deal of control can be exercised over what parts of the distributions are more strongly associated.

Based on the copula theory, Patton [Pat04] constructs models of the time-varying dependence structure that allow for different dependence during bear markets than bull markets. Stock returns appear to be more highly correlated during market downturns than during market upturns. For evaluating asymmetry in dependence, Patton [Pat06] considered an extension of theory of copulas to allow for conditioning variables, and employed it to construct flexible models of conditional dependence structure in the joint density of the DM/USD and

Yen/USD exchange rates. Two different copulas were estimated: the copula associated with the bivariate normal distribution and the symmetrized Joe-Clayton copula, which allows for general asymmetric dependence. Time variation in the dependence structure between the two exchange rates was captured by allowing the parameters of the two copulas to vary over the sample period. He found evidence that the mark-dollar and yen-dollar exchange rates are more correlated when they are depreciating against the dollar than when they are appreciating. On equity returns, Longin and Solnik [Lon01], using extreme value theory to model the multivariate distribution tails, derive the distribution of extreme correlation for a wide class of return distributions. Empirically, they reject the null hypothesis of multivariate normality for the negative tail, but not for the positive tail. They report that correlation increases in bear markets, but not in bull markets.

Cizeau, Potters and Bouchaud [Ciz01] studied the correlations between stock returns, conditioning on absolute market returns, fraction of positive/negative returns, and large individual stock returns-quantile correlations and exceedance correlations as different indicators, all based on a simple but comprehensive non-Gaussian one-factor model. Assuming that the return of every stock is the sum of random independent (non-Gaussian) factors, they decompose a return into a market part and a residual part in the model. In a generic factor model, the residuals are combinations of all the factors except the market and are therefore independent of it. Their model, which accounts for fat tail effects, explains the correlations between stock returns increase in high volatility periods, and in particular explains the level and asymmetry of empirical exceedance correlations. Conditioning on exceedance correlations, they study more specifically how extreme stock returns are correlated between themselves. For this, they consider a pair of individual stocks which their normalized returns are larger as well as smaller than a level. Large and small returns correspond to extreme correlations. They find that the correlation grows with extreme returns. However, correlation between extreme negative pair of returns is larger than correlation between extreme positive pair of returns.

Consistent with the Archimedean copulas and one-factor model, the proposed multivariate absolute-based realized correlations exhibit non-linearity in dependence structure of variables. Evidently, based on the sign of skewness in Table 3.8, the $RACOR_{xy}$ and $TSACOR_{xy}$ correlations are consistent with common sense. The simulation study on correlation in the previous sections suggests that $TSACOR_{xy}$ is an unbiased estimator for the true correlation. As such, this estimator consistent with other approaches mentioned above, approximates the true correlation between equity markets to be negatively asymmetric, while the squared based correlation estimates the true correlation to be symmetric as it has been emphasized by Andersen et al. [And01a] and Andersen et al. [And01b]. The squared based realized correlation with the normal distribution for stock index real data leads to an evident bias, whereas absolute based estimators consistent with the other models of dependence possess the skewed distribution for asymmetric real dependence structure and hence it does not exhibit a bias with this regard.

3.3.3 Dynamic behaviors

The underlying efficient market hypothesis (EMH) has enormous philosophical and mathematical appeal. The strong form of the hypothesis is that investors have access to all relevant information, and that this is fully reflected by the current market price. The random arrival of new (independent and identically Gaussian-distributed) information causes traders' expectations to change. This is then translated into a Brownian motion in a Gaussian distribution of (log) price returns. There are variations upon this reasoning, for example, invoking arbitrageurs or informed investors who quickly exploit any inefficiencies due to noise traders or uninformed investors but the pricing outcome is the same. One of the refutable implications of the EMH is the Gaussian distribution of returns. Actual distributions however are sufficiently non-Gaussian so as to require better explanations and mathematical models than provided by the EMH. For a detailed discussion, see for example Rachev et al. [Rac05b].

With many plausible EMH violations (and the impossibility of performing controlled experiments with real markets), it is extremely difficult to draw conclusions regarding the chain of cause and effect from statistical analysis alone. However, these analysis have identified a set of stylized facts that appear to be prevalent across asset classes independent of trading rules, geography or culture. These include the lack of linear correlations in price returns over all but the shortest timescales, non-Gaussianity, excess kurtosis (fat tails) in the price return distribution, volatility clustering (ARCH-effects) short- and long-range dependence, temporal dependence of the tail behavior, skewed distributions, temporal dependence of the tail behavior, and heteroskedasticity. See for example Rachev and Mittnik [Rach00]. Some finer details have also been revealed, most notably the existence of power-law scalings and estimates of the exponents.

The class of models addressing the stylized facts is an attempt to provide a framework within which to study systematically the effects of various, simple, EMH violations. The hope is that the insights gained will result in a greater theoretical understanding of the operation of markets. Now, issues related to dynamic features of the volatility and correlation estimators are extracted by detailed examinations, with particular focus on the long memory.

Realized volatility estimators : Simply behavior of autocorrelation of financial time series has been studied by many researchers. It has been investigated to see how the autocorrelations decay over the lags. Ding, Granger and Engle [Din93] and Andersen and Bollerslev [And97] argue that the autocorrelations of squared and absolute returns decay at a much slower hyperbolic rate over longer lags. Consistent with this finding, in the figures 3.5 and 3.6, on the left panels, autocorrelation functions based on NASDAQ (figure 3.5) and DAX (figure 3.6) for realized volatilities have been drawn.

The top plot in figure 3.5 belongs to TSRV volatility which differs remarkably from the two others. While the autocorrelation of TSRV volatility, computed from NASDAQ data, is going to be insignificant around 140 lags, this serial correlation for realized absolute volatilities is still significant around 240

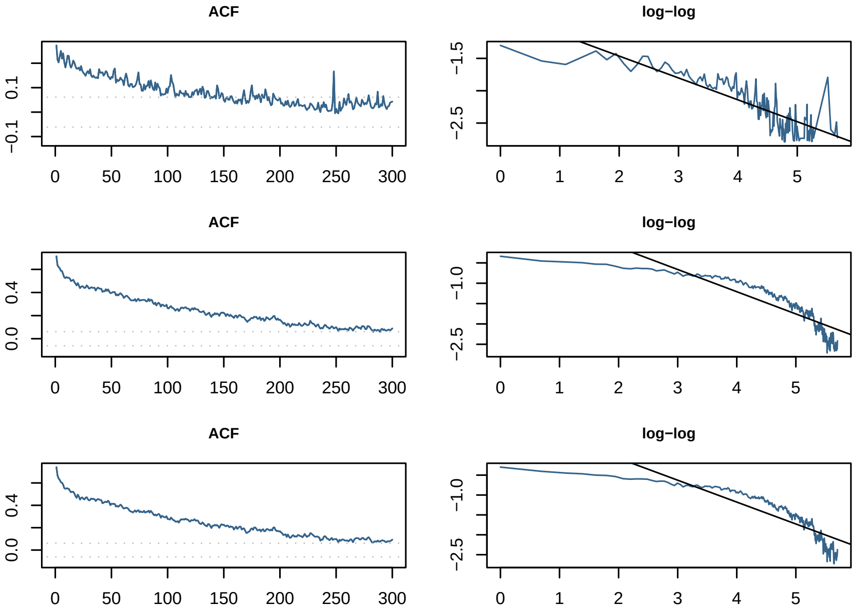


Figure 3.5: Autocorrelation function and long memory autocorrelation function plots (ACF and log-log) of volatilities, computed based on NASDAQ data. For all functions of both kind of autocorrelation function and long memory autocorrelation function, the number of lags is arbitrarily equal to 300. The top row belongs to TSRV measure, the middle to RA, and the bottom to TSAV. Left plots are autocorrelation functions and right ones are long memory autocorrelation functions. The axes on the right plots are log of that of the left plots. Estimated Hurst exponent (self-similarity parameter) in long memory plots for TSRV, RA, and TSAV are respectively equal to 0.65, 0.74, and 0.76.

lags equivalent to almost one calendar year. This exhibits a quite considerable difference. Also in figure 3.6, the difference is observable. Autocorrelation of squared based measure estimated from DAX data in figure 3.6, lasts only up to 170 lags. Instead, autocorrelation of absolute based measures dies away to be insignificant around 260 lags, more than one year. These important findings imply that a shock in the volatility process will have a long-lasting impact.

An autocorrelation may be a sign of long memory process. The autocorrelation function (ACF) can be completed to have more meaningful sense by long memory autocorrelation function. This may be a very interesting signature for series dynamics, and it sets a pretty high hurdle for any financial model to meet. Usually it is spoken of a long memory behavior, if the decay in the ACF is slower than a hyperbolic rate, i.e., the correlation function decreases algebraically with increasing (integer) lag. Thus it makes sense to investigate the decay on a double logarithmic scale and to estimate the decay exponent. Graphically, if the time series exhibits long memory behavior, it can easily be seen as a straight line in the plot on the right panels of figures 3.5 and 3.6. This double logarithmic plot is displayed and a linear regression fit is done from

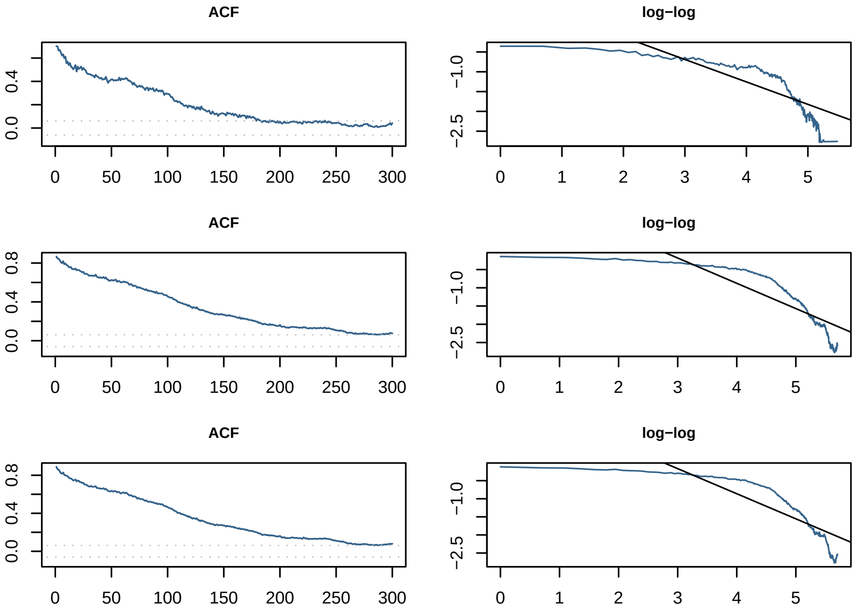


Figure 3.6: Autocorrelation function and long memory autocorrelation function plots of volatilities, computed based on DAX data. The axes on the right plots are log of that of the left plots. For all functions of both kind of autocorrelation function and long memory autocorrelation function, the number of lags is arbitrarily equal to 300. The top row belongs to TSRV measure, the middle to RA, and the bottom to TSAV. Left plots are autocorrelation functions and right ones are long memory autocorrelation functions. Estimated Hurst exponent (self-similarity parameter) in long memory plots for TSRV, RA, and TSAV are respectively equal to 0.72, 0.83, and 0.85.

which the intercept and slope are calculated. Corresponding long memory plots of volatility series in figures 3.5 and 3.6 show a slow decay for the estimators, meanwhile absolute based estimators explicitly indicate longer memory. Log-log plots for RA and TSAV estimators exhibit a more straight and smoother curve. The curve in this plot for TSRV estimator fluctuates around a straight line. So, the volatility estimators include long memory behavior as a dynamic stylized fact of market. Finding long memory in realized volatility of NASDAQ and DAX here is consistent with those empirical experiments on tickers included in NASDAQ by Andersen et al. [And01a] and in DM/US dollar and Yen/US dollar exchange rates by Andersen et al. [And01b] both at a 5-minute frequency.

Andersen et al. [And03] suggest that the long-run dynamics of realized logarithmic volatilities can be well approximated by a fractionally-integrated long-memory process. Following this suggestion, we estimate a multivariate model for the logarithmic realized volatilities. Advocated to Andersen et al. [And03] we estimate the degree of fractional integration, d, obtained using the Geweke and Porter-Hudak (GPH) [Gew83] log-periodogram regression estimator. The GPH estimator is based on the regression equation using the peridogram function as an estimate of the spectral density. For more information see [Gew83]. The estimated d's in volatility series are equal to 0.37 (0.023), 0.41 (0.028), 0.42 (0.029) and to 0.39 (0.021), 0.42 (0.031), 0.43 (0.035) in the structure of TSRV, RA, and TSAV volatility estimators for NASDAQ and DAX respectively. The values in parenthesis are corresponding asymptotic standard error. Andersen et al. [And03] reported estimated d equal to 0.387, 0.413, and 0.43 respectively for DM/USD, Yen/USD, and Yen/DM in daily realized squared volatility.

The presence of slow autocorrelation decay may be an indication of the presence of a unit root, as in the integrated GARCH model of Engle and Bollerslev [Eng86]. In Table 3.7, p-values for Augmented Dickey-Fuller test of nonstationarity with 20 and 30 augmentation lags are reported. The test soundly rejects the null hypothesis for all of the volatility series with 20 lags in NASDAQ, but significantly accepts all volatilities in DAX at 5 percent level. With 30 lags, the test accepts nonstationarity of volatility series except TSRV in NASDAQ. This suggests that almost all volatilities follow a unit root variety.

Taylor [Tay86] analyzes 40 series of returns and observes that the sample autocorrelations of absolute returns seem to be larger than the sample autocorrelations of squares. If Y_t, $t = 1, ..., T$, is the series of returns and $r_\theta(k)$ denotes the sample autocorrelation of order k of $|y_t|^\theta$, $\theta > 0$, the Taylor Effect can be defined as $r_1(k) > r_\theta(k)$ for any $\theta \neq 1$. So, the autocorrelations of absolute returns to the power of theta reach their maximum at $\theta = 1$. In figure 3.7, plots depict autocorrelations as a function of the exponent θ for each lag from 1 to maximum lag (here in this figure, 10 lags). In the case that the above formulated hypothesis is supported, all the curves should peak at the same value around $\theta = 1$. The plots related to the absolute-based volatility estimators exhibit more number of points from 10 points met the vertical line of $\theta = 1$.

Consider again a self-similar process, $X(at) \overset{d}{=} a^H X(t)$, described in section 2.7. Statistically, self-similar means that the statistical properties for the entire data set are the same for sub-sections of the data set. In other words, the self-similar dimension of fractional integration is invariant to the horizon. Estimating the Hurst exponent for a data set provides a measure of whether the data is a pure random walk or has underlying trends. The values of Hurst exponent range between 0 and 1. A Hurst exponent value in range $0.5 < H < 1$ indicates persistent behavior (e.g., a positive autocorrelation). If the Hurst exponent is $0.5 < H < 1$, the process will be a long memory process. Furthermore, the closer H is to 1, the stronger the dependence of the process is. Data sets like this are sometimes referred to as fractional Brownian motion. A value of 0.5 indicates a true random walk (a Brownian time series with no autocorrelation). The fractal dimension is directly related to the Hurst exponent for a statistically self-similar data set. In a random walk there is no correlation between any element and a future element. A small Hurst exponent has a higher fractal dimension and a rougher surface. A larger Hurst exponent has a smaller fractional dimension and a smoother surface. A Hurst exponent value $0 < H < 0.5$ will exist for a time series with anti-persistent behavior (or negative autocorrelation). Here an increase will tend to be followed by a decrease and inversely. This behavior is

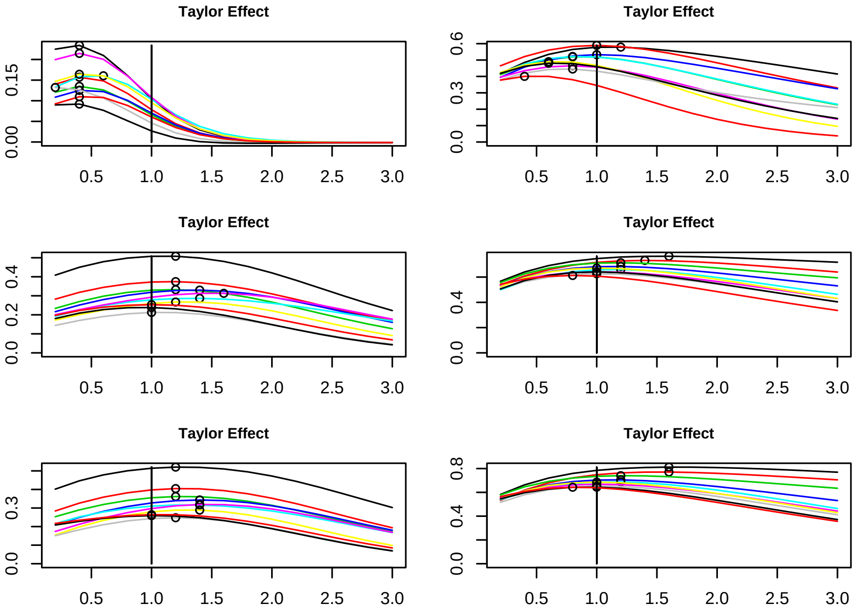


Figure 3.7: The Taylor effect plot indicates that Taylor Effect exists in a series, where the curves peak at the value around $\theta = 1$ which is on the x axis. Left panel belongs to NASDAQ and the right one to DAX. First, middle, and bottom rows belong to TSRV, RA, and TSAV volatilities respectively. The absolute-based estimators show better the effect.

sometimes called mean reversion. There are several estimators that are used to estimate the value of the Hurst parameter. Some more common methods include Absolute value method, Variance method, R/S method, Periodogram method, Whittle estimator, Variance of residuals, and Abry-Veitch method. The Hurst exponents estimated by R/S method are equal to 0.65, 0.74, and 0.76 and to 0.72, 0.83, and 0.85 in the structure of TSRV, RA, and TSAV volatility estimators for NASDAQ and DAX respectively. Estimated based on the Whittle method [Whi63], the values of Hurst are equal to 0.63, 0.72, and 0.75 and to 0.73, 0.79 and 0.83 respectively. Although estimated Hurst exponents by two methods are not exactly the same, differences in various series are meaningfully kept and generally the methods endorse each other. In fact, consistent with Andersen et al. [And01a], there is the strong evidence to suggest that volatility is a long memory process.

Realized correlation estimators : An existence of regularities in the patterns and temporal dependencies of comovements across the stock markets is studied here. The existence of such the regular behaviors implies the dynamics of correlation series. Now we draw regular patterns in correlations series. Considering Figure 3.8, a long autocorrelation (ACF plot) has been completely disappeared for squared based correlation now. Based on the long memory autocorrelation plot in Figure 3.8, a temporal dependence for $RSCOR_{xy}$ estimator can not be reported. The degree of fractional integration, d, is estimated equal to 0.05 (0.004). Of course, the $RACOR_{xy}$ and $TSACOR_{xy}$ estimators seem to keep still their dynamic properties. These results are consistent with the results of Taylor [Tay86]. He observed that the sample autocorrelations of absolute returns seem to be larger than the sample autocorrelations of squares. The absolute based correlation estimators exhibit long memory dependence with Hurst exponents based on R/S method equal to 0.58 and 0.59 and based on the Whittle estimator are 0.57 and 0.58 respectively. The degree of fractional integration, d, here equals 0.31 (0.019) and 0.33 (0.022) respectively.

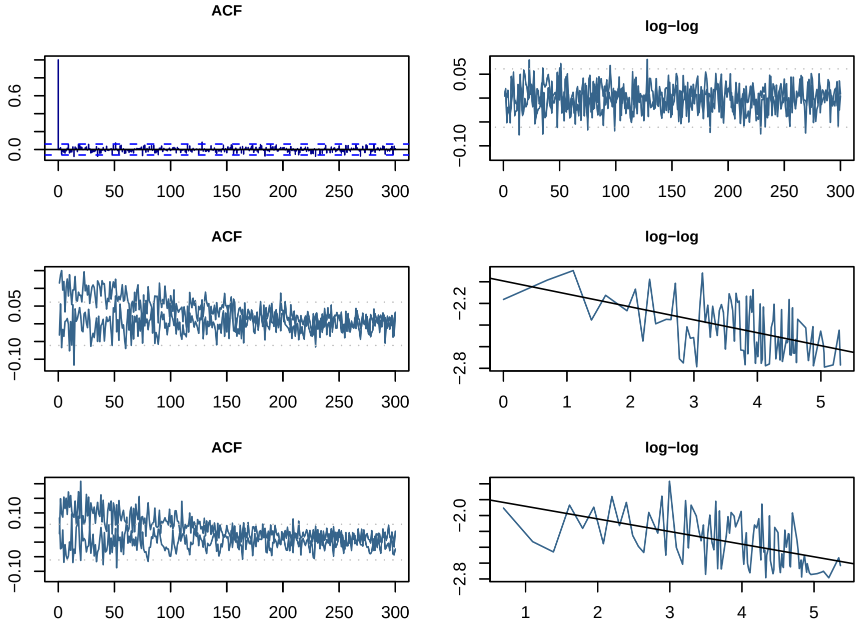
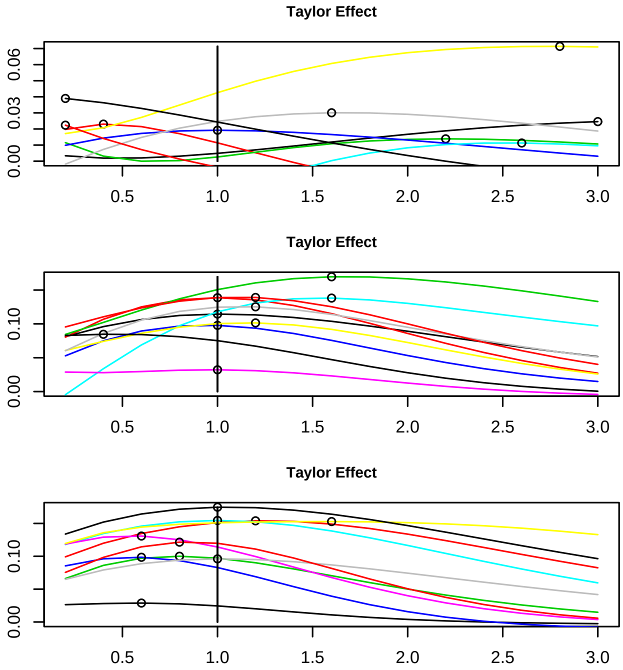


Figure 3.8: Autocorrelation function and long memory autocorrelation function plots (ACF and log-log) of correlations between NASDAQ and DAX. The axes on the right plots are log of that of the left plots. For all functions of both kind of autocorrelation function and long memory autocorrelation function, the number of lags is arbitrarily equal to 300. The top row belongs to $RSCOR_{xy}$, the middle to $RACOR_{xy}$, and the bottom to $TSACOR_{xy}$ correlation. Left plots are autocorrelation functions and right ones are long memory autocorrelation functions. Estimated Hurst exponent (self-similarity parameter) in long memory plots for $RACOR_{xy}$, and $TSACOR_{xy}$ are respectively equal to 0.58 and 0.59. The $RSCOR_{xy}$ exhibits no long memory.

A glance at Figure 3.9 reveals that only one of the curves in the Taylor effect plot of $RSCOR_{xy}$ correlation reaches at its pinnacle around $\theta = 1$. Instead, the Taylor effect appears considerably in $RACOR_{xy}$ and $TSACOR_{xy}$ series.



Taylor Effect

Taylor Effect

Taylor Effect

Figure 3.9: In spite of the Taylor effect plot for the $RSCOR_{xy}$ correlation where one of lags peaks around $\theta = 1$; the effect appears in the $RACOR_{xy}$, and $TSACOR_{xy}$ correlation series remarkably.

3.4 Is volatility really forecastable?

The Hurst exponent promises a gleam of hope for predictability in financial markets which seemingly sound unpredictable at all, under efficient market hypothesis; since it shows well regularity in chaotic and stochastic behaviors of particles or agents. Peters [Pet96] suggests that a Hurst exponent value between $0.5 < H < 1.0$ shows that the efficient market hypothesis is incorrect. Returns are not randomly distributed. There is some underlying predictability.

Poon and Granger [Poo03] discuss that financial market volatility is clearly forecastable. The debate is on how far ahead one could accurately forecast and to what extent could volatility changes be predicted. This conclusion does not violate market efficiency since accurate volatility forecast is not in conflict with underlying asset and option prices being correct. The option implied volatility being a market based volatility forecast has been shown to contain most information about future volatility. The supremacy among historical time series models depends on the type of asset being modeled.

However, the problem of estimating the Hurst exponent itself, involves a complex problem of accurate calculation. Different methods of estimating Hurst exponent do not yield exactly the same result. Moreover, we are not certain about a specific variable of interest to be a representative for predictability of the market. In our investigation here, volatility reflects regularity in the markets. But as reported by many, for an example Ding, Granger, and Engle [Din93], original prices do not show the regularity, at least by Hurst exponent, among statistics. It is now well established that the stock market returns themselves contain little serial correlation which is in agreement with the efficient market theory. But this empirical fact does not necessarily imply that returns are independently identically distributed as many theoretical financial models assume. It is possible that the series is serially uncorrelated but is dependent. The stock market data is especially so, since if the market is efficient, a stock's price should change with the arrival of information. If information comes in bunches, the distribution of the next return will depend on previous returns although they may not be correlated. As the return period increases, the return values reflect longer trends in the time series. Perhaps the higher Hurst exponent value is actually showing the increasing upward or downward trends. This does not, by itself, show that the efficient market hypothesis is incorrect. Even if we accept the idea that a non-random Hurst exponent value does damage to the efficient market hypothesis, estimation of the Hurst exponent seems of little use when it comes to time series forecasting. At best, the Hurst exponent tells us that there is a long memory process. The Hurst exponent does not provide the local information needed for forecasting. Nor can the Hurst exponent provide much of a tool for estimating periods that are less random, since a relatively large number of data points are needed to estimate the Hurst exponent. For example a constant Hurst exponent over time also does not seem a sound and reasonable conclusion. However, this statistic can be useful in analyzing the behavior of market models.

Fama [Fam98] believes that the efficient market hypothesis "survives the challenge from the literature on long-term return anomalies. Consistent with the market efficiency hypothesis that the anomalies are chance results, apparent overreaction to information is about as common as underreaction, and post-event continuation of pre-event abnormal returns is about as frequent as post-event reversal. Most important, consistent with the market efficiency prediction that apparent anomalies can be due to methodology, most long-term return anomalies tend to disappear with reasonable changes in technique".

3.5 Association between volatility and correlation

The realized volatility is modeled according to univariate distributions. However, addressing associations between multivariate distributions may yield some important facts. There may exist an association between estimators of variation and covariation. Key issues relevant in financial economic applications include, for example, whether and how $TSAVstd$, and $TSACORxy$ move together. These questions are difficult to answer using conventional volatility models, but they are relatively easy to address using the realized volatilities and correlations. There is indeed strong evidence that realized volatilities and correlations move together in a manner broadly consistent with latent factor structure. That is, realized correlation is itself correlated with realized volatility, which Andersen et al. [And01a] call volatility effect in correlation (VIC).

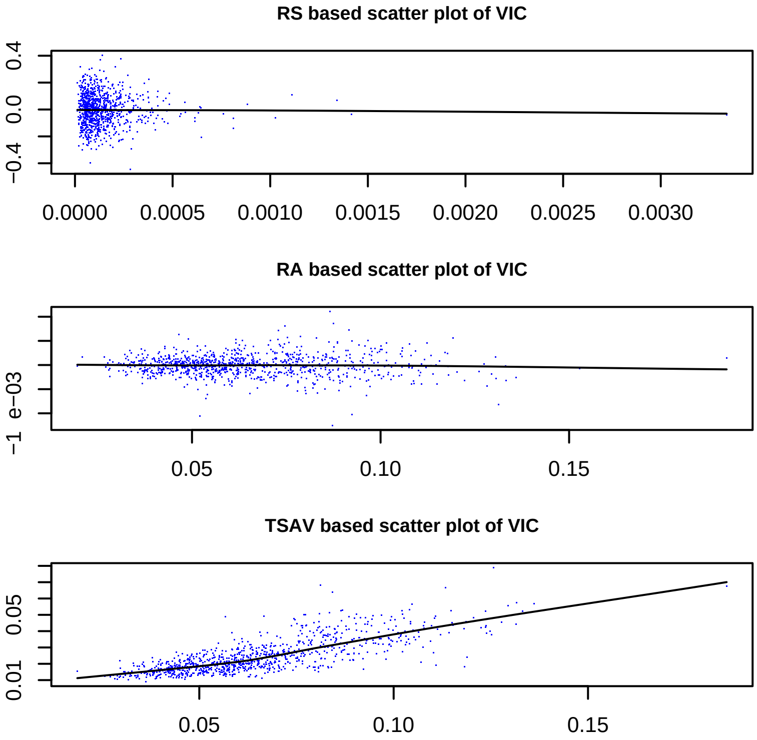


Figure 3.10: Scatter plots of different Volatility Effect in Correlation are uncovered to document association between realized volatility and correlation of NASDAQ and DAX. Axes x and y represent realized volatility and correlation respectively. The smooth line indicates the trend of association.

Now the question is how to model the association. In turn, modeling the association produces the problem of measurement and estimation. Andersen et al. [And01b] estimate the kernel density of correlations between realized correlation and logarithmic realized standard deviation when the medians of both logarithmic realized standard deviations of Deutsche Mark and Yen are less than a threshold equal to -0.46 and when both are greater than -0.46 and they show density distributions of high volatility days differ from that of low volatility days. Huang and Nieh [Hua04] estimate a linear regression and indicate a positive association between realized correlation and volatilities significantly. Sun et al. [Sun08] introduce a copula ARMA-GARCH model for analyzing the co-movement of international equity markets. The model is implemented with an ARMA-GARCH model for the marginal distributions and a copula for the joint distribution. After goodness of fit testing, they find that the Student's t copula ARMA(1,1)-GARCH(1,1) model with fractional Gaussian noise is superior to alternative models investigated in their study where they model the simultaneous co-movement of nine international equity market indexes. They, indeed, studied volatility effect in correlation by their model. In fact, VIC effect is explained by the tail dependence of underlying assets, which exhibits extreme events happening simultaneously by their copula based model. In order to model VIC, unfortunately there is no possibility, or at least there is difficulty, to construct again a conditional multivariate realized volatility in correlation (VIC) here between realized volatility and correlation like the construction of realized volatility or correlation procedure to take advantages of time-varying instantaneous and contemporaneous characteristics of series. We have to turn back to the conventional techniques which fail to formulate directly observable instantaneous and contemporaneous measures. We simply proceed with a scatter plot and an estimated smooth trend of VIC.

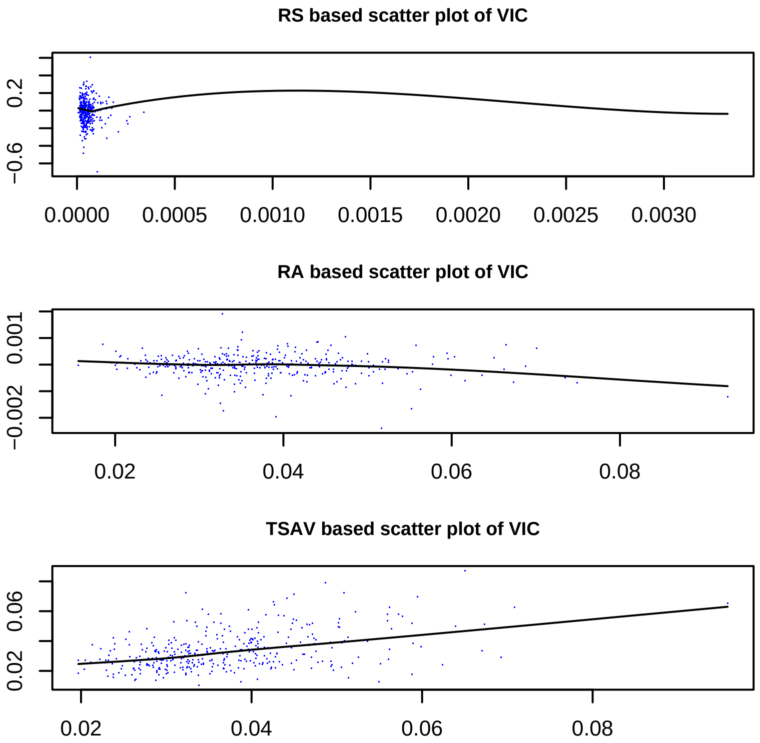


Figure 3.11: Scatter plots of different Volatility Effect in Correlation are unveiled to document association between realized volatility and correlation of NASDAQ and TNT. Axes x and y represent realized volatility and correlation respectively. The smooth line indicates the trend of association.

In Figure 2.10, scatter plots of different realized correlations between NASDAQ and DAX on y axis against their corresponding realized volatilities of NASDAQ are drawn to unveil association between realized volatility and correlation. A simple smooth curve passed through the points reveals lots of important facts flowing between the markets. As a matter of fact, in $TSACORxy$ based VIC plot, which depicts a two-scale absolute based correlation against its corresponding two-scale absolute based volatility, a positive trend is observed, while another VIC plots seem to show almost a constant trend. Applying VIC for studying of relationship between markets, we found that when the leading market (NASDAQ) is highly volatile (measured by realized volatility), the relationship between two markets becomes stronger, and when the leading market goes to calm down, the association (measured by realized correlation) between the markets goes to relax. So, two markets tend to be highly correlated when the leading market is highly volatile and inversely. Of course, this relation between realized volatility and correlation is strongly obvious, when we consider the $TSACORxy$ estimator. The findings are consistent with those in [Saf07b], where Euro/USD and Euro/GBP exchange markets have been studied. A similar correlation effect in volatility was documented for international equity returns by Solnik, Boucrelle, and Le Fur [Sol96]. It is also a common belief that cross-correlations between stocks actually fluctuate in time, and increase substantially in a period of high market volatility [Ciz01]. In other words, the time fluctuations of the measured cross-correlations between stocks are directly related to the fluctuations of the market volatility. Further, Cizeau, Potters and Bouchaud [Ciz01] investigate that how much of these correlations can be explained within a simple non-Gaussian one-factor description, which accounts for fat-tail effects, with time-independent correlations. The one-factor model, conditioning on absolute market returns larger than a given value, predicts an increase of the correlations in high volatility periods. The much discussed exceedance correlations can also be reproduced quantitatively and reflect both the non-Gaussian nature of the fluctuations and the negative skewness of the index, and not the fact that correlations themselves are time dependent.

The volatility effect in correlation can be studied also in relationships within a given market. Summarily the results of association between NASDAQ and TNT stock ticker of 5 minute frequency are reported. American TNT is handled in NYSE where NASDAQ index is provided. Our TNT time series affords a period from 5.04.2005 to 14.09.2006. NASDAQ is also truncated to overlap exactly this period. The simple result is summarized in Figure 3.11. Like result of investigating comovement between the markets, the $TSACORxy$ based measure exhibits a positive trend between the volatility of NASDAQ and the correlation between NASDAQ and TNT equity, where another VIC plots exhibit a nearly constant smooth trend. This result implies that TNT equity returns are viewed so that it has no a constant and stable relationship with NASDAQ index returns. Here VIC association conveys that the so-called Beta or systematic risk in Capital Asset Pricing Model (CAPM) which is conventionally assumed to be constant should vary over the time.

3.6 Concluding remarks and some discussions

Mainly the consistency and unbiasedness of proposed realized volatility and correlation estimators for the corresponding integrated volatilities and correlations have been studied under assumption of Gaussian noise. In fact, the consistency of volatility estimators differs given they are constructed differently. This fact is valid for correlation estimators as well. In addition, the behaviors of introduced estimators on finite samples were studied by simulation experiments. The TSAV volatility estimator eventually is a consistent and unbiased estimator for integrated power variation where $r = 1$ as the frequency increases even under the assumption of existence of microstructure frictions. This suggests that the estimator converges even at high frequency levels, where the noise especially exists. The TSAV estimator is constructed upon the subsampling and averaging approach which corrects for the bias caused by the microstructure noise. All the estimators display lower error under the fractional stable noise which is the most realistic process compared to other noise processes investigated by the simulations.

The empirical study of some important distributional and dynamic aspects of different alternative realized volatility and correlation estimators was another subject of this current part. None of volatility measures exactly pose a normal daily distribution tested by Jarque-Bera test of normality. Some of the volatility and correlation estimators indicate heavy tail in distributions. While squared based volatility shows heavier tail than absolute based volatility estimators, the absolute based correlation estimators show heavier tail than squared based correlation. In our experiments, we found that absolute based volatility estimators include longer memory behavior as a dynamic stylized fact of markets. Self-similarity structures computed by Hurst exponent was documented in the structures of series generated by realized measures.

Consistent with Andersen et al. [And01a] and Andersen et al. [And01b] our results suggest that realized squared correlation is viewed to pose the normal distribution. However, according to our experiment, it seems to fail containing dynamic properties such as long memory. In contrast, tested by Jarque-Bera estimator, the null hypothesis of normality for the proposed absolute based correlation estimators can not significantly be accepted. Consistent with common sense and in particular with Archimedean copulas and one-factor model, we found that the multivariate absolute-based realized correlations exhibit nonlinearity in dependence structure of time-varying correlation series. They indicate negative asymmetry in correlation implying fatter left tail where the extreme values are mainly populated there. It turns out that downside comoves are greater than upside comoves between markets. The autocorrelation and long memory, which have been well documented in many real world financial time series processes, are included in the structure of absolute based correlation estimators.

Realized correlation is itself correlated with realized volatility, which is called the volatility effect in correlation. This fact stimulates one to revise for example systematic risk assumptions in CAPM theory. Using VIC analysis, the

$TSACORxy$ based VIC exhibits a positive trend between the volatility of NASDAQ and the correlation between NASDAQ and TNT equity, where another VIC plots exhibit a nearly constant smooth trend. Also applying this effect on relation between the markets, we found that when NASDAQ is highly volatile, the relationship between NASDAQ and DAX becomes stronger, and when NASDAQ goes to calm down, the association between the indices goes to relax.

Construction of some kind of combined realized measures, for example return per unit of volatility which may be somehow close to Sharpe Ratio, gives a strong analytical tool at hand to study jointly dynamics of two most important criteria for investors, namely volatility and return, as the time in these series varies rather a constant Sharpe Ratio.

Part II

Multiresolution modeling and forecasting volatility

Introduction to the part: When modeling conditional volatility in financial time series, a time series can be decomposed into predictable and unpredictable components. Then consideration can be centered on the determinants of the predictable part. Like many financial variables, volatility can be modeled and estimated parametrically and nonparametrically. Heteroskedastisity is well-known to be modeled by GARCH approach. Meanwhile, motivated by this parametric approach, the Conditional Heteroskedastic Autoregressive Nonlinear (CHARN) model nonparametrically models return and volatility of a time series.

A multiresolution analysis based on wavelet transformations, however, can help to boost estimation performance. The multiresolution analysis can be implemented utilizing maximal overlap discrete wavelet transform (MODWT). The MODWT is in particular useful for analyzing and forecasting time series that exhibit nonstationary property, since time-dependent events at various scales are properly localized by MODWT. It can be investigated how multiresolution analysis can help to enhance the estimation power of the CHARN model. Indeed, applying capabilities of wavelet analysis and advantages, such as locality in time and frequency, ability to handle multiscale information and especially ability to describing heterogeneous data series help to obtain better results. A Multiscale resolution CHARN model to be estimated by the support vector regression machine is proposed in such a way that the capabilities of wavelet transformation are exploited by multiresolution analysis.

As a tool for regression estimation, Support Vector Regression machine is considered to be combined with the multiresolution analysis. This combination is expected to yield higher performance of learning. The combination is designed so as the original time series is decomposed into several scales or resolutions, each scaled time series is approximated separately by a SVR machine and then the fitted values on different scales are linearly summed up to obtain an overall function estimation for the original time series.

Literature review Härdle and Tsybakov [Hae97] consider the class of dynamic model CHARN in which both the conditional mean and the conditional variance (volatility) are unknown functions of the past. They construct an estimator based on local polynomial fitting. They examine the rates of convergence of these estimators and give a result on their asymptotic normality. The local polynomial fitting of the volatility function is applied to different foreign exchange rate series. They find an asymmetric U-shaped smiling face form of the volatility function. Härdle, Tsybakov and Yang [Hae98] again approximate the CHARN model by the local polynomial estimator using foreign exchange rate data. The returns on exchange rates show negative correlation when the two series have opposite lagged values and positive correlation elsewhere. Härdle and Wieu [Hae92] propose to use the ordinary Nadaraya-Watson kernel regression for estimation of the CHARN model. Polzehl and Spokoiny [Pol03] introduce an adaptive weights smoothing (AWS) procedure which is fully adaptive and dimension free. The AWS method is generalized to the case of an arbitrary local linear parametric structure. They illustrate the performance of the procedure in univariate and bivariate situations.

Aussem, Campbell and Murtagh [Aus98] discuss forecasting strategies based on the assumption that the time series exhibits characteristics spanning different time scales. The method is illustrated on the S&P500 daily prices. A wavelet decomposition of the original series is first carried out to decompose a time series into varying scales of temporal resolution, with the aim of underlying temporal structures becoming more tractable. Using the resulting wavelet coefficient, appropriately modified for time series data, as the new input patterns, a recurrent neural network applied on an autoregressive model is successfully trained to provide five days ahead forecasts for S&P500 closing price forecasts. A more sophisticated method that has proved useful is proposed: each individual wavelet series is fitted with a neural network model to output the wavelet forecast. The latter are afterward recombined to form the overall S&P500 forecast. The method is shown to significantly reduce the MSE and allows distinct forecasting techniques to be fruitfully combined. Bashir and El-Hawary [Bas00] report the application of the wavelet neural networks (WNNs) to short-term load forecasting. The wavelet neural network has much higher ability of generalization and fast convergence for learning than a multilayer feedforward neural network. The results of the network have been compared with artificial neural network and show an improved forecast with fast convergence. Lotric [Lot04] adds a denoising unit based on wavelet multiresolution analysis ahead of the multilayered perceptron. Chen et al. [Che06a] present a local linear wavelet neural network (LLWNN). The difference of the network with conventional wavelet neural network (WNN) is that the connection weights between the hidden layer and output layer of conventional WNN are replaced by a local linear model. A hybrid training algorithm of particle swarm optimization (PSO) with diversity learning and gradient descent method is introduced for training the LLWNN. Simulation results for the prediction of time series show the feasibility and effectiveness of the proposed method. In their paper, Soltani et al. [Sol00] deal with the problem of long-term memory time series prediction. The presented method is based on the multiscale filtering which iteratively decomposes a series into a trend and a hierarchy of details that are stationary and contain only short memory. Thus, the obtained series are modeled with classical ARMA models. The advantage of this method is that it overcomes the tricky problem of the fractional integration parameter estimation. The statistical properties of the obtained series are studied and the use of multichannel autoregressive models is justified when the moving average part does not exist. Results obtained through the use of both simulated and real-life series show the efficiency of the approach.

Problem description The problem of regression estimation to be approximated in terms of a finite sample of financial time series is considered. Financial time series samples are considered as very complex data series. They are usually nonstationary and include a combination of signal as well as noise components. Financial estimation and forecasting is an example of a regression estimation which is challenging due to the high noise [Gil01]. In fact, financial data are very noisy, unstable, and non-Gaussian [Cam97]. Even further, financial time series are among the noisiest and most difficult signals to forecast [Abu96]. Modeling and estimation of nonstationary time series is typically more difficult that stationary time series.

As a supervised learning machine, support vector regression provides a valuable framework for the representation of relationships present in data. Nonetheless, the choice of input data is not a trivial matter when difficult noisy data is handled. Data preprocessing and decomposition remain essential steps in the knowledge discovery process for real world application and, when correctly carried out, greatly improve the machine's ability to capture valuable information. Wavelet preprocessing and decomposing for enhancing prediction comes from multiresolution analysis provided by wavelet transform. The wavelet transform can decompose one time series into several time series with different resolutions which have different levels of smoothness. The smoother level is more predictable, whereas the detailed level is less predictable, or more related to the noise.

Motivation An idea in modeling financial markets is the hypothesis of a heterogeneous market where the market agents differ in their perceptions of the market, risk profiles, institutional constraints, degree of information, prior beliefs, and other characteristics such as geographical locations. Müller et al. [Mue97] argue that many differences among market participants translate to a sensitivity to different time horizons.

"The diversity of agents in a heterogeneous market makes volatilities of different time resolutions behave differently" [Mue97]. The long memory of volatility as already found in Dacorogna et al. [Dac93] and Ding et al. [Din93] is explained in terms of different market participants with different time horizons, from short-term dealers to long-term investors.

There is a growing number of studies dealing with different types of traders. For example, Müller et al. [Mue97] focus on the time horizon and the temporal resolution with which different traders are viewing and influencing the market. They believe that the time horizon is one of the most important aspects in which trading behaviors differ and give some evidence for this. They argue there is also a methodological reason for this: the time horizon aspect can be investigated by studying the time series of prices whereas the study of other properties of trader groups often requires some less easily obtainable and quantifiable information.

Inspired by heterogeneous market agents, our basic idea is that different classes of market agents perceive, react to, and in particular cause different resolutions of volatility. Short-term traders evaluate the market at a higher frequency and have a shorter memory than long-term traders. Thus, we divide not only the market agents into different classes but also volatility into different resolutions or scales. In turn, the heterogeneous agents cause different volatility resolutions. A multi-agent approach with the type of noise trader-fundamentalist interaction introduced by Beja and Goldman [Bej80] and Day and Huang [Day90]. It can be supposed that the noise traders or daily brokers who trade daily, cause volatility of finer resolutions and fundamentalists who invest yearly, cause volatility of coarser resolutions.

Objective Applying the multiresolution decomposition of a time series into several scales by wavelets, our specific objective is to improve estimation performance of the CHARN model estimated by support vector regression. Exploiting multiresolution analysis, each separated scale especially the smoother scale can be estimated more accurate than the original time series. Because time-dependent events at various scales are properly localized by MODWT analysis, it is particularly useful for analyzing and forecasting time series that exhibit nonstationary characteristics. The objective comes through in both in-sample and out-of-sample estimations by the multiresolution analysis. In general, improvement of the predictability power is considered as a direct goal for this part.

Contribution This part contributes by combining a multiresolution analysis with support vector regression in order to improve predictability power. In the other side, combination of the multiresolution analysis with the financial time series CHARN model is a new contribution.

Structure of the part The part contains chapters 4, 5, 6, and 7. Chapter 4 discusses theoretically what is a multiresolution analysis. Some kinds of wavelet functions are briefly defined in section 4.1. Section 4.2 is devoted to theories explaining the multiresolution analysis by wavelets. The chapter uses the maximal overlap discrete wavelet transform for decomposition of a time series into scales. It is described in this section. The maximal overlap discrete wavelet transform is suitable for the multiresolution analysis demonstrated in section 4.3. Section 4.4 compares Fourier transform with wavelet transform. The theory behind support vector regression, i.e., statistical learning theory, is illustrated in chapter 5. Then support vector regression is formulated in this chapter. Section 5.1 describes the statistical learning theory. Support vector regression is modeled in the next section. It is seen in section 5.3 that how nonlinearity reality of for example financial time series is dealt with. Section 5.4 summarily introduces some popular implementation algorithms of support vector regression. The financial time series CHARN model to be estimated by support vector regression is explained in chapter 6. It is shown in section 6.2 that how the model is estimated. For estimation, some common nonparametric algorithms are introduced in the next section. Then the results of model estimation are appeared in chapter 7, section 7.1 with a single resolution design and section 7.2 with a multiresolution design. Section 7.3 gives some conclusions on the results of the present part.

Chapter 4

Multiresolution analysis

4.1 Wavelet transformation

Wavelets are mathematical tools for analyzing time series or images (although not exclusively so) [Per00]. Our discussion of wavelets here focuses on their use with time series, which we take to be any sequence of observations associated with an ordered independent variable t (the variable t can be assumed either a discrete set of values such as integers or a continuum of values such as the entire real axis). Broadly speaking, there have been two main waves of wavelets. The first wave resulted in what is known as the continuous wavelet transform (CWT), which is designed to work with time series defined over the entire real axis; the second, in the discrete wavelet transform (DWT), which deals with series defined essentially over a range of integers (usually $t = 0, 1, ..., N - 1$, where N denotes the number of values in the time series).

4.1.1 The essence of a wavelet

What is a wavelet? As the name suggests, a wavelet is a small wave. A small wave grows and decays essentially in a limited time period. The contrasting notion is obviously a big wave. An example of a big wave is the sine function, which keeps on oscillating up and down on a plot of $\sin(u)$ versus $u \in (-\infty, \infty)$. To begin to quantify the notion of a wavelet, let us consider a real-valued[1] function $\psi(.)$ defined over the real axis $(-\infty, \infty)$ and satisfying two basic properties [Per00].

1. The integral of $\psi(.)$ is zero:

$$\int_{-\infty}^{\infty} \psi(u)du = 0. \tag{4.1}$$

[1] Of course, in a wider exposition complex-valued wavelets can be assumed. But they are applied in other areas such as geophysical applications than finance.

2. The square of $\psi(.)$ integrates to unity:

$$\int_{-\infty}^{\infty} \psi^2(u)du = 1 \tag{4.2}$$

(for the sine function, the above integral would be infinite, so $\sin^2(.)$ cannot be renormalized to integrate to unity).

If Equation (4.2) holds, then for any ϵ satisfying $0 < \epsilon < 1$, there must be an interval $[-T, T]$ of finite length such that

$$\int_{-T}^{T} \psi^2(u)du > 1 - \epsilon.$$

If we think of ϵ as being very close to zero, then $\psi(.)$ can only deviate insignificantly from zero outside of $[-T, T]$: its nonzero activity is essentially limited to the finite interval $[-T, T]$. Since the length of the interval $[-T, T]$ is vanishingly small compared to the infinite length of the entire real axis $(-\infty, \infty)$, the nonzero activity of $\psi(.)$ can be considered as limited to a relatively small interval of time. While Equation (4.2) says $\psi(.)$ has to make some excursions away from zero, Equation (4.1) tells us that any excursion, it make above zero, must be canceled out by excursion below zero. So $\psi(.)$ must resemble a wave. Hence Equations (4.1) and (4.2) lead to a small wave or wavelet.

Based on the definitions below, one can verify that these functions indeed satisfy Equations (4.1) and (4.2). A popular wavelet is called the Haar wavelet function:

$$\psi^{(H)}(u) \equiv \begin{cases} -1/\sqrt{2}, & -1 < u \leq 0; \\ 1/\sqrt{2}, & 0 < u \leq 1; \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

The above is arguably the oldest wavelet function, being named after A. Haar, who developed an analysis tool in an article in 1910 [Haa10]. To form other two wavelets, we start with the Gaussian probability density function (PDF) for a random variable with mean zero and variance σ^2:

$$\phi(u) \equiv \frac{e^{-u^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}, \quad -\infty < u < \infty.$$

The first derivative of $\phi(.)$ is

$$\frac{d\phi(u)}{d(u)} = -\frac{ue^{-u^2/2\sigma^2}}{\sigma^3\sqrt{2\pi}}.$$

If we renormalize the negative of the above to satisfy Equation (4.2), we obtain the wavelet

$$\psi^{(\mathrm{fdG})}(u) \equiv \frac{\sqrt{2} u e^{-u^2/2\sigma^2}}{\sigma^{3/2}\pi^{1/4}}. \tag{4.4}$$

Equation (4.4) represents the first derivative Gaussian (fdG) wavelet transform. With proper renormalization again, the negative of the second derivative of $\phi(.)$ also yields a wavelet, usually referred to as the Mexican hat:

$$\psi^{(\mathrm{Mh})}(u) \equiv \frac{2(1-\frac{u^2}{\sigma^2})e^{-u^2/2\sigma^2}}{\pi^{1/4}\sqrt{3\sigma}}. \tag{4.5}$$

To introduce Daubechies' wavelet system, let us describe here the basic idea and the principal characteristics of the multiresolution wavelet decomposition. The multiresolution analysis will be followed in details later in MODWT framework. The main equation of the multiresolution theory is the scaling equation which establishes a connection between the two symmetries underlying the wavelet theory: dilations and translations [Gag94] [Lin93]. Given a set of coefficients $a_k, k \in Z$, the scaling equation

$$\varphi(x) = 2\sum_k a_k \varphi(2x-k),\ x \in R$$

and the normalization

$$\int \varphi(x)dx = \sum_k a_k = 1,$$

define a scaling function $\varphi(x)$. By defining the set of translates of the dilated function $\varphi(x)$,

$$\varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k),\ j \in Z, \tag{4.6}$$

the multiresolution analysis of $L^2(R)$ consists of the decomposition of the Hilbert space $L^2(R)$ (the space of square-integrable functions) into the chain of closed subspaces

$$... \subset V_{j-1} \subset V_j \subset V_{j+1} \subset ...$$

where

$$V_j = \mathrm{Span}\left\{\varphi_{j,k}(x),\ k \in Z\right\}$$

and such that

$$\bigcap_j V_j = \{0\},$$

$$\bigcup_j V_j = L^2(R).$$

The set of functions $\varphi_{j,k}(x)$ is called a Riesz basis of V_j [Che06b] [Res97]. According to above chain of closed subspaces, multiresolution property means that V_j is a subset of V_{j+1} [Che06b]. In fact a multiresolution analysis of $L_2(R)$ is defined as a sequence of nested subspaces V_j with scaling function $\varphi(x)$ if the above properties hold. The multiresolution analysis aims to decompose $L^2(R)$ as

$$L^2(R) = V_{j_0} \oplus \sum_{j \geq j_0} W_j, \tag{4.7}$$

where W_j is defined as the orthogonal complement of V_j in V_{j+1}, that is

$$V_{j+1} = V_j \oplus W_j. \tag{4.8}$$

So, each element of V_{j+1} can be uniquely written as the orthogonal sum of an element in V_j and an element in W_j that contains the complementing details, i.e., $V_{j+1} = V_j \oplus W_j$. For a given scale j,

$$W_j = \text{Span}\left\{\psi_{j,k}(x),\ k \in Z\right\},$$

where

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \tag{4.9}$$

and $\psi(x)$ is the wavelet of the multiresolution analysis which satisfies

$$\psi(x) = 2\sum_k b_k \psi(2x - k)\ x \in R.$$

The function $\psi_{j,k}(x)$ in (4.9) is called wavelet function. Following (4.7), any function of $L^2(R)$ can be expanded as a linear combination of translates of the scaling function $\varphi(x)$ at some fixed scale and the translates of the wavelet $\psi(x)$ expressed at finer scales as

$$f(x) = \sum_k v_{j_0,k}\varphi_{j_0,k}(x) + \sum_{j \geq j_0}\sum_k w_{j,k}\psi_{j,k}(x).$$

Thanks to the orthonormal decomposition (4.8), we then have

$$V_N = V_{N_0} \oplus W_{N_0} \oplus W_{N_0+1} \cdots \oplus W_{N-1}$$

for some larger scale $N_0 < N$. This decomposition amounts to consider the equivalent finite expansion for $f(x)$,

$$f(x) = \sum_k v_{N_0,k} \varphi_{N_0,k}(x) + \sum_{j=N_0}^{N-1} \sum_k w_{j,k} \psi_{j,k}(x).$$

In this expansion which is called multiscale expansion [Mcc94], the first term represents the approximation of $f(x)$ at a given coarse scale. The remaining terms are the corrections at finer scales. The expansion completely describes the function $f(x)$.

The basic ingredient of the multiresolution analysis, i.e., the scaling function $\varphi(x)$ entails four important constraints including: i) compactness of its support, ii) orthogonality of its translates, iii) regularity, and iv) symmetry.

The first condition insures an exact local description of the functions of $L^2(R)$. As a consequence, there are a finite number of non-vanishing scaling coefficients a_k, and it is considered $a_k \neq 0$, for $k = -J, -J+1, ..., J, J+1$, where J is an arbitrary integer. It is straightforward to show that both $\varphi_{j,k}(x)$ and $\psi_{j,k}(x)$ have a support in the interval $\left[2^{-j}(-J+k), 2^{-j}(J+k+1)\right]$. The first three conditions define the so-called Daubechies' wavelet analysis [Dau88] for which the regularity condition sets the polynomial content of the V spaces (scaling functions of regularity $R(R \leq J)$ allow exact representations of polynomials of order R in the V spaces). The scaling function and its translations thus define a polynomial interpolation scheme up to order J.

Daubechies [Dau88] found the system of the wavelet functions with $2N$ coefficients. Haar wavelet functions themselves are defined not to overlap each other. On the other hand, Daubechies' wavelet functions overlap each other to some degree and interpolate together. Still they are orthogonal.

In summary, a wavelet by definition is any function that integrates to zero and is square integrable. Here, we have intentionally given just a bare bones definition of a wavelet so that we can focus on presenting the key concepts behind the subject.

4.1.2 Wavelet applications

Wavelet transforms are now being adopted for a vast number of different applications. Many areas of physics including molecular dynamics, astrophysics, density-matrix localization, seismic geophysics, optics, turbulence and quantum mechanics. Other areas like engineering, industry, economics and finance, chemie, bioinformatic, medicine, geophysics, computer science apply wavelet analysis to solve related problems. In some of these problems like image processing, blood-pressure, heart-rate, DNA analysis, protein analysis, climatology, general signal processing, speech recognition, computer graphics, signal denoising and multifractal analysis wavelet transforms is effectively used. Document and texture analysis, character recognition, face and gesture recognition, computer vision, biomedical image application, remote sensing, geophysics exploration, regression function approximation are among those applications of wavelets. One use of wavelets is in data compression. Like several other transforms, the wavelet transform can be used to transform raw data (like image, audio and video), then encode the transformed data, resulting in effective compression.

However, financial applications of wavelets in time series analysis are relatively new but emerging.

A discrete wavelet transform (DWT) provides a background for the Maximal Overlap Discrete Wavelet Transform (MODWT). The Maximal Overlap Discrete Wavelet Transform is suitable for multiresolution analysis applied in this part.

4.1.3 The Discrete Wavelet Transform

The key feature in the discrete wavelet transform (DWT) is that the translation parameter t is not continuous, but instead is integer. In practical applications, we only have a finite number N of sampled values. If we only have these samples, it is not possible to compute a continuous wavelet exactly, but we can resort to approximations. Generally, the discrete wavelet transform of a time series X of the length N is a linear transformation which can therefore be represented in matrix form

$$\mathbf{W} = WX.$$

Here $\mathbf{W}$ is the column vector of DWT coefficients and W is an orthonormal DWT matrix constructed according to the type of wavelet we choose to use. Constantine, Percival, and Reinhall [Con01] mention a number of advantages in using the discrete wavelet transform on turbulence data “as follows

- *Decomposition based on scale:* Turbulence is known to exhibit fluctuations at various spatial scales, and hence the DWT is a natural analyzer.

- *Decorrelation of time series:* While turbulence data are typically highly correlated, their wavelet coefficients are approximately uncorrelated. This property is crucial for obtaining viable approximate maximum likelihood estimates of fractionally differenced parameters.

- *Localized time and scale content:* Each wavelet coefficient is localized in time, allowing us to track changes in the characteristics of a time series at a particular scale as a function of time.

- *Separation of nonlinear trends from noise:* The wavelet coefficients are inherently blind (invariant) to nonlinear polynomial trend contamination in the original time series”.

Further details on discrete wavelet transform prescriptions can be found in Appendix A where it is explained in a pyramid algorithm framework.

4.2 The Maximal Overlap Discrete Wavelet Transform

Here a modified version of the discrete wavelet transform called the maximal overlap DWT (MODWT) is described. Essentially the same transform has been discussed in the wavelet literature in the context of infinite sequences under the name undecimated DWT (Shensa [She92]) and in the context of power of two sequences under the names stationary DWT (Nason and Silverman [Nas94]), translation-invariant DWT (Coifman and Donoho [Coi95]; Liang and Parks [Lia96]), and time-invariant DWT (Pesquet, Krim, and Carfantan [Pes96]).

The maximal overlap discrete wavelet transform (MODWT) of Percival and Walden [Per00] is basically a nondecimated version of the discrete wavelet transform (DWT) of Mallat [Mal89]. Cornish, Bretherton, and Percival [Cor05] define the MODWT. "The MODWT is a linear filtering operation that transforms a series into coefficients related to variations over a set of scales. It is similar to the DWT in that both are linear filtering operations producing a set of time-dependent wavelet and scaling coefficients. Both have basis vectors associated with a location t and a unitless scale $\tau_j = 2^{j-1}$ for each decomposition level $j = 1, ..., J_0$. Both are suitable for analysis of variance (ANOVA) and multiresolution analysis (MRA)".

As Percival and Mofjeld [Per97] explain, "the MODWT of a time series leads to two types of analysis. The first is an additive decomposition known as multiresolution analysis, which breaks up the series into a number of details and a single smooth. Each detail is a time series describing variations at a particular time scale, whereas the smooth describes the low-frequency variations. The second type of analysis decomposes the sample variance of the time series across different time scales and over time". The decomposition across time is facilitated by a compactly supported least asymmetric (LA) wavelet filter due to Daubechies [Dau92] that helps events align in the analysis with events in the original series.

Because time-dependent events at various scales are properly localized by MODWT, it is particularly useful for analyzing and forecasting time series that exhibit nonstationary characteristics.

Let X be a column vector containing a sequence $X_0, X_1, ..., X_{N-1}$ of N observations of a real-valued time series. We assume that the observation X_t was collected at time t, where Δt is the time interval between adjacent observations (e.g., $\Delta t = \frac{1}{2}$ day for a time series). It is also assumed that the sample size N is an integer multiple of 2^J, where J is a positive integer.

Decomposing an infinite sequence $\{X_t\}$ of Gaussian random variables using the MODWT to J_0 levels theoretically involves the application of J_0 pairs of filters. The filtering operation at the jth level consists of applying a wavelet (high-pass) filter $\left\{\tilde{h}_{j,l}\right\}$ to yield a set of wavelet coefficients

$$\bar{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l} \tag{4.10}$$

and a scaling (low-pass) filter $\{\tilde{g}_{j,l}\}$ to yield a set of scaling coefficients

$$\bar{V}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{g}_{j,l} X_{t-l} \tag{4.11}$$

for all time $t = ..., -1, 0, 1, ...$ [Per00]. The equivalent wavelet $\left\{\tilde{h}_{j,l}\right\}$ and scaling $\{\tilde{g}_{j,l}\}$ filters for the jth level are a set of scale-dependent localized differencing and averaging operators, respectively, and can be regarded as stretched versions of the base ($j = 1$) filters. The jth level equivalent filter coefficients have a width $L_j = (2^j - 1)(L - 1) + 1$, where L is the width of the $j = 1$ base filter. In practice, the filters for $j > 1$ are not explicitly created because the wavelet and scaling coefficients can be generated sequentially using an elegant algorithm that involves just the $j = 1$ filters operating on the jth level scaling coefficients to generate the $j + 1$ level wavelet and scaling coefficients [Per00]. The jth level wavelet coefficients characterize those components of the signal with fluctuations matching the unitless scale $\tau_j = 2^{j-1}$. If $\{X_t\}$ is either a stationary process or a non-stationary process with stationary backward differences, and L is suitably chosen, then $\bar{W}_{j,t}$ is a Gaussian stationary process with zero mean and known power spectral density [Per00].

In addition, MODWT coefficients for different scales are approximately uncorrelated and are hence useful statistical measures for partitioning variability by scale.

The real world financial time series or signals are usually sampled over a finite interval at discrete times. To complete the filtering operation at each level for a finite time series $\{X_t\}$, $t = 0, ..., N - 1$, the MODWT treats the series as if it were periodic, whereby the unobserved samples $X_{-1}, X_{-2}, ..., X_{-N}$ are assigned the observed values at $X_{N-1}, X_{N-2}, ..., X_0$. The MODWT coefficients are thus given by

$$\tilde{W}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} X_{t-l} \text{ mod N}, \tag{4.12}$$

and

$$\tilde{V}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{g}_{j,l} X_{t-l} \text{ mod N}, \tag{4.13}$$

for $t = 0, ..., N-1$.

When considering the statistical properties of DWT coefficients, it is useful to divide the wavelet and scaling coefficients into boundary and interior coefficients. Boundary coefficients are those subject to change if the mod operator were to be dropped in (4.12) and (4.13) [Con01]. Obviously, MODWT coefficients generated by both beginning and ending components could be spurious. Hence, we will adjust the boundary-affected coefficients later. This periodic extension of the time series is known as analyzing $\{X_t\}$ using circular boundary conditions.

Applying the MODWT to a time series requires specification of a wavelet filter and of the index J_0 for the maximum scale of interest. To make appropriate selections we must take into account the goals of the analysis and time series being analyzed.

4.2.1 Wavelet filter

There are two considerations about the filter choice: the type and the length. Percival and Walden [Per00] demonstrate the artifacts in some of the DWT filters, but the problem is much mitigated in MODWT case. Daubechies least asymmetric (LA) MODWT filters, which are also called as symlets, are among the popular choices, because LA filters provide most accurate synchronization between wavelet coefficients and the original series. The Daubechies class of wavelets possesses appealing regularity characteristics and produces transforms that are effectively localized differences of adjacent weighted averages [Dau92]. The least asymmetric (LA) subclass has approximate linear phase and exhibits near symmetry about the filter midpoint. This linear phase property means that events and sinusoidal components in the wavelet and scaling coefficients at all levels can be aligned with the original time series. For the MODWT, this alignment is achieved by circularly shifting the coefficients by an amount dictated by the phase delay properties of the basic filter.

The MODWT coefficients can be calculated from the Daubechies family of compactly supported wavelet filters, which are well localized in time. Using Daubechies least asymmetric family of wavelet filters (LA), the MODWT is constructed via approximate linear-phase filtering operations, thus allowing wavelet coefficients at various scales to be aligned in time with the events of the original series. This property makes the MODWT a particularly useful tool in the analysis of time-dependent processes [Jen00]. Gencay et al. [Gen04] express that "the least asymmetry wavelet of length 8, i.e., LA(8) is a widely used wavelet and is applicable in a wide variety of data types. In practice, if one wants to have the MODWT coefficients be alignable in time, the optimal choice is often LA(8)".

Percival and Walden [Per00] argues that LA(8) often provides "a good tradeoff between the width of the wavelet function and its smoothness. Being relatively short, and therefore providing a narrower cone of influence in the wavelet decomposition, its shape is still a good match to the characteristic features for most of the time series". Least asymmetric means that the associated wavelet filter has nearly zero phase property, i.e., the resulting features in the wavelet decomposition will be aligned in time with the features in the time series being analyzed [Div07].

The choice of filters' length is based on the trade-off between leakage and the number of boundary affected coefficients. If the length (L) is larger, the filters are much closer to the ideal high (low) pass only filters. However, the number of boundary affected coefficients will increase, reducing the size of unaffected coefficients. The LA filters are available in even widths L. The optimal filter width is dependent on the characteristics of the signal and problem domain of interest. A wider filter is smoother in appearance and reduces the possible appearance of artifacts in multiresolution analysis (MRA) due to the filter shape. It also results in better uncorrelatedness between wavelet coefficients across scales for certain time series, which is useful for deriving confidence bounds from certain wavelet-based estimates [Cra05]. However, using a wider filter results in many more boundary coefficients, especially at higher levels.

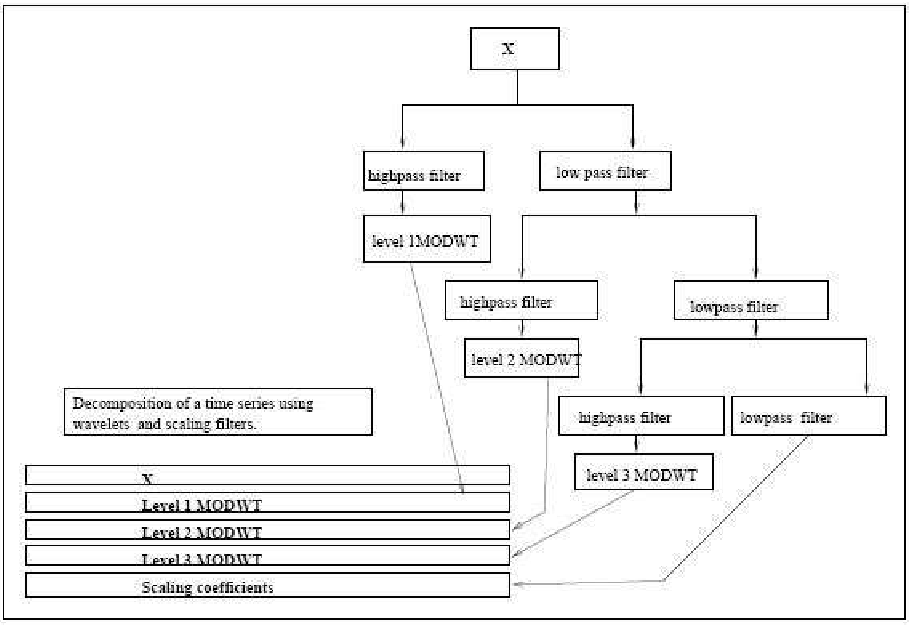


Figure 4.1: The pyramid algorithm is an iterative filtering algorithm to transform a time series into a collection of wavelet coefficients.

4.2.2 Number of scales

A time series can be completely or partially decomposed into a number of scales or levels. For complete decomposition of a series of length $N = 2^J$ using the DWT, the maximum number of scales in the decomposition is J. In practice, a partial decomposition of level $J_0 \leq J$ suffices for many applications. A J_0 level DWT decomposition requires that N be an integral multiple of 2^{J_0}. The MODWT can accommodate any sample size N and, in theory, any J_0. In practice, the largest level is commonly selected such that $J_0 \leq 2(N)$ in order to preclude decomposition at scales longer than the total length of the time series. In particular, for alignment of wavelet coefficients with the original series, the condition $L_{J_0} < N$, i.e., the width of the equivalent filter at the J_0th level is less than the sample size, should be satisfied to prevent multiple wrappings of the time series at level J_0. Selection of J_0 determines the number of octave bands and thus the number of scales of resolution in the decomposition.

The wavelet and scaling filters are used in a pyramid algorithm (an iterative filter algorithm) to transform $\{X_t\}$ into a collection of wavelet coefficients $W_{j,t}$ and scaling coefficients $V_{j,t}$ that can be associated with scales of, respectively, $\tau_j \equiv 2_{j-1}$ and τ_j, $j = 1, ..., J$ [Con01]. Figure 4.1 illustrates the pyramid algorithm. Appendix A explains the pyramid algorithm in more details. The pyramid algorithm can also be interpreted as a cascade filter bank operation [Con01].

4.3 Multiresolution analysis (MRA) by MODWT

Actually, the level-j wavelet coefficients, $\tilde{W}_{j,t}$, are associated with changes of X_t on the scale $\tau_j \equiv 2^{j-1}$ and the level-j scaling coefficients $\tilde{V}_{j,t}$ are associated with average of X_t on the scale $2\tau j$. In multiresolution analysis, the level-j wavelet details are defined by

$$\tilde{D}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{h}_{j,l} \tilde{W}_{j,t} \bmod t = 0, ..., N, \tag{4.14}$$

and the level-j wavelet smooths by

$$\tilde{S}_{j,t} = \sum_{l=0}^{L_j-1} \tilde{g}_{j,l} \tilde{V}_{j,t} \bmod t = 0, ..., N. \tag{4.15}$$

As with the wavelet and scaling coefficients, the level-j wavelet details $\tilde{D}_{j,t}$ are associated with changes of X_t on the scale τ_j and level-j wavelet smooths $\tilde{S}_{j,t}$ are associated with average of X_t on the scale $2\tau j$. A wavelet basis consists of a father wavelet that represents the smooth baseline trend and a mother wavelet that is dilated and shifted to construct different levels of detail. At high scales, the wavelet has a small time support, enabling it to zoom in on details such as spikes and cusps, and on short-lived phenomena. At low scales, wavelets capture long-run phenomena. $\tilde{D}_{j,t}$ is mother wavelet. A mother wavelet is a source function, from which translated and scaled wavelet functions (with different regions of support) are constricted. $\tilde{S}_{j,t}$ is the father wavelet, also referred to as the scaling function that represents the coarsest components or the smooth baseline trend of the function. It is just a horizontal line equal to one. While the father wavelet integrates to one, the mother wavelet integrates to zero, reflecting the fact that it is used to represent differences in the data that average out to zero. The father wavelet covers the whole time support at the lowest scale of resolution, while the mother wavelet is dilated and translated to capture different levels of fineness. An alternative way to view the difference is that the father wavelet acts as a low pass filter, whereas the mother wavelets act as high pass filters. Different scales translate into different frequency bands that are passed.

The MODWT is equivalent to the original time series in the sense that, given the MODWT coefficients, we can reconstruct the original time series, X. This leads to the following additive decomposition, which is known as a multiresolution analysis (MRA)

$$X_t = \sum_{j=1}^{J} \tilde{D}_{j,t} + \tilde{S}_{j,t}. \tag{4.16}$$

In the above, the $D_{j,t}$ is an N dimensional vector that depends upon just $\tilde{W}_{j,t}$ and hence is constructed using just those MODWT wavelet coefficients that are associated with changes of averages on a scale of τ_j. This vector is called a detail series and is the part of the MRA of X that can be attributed to variations on a scale of τ_j. The final term in the MRA is $\tilde{S}_{j,t}$, which again is an N dimensional vector, but this depends just on the scaling coefficients $\tilde{V}_{j,t}$. The vector $\tilde{S}_{j,t}$ is called the smooth series, because it is associated with averages over scales $2\tau_j$ and longer and hence captures the slowly varying portion of X. Thus an MRA is an additive decomposition that expresses a time series as the sum of several new series, each of which can be associated with variations on a particular scale.

Percival and Mofjeld [Per97] mention four important "properties that distinguish the MODWT from the DWT:

1. While the DWT of level J restricts the sample size to an integer multiple of 2^J, the MODWT of level J is well defined for any sample size N (for convenience, however, we again assume that N is at least as large as the length L_1 of the wavelet filter). When N is an integer multiple of 2^J, the DWT can be computed using $O(N)$ multiplications, whereas the corresponding MODWT requires $O(N \log_2 N)$ multiplications. There is thus a computational price to pay for using the MODWT, but its computational burden is the same as the widely used fast Fourier transform algorithm and hence is usually quite acceptable.

2. As is true for the DWT, the MODWT can be used to form a multiresolution analysis. In contrast to the usual DWT, both the MODWT wavelet

and scaling coefficients and multiresolution analysis are shift invariant in the sense that circularly shifting the time series by any amount will circularly shift by a corresponding amount the MODWT wavelet and scaling coefficients, details and smooths.

3. In contrast to the DWT details and smooths, the MODWT details and smooths are associated with zero phase filters, thus making it possible to meaningfully line up features in a multiresolution analysis with the original time series.

4. As is true for the DWT, the MODWT can be used to form an analysis of variance based upon the wavelet and scaling coefficients. Under a stationarity assumption on the wavelet coefficients, the MODWT yields an estimator of the variance of the wavelet coefficients that is statistically more efficient than the corresponding estimator based on the DWT".

4.4 Short Time Fourier Transform vs. Wavelet Transform

The wavelet transform is often compared with the Fourier transform, in which signals are represented as a sum of sinusoids. Just as Fourier analysis is based upon the notion of representing (or re-expressing) a time series as a linear combination of sinusoids, the idea underlying wavelet analysis is to represent the series as a linear combination of wavelets. In Fourier analysis, each sinusoid is associated with a particular frequency, so what frequencies are important in a particular time series can be deduced by studying the magnitudes of the coefficients of the various sinusoids in the linear combination. In contrast, each wavelet is associated with two independent variables, namely, time and scale, because each wavelet is essentially nonzero only inside a particular interval of times. Within that interval, the wavelet spends roughly an equal amount of time above and below zero, so it appears to be a small wave.

The Short Time Fourier Transform (STFT) is a modified version of the Fourier Transform. The Fourier Transform separates the waveform into a sum of sinusoids of different frequencies and identifies their respective amplitudes. Thus it gives us a frequency-amplitude representation of the signal. In STFT, a nonstationary signal is divided into small portions, which are assumed to be stationary. This is done using a window function of a chosen width, which is shifted and multiplied with the signal to obtain the small stationary signals. The Fourier Transform is then applied to each of these portions to obtain the Short Time Fourier transform of the signal.

The problem with STFT goes back to the Heisenberg uncertainty principle which states that it is impossible for one to obtain which frequencies exist at which time instance, but, one can obtain the frequency bands existing in a time interval. This gives rise to the resolution issue where there is a trade-off between the time resolution and frequency resolution. To assume stationarity, the window is supposed to be narrow, which results in a poor frequency resolution, i.e., it is difficult to know the exact frequency components that exist in the signal; only the band of frequencies that exist is obtained. If the width of the window is increased, frequency resolution improves but time resolution becomes poor, i.e., it is difficult to know what frequencies occur at which time intervals. Also, choosing a wide window may violate the condition of stationarity. Consequently, depending on the application, a compromise on the window size has to be made. Once the window function is decided, the frequency and time resolutions are fixed for all frequencies and all times.

The wavelet transform solves the above problem to a certain extent. In contrast to the STFT, which uses a single analysis window, the wavelet transform uses short windows at high frequencies and long windows at low frequencies. This results in multiresolution analysis by which the signal is analyzed with different resolutions at different frequencies, i.e., both frequency resolution and time resolution vary in the time-frequency plane without violating the Heisenberg inequality.

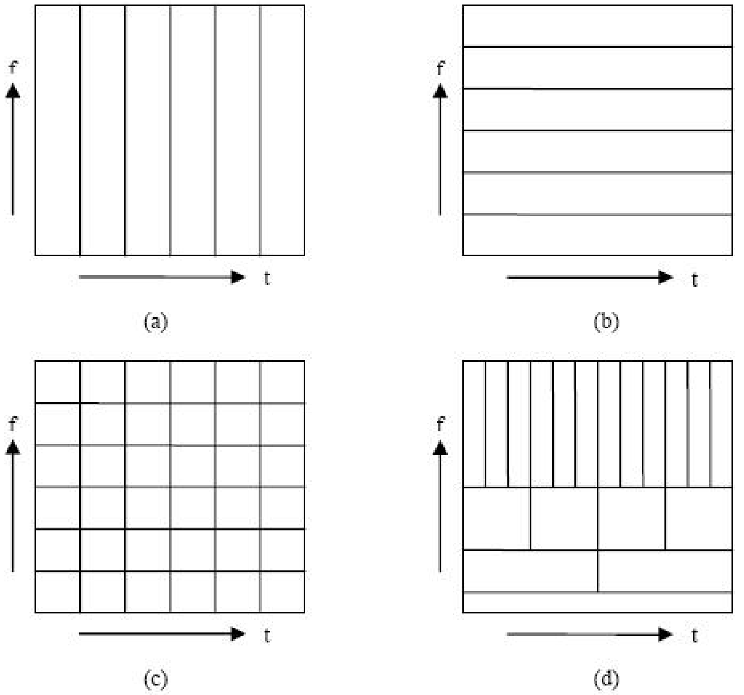


Figure 4.2: The Time-Frequency tiling for (a) Time-Domain (b) Frequency-Domain (c) STFT (d) Wavelet.

In the wavelet transform, as frequency increases, the time resolution increases; likewise, as frequency decreases, the frequency resolution increases.

Thus, a certain high frequency component can be located more accurately in time than a low frequency component and a low frequency component can be located more accurately in frequency compared to a high frequency component.

In summary, the main difference between two transforms is that wavelets are localized in both time and frequency whereas the standard Fourier transform is only localized in frequency. The Short-time Fourier transform (STFT) is also time and frequency localized but there are issues with the frequency time resolution and wavelets often give a better signal representation using multiresolution analysis. Figure 4.2 helps to understand the difference visually. Plot 4.2(a) shows the time-frequency tiling in the time-domain plane and plot 4.2(b) shows the tiling in frequency-domain plane. It is seen that plot 4.2(a) does not give any frequency information and plot 4.2(b) does not give any time information. Similarly plot 4.2(c) shows the tiling in STFT and plot 4.2(d) shows the tiling in wavelet transform. It is seen that STFT gives a fixed resolution at all times, whereas wavelet transform gives a variable resolution.

Chapter 5

Support vector regression machine

5.1 Statistical Learning Theory

Statistical Learning Theory (SLT) addresses a key question [Ped98] that “arises when constructing predictive models from data-how to decide whether a particular model is adequate or whether a different model would produce better predictions”. SLT is a framework in which learning from examples can be studied in a principled way. Whereas classical statistics typically assumes that the form of the correct model is known and the objective is to estimate the model parameters, statistical learning theory presumes that the correct form is completely unknown and the goal is to identify the best possible model from a set of competing models. The models need not have the same mathematical form and none of them need to be correct. The theory provides a sound statistical basis for assessing model adequacy under these circumstances, which are precisely the circumstances encountered in machine learning, pattern recognition, and exploratory data analysis.

5.1.1 Setting a learning problem

Vapnik [Vap99] explains that “the model of learning from examples can be described using three components:

1. a generator of random vectors, drawn independently from a fixed but unknown distribution $P(x)$;
2. a supervisor that returns an output vector y for every input vector x, according to a conditional distribution function[1] $P(y\,|x)$, also fixed but unknown;

[1] This is the general case which includes a case where the supervisor uses a function $y = f(x)$.

3. a learning machine capable of implementing a set of functions $f(x, \alpha)$, $\alpha \in \Lambda$".

The problem of learning is that of choosing from the given set of functions $f(x, \alpha)$, $\alpha \in \Lambda$, the one which predicts the supervisor's response in the best possible way. The selection is based on a training set of random independent identically distributed (i.i.d.) observations drawn according to $P(x, y) = P(x)P(y\,|x)$,

$$(x_1, y_1), ..., (x_\ell, y_\ell). \tag{5.1}$$

Here, (5.1) denotes a training data set.

5.1.2 Problem of risk minimization

Estimating the performance of competing models is the central issue in statistical learning theory. Performance is measured through the use of loss functions. In other words, in order to choose the best available approximation to the supervisor's response, one measures the loss or discrepancy $L(y, f(x, \alpha))$ between the response y of the supervisor to a given input x and the response $f(x, \alpha)$ provided by the learning machine. Consider the expected value of the loss, given by the risk functional

$$R(\alpha) = \int L(y, f(x, \alpha))dP(x, y). \tag{5.2}$$

The goal is to find the function $f(x, \alpha_0)$ which minimizes the risk functional $R(\alpha)$ (over the class of functions $f(x, \alpha)$, $\alpha \in \Lambda$) in the situation where the joint probability distribution $P(x, y)$ is unknown and the only available information is contained in the training set (5.1).

Regression estimation is one of the typical problems such as pattern recognition, regression estimation, and density estimation in learning problems based on the statistical learning theory. Let the supervisor's answer y be a real value, and let $f(x, \alpha)$, $\alpha \in \Lambda$ be a set of real functions which contains the regression function

$$f(x, \alpha_0) = \int y dP(y\,|x).$$

It is known that if $f(x, \alpha) \in L$, then the regression function is the one which minimizes the functional (5.2) with the the following loss function:

$$L(y, f(x, \alpha)) = (y - f(x, \alpha))^2. \tag{5.3}$$

Thus the problem of regression estimation is the problem of minimizing the risk functional (5.2) with the loss function (5.3) in the situation where the probability measure $P(x, y)$ is unknown but the data (5.1) are given.

The general setting of the learning problem can be described as follows. Let the probability measure $P(z)$ be defined on the space Z. Consider the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$. The goal is to minimize the risk functional

$$R(\alpha) = \int Q(z, \alpha) dP(z), \tag{5.4}$$

$\alpha \in \Lambda$ if probability measure $P(z)$ is unknown but an i.i.d. sample

$$z_1, ..., z_\ell \tag{5.5}$$

is given. The learning problems considered above are particular cases of this general problem of minimizing the risk functional (5.4) on the basis of empirical data (5.5), where z describes a pair (x, y) and $Q(z, \alpha)$ is the specific loss function. Below we will describe results obtained for the general statement of the problem. To apply it for specific problems one has to substitute the corresponding loss functions in the formulas obtained.

5.1.3 Empirical Risk Minimization

It is assumed that the expected risk is defined on a large class of functions F and we will denote by f_0 the function which minimizes the expected risk in F. If we allow f_0 to be taken from a very large class of functions F, we can always find a f_0 that leads to a rather small value of risk. The function f_0 is our ideal estimator, and it is often called the target function. This function cannot be found in practice, because the probability distribution $P(x, y)$ in (5.2) that defines the expected risk is unknown, and only a sample of it, the data set (5.1), is available. To overcome this shortcoming, we need an induction principle that we can use to learn from the limited number of training data we have. The SLT, as developed by Vapnik [Vap98], builds on the so-called Empirical Risk Minimization (ERM) induction principle. The ERM method consists in using the data set (5.1) to build a stochastic approximation of the expected risk. Therefore, in order to minimize the risk functional (5.4), for an unknown probability measure $P(z)$, the following induction principle is usually used. The expected risk functional $R(\alpha)$ is replaced by the empirical risk functional

$$R\text{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z, \alpha) \tag{5.6}$$

constructed on the basis of the training set (5.5). The principle is to approximate the function $Q(z, \alpha_0)$ which minimizes risk (5.4) by the function $Q(z, \alpha_\ell)$ which minimizes empirical risk (5.6). This principle is called the empirical risk minimization induction principle (ERM principle). The ERM principle is quite general. The classical methods for solving a specific learning problem, such as the least squares method in the problem of regression estimation or the maximum likelihood method in the problem of density estimation are realizations of the ERM principle for the specific loss functions. Indeed, in order to specify the regression problem, one introduces a $n+1$-dimensional variable $z=(x,y)=(x^1,...,x^n,y)$ and uses loss function (5.3). Using this loss function in the functional (5.6) yields the functional

$$R_{\mathrm{emp}}(\alpha)=\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i-f(x,\alpha))^2,$$

which one needs to minimize in order to find the regression estimate (i.e., the least square method).

5.1.4 Four ingredients of Learning Theory

Learning theory has to address the following four questions [Vap99]:

1. What are the conditions for consistency of the ERM principle? To answer this question one has to specify the necessary and sufficient conditions for convergence in probability of the following sequences of the random values.

   (a) The values of risks $R(\alpha_\ell)$ converging to the minimal possible value of the risk $R(\alpha_0)$ where $R(\alpha_\ell)$, $\ell=1,2,...$ are the expected risks for functions $Q(z,\alpha_\ell)$, each minimizing the empirical risk $R_{\mathrm{emp}}(\alpha_\ell)$,

   $$R(\alpha_\ell)\rightarrow P_{\ell\rightarrow\infty}R(\alpha_0). \tag{5.7}$$

   (b) The values of obtained empirical risks $R_{\mathrm{emp}}(\alpha_\ell)$, $i=1,2,...$ converging to the minimal possible value of the risk $R(\alpha_0)$

   $$R_{\mathrm{emp}}(\alpha_\ell)\rightarrow P_{\ell\rightarrow\infty}R(\alpha_0). \tag{5.8}$$

   Equation (5.7) shows that solutions found using ERM converge to the best possible one. Equation (5.8) shows that values of empirical risk converge to the value of the smallest risk.

2. How fast does the sequence of smallest empirical risk values converge to the smallest actual risk? In other words, what is the rate of generalization of a learning machine that implements the empirical risk minimization principle?

3. How can one control the rate of convergence (the rate of generalization) of the learning machine?

4. How can one construct algorithms that can control the rate of generalization?

The answers to these questions form the four ingredients of learning theory:

1. the theory of consistency of learning processes;
2. the nonasymptotic theory of the rate of convergence of learning processes;
3. the theory of controlling the generalization of learning processes;
4. the theory of constructing learning algorithms.

5.1.5 Consistency of learning processes

The theory of consistency is an asymptotic theory. It describes the necessary and sufficient conditions for convergence of the solutions obtained using the proposed method to the best possible as the number of observations is increased. The question arises: Why do we need a theory of consistency if our goal is to construct algorithms for a small (finite) sample size? We need a theory of consistency because it provides not only sufficient but necessary conditions for convergence of the empirical risk minimization inductive principle. Therefore, any theory of the empirical risk minimization principle must satisfy the necessary and sufficient conditions. The main capacity concept, the so-called Vapnik-Cervonenkis (VC) entropy has been introduced by Vapnik and Chervonenkis [Vap71], [Vap81], [Vap91] which defines the generalization ability of the ERM principle. It is shown that the nonasymptotic theory of learning is based on different types of bounds that evaluate this concept for a fixed amount of observations. The key theorem of the theory concerning the ERM-based learning processes is the following [Vap99].

"The Key Theorem:

The Key Theorem: Let $Q(z, \alpha)$, $\alpha \in \Lambda$ be a set of functions that has a bounded loss for probability measure $P(z)$

$$A \leq \int Q(z, \alpha) dP(z) \leq B,$$

and $\forall \alpha \in \Lambda$. Then for the ERM principle to be consistent it is necessary and sufficient that the empirical risk $R_{\mathrm{emp}}(\alpha)$ converges uniformly to the actual risk $R(\alpha)$ over the set $Q(z, \alpha)$, $\alpha \in \Lambda$ as follows:

$$\lim_{\ell \to \infty} \mathrm{Prob} \left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\mathrm{emp}}(\alpha)) > \varepsilon \right\} = 0, \tag{5.9}$$

and $\forall \varepsilon$. This type of convergence is called uniform one-sided convergence". In other words, according to the Key theorem the conditions for consistency of the ERM principle are equivalent to the conditions for existence of uniform one-sided convergence (5.9). This theorem is called the Key theorem because it asserts that any analysis of the convergence properties of the ERM principle must be a worst case analysis. The necessary condition for consistency (not only the sufficient condition) depends on whether or not the deviation for the worst function over the given set of functions

$$\Delta(\alpha_{\text{worst}}) = \sup_{\alpha\in\Lambda}(R(\alpha) - R\text{emp}(\alpha)),$$

converges in probability to zero. From this theorem it follows that the analysis of the ERM principle requires an analysis of the properties of uniform convergence of the expectations to their probabilities over the given set of functions. To describe the necessary and sufficient condition for uniform convergence (5.9), we explain a concept called the entropy of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$ on the sample of size ℓ.

Let $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ be a set of bounded loss functions. Using this set of functions and the training set (5.5) one can construct the following set of ℓ-dimensional real-valued vectors

$$q(\alpha) = Q(z_1, \alpha), ..., Q(z_\ell, \alpha)), \alpha \in \Lambda \tag{5.10}$$

This set of vectors belongs to the ℓ-dimensional cube with the edge $B-A$ and has a finite ε-net in the metric C. Let $N = N^\Lambda(\varepsilon; z_1, ..., z_\ell)$ be the number of elements of the minimal[2] ε-net of the set of vectors $q(\alpha), \alpha \in \Lambda$. The logarithm of the (random) value $N^\Lambda(\varepsilon; z_1, ..., z_\ell)$

$$H^\Lambda(\varepsilon; z_1, ..., z_\ell) = \ln N^\Lambda(\varepsilon; z_1, ..., z_\ell),$$

is called the random VC-entropy of the set of functions $A \leq Q(z, \alpha) \leq B$ on the sample size $z_1, ..., z_\ell$. The random entropy describes the diversity of the set of functions on the given data. It is a random variable since it is constructed using random i.i.d. data. The expectation of the random VC-entropy

$$H^\Lambda(\varepsilon; \ell) = EN^\Lambda(\varepsilon; z_1, ..., z_\ell)$$

is called the VC-entropy of the set of functions $A \leq Q(z, \alpha) \leq B$, $\alpha \in \Lambda$ on the sample of the size ℓ. Here expectation is taken with respect to product-measure $P(z_1, ..., z_\ell) = P(z_1), ..., P(z_\ell)$. The main results of the theory of uniform convergence of the empirical risk to actual risk for bounded loss function includes the following theorem [Vap71].

Theorem: For uniform two-sided convergence of the empirical risks to the actual risks

$$\lim_{\ell\to\infty}\text{Prob}\left\{\sup_{\alpha\in\Lambda}\left|R(\alpha) - R\text{emp}(\alpha)\right| > \varepsilon\right\} = 0, \forall\varepsilon, \tag{5.11}$$

[2] The set of vectors $q(\alpha), \alpha \in \Lambda$ has minimal ε-net $q(\alpha_1), ..., q(\alpha_N)$ if there exist $N = N^\Lambda(\varepsilon; z_1, ..., z_\ell)$ vectors $q(\alpha_1), ..., q(\alpha_N)$, such that for any vector $q(\alpha^*), \alpha^* \in \Lambda$ one can find among these N vectors one $q(\alpha_r)$ which is ε-close to this vector (in a given metric). For a C metric that means

$$\rho(q(\alpha^*), q(\alpha_r)) = \max_{1\leq i\leq\ell}|Q(z_i, \alpha^*) - Q(z_i, \alpha_r)| \leq \varepsilon.$$

N is minimal number of vectors which possess this property.

it is necessary and sufficient that the equality

$$\lim_{\ell\to\infty}\frac{H^{\Lambda}(\varepsilon,\ell)}{\ell}=0,\forall\varepsilon>0 \tag{5.12}$$

be valid. According to the key assertion, this implies the necessary and sufficient conditions for consistency of the ERM principle.

Under which conditions is the asymptotic rate of convergence fast? The asymptotic rate of convergence is fast, if for any $\ell>\ell_0$ the exponential bound

$$P\left\{R(\alpha_\ell)-R(\alpha_0)>\varepsilon\right\}<e^{-c\varepsilon^2\ell}$$

holds true, where $c>0$ is some constant. The equation

$$\lim_{\ell\to\infty}\frac{H^{\Lambda}_{\text{ann}}(\ell)}{\ell}=0$$

describes the sufficient condition for fast convergence. It guarantees a fast asymptotic rate of convergence. Note that both the equation describing the necessary and sufficient condition for consistency and the one that describes the sufficient condition for fast convergence of the ERM method are valid for a given probability measure $P(z)$ (both VC-entropy $H^{\Lambda}(\ell)$ and VC-annealed entropy $H^{\Lambda}_{\text{ann}}(\ell)$ are constructed using this measure). However our goal is to construct a learning machine for solving many different problems (i.e., for many different probability measures).

Under what conditions is the ERM principle consistent and rapidly converging, independently of the probability measure? The following equation describes the necessary and sufficient conditions for consistency of ERM for any probability measure

$$\lim_{\ell\to\infty}\frac{G^{\Lambda}(\ell)}{\ell}=0,$$

where $G^{\Lambda}(\ell)$ is a growth function. This condition is also sufficient for fast convergence. It describes the conditions under which the learning machine implementing ERM principle has an asymptotic high rate of convergence independently of the problem to be solved.

5.1.6 Bounds on rate of convergence of learning machine

In order to estimate the quality of the ERM method for a given sample size, it is necessary to obtain nonasymptotic bounds on the rate of uniform convergence. A nonasymptotic bound of the rate of convergence can be obtained using a new capacity concept, called the VC dimension, which allows us to obtain a constructive bound for the growth function. The concept of VC-dimension is based on a remarkable property of the growth function $G^{\Lambda}(\ell)$.

Theorem: Any growth function either satisfies the equality

$$G^{\Lambda}(\ell) = \ell \text{ln} 2$$

or is bounded by the inequality

$$G^{\Lambda}(\ell) < h\left(\text{ln}\frac{\ell}{h} + 1\right)$$

where h is an integer for which

$$G^{\Lambda}(h) = h\text{ln}2,$$

$$G^{\Lambda}(h+1) \neq (h+1)\text{ln}2.$$

In other words, the growth function will be either a linear function or will be bounded by a logarithmic function. (For example, it cannot be of the form $G^{\Lambda}(\ell) = c\sqrt{\ell}$).

Equivalently to define, the VC-dimension of a set of indicator functions $Q(z, \alpha), \alpha \in \Lambda$, is the maximum number h of vectors $z_1, ..., z_h$ which can be separated in all 2^h possible ways using functions of this set[3] (shattered by this set of functions). If for any n there exists a set of n vectors which can be shattered by the set $Q(z, \alpha), \alpha \in \Lambda$, then the VC-dimension is equal to infinity. Let $a \leq Q(z, \alpha) \leq A, \alpha \in \Lambda$ be a set of real-valued functions bounded by constants a and A (a can approach $-\infty$ and A can approach ∞). Let us consider along with the set of real-valued functions $Q(z, \alpha), \alpha \in \Lambda$, the set of indicator functions

$$I(z, \alpha, \beta) = \theta\left\{Q(z, \alpha) - \beta\right\}, \alpha \in \Lambda \tag{5.13}$$

where $a < \beta < A$ is some constant, $\theta(u)$ is the step function

$$\theta(u) = \begin{cases} 0, & \text{if} \quad u < 0 \\ 1, & \text{if} \quad u \geq 0. \end{cases}$$

The VC-dimension of the set of real valued functions $Q(z, \alpha), \alpha \in \Lambda$ is defined to be the VC-dimension of the set of indicator functions (5.13).

In following, distribution independent bounds for the rate of convergence of learning processes is addressed. Consider sets of functions which possess a finite VC-dimension h. Two cases are distinguished: 1) the case where the set of loss functions $Q(z, \alpha), \alpha \in \Lambda$ is a set of totally bounded functions; 2) the case where the set of loss functions $Q(z, \alpha), \alpha \in \Lambda$ is not necessarily a set of totally bounded functions.

Case 1- The Set of Totally Bounded Functions: Without restriction in generality, we assume that

[3] Any indicator function separates a set of vectors into two subsets: the subset of vectors for which this function takes value zero and the subset of vectors for which it takes value one.

$$0 \leq Q(z,\alpha) \leq B, \;\; \alpha \in \Lambda \tag{5.14}$$

The main result in the theory of bounds for sets of totally bounded functions is the following [Vap98],[Vap95].

Theorem: With probability at least $1-\eta$, the inequality

$$R(\alpha) \leq R\mathrm{emp}(\alpha) + \frac{B\varepsilon}{2}\left(1+\sqrt{1+\frac{4R\mathrm{emp}(\alpha)}{B\varepsilon}}\right) \tag{5.15}$$

holds true simultaneously for all functions of the set (5.14), where

$$\varepsilon = 4\frac{h\left(\ln\frac{2\ell}{h}+1\right)-\ln\eta}{\ell}. \tag{5.16}$$

This theorem provides bounds for the risks of all functions of the set (5.13), including the function $Q(z,\alpha_\ell)$ which minimizes empirical risk (5.6). The bounds follow from the bound on uniform convergence (5.11) for sets of totally bounded functions that have finite VC dimension.

Case 2- The Set of Unbounded Functions: Consider the set of (nonnegative) unbounded functions $0 \leq Q(z,\alpha), \alpha \in \Lambda$. Without additional information about the set of unbounded functions and/or probability measures, it is impossible to obtain an inequality of type (5.15). Below we use the following information:

$$\mathrm{sup}_{\alpha\in\Lambda}\frac{\left(\int Q^p(z,\alpha)dP(z)\right)^{1/P}}{\int Q(z,\alpha)dP(z)} \leq \tau < \infty \tag{5.17}$$

where $p > 1$ is some fixed constant[4]. The main result for the case of unbounded sets of loss functions is the following [Vap98],[Vap95].

Theorem: With probability at least $1-\eta$ the inequality

$$R(\alpha) \leq \frac{R\mathrm{emp}(\alpha)}{(1-a(p)\tau\sqrt{\varepsilon})_+}, a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}} \tag{5.18}$$

holds true simultaneously for all functions of the set, where ε is determined by (5.17), $(a)_+ = \max(a,0)$. The theorem bounds the risks for all functions of the set (including the function $Q(z,\alpha_\ell)$).

[4] This inequality describes some general properties of distribution functions of the random variables $\xi_\alpha = Q(z,\alpha)$, generated by the $P(z)$. It describes the tails of distributions (the probability of big values for the random variables ξ_α). If the inequality (5.17) with $p > 2$ holds, then the distributions have so-called light tails (large values do not occurs very often). In this case rapid convergence is possible. If, however, (5.17) holds only for $p < 2$ (large values of the random variables ξ_α occur rather often) then the rate of convergence will be small (it will be arbitrarily small if p is sufficiently close to one).

5.1.7 Structural Risk Minimization Induction Principle

The core results in statistical learning theory are a series of probability bounds developed by Vapnik and Chervonenkis [Vap71], [Vap81], and [Vap91] that define small-sample confidence regions for the maximum difference between expected or true risk and empirical risk. The confidence regions differ from those obtained in classical statistics in three respects. First, they do not assume that the chosen model is correct. Second, they are based on small-sample statistics and are not asymptotic approximations. Third, a uniform method is used to take into account the degree to which overfitting can occur for a given set of competing models. This method is based on a measurement known as the Vapnik-Chervonenkis (VC) dimension. Conceptually speaking, the VC dimension of a set of models is the maximum number of data vectors for which overfitting is virtually guaranteed in the sense that one can always find a specific model that fits the data exactly [Ped98]. Hence, the SLT provides probabilistic bounds on the distance between the empirical and expected risk of any function (therefore including the minimizer of the empirical risk in a function space that can be used to control overfitting). Hence, to avoid overfitting (to get a small confidence interval) one has to construct networks with small VC-dimension.

The theory for controlling the generalization of a learning machine is devoted to constructing an induction principle for minimizing the risk functional which takes into account the size of the training set (an induction principle for a small sample size[5]). The goal is to specify methods which are appropriate for a given sample size.

The ERM principle is intended for dealing with a large sample size. Indeed, the ERM principle can be justified by considering the inequality (5.15). When ℓ/h is large, the second summand on the right hand side of inequality (5.15) becomes small. The actual risk is then close to the value of the empirical risk. In this case, a small value of the empirical risk provides a small value of (expected) risk. However, if ℓ/h is small, then even a small $R_{\text{emp}}(\alpha_\ell)$ does not guarantee a small value of risk.

In general, straight minimization of the empirical risk in F can be problematic. First, it is usually an ill-posed problem [Tik77] in the sense that there might be many, possibly infinitely many, functions minimizing the empirical risk. Second, it can lead to overfitting, meaning that although the minimum of the empirical risk can be very close to zero, the expected risk-which is what we are really interested in-can be very large. In fact, overfitting occurs when the best model relative to the training data tends to perform significantly worse when applied to new data.

In such the conditions, the minimization for $R(\alpha)$ requires a new principle, based on the simultaneous minimization of two terms in (5.15) one of which depends on the value of the empirical risk while the second depends on the VC-dimension of the set of functions. To minimize risk in this case, it is necessary to find a method which, along with minimizing the value of empirical risk, controls the VC-dimension of the learning machine. The following principle,

[5] The sample size ℓ is considered to be small if ℓ/h is small, say $\ell/h < 20$ [Vap99].

which is called the principle of structural risk minimization (SRM), is intended to minimize the risk functional with respect to both empirical risk and VC-dimension of the set of functions.

Let S the set of functions $Q(z, \alpha), \alpha \in \Lambda$ be provided with a structure: so that S is composed of the nested subsets of functions $S_k = \{Q(z, \alpha), \alpha \in \Lambda\}$ such that

$$S_1 \subset S_2 \subset, ..., S_n \tag{5.19}$$

and $S^* = \cup_k S_k$. An admissible structure is one satisfying the following three properties.

1. The set S^* is everywhere dense in S.
2. The VC-dimension h_k of each set S_k of functions is finite.
3. Any element S_k of the structure contains totally bounded functions $0 \leq Q(z, \alpha) \leq B_k, \alpha \in \Lambda_k$.

The SRM principle suggests that for a given set of observations $z_1, ..., z_\ell$ choose the element of structure S_n, where $n = n(\ell)$ and choose the particular function from S_n for which the guaranteed risk (5.15) is minimal. The SRM principle actually suggests a trade-off between the quality of the approximation and the complexity of the approximating function (as n increases, the minima of empirical risk are decreased; however, the term responsible for the confidence interval (summand in (5.15)) is increased. The SRM principle takes both factors into account.). The main results of the theory of SRM are the following [Dev96], [Vap98].

Theorem: For any distribution function, the SRM method provides convergence to the best possible solution with probability one.

In other words, SRM method is universally strongly consistent.

Theorem: For admissible structures the method of structural risk minimization provides approximations $Q(z, \alpha_\ell^{n(\ell)})$ for which the sequence of risks $R(\alpha_\ell^{n(\ell)})$ converges to the best one $R(\alpha_0)$ with asymptotic rate of convergence[6]

$$V(\ell) = r_{n(\ell)} + B_{n(\ell)} \sqrt{\frac{h_{n(\ell)} \ln \ell}{\ell}}, \tag{5.20}$$

if the law $n = n(\ell)$ is such that

[6] We say that the random variables $\xi_\ell, \ell = 1, 2, ...$ converge to the value ξ_0 with asymptotic rate $V(\ell)$, if there exists constant C such that

$$V^{-1}(\ell) \left|\xi_\ell - \xi_0\right| \to P_{\ell \to \infty} C.$$

$$\lim_{\ell\to\infty} \frac{B^2_{n(\ell)} h_{n(\ell)} \ln \ell}{\ell} = 0. \tag{5.21}$$

In (5.20), B_n is the bound for functions from S_n and $r_n(\ell)$ is the rate of approximation

$$r_n = \inf_{\alpha \in \Lambda_n} \int Q(z, \alpha) dP(z) - \inf_{\alpha \in \Lambda} \int Q(z, \alpha) dP(z).$$

To implement the SRM induction principle in learning algorithms one has to control two factors that exist in the bound (5.15) which has to be minimized: 1) the value of empirical risk; 2) the capacity factor (to choose the element S_n with the appropriate value of VC dimension).

Support Vector Machine as a well-known learning algorithm is firmly grounded in the framework of statistical learning theory, or VC theory. One of the main practical problems which can be solved in the Statistical Learning Theory framework by the Support Vector Machine is regression estimation. In the rest of the current chapter, we address the problem of regression estimation by the Support Vector Machine formulation.

5.2 The ε-insensitive support vector regression

Two sets of random variables $x \in X \subseteq R^d$ and $y \in Y \subseteq R$ related by a probabilistic relationship are considered. The relationship is probabilistic because generally an element of X does not determine uniquely an element of Y, but rather a probability distribution on Y. This can be formalized assuming that an unknown probability distribution $P(x, y)$ is defined over the set $X \times Y$. We are provided with examples of this probabilistic relationship, that is with a data set $D_\ell \equiv \{(x_i, y_i) \in X \times Y\}_{i=1}^{\ell}$ called training set, obtained by sampling ℓ times the set $X \times Y$ according to $P(x, y)$. The problem of learning consists in, given the data set D_ℓ, providing an estimator, that is a function $f : X \to Y$, that can be used, given any value of $x \in X$, to predict a value y. For an example consider a case where x is a set of parameters, such as pose or facial expressions, y is a motion field relative to a particular reference image of a face, and $f(x)$ is a regression function which maps parameters to motion. For an example from finance, the training data set might be the exchange rates for some currency or stock index measured at subsequent days together with corresponding econometric indicators.

5.2.1 Standard formulation

In ε-SV Regression of Vapnik, the goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time, is as flat as possible. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this constant. This may be important if one wants to be sure not to lose more than ε money when dealing with exchange rates, for instance. For simplicity reasons, we begin by describing the case of linear functions f, taking the form

$$f(x) = \langle w, x \rangle + b \tag{5.22}$$

with

$$w \in X, b \in R$$

where $\langle ., . \rangle$ denotes the dot product in X. Flatness in the case of (5.22) means that one seeks small w. One way to ensure this, is to minimize the Euclidean norm, i.e., $\|w\|^2 = \langle w, w \rangle$. Formally, we can write this problem as a convex optimization problem by requiring:

$$\texttt{minimize } \frac{1}{2} \|w\|^2 \tag{5.23}$$

subject to

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon. \end{cases}$$

The implicit assumption in (5.23) is that such a function f actually exists that approximates all pairs (x_i, y_i) with ε precision, or in other words, that the convex optimization problem is feasible. Sometimes, however, this may not be the case, or we also may want to allow for some errors. One can introduce slack variables ξ_i, ξ_i^* to cope with otherwise infeasible constraints of the optimization problem (5.23). Hence, we arrive at the formulation (Smola and Schölkopf [Smo04]):

$$\texttt{minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \tag{5.24}$$

subject to

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \end{cases}$$

and $\xi_i, \xi_i^* \geq 0$. The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. This corresponds to dealing with a so called ε-insensitive loss function, $|y - f(x)|_\varepsilon$, described by a binary loss function, where only the corresponding input points with error larger than ε contribute to the cost insofar, as the deviations are penalized in a linear fashion. The ε-insensitive loss function, $|y - f(x)|_\varepsilon$, is specified by

$$|y - f(x)|_\varepsilon = \begin{cases} 0, & \text{if} \quad |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon, & \text{if} \quad \text{otherwise.} \end{cases} \tag{5.25}$$

Figure 5.1 depicts graphically the ε-insensitive loss function. It turns out that in most cases the optimization problem (5.24) can be solved more easily in its dual formulation. Moreover, support vector machine can be solved based on nonlinear functions utilizing kernel functions in high dimensional space. Hence, a standard dualization method utilizing Lagrange multipliers is used. The key idea is to construct a Lagrange function from the objective function and the corresponding constraints, by introducing a dual set of variables. The Lagrange function has a saddle point with respect to the primal and dual variables at the solution. We have

$$\begin{aligned} L := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) - \sum_{i=1}^{\ell}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ - \sum_{i=1}^{\ell}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ - \sum_{i=1}^{\ell}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b). \end{aligned} \tag{5.26}$$

Here L is the Lagrangian and $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ are Lagrange multipliers. Hence, the dual variables have to satisfy positivity constraints, i.e., $\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \geq 0$. It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality.

$$\partial_b L = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i) = 0 \tag{5.27}$$

$$\partial_w L = w - \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)x_i = 0 \tag{5.28}$$

$$\partial_{\xi_i} L = C - \alpha_i - \eta_i = 0 \tag{5.29}$$

$$\partial_{\xi_i^*} L = C - \alpha_{i=1}^* - \eta_i^* = 0 \tag{5.30}$$

Substituting (5.27), (5.28), (5.29), and (5.30) into (5.26) yields the dual optimization problem. Finally, we have a dual optimization problem as

$$\texttt{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\varepsilon\sum_{i=1}^{\ell}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i(\alpha_i - \alpha_i^*) \end{cases} \tag{5.31}$$

subject to

$$\begin{cases} \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C]. \end{cases}$$

In deriving (5.31) we already eliminated the dual variables η_i, η_i^* through the conditions (5.29) and (5.30) which can be reformulated as $\eta_i = C - \alpha_i$ and $\eta_i^* = C - \alpha_i^*$. The Equation (5.28) can be rewritten as

$$w = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)x_i. \tag{5.32}$$

Therefore, the estimated regression function can be obtained by

$$f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b, \tag{5.33}$$

where α_i and α_i^* are unknown variables of interest to be found by solving the optimization problem. The estimated regression (5.33) is the so-called Support Vector expansion, i.e., w can be completely described as a linear combination of the training patterns x_i. In a sense, the complexity of a function's representation by SVs is independent of the dimensionality of the input space X, and depends only on the number of SVs. Moreover, note that the complete algorithm can be described in terms of dot products between the data. Even when evaluating $f(x)$, we need not compute w explicitly. These observations will come in handy for the formulation of a nonlinear extension.

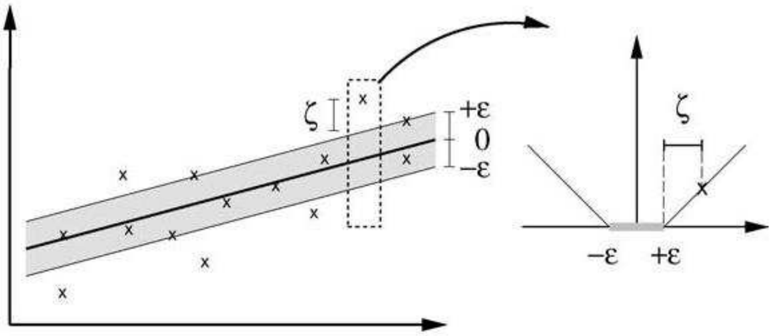


Figure 5.1: Figure is adopted from Smola and Schölkopf [Smo04]. The left panel represents fitting a linear regression by SVR machine. Errors equal to or smaller than $\pm\varepsilon$ are ignored in fitting the curve. The right panel represents corresponding ε-insensitive loss function.

Computing b can be done by exploiting the so called Karush-Kuhn-Tucker (KKT) conditions. These state that at the point of the solution, the product between dual variables and constraints has to vanish, i.e.,

$$\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) = 0 \tag{5.34}$$

$$\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b) = 0, \tag{5.35}$$

and

$$(C - \alpha_i)\xi_i = 0 \tag{5.36}$$

$$(C - \alpha_i^*)\xi_i^* = 0. \tag{5.37}$$

This allows us to make several useful conclusions. Firstly only samples (x_i, y_i) with corresponding $\alpha_i, \alpha_i^* = C$ lie outside the ε-insensitive tube. Secondly, $\alpha_i \alpha_i^* = 0$, i.e., there can never be a set of dual variables α_i, α_i^* which are both simultaneously nonzero. This allows us to conclude [Smo04] that

$$\varepsilon - y_i + \langle w, x_i \rangle + b \geq 0 \text{ and } \xi_i = 0 \text{ if } \alpha_i < C \tag{5.38}$$

$$\varepsilon - y_i + \langle w, x_i \rangle + b \leq 0 \quad \xi_i = 0 \text{ if } \alpha_i > C. \tag{5.39}$$

In conjunction with an analogous analysis on α_i^* we have

$$\max \{-\varepsilon + y_i - \langle w, x_i \rangle \,|\alpha_i < C \text{ or } \alpha_i^* > 0\} \leq b \leq$$

$$\min \{-\varepsilon + y_i - \langle w, x_i \rangle \,|\alpha_i > 0 \text{ or } \alpha_i^* < C\}. \tag{5.40}$$

If some $\alpha_i^* \in (0, C)$ the inequalities become equalities. See also Keerthi et al. [Kee01] for further means of choosing b. From (5.34) and (5.35) it follows that only for $|f(x_i) - y_i| \geq \varepsilon$ the Lagrange multipliers may be nonzero, or in other words, for all samples inside the ε-tube the α_i, α_i^* vanish: for $|f(x_i) - y_i| < \varepsilon$ the second factor in (5.34) and (5.35) is nonzero, hence α_i, α_i^* have to be zero such that the KKT conditions are satisfied. Therefore, we have a sparse expansion of w in terms of x_i (i.e., we do not need all x_i to describe w). The examples that come with nonvanishing coefficients are called Support Vectors.

5.3 Nonlinearity

In many real world cases one has to make the SV algorithm nonlinear. This, for instance, could be achieved by simply preprocessing the training patterns x_i by a map $\Phi : X \rightarrow F$ into some feature space F and then applying the standard SV regression algorithm.

Consider the map $\Phi : R^2 \rightarrow R^3$ with $\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$, for an example. It is understood that the subscripts in this case refer to the components of $x \in R^2$. Training a linear SV machine on the preprocessed features would yield a quadratic function. While this approach seems reasonable in the particular example above, it can easily become computationally infeasible for both polynomial features of higher order and higher dimensionality, as the number of different monomial features of degree p is $\binom{d+p-1}{p}$ where $d = \dim(X)$. Typical values for OCR tasks with good performance are $p = 7, d = 28 \times 28 = 784$, corresponding to approximately 3.7×10^{16} features.

Clearly this approach is not feasible and we have to find a computationally cheaper way. The key observation in Boser, Guyon and Vapnik [Bos92] is that for the feature map of the abovementioned example, we have

$$\left\langle (x_1^2, \sqrt{2}x_1x_2x_2^2), (x_1^{'2}, \sqrt{2}x_1^{'}x_2^{'}, x_2^{'2}) \right\rangle = \left\langle x, x^{'} \right\rangle^2 .$$

In fact, the SV algorithm only depends on dot products between patterns x_i. Hence, it suffices to know $k(x, x^{'}) := \left\langle \Phi(x), \Phi(x^{'}) \right\rangle$ rather than Φ explicitly which allows us to restate the SV optimization problem:

$$\texttt{maximize} \left\{ \begin{array}{l} -\frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) \\ -\varepsilon \sum_{i=1}^{\ell}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i(\alpha_i - \alpha_i^*) \end{array} \right. \tag{5.41}$$

subject to

$$\left\{ \begin{array}{l} \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C]. \end{array} \right.$$

Likewise the expansion of f in (5.32) may be written as

$$w = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)\Phi(x_i), \tag{5.42}$$

and therefore

$$f(x) = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)k(x_i, x) + b. \tag{5.43}$$

The difference to the linear case is that w is no longer given explicitly. Also note that in the nonlinear setting, the optimization problem corresponds to finding the flattest function in feature space, not in input space. The question that arises now is, which functions $k(x, x^{'})$ correspond to a dot product in some feature space F. In 1909, Mercer [Mer09] proved a theorem which defines the general form of inner products in Hilbert spaces. The following theorem characterizes these functions (defined on X).

Theorem: The general form of the inner product in Hilbert space is defined by the symmetric positive definite function $k(x, y)$ that satisfies the condition

$$\int k(x,y)z(x)z(y)dxdy \geq 0,$$

for all functions $z(x), z(y)$ satisfying the inequality

$$\int z^2(x)dx \leq \infty.$$

Therefore, any function satisfying Mercer's condition can be used for constructing rule (5.41) which is equivalent to constructing an optimal regression function in some feature space [Vap99]. The learning machines which construct decision functions of the type (5.41) are called support vectors networks or Support Vector Machines (SVM's) [Vap99]. This name stresses that for constructing this type of machine, the idea of expanding the solution on support vectors is crucial. In the SVM machines, the complexity of construction depends on the number of support vectors rather than on the dimensionality of the feature space. Using different expressions for inner products $k(x, x^{'})$ one can construct different learning machines with arbitrary types of (nonlinear in input space) decision surfaces.

In summary, by replacing the inner product with an appropriately chosen kernel function, one can implicitly perform a nonlinear mapping to a high dimensional feature space [Cri00]. Typical choices of kernels include, for example, Gaussian, polynomial, and Radial Basis Function (RBF) kernels [Sch00]. The SVM possesses "some useful properties [Vap99]:

- The optimization problem for constructing an SVM has a unique solution.
- The learning process for constructing an SVM is rather fast.
- Simultaneously with constructing the decision rule, one obtains the set of support vectors.
- Implementation of a new set of decision functions can be done by changing only one function (kernel $K(x_i, x)$), which defines the dot product in Z-space".

5.4 Implementation algorithms

While there has been a large number of implementations of SV algorithms in the past years [Smo04], we focus on some of the most effective ones which are useful for practitioners who would like to actually code a SV machine by themselves. For doing so, we very briefly cover some major implementation algorithms and optimization packages. There are, however, many other softwares implementing a QP programming we do not enumerate here. First, we summarily concentrate on the most effective and popular algorithms for implementing the QP programming namely the interior point algorithm and sequential minimal optimization.

Interior point algorithm: In a nutshell, the idea of an interior point algorithm is to compute the dual of the optimization problem and solve both primal and dual simultaneously. This is done by only gradually enforcing the KKT conditions to iteratively find a feasible solution and to use the duality gap between primal and dual objective function to determine the quality of the current set of variables. For both feasible primal and dual variables, the primal objective function (of a convex minimization problem) is always greater or equal than the dual objective function. Since SVMs have only linear constraints, the constraint qualifications of the strong duality theorem (Bazaraa, Sherali and Shetty [Baz93], Theorem 6.2.4) are satisfied and it follows that the gap vanishes at optimality. Thus the duality gap is a measure how close (in terms of the objective function) the current set of variables is to the solution.

Sequential minimal optimization: The Sequential Minimal Optimization (SMO) algorithm was proposed by Platt [Pla99] that puts chunking to the extreme by iteratively selecting subsets only of size 2 and optimizing the target function with respect to them. In fact, the working set is restricted to only two elements. The main advantage is that each two-variable sub-problem can be analytically solved, so numerical optimization software are not needed. For this method, at least two elements are required for the working set. Otherwise, the equality constraint leads to a fixed optimal objective value of the sub-problem. Then, the decomposition procedure stays at the same point. Therefore, the SMO algorithm solves the SVM quadratic problem (QP) without using any numerical QP optimization steps. Instead, it chooses to solve the smallest possible optimization problem involving two elements of α_i. At every step, SMO chooses two α_i to jointly optimize and finds the optimal values for these α_i analytically, thus avoiding numerical QP optimization, and updates the SVM to reflect the new optimal values.

Most commercially available packages for Quadratic Programming (QP) can also be used to train SV machines. These are usually numerically very stable general purpose codes, with special enhancements for large sparse systems. While the latter is a feature that is not needed at all in SV problems (there the dot product matrix is dense and huge) they still can be used with good success.

OSL: This package was written by IBM-Corporation (1992). It uses a two-phase algorithm. The first step consists of solving a linear approximation of the QP problem by the simplex algorithm of Dantzig [Dan62]. Next a related very simple QP problem is dealt with. When successive approximations are close enough together, the second subalgorithm, which permits a quadratic objective and converges very rapidly from a good starting value, is used. Recently an interior point algorithm was added to the software suite.

CPLEX: This package has been provided by CPLEX-Optimization-Inc. (1994). It uses a primal-dual logarithmic barrier algorithm introduced by Megiddo [Meg89] instead with predictor-corrector step (see, e.g., Lustig, Marsten and Shanno [Lus92], Mehrotra and Sun [Meh92]).

MINOS: Written by the Stanford Optimization Laboratory (Murtagh and Saunders [Mur83]) uses a reduced gradient algorithm in conjunction with a quasi-Newton algorithm. The constraints are handled by an active set strategy. Feasibility is maintained throughout the process. On the active constraint manifold, a quasi-Newton approximation is used.

MATLAB: The large-scale algorithm is a subspace trust-region method based on the interior-reflective Newton method described in Coleman and Li [Col94] and [Col96]. Each iteration involves the approximate solution of a large linear system using the method of preconditioned conjugate gradients (PCG). For medium-scale optimization, MATLAB uses an active set method, which is also a projection method, similar to that described in Gill et al. [Gil81]. It finds an initial feasible solution by first solving a linear programming problem.

LOQO: Developed by Vanderbei [Van94], it is another example of a primal-dual interior point code which preserves the primal-dual symmetry. It is a system for solving smooth constrained optimization problems. The problems can be linear or nonlinear, convex or nonconvex, constrained or unconstrained. The only real restriction is that the functions defining the problem be smooth (at the points evaluated by the algorithm).

Chapter 6

Modeling volatility

6.1 CHARN modeling volatility

Financial time series are of extremely complex nature. However, there exist some universal phenomena that are called the stylized facts. The study of statistical properties of financial time series has revealed a wealth of interesting the stylized facts which seem to be common to a wide variety of markets, instruments and periods. Rachev and Mittnik [Rach00] argue that "a complete model should be rich enough to encompass relevant stylized facts, such as

- non–Gaussian, heavy–tailed and skewed distributions
- volatility clustering (ARCH–effects)
- temporal dependence of the tail behavior
- short– and long–range dependence".

Among these properties, volatility clustering is one of the most important stylized facts in financial time series data [Gau00]. Whereas price changes themselves appear to be unpredictable, the magnitude of those changes appear to be predictable in the sense that large changes tend to be followed by large changes-of either sign- and small changes tend to be followed by small changes.

In parametric modeling, the specification and estimation of any econometric relationship possesses many significant challenges. This is especially true when it comes to the choice of functional form, as the latter is not always suggested or prescribed by the underlying economic theory. Any misspecification of the functional form of an econometric model can have serious consequences for statistical inference, for example, the parameter estimates may be inconsistent. Moreover, an important issue that arises in all estimation strategies for dealing with volatility is exactly how anticipated values and volatility terms should be related to the information available to agents. Most strategies presume linearity, or perhaps a quadratic relationship [Pag88], normal or another specific distribution of innovations and there are serious consequences for estimation, if these assumptions are invalid. A natural response to the overwhelming variety of parametric ARCH or GARCH models involving misspecification, linearity or quadratic, and distributional assumptions, is to consider and estimate more flexible nonparametric models as a way of circumventing these difficulties as Bollerslev, Engle, and Nelson state [Bol94]. Härdle and Yang [Hae96] stress that nonparametric procedures are an interesting alternative to classical time series analysis. The nonparametric technique follows the principle of "letting the data speak for themselves", and provides guidance in choosing parametric models.

In fact, the nonparametric modeling of mean and variance function does not depend on specific structures of any quantity. In the framework of ARCH models, Gourieroux and Monfort [Gou92] model both the conditional mean and the conditional variance nonparametrically. They specify

$$Y_i = \sum_{j=1}^{J} \alpha_j I(X_i \in A_j) + \sum_{j=1}^{J} \beta_j I(X_i \in A_j)\xi_i, \tag{6.1}$$

$$X_i = (Y_{i-1}, Y_{i-2}, ..., Y_{i-m}) \in R^{md}, \ Y_i \in R^d,$$

which is called Qualitative Threshold ARCH model. Here $\{A_j\}_{j=1}^{J}$ with fixed J denotes a partition of the set of lagged values for Y, and $(\alpha_j), (\beta_j)$ are unknown parameter vectors and matrices respectively and ξ_i is white noise [Hae96]. A generalization of model (6.1) to a wider class of conditional mean and variance functions can be seen as a limit of (6.1) for $J \to \infty$ thus allowing J to be unknown

$$Y_i = f(X_i) + \sigma(X_i)\xi_i, \tag{6.2}$$

where $\xi_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{id}) \in R^d, i = m, m+1, ..., n$ are random vector variables and ξ_i are i.i.d. with $E(\xi_{1j}) = 0$, for any $1 \leq j \leq d$, $E(\xi_{1j}^2) = 1$. The mean vector function $f : R^{md} \to R^d$ and volatility function $\sigma : R^{md} \to R^d \times R^d$ are unknown, $\sigma(x)$ is positive definite for any $x \in R^{md}$, and the initial value $X_m = (Y_{m-1}, Y_{m-2}, ..., Y_0)$ is a random vector variable independent of $\{\xi_i\}$. Here $f(X_t)$ is the conditional mean function and $\sigma(X_t)$ is the conditional variance function. The model neither makes structural assumptions on f and σ, nor distributional assumptions on ξ. According to Härdle and Yang [Hae96], and Härdle et al. [Hae98], the model (6.2) is called a Conditional Heteroskedastic AutoRegressive Nonlinear (CHARN) model[1] [2], somehow as a nonparametric

[1] In 1993, Diebolt and Guegan [Die93] derive new bounds for the tail of the stationary density of certain non-linear d-dimensional processes $\{X_t; t \in N\}$ defined by the recursive scheme $X_t = T(X_{t-1}) + \sigma(X_{t-1})\varepsilon_t$, where $T(x)$ is a function $R^d \to R^d (d \geq 1), \sigma(x)$ is $d \times d$ regular matrix-valued function defined on R^d and ε_t is a sequence of i.i.d. random variables with mean 0 and variance 1 whose common distribution has a positive density $\mu(x)$. They assume that $\sigma(x)$ is invertible for all x. They do not assume that $T(x)$ or $\sigma(x)$ is continuous, nor that the density $\mu(x)$ is Gaussian, or even continuous.

[2] Assuming that the observed process does not have the same trend functions and the same volatility functions at each time instant, recently the CHARN model has been generalized to

alternative to GARCH family of models. Estimation of the CHARN model is relatively simple. Theoretical results about stability properties of thess processes are available [Fra07]. Both the conditional mean and the conditional variance (volatility) matrix are unknown functions of the past.

The CHARN model provides a generalization for the popular GARCH(1,1) model in that $f(X)$ is a nonparametric function, and most importantly $\sigma^2(x)$ is not a linear function of X_{t-1}^2. The symmetry in X_{t-1} of the conditional variance in GARCH models is a particularly undesirable restriction when modeling financial time series due to the empirically well documented leverage effect. However, the CHARN model is more restrictive than traditional GARCH models in that its markov property restricts its ability to effectively model the longer memory that is commonly observed in return processes.

Mandelbrot, Fisher and Calvet [Man97] state that the common strand in GARCH-type representations is a conditional distribution of returns that has a finite, time-varying second moment. This directly addresses volatility clustering in the data, and mitigates the problem of fat tails.

There are some theoretical reasons to believe that the return, X_i, may be a (first order) Markov chain [Hei96]. If financial markets worked efficiently, then all relevant information would be included in present returns[3]. In this case, a forecast based on all available information is not better than a forecast based solely on today's returns. More precisely, the conditional distribution of future returns based on the whole information set is equal to the conditional distribution given today's return. However, the assumption of efficiency is not convincing if investors have to bear some cost for acquiring information. As in Grossman and Stiglitz's model [Gro80] returns may then reveal information rather slowly. Whatever is the case, the ξ_i in the CHARN model (6.2) is explicitly assumed to be strict white noise.

6.2 Nonparametric estimation

Recall the CHARN model (6.2), in order to get the estimator for the conditional mean and variance functions, Härdle and Tsybakov [Hae97] present local polynomial estimators. As Martins-Filho and Yao [Mar06] discuss, the estimators described by Härdle and Tsybakov [Hae97] for estimating the conditional variance suffers from significant bias and does not produce estimators that are constrained to be positive. Furthermore, the estimator is not asymptotically design adaptive to the estimation of $f(X_i)$, i.e., the asymptotic properties of their estimator for conditional volatility is sensitive to how well $f(X_i)$ is estimated.

a model called Conditional Heteroscedastic Autoregressive Mixture of Experts (CHARME) which is useful for modeling time series data that are piecewise stationary such that their dynamics switch sometimes from one state to another. A typical example is given by stock returns if the market changes from a quiescent to a volatile phase. For more detail, see Franke, Stockis and Kamgaing [Fra07].

[3] This notion of efficiency is not to be confused with Pareto-efficiency. A suitable term would be information-efficiency.

Following [Mar06] and [Jul05], we therefore consider alternative estimation procedures due to Fan and Yao [Fan98], which is described as follows.

The first alternative is based on the following steps: i) Estimate $\hat{f}(X_i)$, ii) Estimate the equation $Y_i^2 = g(X_i) + \xi_i$, yielding an estimator $\hat{g}(X_i)$ for the second moment, iii) Estimate the conditional variance function $\hat{\sigma}^2(X_i) = \hat{g}(X_i) - \hat{f}^2(X_i)$. The only possible problem that may arise is the presence of negative values for $\hat{\sigma}^2(X_i)$. The second alternative is as follows: i) Estimate $\hat{f}(X_i)$ using some nonparametric technique, ii) Estimate the Heteroskedastic residuals $\hat{\epsilon_i} = Y_i - \hat{f}(X_i)$, and demean them, $e_i = \hat{\epsilon_i} - \bar{\epsilon}$. Then estimate $e_i^2 = \sigma^2(X_i) + \eta_i$, leading to the estimator $\hat{\sigma}^2(X_i)$ which characterizes for having $\hat{\sigma}^2(X_i) > 0$ for all i.

An important feature of nonparametric strategies based on interval specific information is that, they provide asymptotically unbiased measures, and therefore approximately serially uncorrelated measurement errors. See Andersen et al. [And02].

The CHARN model has usually been estimated by means of local polynomial regression [Hae97], [Hae98], [Jul05], Nadaraya-Watson kernel regression [Hae92] and local polynomial smoothing [Pol03] regression techniques. The techniques have a good performance for approximating CHARN model. It is shown how good they fit a regression function, regarding to Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) criterion. The MAE is similar to the RMSE but is less sensitive to large forecast errors. On the other hand, the result of MAE tends to place less emphasis on the larger errors and therefore, gives a more conservative measure than the RMSE. Generally both of them measure the deviation between actual and forecasted value. The forecasting powers of these techniques for estimating CHARN model are compared. In addition to the common techniques used in estimating the CHARN model, Neural Network because of its popularity and the Support Vector Regression are considered for estimation.

6.3 Algorithms for estimation

Before getting the functions, it is needed to briefly enumerate those algorithms which would be used for fitting local polynomial, kernel and local polynomial smoothing regression functions. These algorithms are alternative techniques as benchmarks. The algorithms are run in software R which is an open source system for statistical computation and graphics[4] including stats, aws, e1071, lpridge, AMORE and locfit libraries. We note that except loess algorithm, conventional functions have no prediction method. In the next parts where the results of running above functions would be presented, we will see this problem is not crucial.

[4] More information about included packages, documents and downloading source codes can be found on: http://www.r-project.org

Lpepa: Local polynomial regression fitting with Epanechnikov weights is a fast and stable algorithm for nonparametric estimation of regression functions and their derivatives via local polynomials with Epanechnikov weight function. This algorithm known as lpepa, can be run in the software R with package Lpridge. More details can be found in [Sei94] and other related sources.

Ksmooth: It runs Nadaraya-Watson kernel regression estimate found in package of stats within software R. In fact, it is a kernel regression smoother which is based on selecting an optimal bandwidth for smoothness level.

Loess: Local polynomial regression fitting in package stats, is possible also using loess function. It fits a polynomial surface determined by one or more numerical predictors, using local fitting. The fit is made using points in a neighborhood of x, weighted by their distance from x (with differences in parametric variables being ignored when computing the distance). The size of the neighborhood is controlled by span parameter α. More details are found in [Cle92].

Aws: A local polynomial adaptive weights smoothing for regression with additive errors can be run using aws algorithm in the Package aws in software R. This function implements a local polynomial adaptive weights smoothing procedure for regression problems with additive errors as described in [Pol03]. Adaptive weights smoothing is an iterative data adaptive smoothing technique that is designed for smoothing in regression problems with discontinuous regression function. The basic assumption is that the regression function can be approximated by a simple local constant or local polynomial model.

Amore: The AMORE package would represent a neural network. This package offers a highly flexible environment and provides more control over the learning details, allowing the user to customize the available functions. The package is capable of training a multilayer feedforward network according to both the adaptive and the batch versions of the gradient descent with momentum backpropagation algorithm.

Svm: This algorithm implements support vector machine. It can be used to carry out general regression and classification (of nu and epsilon-type), as well as density estimation. It is nested in the e1071 package of the software R. The original codes come from LibSVM codes. It works with different kernels.

A crucial problem in some of these methods, i.e., ksmooth, and lpepa used here is the choice of the local bandwidth array. Too small bandwidths will lead to a wiggly curve, too large ones will smooth away important details. Of course, the performance of model depends strongly on choosing an optimal bandwidth. Finding an optimal bandwidth using related methods like as kdeb in package Locfit is really time expensive. Choosing an optimal architecture and topology for the network in neural network learning seems as a typical problem too.

Chapter 7

Experiments on different resolution models

7.1 Experiments with single resolution model

7.1.1 Estimation schemes

Some algorithms, i.e., ksmooth and lpepa previously described need an optimally chosen bandwidth. To do this, an automatically bandwidth selection using the function kdeb in library locfit in software R is conducted and the value is directly passed to the related functions. For the svm function, which runs support vector regression included in library e1071, a radial basis kernel function is applied by default. For the AMORE function, a MLP feedforward network according to the adaptive type of the gradient descent with momentum backpropagation algorithm is adopted. The activation function would be sigmoid as usual kind of function. In order for uniformity, we keep all these features and properties while running all functions for all data sets.

7.1.2 Data description

For experiments, real world daily data sets are exploited. Daily stock exchange indices including S&P500, Nikkei225, Hang Seng, FTSE100, and DOW30 series would be analyzed in our experiments. These real world data, all which are close type, were extracted from Karlsruhe Capital Market Data Bank (Karlsruher Kapitalmarktdatenbank (KKMDB)) at the University of Karlsruhe[1]. S&P500 covers a period consisting from 13. Nov. 1981 to 30. Dec. 2005 daily. Nikkei225, Hang Seng, FTSE100, and DOW30 also cover duration respectively from 4. Jan. 1984, 31. Dec. 1986, 2. Apr. 1984, and 27. Jul. 1988 all to 30. Dec. 2005 daily. We split all series into two sets: Two-third as in-sample data sets and one-third as out-of-sample data sets. Crashes and bubbles may be included in data, for

[1] More information can be found on page: http://fmi.fbv.uni-karlsruhe.de/149.php

example the crash of October 1987 is present in S&P500, but we do not intend to address this problem here. Dealing with such a problem needs another kind of financial time series modeling. The variables of interest in our analysis are returns defined from daily abovementioned index values, p_t. We define return of an index by $X_t = log(p_t) - log(p_{t-1})$, $t = -R+1, ..., n$, which is the return from holding the index from time t_{-1} to time t.

Table 7.1: Descriptive statistics of indices

Statistic	S&P500	Nikkei225	Hang Seng	FTSE100	DOW30
Minimum	-0.229	-0.161	-0.405	-0.130	-0.074
Maximum	0.087	0.124	0.172	0.076	0.062
Mean	0.00038	0.00009	0.00037	0.00030	0.00038
Median	0.00046	0.00037	0.00055	0.00062	0.00050
Sum	2.328	0.484	1.756	1.623	1.652
Variance	0.00011	0.00019	0.00030	0.00011	0.00010
Skewness	-1.830	-0.119	-3.386	-0.547	-0.291
Kurtosis	41.152	7.737	76.576	8.152	4.970

Table 7.1 contains some basic descriptive statistics of our time series. Positive mean and median returns explain an average positive return trend. In particular, excess kurtosis (peakedness) and skewness (asymmetry) show obviously our time series depart from normality. Among those of all indices, higher kurtosis (76.6 and 41.2) and skewness (-3.38 and -1.83) coefficients for Hang Seng and S&P500, for example, explain more distant distribution of these time series from that of a normal. Since kurtosis coefficients of series are higher than 3 (for a normal distribution), we find immediately that they are leptokurtic with fat tail. Negative coefficients of skewness for all series describe that our probability density functions are negatively skewed and therefore they are asymmetric with longer negative tail. Since the probability densities are skewed, median values of all series are higher than mean values.

These findings are confirmed by Figure 7.1 in which plots display kernel density estimation of sample distributions for time series. Negative tails on plots are evident. These findings are in line partly with those of Hoechstoetter, Rachev and Fabozzi [Hoe05]. All of the tests performed in their study reject the Gaussian hypothesis for the logarithmic returns of the German blue chip stocks. Excess kurtosis here may be related to volatility clustering. From these findings, we are empirically convinced to use nonparametric models which are distribution free and need no assumption about distribution.

7.1.3 Results

Consider now the CHARN model (6.2) to be estimated by algorithms described above. Table 7.2 indicates the results [Saf08a] of running model with regard to RMSE and MAE criterion applying various techniques based on in-sample

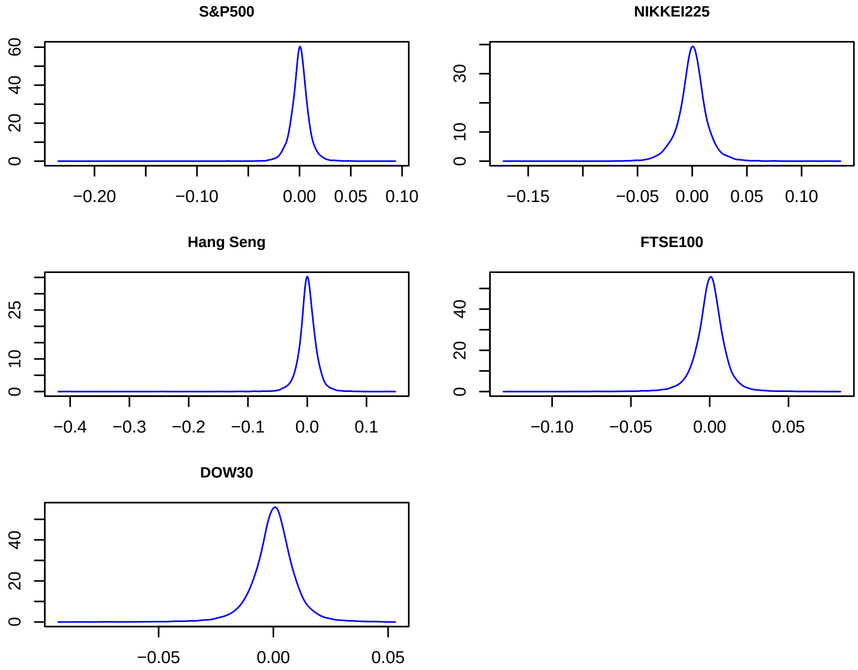


Figure 7.1: Kernel densities of the return data sets are graphically depicted. Leptokurtosis with fat negative tails is obviously evident in distributions, in particular in S&P500 and Hang Seng.

data set for fitting functions or training machines. Clearly, a nearly similar performance is seen, for example in terms of RMSE, between loess (0.01, 0.014, 0.0094, and 0.014) and AMORE (0.011, 0.0159, 0.0106, and 0.01309) for all data sets, excluding Hang Seng, but a dramatic discrepancy between both of them with lpepa (0.024, 0.033, 0.022, and 0.031) and aws (0.015, 0.023, 0.016, and 0.022) which have a relative poor performance to capture fully data points. This result is also valid in terms of MAE. In some cases, for example S&P500, FTSE100 and DOW30, discrepancies amount even to several times. More important, svm outperforms all other techniques with a remarkable difference for all indices. For example, loess has mean absolute error 0.007, 0.00989, 0.01478, 0.0071, and 0.01, while svm has mean absolute error equal to 0.0064, 0.009, 0.012, 0.0068, and 0.0093, in fitting all time series respectively. Different performance between svm with AMORE and loess, particularly in case of FTSE100, is highly competitive.

Table 7.2: Estimated RMSE and MAE (In-sample data set)

Technique	Criteria	S&P500	Nikkei225	Hang Seng	FTSE100	DOW30
ksmooth	RMSE	0.01058	0.01349	0.06580	0.00980	0.01549
	MAE	0.00742	0.00930	0.04709	0.00734	0.01168
lpepa	RMSE	0.02387	0.03317	0.04648	0.02258	0.03130
	MAE	0.01672	0.02417	0.03045	0.01692	0.02370
loess	RMSE	0.01000	0.01414	0.02191	0.00938	0.01367
	MAE	0.00700	0.00989	0.01478	0.00715	0.01029
aws	RMSE	0.01516	0.02345	0.03049	0.01612	0.02191
	MAE	0.01181	0.01732	0.02101	0.02430	0.01693
AMORE	RMSE	0.01106	0.01597	0.01986	0.01060	0.01309
	MAE	0.00833	0.01222	0.01270	0.00768	0.01002
svm	RMSE	0.00943	0.01330	0.01924	0.00894	0.01257
	MAE	0.00640	0.00907	0.01188	0.00684	0.00933

According to Table 7.2, support vector regression clearly affords better performance in terms of RMSE and MAE metrics than local polynomial regression, Nadaraya-Watson kernel regression, neural network and local polynomial adaptive weights smoothing regression estimators.

To see the generalization performance of different estimators, out-of-sample experiments have also been conducted in our investigation. Table 7.3 presents the results. Since loess was a considerable competitor for svm and also the only algorithm which provides a prediction method, among conventional techniques, it was led into out-of-sample round. Based on the table, svm also outperforms loess absolutely in case of all indices in prediction phase. Superiority of support vector machine in prediction of financial time series has been reported by several authors. For example, Ullrich, Seese and Chalup [Ull05] conclude that support vector machines consistently perform well relative to traditional forecasting techniques in terms of forecasting accuracy and in terms of trading performance via a simulated strategy. Application of SVMs is not restricted only to the financial time series forecasting. SVMs have been successfully applied on for example bankruptcy prediction problem. The experiment results of Min and Lee [Min05] show that SVM outperforms the other methods such as multiple discriminant analysis (MDA), logistic regression analysis (Logit), and three-layer fully connected back-propagation neural networks (BPNs). Härdle, Moro, and Schäfer [Hae06] describe the rating methodology that is based on the nonparametric nonlinear classification method, the support vector machine, and a proposed nonparametric technique for mapping rating scores into probabilities of default. They give an introduction to underlying statistical models and introduce the results of testing their approach on the German Central Bank data.

Table 7.3: Estimated RMSE and MAE (Out-of-sample data set)

Technique	Criteria	S&P500	Nikkei225	Hang Seng	FTSE100	DOW30
loess	RMSE	0.01304	0.01581	0.01761	0.01233	0.00949
	MAE	0.00492	0.01203	0.01344	0.00901	0.00720
svm	RMSE	0.01183	0.01449	0.01342	0.01204	0.00848
	MAE	0.00486	0.01092	0.00997	0.00871	0.00637

For simplicity, we have not carried out sensitivity analysis related to the free parameters. However, a sensitivity analysis can reveal more details and may improve performance. We summarize in the following important points:

- Tuning bandwidth for those algorithms which need this free parameter is not only time-consuming but also is hard to find a rather well-tuned level even using automatically kdeb function. The results critically depend on that level. Choosing the optimal number of layers and neurons in neural network is a difficult problem too.
- The SVR is faster than those techniques which need tuning bandwidth. The regression function in SVR is only determined by the support vectors, and the number of support vectors is smaller compared to the number of training samples.
- Both training and testing stages for SVR show better results than other algorithms. As explained previously, SVR machine provides smaller RMSE and MAE than those of other alternative benchmark techniques. This is because SVR adopt the structural risk minimization principle, eventually leading to better generalization than conventional techniques. In addition, the SVR machine is eventually solved by a quadratic optimization formulation by which a unique solution can be obtained.

In the next section, we investigate whether the wavelet decomposition can still improve estimation accuracy of the regarded CHARN model. A multiscale resolution approach is compared to the traditional single resolution one previously done. The objective is to improve previous results in estimation performance of the model by following a wavelet preprocessing procedure, although studies around heterogeneous financial agents present interesting findings which deepen our proposition in theory.

7.2 Experiments with multiresolution model

Support vector regression, as a supervised learning machine provides a strong framework for the representation of relationships present in data structure. In previous section, it was experienced that it performs analogously well. Nonetheless, the choice of input data is not a trivial matter when difficult noisy and nonstationary data is handled. Data preprocessing and decomposition remain essential steps in the knowledge discovery process for real world application and, when correctly carried out, greatly improve the machine's ability to capture valuable information. Wavelet preprocessing and decomposing for enhancing prediction power comes from multiresolution analysis provided by wavelet transform. The wavelet transform can decompose one time series into several time series with different resolutions which have different levels of smoothness. The smoother level is more predictable, whereas the detailed level is less predictable, or more related to the noise. In this section, it is explored how the use of nonlinear regression fed with decomposed data can aid in better capturing useful information on various time scales.

By applying the wavelet representation, a multiresolution representation is built based on the differences of information available at two successive resolutions 2^j and 2^{j+1}. In fact, the multiresolution analysis is applied in order to obtain a further information from the signal that is not readily available in the row signal. Such a representation can be computed by decomposing the signal using a wavelet orthonormal basis. Therefore, the multiresolution representations are very effective for analyzing the information content of signal. The decomposition defines a multiresolution representation called a wavelet representation [Mal89].

Among all wavelets proposed in the literature, Daubechies and Morlet wavelet transforms have been increasingly adopted by signal and image processing researchers [Mur04]. Daubechies discrete wavelets exhibit a good trade-off between parsimony and information richness. Haar wavelet, for example, as one of popular wavelet transform, has some serious limitations because of its discontinuity. Daubechies wavelets are orthogonal wavelets and have less information redundancy than other wavelet transforms [Akh05]. They are nearly symmetric, a necessary property for compactly supported wavelets [Sch02].

7.2.1 Estimation schemes

Now we keep all estimation conditions previously used for a single resolution model estimation to suitably compare the results. Hence, the same data described in Table 7.1, the same multi-step procedure of estimation, and the same performance metrics (RMSE and MAE) are preserved for the multiresolution model estimation. Since a neural network has been successfully trained to provide five days ahead forecasts for S&P500 as the initial idea of wavelet preprocessing for enhancing prediction by Aussem, Campbell and Murtagh [Aus98], we also estimate the Multiscale CHARN model by the AMORE package here.

In summary, using the MODWT transform (4.12) and (4.13) the wavelet and scaling coefficients for time series are calculated, and then we feed several SVR machines with the whole coefficients of each scale, obtained from (4.12) and (4.13) equations, to run the CHARN model for each scale separately. On each scale, a machine runs separately a CHARN model. Then we additively reconstruct the wavelet details and smooths obtained by (4.14) and (4.15) using the multiresolution analysis Equation (4.16). This multiresolution analysis strategy is identically implemented for both training and forecasting stages. The type of wavelet system including the scaling filter and the wavelet filter is Daubechies least asymmetric family of wavelet filters (LA) of the length or order 8 which is particularly useful tool in the analysis of time-dependent processes. Usually Daubechies's least asymmetric wavelet of order 8 yields markedly better results [Per93]. The least asymmetric LA(8) wavelet filter, based on eight non-zero coefficients, yields coefficients that exhibit better uncorrelatedness across scales than the Haar filter and is better suited for decomposition of broadband turbulent signals [Cor05] [Mcc96]. The Daubechies least asymmetric scaling wavelet filter (LA8) looks like the Mexican Hat, but is also weakly asymmetric; a fact that makes LA(8) filter more malleable than the Mexican Hat [Wes04].

Aussem, Campbell, and Murtagh [Aus98] considered two types of wavelet feature as follows:

1. Decomposition-based approach: Wavelet coefficients at a particular time point are taken as a feature vector.

2. Scale-based approach: Modeling and predicting are run independently at each resolution level, and the results were combined.

If we follow the scale-based approach, a crucial task remained is to know how many and which wavelet coefficients will be used at each scale. A sparse representation of the information contained in the decomposition is the key to address this. There is no clear and efficient method to gain sparsity. However, the standard support vector regression machine solves the problem efficiently, since its solution is eventually sparse. Therefore, our scheme is selecting the whole coefficients.

Utilizing the function modwt in package waveslim included in software R and choosing arbitrarily 5 levels of scale, J=5, like that in Figure 7.2 for instance, we would be able to decompose a time series into 5 scaled series plus a smooth level. Then 5 scaled series plus a smooth series as input variables can feed 6 SVR machines. After training, we additively reconstruct an estimated series (returns on index) from the output of 6 machines. Test set phase follows exactly the same procedure as that for train phase.

7.2.2 Results

The original training data points can be compared against 5 level decomposed time series in Figure 7.2. Of course, if more level resolutions are selected, more smoothed scales could be seen. The results of estimation by singlescale svm and AMORE are repeated from Table 7.2 in Table 7.4 solely for comparison with the results of the multiscale model. In all series, a multiscale svm model shows better results than a singlescale resolution svm in training phase. Table 7.4 [Saf08a] reports higher than 5% improvement in accuracy of multiscale model relative to the singlescale model for Hang Seng, FTSE and DOW series in terms of RMSE. It is also found that the results of model estimation by neural network have been considerably improved from the singlescale to the multiscale model. Although Support Vector Regression still outperforms absolutely Neural Network in our experiments, in cases of S&P500 and Hang Seng the Neural Network reveals

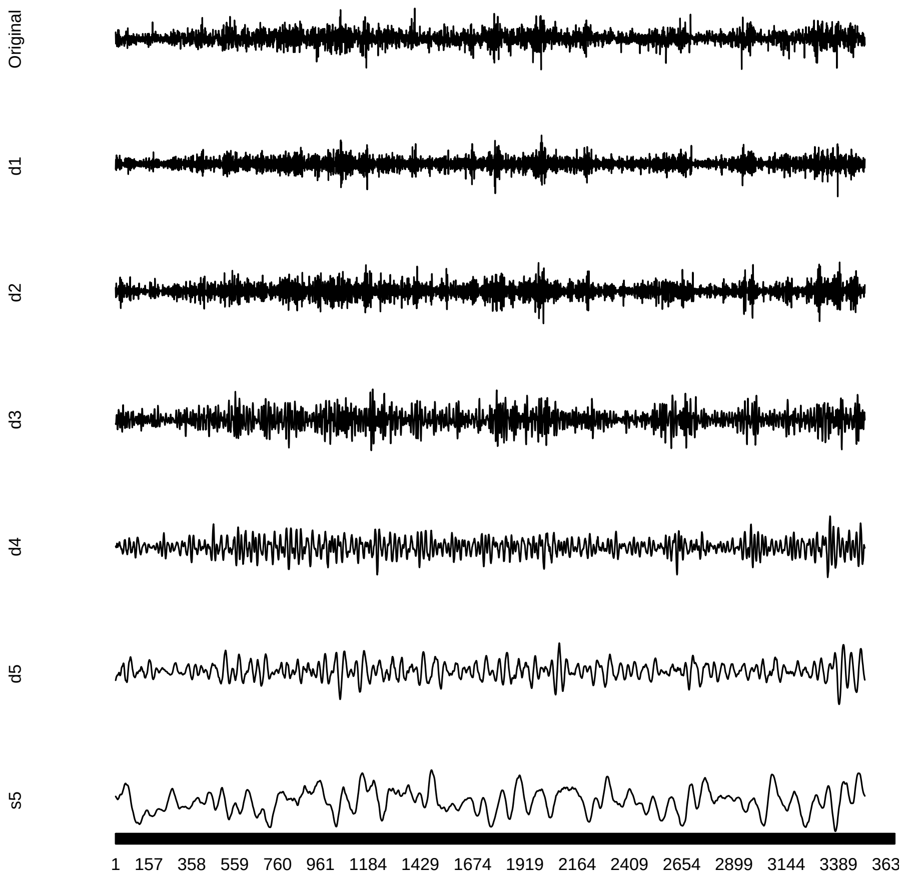


Figure 7.2: Original training data and 5 levels of resolution decomposition: Nikkei.

even better improvement. This implies that no matter what is the estimation tool, the multiscale setup constructed based on decomposed time series models better reality. Even better results for the test phase is observed. Since AMORE dose not have prediction method, table 7.5 reports only the results of predicted model by svm. Interestingly, it can be observed much more improvement in out-of-sample data sets (Table 7.5). For an example, a multiresolution model run by support vector machine can reach to 28.1% and 26.8% improvement of accuracy in terms of RMSE and MAE relative to a single resolution model for FTSE100 data set. Residuals of fitting and predicting functions are depicted in Figure 7.3 to represent how much distance between real and captured points is still residue. Smaller residuals in plots of svm and even better in plots of multiresolution svm for both train and test sets are manifest.

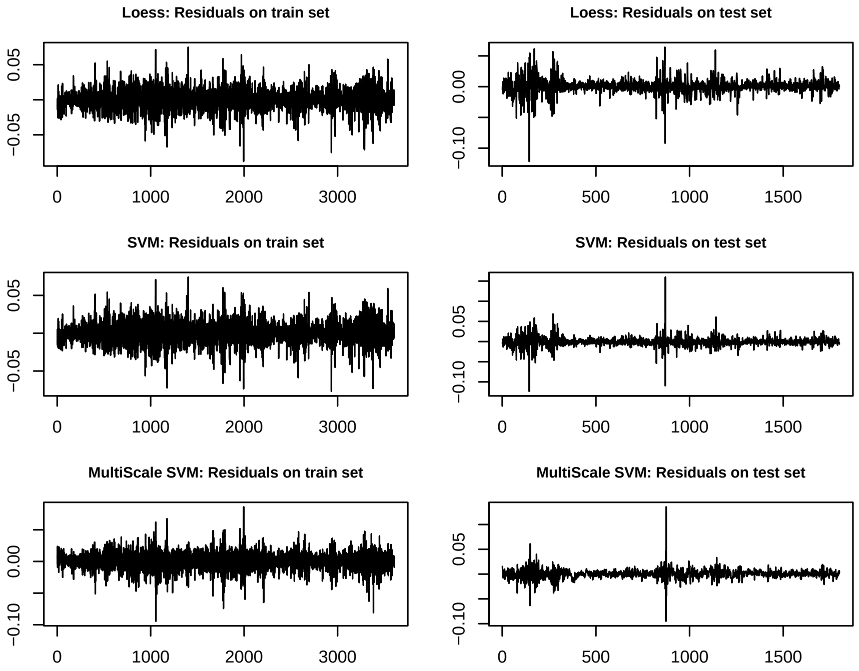


Figure 7.3: Residuals on fitting and predicting curves for Nikkei time series. Smaller residuals for svm and rather for multiscale svm are obviously evident.

Table 7.4: Estimated RMSE and MAE (In-sample data set)

Technique	Criteria	S&P500	Nikkei225	Hang Seng	FTSE100	DOW30
AMORE	RMSE	0.01106	0.01597	0.01986	0.01060	0.01309
	MAE	0.00833	0.01223	0.01270	0.00768	0.01002
Multiscale AMORE	RMSE	0.01037	0.01544	0.01837	0.01021	0.01236
	MAE	0.00796	0.01174	0.01202	0.00734	0.00977
Improvement (%)	RMSE	6.21	3.26	7.51	3.71	5.54
	MAE	4.52	3.98	5.30	4.47	2.46
svm	RMSE	0.00943	0.01330	0.01924	0.00894	0.01257
	MAE	0.00642	0.00907	0.01188	0.00684	0.00933
Multiscale svm	RMSE	0.00911	0.01265	0.01811	0.00848	0.01183
	MAE	0.00626	0.00880	0.01162	0.00640	0.00893
Improvement (%)	RMSE	3.43	4.92	5.85	5.13	5.87
	MAE	2.50	3.30	2.23	6.83	4.20

Exploitation from multiscale decomposition abilities and advantages can help to enhance performance of time series modeling. Wavelets offer advantages over traditional statistical analysis techniques, for example apart from other advantages, ability to minimize correlation and time-dependency of data, and in particular locality of the analysis and ability to handle multiscale information. The interest in wavelets is their speed and locality. Locality is the most important, because many economic time series and even natural phenomena are nonstationary and very local. Locality was realized in estimation procedure by wavelet, and therefore one of reasons behind outperforming the Multiscale CHARN model may be exploiting this ability. For some phenomena, it would be impossible to make sense of the data without wavelets. Using their ability to analyze data sets can help to understand difficult, chaotic and nonstationary data sets. Wavelets are able to economically describe phenomena that are heterogeneous.

Table 7.5: Estimated RMSE and MAE (Out-of-sample data set)

Technique	Criteria	S&P500	Nikkei225	Hang Seng	FTSE100	DOW30
svm	RMSE	0.01183	0.01449	0.01342	0.01204	0.00849
	MAE	0.00886	0.01092	0.00997	0.00872	0.00637
Multiscale svm	RMSE	0.00894	0.01095	0.01049	0.00866	0.00663
	MAE	0.00666	0.00825	0.00800	0.00637	0.00493
Improvement (%)	RMSE	24.4	24.4	21.82	28.1	21.83
	MAE	24.73	24.45	19.78	26.8	22.65

In recent years a number of studies around financial markets based on interacting heterogeneous agents with different time horizons have been developed. These studies and their interesting results support theoretically advocating scale based study of financial markets and therefore the Multiscale CHARN model. Each scale seems to correspond to each agent or class of agents in a financial market. As a relative initial work, Müller et al. [Mue97] focused on time horizons of investment by heterogeneous agents which trade at different frequencies of prices. That is, the population of traders often consists of both long-term traders and short-term traders. The diversity of agents in a heterogeneous market makes volatilities of different time resolutions behave differently. Heterogeneity in agent's time scale is believed to be responsible for a number of stylized facts. Long term traders naturally focus on long term behavior of prices thereby neglecting fluctuations at the smallest time scale, whereas short term traders are not concerned with price movements on the long run but rather aim to exploit short term predictability. A lagged correlation study in Dacorogna et al. [Dac01] reveals that statistical volatility defined over a coarse time grid significantly predicts volatility defined over a fine grid. It has been shown that there is an asymmetry where the coarse volatility predicts fine volatility better than the other way around.

7.3 Concluding remarks and some discussions

The SVR machine was applied for approximating volatility in CHARN framework. An estimation by SVR was conducted and the results were compared with those of benchmark techniques to estimate the CHARN model. The best performance belongs to SVR among technical benchmarks. However, further works may reveal more improvement in performance of SVR with regard to choosing more appropriate kernels, and tuning free parameters.

Moreover, applying the multiresolution analysis by wavelet transformation utilizing the CHARN model studied. Locality capability and ability of explaining phenomena that are heterogeneous by wavelet analysis were exploited through the multiscale decomposition in order to improve performance of estimation and forecasting so that the multiscale resolution model can capture smooth and noise levels of time series separately using several SVR machines more accurate than a singlescale resolution model. The results of the multiscale model are remarkably promising.

A relevant question here may be if the scale time series resulted from multiresolution decomposition have the same properties that different frequencies of an asset return have. In other words, if we decompose for an example a frequency of 20 minute asset return to some scales by wavelets, then each of decomposed scales would have those properties and stylized facts which lower frequencies of 30, and 60 minute and so on have. The question is important, since any time series model should capture any property existing in structure of the data. We advocate however a wavelet decomposition to reach different time scales which represent heterogeneous agents more reasonable and clear rather the raw different finer frequencies based on reason that orthogonal wavelet functions, have no overlap or projection to each other. This means that in a discrete wavelet transform which is orthogonal, each scale does not overlap the next one and therefore does not give redundant information. Daubechies wavelet, has such the property. Instead a raw time series at, for example, 20 minute frequency includes a 20 minute frequency itself, 30 minute, hourly and lower frequencies. So, a higher frequency includes, and therefore overlaps, lower horizon and lower frequencies. Thus using raw frequencies, representing heterogeneous agents, could be misleading, whereas orthogonal wavelet decomposition seems to represent heterogeneity clear, however this should be also studied further. In wavelet domain, different decomposed scales imply the different frequencies of a signal over the time.

Further study may reveal how many levels or scales in multiresolution analysis can have the best estimation performance or optimality in CHARN model estimation or may mathematically reveal irregularity. But back to the reality and heterogeneous agent models, this optimal number should be somehow picked out around the number of agents or of class of agents who act differently in an actual market, if we believe each scale corresponds to each agent or class of agents, i.e., each class of agents causes volatility on its corresponding scale.

The conditional variance, $\sigma(X_i)$, in the CHARN model is not a linear function of X_i. The symmetry in X_i of the conditional variance in GARCH models is a particularly undesirable restriction when modeling financial time series due to the empirically well documented leverage effect. However, CHARN model is more restrictive than traditional GARCH models in that its markov property restricts its ability to effectively model the longer memory that is commonly observed in return processes [Mar06]. But in fact, as Martins-Filho and Yao concluded [Mar06], their simulations indicate that accounting for nonlinearities may be more important than richer modeling of dependency. This was why CHARN model, rather than the GARCH model has been selected, in addition to those reasons latent in nonparametric modeling privileges.

Trading of the stock papers includes not only explicit but also implicit transaction costs which determine a decision of shareholders [Lue96]. Also dividends are remarkable. In both GARCH and CHARN models, these variables are assumed to be zero. Returns in these models refer to those from price's differences. So, returns stemmed from price's increments are not all of what an asset holder obtains. The CHARN model however has still some another shortcomings. Since volatility clustering implies that volatility comes and goes, thus a period of high volatility will eventually give way to more normal volatility and similarly, a period of low volatility will be followed by a rise. Mean reversion in volatility is generally interpreted as meaning that there is a normal level of volatility to which volatility will eventually return [Eng01]. Most evidences in empirical finance indicate that returns on financial assets seem unforecastable at short horizons as Granger claimed [Gra92]. Even mean reversion in volatility, as a further stylized fact of volatility clustering, implies that current information has no effect on the long run forecast. The CHARN model in first term is not able to capture all levels of mean function. So need for a CHARN model including mean shift function motivates future works.

One of the most important feature of return data is the persistence of volatility, which is interconnected to fat tailed returns. It is well known that models in GARCH class generally will possess fat tails, but they are only an empirical description, and not a true behavioral mechanism explaining the existence of fat tailed distributions. Also the distributions of estimated residuals of these models are often fat tailed themselves, suggesting that changing variances alone do not give the whole story [Leb06].

Chapter 8

Conclusions and discussions

The current chapter summarizes main results, conclusions and discussions of the dissertation. They are presented in a more general level here. Details can be found in the corresponding part.

Realized volatility and correlation estimators were addressed. The conditional volatility is latent, and hence is not directly observable. It can be estimated by several univariate, multivariate, conditional and stochastic approaches of volatility such as GARCH family of models, stochastic volatility (SV) models and exponentially weighted moving averages (EWMA) model. However, as it has been observed most of the latent volatility models fail to describe satisfactorily several stylized facts that have been observed in financial time series. More important, realized volatility provides more precise ex-post observations of the actual volatility compared to the other approaches based on daily or coarser frequency data. In fact, the availability of high frequency data has sown seeds for realized volatility modeling. Therefore, realized volatility as a model-free and observable measure of volatility, which needs the analysis of high frequency intraday data, has attracted lots of attention.

However, the biggest challenge to the realized volatility approach is the microstructure noise. It undermines consistency of the realized volatility estimators. Several methods have recently been proposed in the ultra high frequency financial literature to remove the effects of microstructure noise and to obtain consistent estimates of the integrated volatility as a true measure of daily volatility. Even bias-corrected and consistent realized volatility estimates of the integrated volatility can contain residual microstructure noise and other measurement errors that should not be neglected.

The consistency of proposed realized volatility and correlation estimators for integrated volatilities and correlations have been studied under different assumption of Gaussian noise. Naturally the consistency of different volatility estimators differs, given they are constructed differently. It was observed that the TSAV volatility estimator converges faster for integrated power variation as the frequency increases even under the assumption of existence of microstructure frictions. This implies that the estimator converges even at high frequency levels, where the noise especially exists. The TSAV estimator is constructed based on the subsampling and averaging approach which corrects for the bias caused by the microstructure noise.

Given a number of different realized volatility and correlation estimators, it could be interesting to assess their ability to reproduce the stylized facts. It was observed that absolute based volatility and correlation estimators can empirically repeat long memory behavior and reproduce some dynamic stylized facts of financial markets stronger than squared based volatility and correlation estimators. Self-similarity structure computed by Hurst exponent was documented in the structure of series generated by realized measures. None of volatility measures exactly pose a normal daily distribution tested by Jarque-Bera test of normality. Some of the estimators indicate heavy tail in distributions. Tested by Jarque-Bera estimator, the null hypothesis of normality for the absolute based correlation estimators can not significantly be accepted. While squared based volatility shows heavier tail than absolute based volatility estimators, the absolute based correlation estimators show heavier tail than squared based correlation.

Consistent with common sense and in particular with Archimedean copulas and one-factor model, it was empirically found that the multivariate absolute-based realized correlations exhibit negative asymmetry in dependence structure implying fatter left tail where the extreme values are mainly populated there.

In general, using intraday high frequency data yields better volatility estimation. The question that if more information contained in higher frequency data leads to more precisely volatility estimation is an open question. Moreover, the assumption of continuous and unlimited data is not of practical case.

Construction of some time-varying statistics such as Beta or systematic risk, regression with time-varying parameters, or even some kinds of combined estimators, for example return per unit of volatility which may be somehow close to Sharpe Ratio, based on the concepts of realized estimators using absolute values of high frequency data may yield more realistic analytical tool.

There are several methods, mentioned here, to correct the effects of microstructure noise. But what is actually missing, is a model or estimator that can capture the more complex time-dependent characteristics of market microstructure noise. What is called noise, is likely a part of useful and informative data that our model cannot capture or explain. Further attempts may yield some estimators which are able to capture but not correct the market microstructure noise.

The problem of nonparametrically volatility function approximation has been also addressed. To improve upon predictability performance of the SVR machine, multiresolution analysis is applied in conjunction with the SVR machine. A multiresolution analysis (MRA) or multiscale approximation (MSA) is the design method of most of the practically relevant discrete wavelet transforms. Multiresolution analysis using the wavelet transform is an efficient way to span the information contained in a signal.

An approach for forecasting using a wavelet-based multiresolution (multiscale) analysis has been described. Applying the multiresolution analysis, a signal or time series is decomposed into an arbitrary number of different scaled series. Each individual scale is fitted using a SVR machine applying the CHARN volatility model and the aggregate forecast is then obtained by adding up the individual forecasts.

Locality capability and ability of explaining phenomena that are heterogeneous by wavelet analysis were exploited through the multiscale decomposition in order to improve performance of estimation and forecasting so that the multiscale resolution model can capture smooth and noise levels of time series separately using several SVR machines more accurate than a singlescale resolution model. The reason behind outperformance of the mutiresolution approach stems from the fact that the smooth levels are more predictable and in contrast the detailed levels are less predictable and correspond to the noise. In other words, the wavelet transform can decompose one time series into several time series with different resolutions which have different levels of smoothness. The smoother level is more predictable, whereas the rougher (detailed) level is less predictable, or more related to the noise. As a matter of fact, a nonstationary system is dealt with suitably applying multiresolution analysis based on proper wavelet systems.

The support vector regression machine as an application of the statistical learning theory has been successfully exploited. The support vector regression was successfully utilized for estimation and forecasting of volatility in CHARN framework. An estimation by SVR was implemented and the results were compared with those of benchmark techniques to estimate the CHARN model. The best performance of accuracy belongs to the SVR among technical benchmarks. More important, it was empirically shown that the multiresolution approach improves upon the accuracy and precision performances in analogous to a single resolution forecasting approach.

Indeed, the problem which droves the initial development of SVMs occurs in several guises- the bias variance trade-off, capacity control, overfitting- but the basic idea is the same. Roughly speaking, for a given learning task, with a given finite amount of training data, the best generalization performance will be achieved if the right balance is struck between the accuracy attained on that particular training set, and the capacity of the machine, that is, the ability of the machine to learn any training set without error. “A machine with too much capacity is like a botanist with a photographic memory who, when presented with a new tree, concludes that it is not a tree because it has a different number of leaves from anything she has seen before; a machine with too little capacity is like the botanist’s lazy brother, who declares that if it’s green, it’s a tree. Neither can generalize well. The exploration and formalization of these concepts has resulted in one of the shining peaks of the theory of statistical learning” [Bur98].

The distributional properties of scaled series and the whole fitted function in this new forecast approach were not investigated. How the smoother and detailed or even noise series are related to each other; and how each individual scaled series shows stylized facts are another open questions. If the distributional properties of and stylized facts of scaled series are the same as those of corresponding lower frequencies for a given frequency. These subjects are delegated to further investigations. How to model CHARN approach so that it would be able to capture stylized facts beside volatility clustering and heavy tail is also a subject of more investigations.

For improving accuracy and precision in estimation, advantages of aggregation in part I and inversely advantages of disaggregation in part II were somehow benefited. The goal was almost the same, while two opposite methods were adopted. Realized volatility is an aggregation of higher frequency data to gain higher accuracy of volatility estimation. Multiscale analysis was exploited for disaggregation of lower frequency to obtain several scales corresponding to different frequencies to yield higher accuracy of volatility estimation.

Appendix A

Pyramid algorithm

Explicitly, the DWT and MODWT coefficients are computed by way of an algorithm that allows W to be factored in terms of very sparse matrices. The algorithm known as the pyramid algorithm was introduced by Mallat [Mal89]. Figure 4.1 visualizes the pyramid algorithm for the MODWT.

For discrete compactly supported filters of the wavelet family, denote the even-length L of the wavelet filter $\{h_l : l = 0, ..., L-1\}$ and the scaling filter $\{g_l : l = 0, ..., L-1\}$. By definition, the wavelet filters satisfy

$$\sum_{l=0}^{L-1} h_l = 0, \tag{A.1}$$

and

$$\sum_{l=0}^{L-1} h_l^2 = 1, \tag{A.2}$$

and

$$\sum_{l=0}^{L-1} h_l h_{l+2n} = \sum_{-\infty}^{\infty} h_l h_{l+2n} = 0, \tag{A.3}$$

for non-zero integers n. The scaling filters satisfy the conditions in (A2) and (A.3) except (A.1). But, additionally the filters are chosen to satisfy

$$h_l = (-1)^l g_{L-1-l} \tag{A.4}$$

$$g_l = (-1)^{l+1} h_{L-1-l}$$

for $l = 0, ..., L-1$. For example, the Haar wavelet filter $\left\{h_0 = \frac{1}{\sqrt{2}}, h_1 = -\frac{1}{\sqrt{2}}\right\}$ has length $L = 2$. The corresponding scaling filters are $\left\{g_0 = \frac{1}{\sqrt{2}}, g_1 = -\frac{1}{\sqrt{2}}\right\}$. Implementation of the DWT begins by defining the zeroth level scaling coefficients to be the original time series $V_{0,t} \equiv X_t$ [Con01]. When we denote the time series to be transformed by $\{X_t : t = 0, ..., N-1\}$, with $V_{0,t} \equiv X_t$, the jth stage input to the pyramid algorithm is $\{V_{j-1,t} : t = 0, ..., N_{j-1} - 1\}$, where $N_j = \frac{N}{2^j}$. The level j wavelet coefficients $W_{j,t}$ and scaling coefficients $V_{j,t}$ are then formed recursively by

$$W_{j,t} = \sum_{l=0}^{L-1} h_l V_{j-1(2t+1-l)} \text{ mod } \mathrm{N}_{j-1}, \tag{A.5}$$

and

$$V_{j,t} = \sum_{l=0}^{L-1} g_l V_{j-1(2t+1-l)} \text{ mod } \mathrm{N}_{j-1}, \tag{A.6}$$

for $t = 0, ..., N_j - 1$. Letting $\{W_{j,t}\}$ be $\mathbf{W}_j$ and $\{V_{j,t}\}$ be $\mathbf{V}_j$, then $N = 2^J$ and the pyramid algorithm is completed after J repetitions giving $\mathbf{W}_1, ..., \mathbf{W}_J, \mathbf{V}_J$, with the latter two vectors containing only one coefficient each. This gives the definition of the full discrete wavelet transform. If N is an integer multiple of 2^{J_0},$J_0 < J$, then we carry out a partial discrete wavelet transform to level J_0. Therefore, the discrete wavelet transform is an orthonormal transform of $\{X_t\}$. We can relate the wavelet and scaling coefficients at any level directly to the time series $\{X_t\}$. Let $\{h_{j,t}\}$ be the jth level wavelet filter with length $L_j = (2^j - 1)(L-1) + 1$. The jth level scaling coefficients $\{g_{j,t}\}$ are similarly defined. Then

$$W_{j,t} = \sum_{l=0}^{L_j-1} h_l X_{2^j(t+1)-1-l} \text{ mod } \mathrm{N}, \tag{A.7}$$

and

$$V_{j,t} = \sum_{l=0}^{L_j-1} g_l X_{2^j(t+1)-1-l} \text{ mod } \mathrm{N}, \tag{A.8}$$

with the filters satisfying (A.2) and (A.3). The nominal frequency band at every level with which the corresponding wavelet coefficient $\{W_{j,t}\}$ is associated, is given by $|f| \in (\frac{1}{2^{j+1}}, \frac{1}{2^j}]$. For example, $\{W_{1,t}\}$ has nominal frequency band of $(\frac{1}{4}, \frac{1}{2}]$. However, the discrete wavelet transform has a number of limitations. The limitations in the discrete wavelet transform can be overcome by avoiding downsampling. This can be achieved using the maximum overlap discrete wavelet transform MODWT.

Bibliography

[Abu96] Abu-Mostafa, Y. S., Atiya, A. F., 1996, "Introduction to financial forecasting", *Applied Intelligence, 6*, pp. 205-213.

[Ait05] Aït-Sahalia, Y., Mykland, P. A., Zhang, L., 2005,"How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise", *The Review of Financial Studies Vol. 18, No. 2*, pp. 351-416.

[Akh05] Akhbardeh, A., Koivuluoma, M., Koivistoinen, T., Vaerri, A., 2005, "BCG data discrimination using Daubechies compactly supported wavelet transform and neural networks towards heart disease diagnosing," *Proceedings of the 2005 IEEE*, International Symposium on Intelligent Control, Limassol, Cyprus, June 27-29.

[Alo93] Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D., 1993, "Scale-sensitive dimensions, uniform convergence, and learnability", *Symposium on Foundations of Computer Science.*

[And97] Andersen, T.G., Bollerslev, T., 1997,"Heterogeneous information arrivals and return volatility dynamics: Uncovering the long-run in high frequency returns", *Journal of Finance, 52*, pp. 975-1005.

[And98] Andersen, T.G., Bollerslev, T., 1998,"Answering the skeptics: Yes, standard volatility models do provide accurate forecasts", *International Economic Review, 39*, pp. 885-905.

[And99] Andersen, T.G., Bollerslev, T., Lange, S., 1999,"Forecasting financial market volatility: Sample frequency vis-a-vis forecast horizon", *Journal of Empirical Finance, 6*, pp. 457-477.

[And01a] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens H., 2001,"The distribution of realized stock return volatility", *Journal of Financial Economics, 61*, pp. 43-76.

[And01b] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P., 2001,"The distribution of realized exchange rate volatility", *Journal of the American Statistical Association, Vol. 96, No. 453*, pp. 42-55.

[And03] Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P., 2003,“Modeling and forecasting realize d volatility”, *Econometrica, Vol. 71, No. 2*, pp. 579-625.

[And06] Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X., 2006,“Volatility and correlation forecasting”, *Elsevier B.V.*, Handbook of Economic Forecasting, Volume 1, Chapter 15.

[Aus98] Aussem, A., Campbell, J., Murtagh, F., 1998, “Wavelet-based feature extraction and decomposition strategies for financial forecasting,” *Journal of Computational Intelligence in Finance 6(2)*, pp. 5-12.

[Bai92] Baillie, R. T., Bollerslev, T., 1992,“Prediction in dynamic models with time dependent conditional variances”, *Journal of Econometrics, 52*, pp. 91-113.

[Bal84] Ball, C. A., Torous, W. N., 1984, “The Maximum Likelihood Estimation of Security Price Volatility: Theory, Evidence and Application to Option Pricing”, *J. Bus. 57 (1)*, pp. 97-113.

[Ban05a] Bandi, F., Russell, J., 2005,“Volatility”, *Forthcoming in the Handbook of Financial Engineering, Elsevier. Edited by J. R. Birge and V. Linetsky.*

[Ban05b] Bandi, F., Russell, J., 2005,“Microstructure noise, realized volatility, and optimal sampling”, *Working paper, Graduate School of Business, University of Chicago.*

[Bar03a] Barndorff-Nielsen, O.E., Shephard, N., 2003,“Realized power variation and stochastic volatility models”, *Bernoulli, 9(2)*, pp. 243-265.

[Bar03b] Bardet, J., Lang, G., Oppenheim, G., Philippe, A., Taqqu, M., 2003,“Generators of long-range dependent processes: A survey”, *In P. Doulkhan, G. Oppenheim, M. Taqqu (Eds.), Theory and applications of long-range dependence*, Birkhäser: Boston.

[Bar04a] Barndorff-Nielsen, O.E., Shephard, N., 2004,“Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics”, *Econometrica, Vol. 72, No. 3*, pp. 885-925.

[Bar04b] Barndorff-Nielsen, O.E., Shephard, N., 2004,“Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics”, *Econometrica, Vol. 72, No. 3*, pp. 885-925.

[Bar04c] Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2004,“Regular and modified kernel-based estimators of integrated variance: The case with independent noise”, *CAF: Center for Analytical Finance, University of Aarhus, Working Paper No. 196.*

[Bar04d] Barndorff-Nielsen, O.E., Shephard, N., 2004,“Power and bipower variation with stochastic volatility and jumps”, *Journal of Financial Econometrics, 2*, pp. 1-37.

[Bar06] Barndorff-Nielsen, O.E., Lunde, A., Hansen, P. R., Shephard, N., 2006,“Designing Realized Kernels to Measure the Ex-post Variation of Equity Prices in the Presence of Noise”, *Draft paper.*

[Bas00] Bashir, Z., El-Hawary, M. E., 2000, “Short term load forecasting by using wavelet neural networks,” *In Canadian Conference on Electrical and Computer Engineering*, pp. 163-166.

[Baz93] Bazaraa, M. S., Sherali, H. D., Shetty, C. M., 1993, *Nonlinear Programming: Theory and Algorithms*, 2nd edition, Wiley.

[Bej80] Beja, A., Goldman, M. B., 1980, “On the dynamic behavior of prices in disequilibrium,” *Journal of Finance, 35*, pp. 235-248.

[Bol94] Bollerslev, T., Engle, R. F., Nelson, D. B., 1994, “ARCH models,” The Handbook of Econometrics, Volume 4, Amsterdam: North-Holland, pp. 2959-3038.

[Bol02] Bollen, B., Inder, B., 2002,“Estimating daily volatility in financial markets utilizing intraday data”, *Journal of Financial Economics, 9*, pp. 551-562.

[Bol07] Bollerslev, T., Zhou, H., 2007, “Expected Stock Returns and Variance Risk Premia,” *Working Paper.*

[Bos92] Boser, B. E., Guyon, I. M., Vapnik V. N., 1992, “Atraining algorithm for optimal margin classifiers,” In: Haussler D. (Ed.) *Proceedings of the Annual Conference on Computational Learning Theory. ACM Press*, Pittsburgh, PA, pp. 144-152.

[Bos96] Bossaerts, P., Härdle, W., Hafner, Ch., 1996, “A new method for volatility estimation with application in foreign exchange rate series,” *in Finanzmarktanalyse und -prognose mit innovativen quantitativen Verfahren, G. Bol, G. Nakhaeizadeh and K.-H. Vollmer, eds., Physica Verlag*, 71-84.

[Bou02] Bouchaud, J. P., 2002,“An introduction to statistical finance”, *Physica A, 313*, pp. 238-251.

[Bur98] Burges, C.J.C., 1998,“A Tutorial on Support Vector Machines for Pattern Recognition”, *Data Mining and Knowledge Discovery, 2*, pp. 121-167.

[Cam97] Campbell, J., Lo, A., MacKinlay, A., 1997, *The Econometrics of Financial Markets*, Princeton University Press: Princeton, NJ.

[Che02] Cherkassky, V., Ma, Y., 2002, “Selecting the Loss Function for Robust Linear Regression”, *Neural Computation, Under Review, NC 2569.*

[Che04] Cherkassky, V., Ma, Y., 2004, “Comparison of loss functions for linear regression”, *In Proc. of the International Joint Conference on Neural Networks*, Piscataway, NJ: IEEE, pp. 395-400.

[Che06a] Chen, Y., Yang, B., Dong, J., 2006, "Time-series prediction using a local linear wavelet neural network," *Neurocomputing, 69*, pp. 449-465.

[Che06b] Chen, X., He, Z., Xiang, J., Li, B., 2006, "A dynamic multiscale lifting computation method using Daubechies wavelet," *Journal of Computational and Applied Mathematics, 188*, pp. 228-245.

[Ciz01] Cizeau, P., Potters, M., Bouchaud, J.-P., 2001,"Correlation structure of extreme stock returns", *Quantitative Finance, Vol. 1*, pp. 217-222.

[Cle92] Cleveland, W. S., Grosse E., Shyu, W. M., 1992, "Local regression models," Chapter 8 of Statistical Models in S. eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.

[Coi95] Coifman, R. R., Donoho, D., 1995, "Translation-Invariant De-Noising," In: Wavelets and Statistics, Lecture Notes in Statistics 103, eds. A. Antoniadis and G. Oppenheim, Springer-Verlag: New York.

[Col94] Coleman, T. F., Li, Y., 1994, "On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds," *Math. Programming, 67,*, pp. 189-224.

[Col96] Coleman, T. F., Li, Y., 1996, "A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on some of the Variables," *SIAM Journal on Optimization, Vol. 6, No. 4,*, pp. 1040-1058.

[Con01] Constantine, W., Percival, D. B., Reinhall, P. G., 2001, "Inertial range determination for aerothermal turbulence using fractionally differenced processes and wavelets," *Physical Review, Vol. 64*.

[Cor05] Cornish, C. R., Bretherton, C. S., Percival, D. B., 2005, "Maximal Overlap Wavelet Statistical Analysis with Application to Atmospheric Turbulence," *Kluwer Academic Publishers, Printed in the Netherlands*, pp. 2-37.

[Cor07] Corsi, F., 2007,"Realized Correlation Tick-by-Tick", *Discussion Paper 2007-02, Department of Economics, University of St. Gallen.*

[Cox85] Cox, J. C., Rubinstein, M., 1985, *Options Markets*, Prentice Hall: NJ.

[Cra05] Craigmile, P. F., Percival, D. B., 2005, "Asymptotic Decorrelation of Between-Scale Wavelet coefficients", *IEEE Transactions on Information Theory, 51(3),* pp. 1035-1048.

[Cri00] Cristianini, N., Shawe-Taylor, J., 2000, *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press: Cambridge.

[Dac93] Dacorogna, M. M., Nagler, R. J., Olsen, R. B., Pictet, O. V., 1993 , "A geographical model for the daily and weekly seasonal volatility in the FX market," *Journal of International Money and Finance, 12 (4)*, pp. 413-438.

[Dac01] Dacorogna, M., Gencay, R., Mueller, U., Pictet, O., Olsen, R., 2001, *An Introduction to High-Frequency Finance,* Academic Press: San Diego.

[Dan62] Dantzig, G. B., 1962, *Linear Programming and Extensions*, Princeton University Press: NJ.

[Dau88] Daubechies, I., 1988, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math., 41*, pp.906-966.

[Dau92] Daubechies, I., 1992, *Ten Lectures on Wavelets*, SIAM: Philadelphia.

[Dav87] Davidian, M., Carroll, R. J., 1987, "Variance Function Estimation," *J. Amer. Statist. Assoc., 82,* pp. 1079-91.

[Day90] Day, R. H., Huang, W., 1990 , "Bulls, bears, and market sheep," *Journal of Economic Behavior and Organization, 14*, pp. 299-329.

[Dev96] Devroye, L., Györfi, L., Lugosi, G., 1996, "A Probabilistic Theory of Pattern Recognition," *Applications of Mathematics, 31*, Springer: New York.

[Die93] Diebolt, J., Guegan, D., 1993,"Tail behaviour of the stationary density of general non-linear autoregressive processes of order 1", *J. Appl. Prob., 30*, pp. 315-329.

[Din93] Ding, Z., Granger, C. W. J., Engle, R. F., 1993,"A long memory property of stock market returns and a new model", *Journal of Empirical Finance, 1*, pp. 83-106.

[Div07] Divine, D.V., Godtliebsen, F., 2007,"Bayesian modeling and significant features exploration in wavelet power spectra", *Nonlin. Processes Geophys., 14*, pp. 79-88.

[Dro96] Drost, F.C., Werker, B.J.M., 1996,"Closing the GARCH gap: Continuous time GARCH modeling", *Journal of Econometrics, 74*, pp. 31-57.

[Ede00] Ederington, L. H., Guan, W., 2000, "Forecasting Volatility," *Working paper*, University of Oklahoma.

[Emb03] Embrechts, P., Lindskog, F., McNeil, A., 2003,"Modeling dependence with copulas and applications to risk management," in Rachev, S., T., *Handbook of heavy tailed distributions in finance*, Elsevier North-Holland.

[Eng86] Engle, R. F., Bollerslev, T., 1986,"Modeling the Persistence of Conditional Variances", *Econometric Reviews, 5*, pp. 1-50.

[Eng01] Engle R. F., Patton, A. J., 2001, "What good is a volatility model?," *Quantitative Finance, 1(2)*, pp. 237-245.

[Eng02] Engle R. F., 2002, "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business Economic Etatistics, Vol. 20, No. 3*, pp. 339-350.

[Epp79] Epps, T., 1979, "Comovements in stock prices in the very short run," *Journal of the American Statistical Association 74(366),* pp. 291-298.

[Evg99] Evgeniou, T., Pontil, M., Poggio, T., 1999, "A unified framework for Regularization Networks and Support Vector Machines," *A. I. Memo No. 1654, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.*

[Evg00] Evgeniou, T., Pontil, M., Poggio, T., 2000, "Statistical Learning Theory: A Primer," *International Journal of Computer Vision 38(1)*, pp. 9-13.

[Fam98] Fama, E., F., 1998, "Market efficiency, long-term returns, and behavioral finance," *Journal of Financial Economics, 49*, pp. 283-306.

[Fan98] Fan, J., Yao, Q., 1998, "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika, 85*, pp. 645-660.

[Fig97] Figlewski, S., 1997, "Forecasting Volatility," *Finan. Markets, Inst. Instruments. NYU, Salomon Center, 6 (1)*, pp. 1-88.

[Fle98] Fleming, J., 1998, "The quality of market volatility forecasts implied by SP 100 index option prices," *J. Empirical Finance, 5, 4*, pp. 317-345.

[For05] Forsberg, L., Ghysels, E., 2005,"Why do absolute returns predict volatility so well?", *Princeton-Chicago Conference on the Econometrics of High Frequency Financial Data, Bendheim Center for Finance, Princeton University, 2005.*

[Fra07] Franke, J., Stockis, J.-P., Kamging, J. T., 2007, "On Geometric Ergodicity of CHARME Models," *Working Paper, University of Kaiserslautern.*

[Gag94] Gagnon, L., Lina, J. M., 1994, "Symmetric Daubechies' wavelets and numerical solution of NLS equations", *J. Phys. A: Math. Gen. 27*, pp. 8207-8230.

[Gar80] Garman, M. B., Klass, M. J., 1980, "On the Estimation of Security Price Volatilities from Historical Data", *J. Bus. 53 (1)*, pp. 67-78.

[Gau00] Gaunersdorfer, A., Hommes, C. H., 2000, "A nonlinear structural model for volatility clustering," *Working Paper No. 63*, Vienna University of Economics and Business Administration.

[Gen04] Gencay, R., Selcuk, F., Whitcher, B., 2004, "Information Flow Between Volatilities Across Time Scales", *Working Paper.*

[Gew83] Geweke, J., Porter-Hudak, S., 1983, "The Estimation and Application of Long Memory Time Series Models", *Journal of Time Series Analysis, 4,*, pp. 221-238.

[Ghy06] Ghysels, E., Santa-Clara, P., Valkanov, R., 2006,"Predicting volatility: Getting the most out of return data sampled at different frequencies", *Journal of Econometrics, 131*, pp. 59-95.

[Ghy07] Ghysels, E., Sinko, A., 2006,"Volatility forecasting and microstructure noise", *Colloque CIREQ Conference: Realized Volatility, 22-23 April 2006, Montreal.*

[Gil81] Gill, P. E., Murray, W., Wright, M. H., 1981, *Practical Optimization*, Academic Press: London.

[Gil01] Giles, C. L., Lawrence, S., Tsoi, A. C., 2001, "Noisy Time Series Prediction using a Recurrent Neural Network and Grammatical Inference," *Machine Learning, Vol. 44, No. 1/2,* pp. 161-183.

[Gir95] Girosi, F., Jones, M., Poggio, T., 1995, "Regularization theory and neural networks architectures", *Neural Computation, 7*, pp. 219-269.

[Gou92] Gourieroux, C., Monfort, A., 1992, "Qualitative Threshold ARCH Models", *Journal of Econometrics, 52,*, pp. 159-199.

[Gra92] Granger, C. W. J., 1992, "Forecasting stock market prices: Lessons for forecasters," *International Journal of Forecasting, 8*, pp. 3-13.

[Gro80] Grossman, S., Stiglitz, J., 1980, "On the impossibility of informationally efficient markets", *American Economic Review, 70*, pp. 398-408.

[Haa10] Haar, A., 1910, "Zur Theorie der Orthogonalen Funktionensysteme", *Mathematische Annalen, 69*, pp. 331-371.

[Hae92] Härdle W., Vieu, P., 1992, "Kernel regression smoothing for time series," *Journal of Time Series Analysis, 13*, pp. 209-232.

[Hae96] Härdle, W., Yang, L., 1996, "Nonparametric time series model selection," *Interface 96, Computing Science and Statistics*, pp. 407-412.

[Hae97] Härdle, W., Tsybakov, A., 1997, "Local polynomial estimators of the volatility function in nonparametric autoregression," *Journal of Econometrics, 81*, pp. 223-242.

[Hae98] Härdle, W., Tsybakov, A., Yang, L., 1998, "Nonparametric vector autoregression," *Journal of Statistical Planning and Inference, 68*, pp. 221-245.

[Hae06] Härdle, W., Moro, R., Schäfer, D., 2006, "Bankruptcy analysis with support vector machines ," *European Finance Association, 33rd Annual Meeting, 23.26 August 2006, Zuerich.*

[Han06] Hansen, P., Lunde, A., 2006,"Realized Variance and Market Microstructure Noise", *Journal of Business and Economic Statistics, 24*, pp. 127-161.

[Haw80] Hawkins, D.M., 1980, *Identification of Outliers*, London, UK: Chapman and Hall.

[Hay05] Hayashi, T., Yoshida, N., 2005, "On covariance estimation of non-synchronously observed diffusion processes," *Bernoulli, 11*, pp. 359-379.

[Hei96] Heid, F., 1996, "Non-Parametric volatility Estimation of Exchange rate and Stock Prices," *Discussion Paper No. A-533, Rheinische Friedrich-Wilhelms-Universität Bonn.*

[Hoe05] Hoechstoetter, M., Rachev, S., Fabozzi, F., 2005, "Distributional analysis of the stocks comprising the DAX 30," *Probability and mathematical statistics, 25(2)*, pp. 363-383.

[Hua04] Huang, C.-H., Nieh, C.-C., 2004,"Realize the realized stock index volatility", *Asian Economic Journal, Vol. 18, No. 1*, pp. 59-80.

[Jen00] Jensen, M. J., Whitcher, B., 2000, "Time-varying long-memory in volatility: Detection and estimation with wavelets," *Working Paper.*

[Jul05] Julio, J. M., Rodriguez, N., Zarate, H. M., 2005, "Estimating the COP exchange rate volatility smile and the market effect of central bank interventions: A CHARN approach," Borradores de Economia 001901, Banco de la Republica.

[Kea94] Kearns, M., Shapire, R., 1994,"Efficient distribution-free learning of probabilistic concepts", *Journal of Computer and Systems Sciences, 48(3)*, pp. 464-497.

[Kee01] Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., Murty, K. R. K., 2001,"Improvements to platt's SMO algorithm for SVM classifier design", *Neural Computation, 13*, pp. 637-649.

[Lam62] Lamperti, J. W., 1962,"Semi-stable stochastic processes", *Transactions of the American Society, 104 (62)*, pp. 62-78.

[Leb06] LeBaron, B., 2006, "Time scales, agents, and empirical finance," *Manuscript*, Brandeis University.

[Lia96] Liang, J., Parks, T. W., 1996,"A Translation-Invariant Wavelet Representation Algorithm With Applications", *IEEE Transactions on Signal Processing, 44*, pp. 225-232.

[Lin93] Lina, J.-M., Mayrand, M., 1993,"Parameterizations for Daubechies wavelets", *Physical review E, Vol. 48, No. 6*, pp. 4160-4163.

[Lon01] Longin, F., Solnik, B., 2001,"Extreme Correlation of International Equity Markets", *Journal of Finance, 56*, pp. 649-76.

[Lot04] Lotric, U., 2004, "Wavelet based denoising integrated into multilayered perceptron," *Neurocomputing 62*, pp. 179-196.

[Lue96] Lüdecke, T., 1996, *Struktur und qualitaet von finanzmaerkten*, DUV Deutscher Universitäts Verlag.

[Lun07] Lunde, A., Voev, V., 2007, "Integrated Covariance Estimation using High-Frequency Data in the Presence of Noise," *Journal of Financial Econometrics, Vol. 5, No. 1*, pp. 68-104.

[Lus92] Lustig, I. J., Marsten, R. E., Shanno, D. F., 1992,"On implementing Mehrotra's predictor-corrector interior point method for linear programming", *SIAM Journal on Optimization, 2(3)*, pp. 435-449.

[Mae87] Maejima, M., Rachev, S., 1987, "An ideal metric and the rate of convergence to a self-similar process", *Annals of Probability, vol.15*, pp. 702-727.

[Mag98] Magdon-Ismail, M., Nicholson, A., Abu-Mostafa, Y. S., 1998, "Financial markets: Very noisy information processing", *Proceedings of the IEEE, Vol. 86, No. 11*, pp. 2184-2195.

[Mah02] Maheu, J. M., McCurdy, T. H., 2002, "Nonlinear features of realized FX volatility", *Review of Economic Statistics, 84*, pp. 668-681.

[Mal85] Malkeil, B., 1985, *A Random Walk Down Wall Street*, New York: Norton.

[Mal89] Mallat, S. G., 1989,"A Theory of Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*, 674-693.

[Man83] Mandelbrot, B., B., 1983, *The Fractal Geometry of Nature*, San Francisko: W.H. Freeman.

[Man97] Mandelbrot, B. B., Fisher A., Calvet, L., 1997, "A Multifractal Model of Asset Returns," *Cowles Foundation Discussion Papers, No. 1164, Cowles Foundation, Yale University.*

[Mar91] Markowitz, H., 1991, *Portfolio Selection*, Blackwell Publishers.

[Mar06] Martins-Filho, C., Yao, F., 2006,"Estimation of Value-at-Risk and Expected Shortfall based on Nonlinear Models of Return Dynamics and Extreme Value Theory", *Studies in Nonlinear Dynamics & Econometrics, Vol. 10, Issue 2,* Article 4.

[Mcc94] McCormick, K., Raymond, O. Wells, JR., 1994,"Wavelet calculus and finite difference operators", *Mathematics of Computation, Vol. 63, No. 207*, pp. 155-173.

[Mcc96] McCoy, E.J., Walden, A.T., 1996,"Wavelet Analysis and Synthesis of Stationary Long-Memory Processes", *Journal of Computational and Graphical Statistics, Vol. 5, No. 1.*, pp. 26-56.

[Mck99] McKenzie, M. D., 1999, “Power Transformation and Forecasting the Magnitude of Exchange Rate Changes”, *Int. J. Forecast., 15*, 49-55.

[Med02] Meddahi, N., 2002,“A theoretical comparison between integrated and realized volatility”, *Journal of Applied Econometrics, 17*, pp. 479-508.

[Meg89] Megiddo, N., 1989,“Progress in Mathematical Programming”, *chapter Pathways to the optimal set in linear programming, Springer*, New York, NY, pp. 131-158.

[Meh92] Mehrotra, S., Sun, J., 1992, “On the implementation of a (primal-dual) interior point method”, *SIAM Journal on Optimization, 2(4)*, pp. 575-601.

[Mer09] Mercer, J., 1909, “Functions of positive and negative type and their connection with the theory of integral equations”, *Philosophical Transactions of the Royal Society, London A 209*, pp. 415-446.

[Mer80] Merton, R. C., 1980,“On estimating the expected return on the market: An exploratory investigation”, *Journal of Financial Economics, 8*, pp. 323-361.

[Mer03] Mercik, S., Weron, K., Burnecki, K., Weron, A., 2003,“Enigma of self-similarity of fractional levy stable motions”, *Acta Physica Polonica B, 34*, 3773.

[Min05] Min, J. H., Lee, Y.-C., 2005, “Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters,” *Expert Systems with Applications, 28*, pp. 603-614.

[Mue97] Müller, U. A., Dacorogna, M. M., Dave, R. D., Olsen, R. B., Pictet, O. V., Weizsäcker, J. E. von, 1997, “Volatilities of different time resolutions-Analyzing the dynamics of market components”, *journal of Empirical Finance, V. 4, Issues 2-3*, pp. 213-239.

[Mur83] Murtagh, B. A., Saunders, M. A., 1983, “MINOS 5.1 user’s guide”, *Technical Report SOL 83-20R, Stanford University, CA, USA*, Revised 1987.

[Mur04] Murtagh, F., Starck, J. L., Renaud, O., 2004 “On neuro-wavelet modeling,” *Decision Support Systems Journal, Vol. 37*, pp. 475-484.

[Nas94] Nason, G. P., Silverman, B. W., 1994, “The Discrete Wavelet Transform in S”, *Journal of Computationa1 and Graphical Statistics, 3*, pp. 163-191.

[Nel90] Nelson, D. B., 1990,“ARCH models as diffusion approximations”, *Journal of Econometrics, 45*, pp. 7-38.

[Pag88] Pagan, A., Ullah, A., 1988, “The econometric analysis of models with risk terms,” *Journal of applied econometrics, Vol. 3*, pp. 87-105.

[Par80] Parkinson, M., 1980, “The Extreme Value Method for Estimating the Variance of the Rate of Return,” *J. Bus. 53*, pp. 61-65.

[Pat04] Patton, A. J., 2004,“On the out-of-sample importance of skewness and asymmetric dependence for asset allocation”, *Journal of Financial Econometrics, Vol. 2, No. 1*, pp. 130-168.

[Pat06] Patton, A. J., 2006,“Modeling asymmetric exchange rate dependence”, *International Economic Review, Vol. 47, No. 2*, pp. 527-556.

[Pax97] Paxson, V., 1997,“Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic”, *Computer Communication Review, 27 (5)*, pp. 5-18.

[Ped98] Pednault, E. P. D., 1998, “Statistical Learning Theory,” *in The MIT Encyclopedia of the Cognitive Sciences, R.A. Wilson and F.C. Keil (Editors), MIT Press: Cambridge, MA.*, pp. 798-801.

[Per93] Percival, D. B., Guttorp, P., 1993, “Long-Memory Processes, the Allan Variance and Wavelets,” *Working paper.*

[Per97] Percival, D. B., Mofjeld, H. O., 1997, “Analysis of Subtidal Coastal Sea Level Fluctuations Using Wavelets,” *Journal of the American Statistical Association, Vol. 92, No. 439*, pp. 868-880.

[Per00] Percival, D. B., Walden, A. T., 2000, *Wavelet Methods for Time Series Analysis*, Cambridge University Press: Cambridge.

[Pes96] Pesquet, J.-C., Krim, H., Carfantan, H., 1996, “Time-Invariant Orthonormal Wavelet Representations”, *IEEE Transactions on Signal Processing, 44,* pp. 1964-1970.

[Pet96] Peters, E. E., 1996, *Chaos and order in the capital markets: A new view of cycles, prices, and market volatility* , second edition, Wiley Finance.

[Pla99] Platt, J., 1999, “Fast training of support vector machines using sequential minimal optimization,” In: Schölkopf, B., Burges. C. J. C., Smola. A. J. (Eds.), *Advances in Kernel Methods-Support Vector Learning*, Cambridge, MA, MIT Press, pp. 185-208.

[Pol03] Polzehl, J., Spokoiny, V., 2003, “Varying coefficient regression modeling by adaptive weights smoothing,” WIAS-Preprint 818.

[Poo03] Poon, S.-H., Granger, C. W. J., 2003, “Forecasting Volatility in Financial Markets: A Review”, *Journal of Economic Literature, Vol. XLI,* pp. 478-539.

[Rac00] Rachev, S., Mittnik, S., 2000, *Stable Paretian models in finance*, Wiley: New York.

[Rac01] Rachev, S., Samorodnitsky, G., 2001, “Long strange segments in a long range dependent moving average”, *Stochastic Processes and their Applications, vol. 93*, pp. 119-148.

[Rac05a] Rachev, S., Stoyanov, S.V., Biglova, A., Fabozzi, D.J., 2005, "An Empirical Examination of Daily Stock Return Distributions for U.S. Stocks", *In Baier, D., Decker, R., Schmidt-Thieme, L. (Eds.)* , Data Analysis and Decision Support, Springer.

[Rac05b] Rachev, S., T., Menn, C., Fabozzi, F. J., 2005, *Fat-tailed and skewed asset return distributions: Implications for risk management, portfolio selection, and option pricing*, Wiley Finance.

[Rac07] Rachev, S., Stoyanov, S.V., Wu, C., Fabozzi, D.J., 2007, "Empirical Analyses of Industry Stock Index Return Distributions for the Taiwan Stock Exchange", *ANNALS OF ECONOMICS AND FINANCE Vol. 8, No. 1.*

[Ren03] Reno, R., 2003, "A closer look at the Epps effect", *International Journal of Theoretical and Applied Finance, 6*, pp. 87-102.

[Res97] Restrepo, J. M., Leaf, G. K., 1997, "Inner product computations using periodized Daubechies wavelets", *International Journal for Numerical Methods in Engineering, 40*, pp. 3557-3578.

[Rol84] Roll, R., 1984,"A simple measure of the effective bid-ask spread in an efficient market", *Journal of Finance, 39*, pp. 1127-1139.

[Saf07a] Safari, A., Seese, D., 2007, "Distributional and dynamical properties of realized volatility and correlation," *Quantitative Finance*, Under second review.

[Saf07b] Safari, A., Seese, D., 2007, "Behavior of realized volatility and correlation in exchange markets," *Euroasian Review of Econometrics*, Under review.

[Saf08a] Safari, A., Seese, D., 2008, "Nonparametric estimation of a Multiscale CHARN model using SVR," *Quantitative Finance*, To appear.

[Saf08b] Safari, A., Seese, D., Sun, W., Rachev, S., 2008, "Realized Volatility and Correlation Estimators under Non-Gaussian Microstructure Noise," *In: Frank Columbus (ed.), Economic dynamics: Theory, Games and Empirical Studies. NOVA: New York*, To appear.

[Saf08c] Safari, A., Seese, D., Chalup, S., 2008, "An ε-E-insensitive support vector regression," *ICML 2008*, Submitted.

[Sam94] Samorodnitsky, G., Taqqu, M. S., 1994, *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman and Hall, London.

[Sch77] Scholes, M., Williams, J., 1977, "Estimating betas from nonsynchronous trading," *Journal of Financial Economics, 5*, pp. 309-327.

[Sch00] Schölkopf, B., Smola, A. J., Williamson, R. C., Bartlett, P. L., 2000, "New Support Vector Algorithms," *Neural Computation: Massachusetts Institute of Technology, 12*, pp. 1207-1245.

[Sch02] Schleicher, Ch., 2002, "An introduction to wavelets for economists," *Working Paper 2002-3*, Monetary and Financial Analysis Department, Bank of Canada.

[Sei94] Seifert, B., Brockmann, M., Engel, J., Gasser, T., 1994, "Fast algorithms for nonparametric curve estimation," *Journal of Computational and Graphical Statistics, 3*, pp. 192-213.

[She92] Shensa, M. J., 1992, "The Discrete Wavelet Transform: Wedding the A'Trous and Mallat Algorithms," *IEEE Transactions on Signal Processing, 40*, pp. 2464-2482.

[Smo98] Smola, A. J., Murata, N., Schölkopf, B., Müller, K.-R., 1998,"Asymptotically optimal choice of ε-loss for support vector machines," *In Proceedings of the International Conference on Artificial Neural Networks*, Berlin: Springer-Verlag, pp. 105-110.

[Smo04] Smola, A. J., Schölkopf, B., 2004, "A tutorial on Support Vector regression," *Statistics and Computing, 14*, pp. 199-222.

[Sol96] Solnik, B., Boucrelle, C., and Le Fur, Y., 1996,"International Market Correlation and Volatility", *Financial Analysts Journal, September-October*, pp. 17-34.

[Sol00] Soltani, S., Boichu, D., Simard, P., Canu, S., 2000, "The long-term memory prediction by multiscale decomposition," *Signal Processing, 80*, pp. 2195-2205.

[Sto04] Stoev, S., Taqqu, M. S., 2004,"Simulation methods for linear fractional stable motion and FARIMA using the fast Fourier transform", *Fractals, Vol. 12, No. 1*, pp. 95-121.

[Sun06] Sun, W., Rachev, S., Fabozzi, F. J., 2006,"Long-Range Dependence, Fractal Processes, and Intra-Daily Data", *Handbook of IT and Finance*, Springer, forthcoming.

[Sun07] Sun, W., Rachev, S., Fabozzi, F. J., 2007,"Fractals or I.I.D. : Evidence of long-range dependence and heavy tailedness from modeling German market returns", *Journal of Economics and Business, 59*, pp. 575-595.

[Sun08] Sun, W., Rachev, S., Fabozzi, F. J., Kalev, P.S., 2008,"Unconditional Copula-Based Simulation of Tail Dependence for Co-movement of International Equity Markets", *Empirical Economics, Springer*, forthcoming.

[Tay86] Taylor, S., 1986, *Modeling Financial Time Series*, Wiley: New York.

[Tik77] Tikhonov, A. N., Arsenin, V. Y., 1977, *Solutions of Ill-posed Problems*, Winston: Washington D. C.

[Ull05] Ullrich, Ch. M., Seese, D., Chalup, S., 2005, “Predicting foreign exchange rate return directions with Support Vector Machines,” *Proceedings of the 4th Australasian Data Mining Conference, Sydney, Australia*, pp. 221-240.

[Van94] Vanderbei, R. J., 1994, *LOQO: An interior point code for quadratic programming*, TR SOR-94-15, Statistics and Operations Research, Princeton University, NJ.

[Vap71] Vapnik, V. N., Chervonenkis, A. Y., 1971,“On the uniform convergence of relative frequencies of events to their probabilities”, *Theory Probab. Appl., 16*, pp. 264-280.

[Vap81] Vapnik, V. N., Chervonenkis, A. Y., 1981, “Necessary and sufficient conditions for the uniform convergence of means to their expectations”, *Theory of Probability and its Applications, 26*, pp. 532-553.

[Vap91] Vapnik, V. N., Chervonenkis, A. Y., 1991, “The necessary and sufficient conditions for consistency of the method of empirical risk minimization”, *Pattern Recognition and Image Analysis, 1, (3), pp. 284-305.* Originally published in (1989) Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting, 2.

[Vap95] Vapnik, V. N., 1995, *The Nature of Statistical Learning Theory*, Springer-Verlag: New York.

[Vap98] Vapnik, V. N., 1998, *Statistical Learning Theory*, Wiley: New York.

[Vap99] Vapnik, V. N., 1999,“An Overview of Statistical Learning Theory”, *IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 5*, pp. 988-999.

[Wah90] Wahba, G., 1990, “Splines Models for Observational Data”, *Series in Applied Mathematics, Vol. 59*, SIAM: Philadelphia.

[Wer05] Weron, A., Burnecki, K., Mercik, S., Weron, K., 2005,“Complete description of all self-similar models driven by Levy stable noise”, *Physical Review E, 71.*

[Wes04] West, B.J., Scafetta, N., Cooke, W.H., Balocchi, R., 2004,“Influence of Progressive Central Hypovolemia on Hölder Exponent Distributions of Cardiac Interbeat Intervals”, *Annals of Biomedical Engineering, Vol. 32, No. 8*, pp. 1077-1087.

[Whi63] Whittle, P., 1963,“On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral matrix”, *Biometrika, 40*, pp. 129-134.

[Yan02] Yang, H., Chan, L., King, I., 2002,"Support vector machine regression for volatile stock market prediction", *Springer-Verlag, Berlin Heidelberg, IDEAL*, pp. 391-396.

[Yan04] Yang, H., Huang, K., Chan, L., King, I., Lyu, M.R., 2004,"Outliers treatment in support vector regression for financial time series prediction", *Springer-Verlag, Berlin Heidelberg, ICONIP*, pp. 1260-1265.

[Zha03] Zhang, J., Tsui, F. C., Wagner, M. M., Hogan, W. R., 2003, "Detection of Outbreaks from Time Series Data using Wavelet Transform," *Journal of the American Medical Informatics Association: AMIA Annu. Symp. Proc. 2003*, pp. 748-752.

[Zha05] Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005,"A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data", *Journal of the American Statistical Association, Vol. 100, No. 472*, pp. 1394-1411.

[Zha06] Zhang, L., 2006,"Estimating Covariation: Epps Effect, Microstructure Noise", *Unpublished paper: Department of Finance, University of Illinois at Chicago.*

[Zho96] Zhou, B., 1996,"High-frequency data and volatility in foreign-exchange rates", *Journal of Business and Economic Statistics, 14 (1)*, pp. 45-52.