PhD Thesis

Financial Risk Management and Portfolio Optimization Using Artificial Neural Networks and Extreme Value Theory

Author: MABOUBA DIAGNE [1]

Supervisor: Prof.Dr Juergen Franke [2]
Reporter: Prof.Dr Marlene Mueller

University of Kaiserslautern

Mathematics Department/Financial Mathematics

10th October 2002

[1] Diagne Mabouba, Dresdner Bank AG, Corporate Center Revision, Investment Banking, Mainzer Landstrae 27 - 31, D-60329 Frankfurt am Main, Germany

[2] Prof.Dr Juergen Franke, Fachbereich Mathematik, Universitaet Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany.

Contents

Acknowledgements

I particular thank Prof. Dr Juergen Franke my Supervisor, Prof.Dr Marlene Mueller for the correction of the PhD, Prof.Dr Ivar Ekeland of the Finance Institute of Paris Dauphine, Prof. Dr Mary Teuw Niane Director of the E.U.R of Gaston Berger in Senegal, Martin Schreiter and Jan Muench of CC Revision/Risk Control/Dresdner Bank who over the last years generously shared with me many fine insights into the subjects of financial mathematics and financial risk management. I'm also very thankful to Prof. Halbert White of the university of San Diago who did provide to me some of his scientific papers about neural networks. Lastly I thank my wife Mbodj Maram and my mother Diagne Asta for their unconditional support over the years.

Introduction

During the last two decades, numerous important results on feed-forward artificial neural networks (ANN) have been developed (see White-Hornik and Stinchcombe [1988]: a, b), Caroll and Dickinson [1989], Funahaschi [1989]). Beyond the important fact that the output functions of feed-forward ANN help to build non-parametric approximates for arbitrary measurable functions, the theory of ANN also provides a broad range of financial applications more specifically in market risk quantification, portfolio optimisation (see Franke [1998]) and financial forecasting.
A highly comprehensive market risk measures is the so-called Value-at-Risk(VaR). VaR summarizes throughout one single quantitative parameter the whole market risk exposure. VaR approaches might be named differently e.g. Bankers Trusts Capital at Risk (CaR), J.P Morgan's Value-at-Risk, Daily Earning at Risk (DEaR), Dollar at Risk (DaR), Money at Risk (MaR)). Its technical implementations differ among the financial institution where it is used. But there is some convergence in terms of high-level approaches for measuring aggregate market risk exposure (see Alexander [1999]). This is one of the main reason why on April 1995, the B.I.S [1] recommended its use and consider the VaR as a standard risk-controlling tool for aggregating market risk exposure.
Most of the widely used VaR methods and their underlying financial returns models, from Monte Carlo Algorithm, Delta Normal, Delta Gamma VaR, Variance-Covariance, all display some fundamental weaknesses by assuming that the dynamics of portfolio changes and the absolute or logarithmic returns of financial assets are described using the following questionable assumptions and drawbacks:

- Distributions of security returns assumed to be multivariate normal;
- Constant volatility of the financial returns;
- Linear or quadratic assumptions of portfolio pricing functions;
- Disability to handle or control extreme market events.

[1] B.I.S=Bank of International Settlement

Despite its conceptual simplicity, the measurement of the VaR is a very challenging statistical problem and none of the methodologies developed so far give satisfactory solutions. In reality, logarithmic returns and the market values of portfolio changes usually display some patterns of:

- Heavy tailedness;
- Skewness;
- Heteroscedasticity;
- Strong Non-Linearity.

Therefore the classical normality assumptions with constant volatility or the linear and quadratic hypothesis are not appropriate for modelling the financial returns that requires the VaR calculation. The payoff profile of portfolios containing complex derivatives such as barrier options or digital ones do not usually cope with the quadratic assumption of the portfolio values. The constant volatility assumption usually made by most of the existing financial returns model and VaR methodologies do not really provide a fair description of the stochastic behaviour of the underlying risk elements. Therefore they cannot be used for a precise valuation of contigent claims like interest rate derivatives such as American-style swap options, callable bonds or structured notes.
In practical applications, focus is frequently on higher-order statistics such as the skewness[2] and the kurtosis[3] and the stochastic volatility of financial returns that serve as the basis for hedging strategies or risk control. The increasing globalization and complexity of capital markets and the expanding range of exotic financial instruments have made trading-risk management more difficult to accomplish and evaluate. Therefore risk management systems need to be more and more sophisticated due the complexity of certain range of non-linear traded contingents claims, special methods such ANN combined with Extreme Value Theory(EVT) can provide considerable insights in order to correct the fundamental weaknesses of existing financial returns models and VaR methodologies. Throughout this thesis, the main focus consists of correcting such drawbacks by using ANN combined with EVT. A new VaR methodology that overcomes several well-known deficiencies of existing VaR approaches will be the main focus. Such a methodology enables to avoid potentially disastrous clustering in predicted tail events by accurately estimating the conditional distribution of asset

[2] Skew(X):= $E\left[\frac{X-E(X)}{\sigma(X)}\right]^3$ is the skewness of the random variable X. Skew(X) says how symmetric is X.

[3] Kurtosis(X):=$E\left[\frac{X-E(X)}{\sigma(X)}\right]^4$ is the tail measure of the random variable X.

returns by using ANN and EVT. To correct the normality assumption as well as the constant volatility hypothesis, one can model the changes of the market values of financial assets and their corresponding unexpected returns using heavy tailed autoregressive and conditionally heteroskedastic time series. The models used for the innovation of such models can be either Cauchy, Pareto, Weibull, Student, Log-gamma or Frechet distributions. The heavy tailedness imposed on the innovations of such autoregressive financial time series represents also another form of correction of the classical unwarrant normality assumptions usually made for getting a quick and easy computable form of the VaR. In order to estimate the future market value of a given financial instrument using such heavy tailed, heteroskedastic and autoregressive models, one needs to quantify the expected return, the conditional stochastic volatility of the market value of the instrument and the conditional quantile of the corresponding innovations that can also be equated as the unexpected returns (see Engles, EGARCH[4][1982]). As stated in Franke [1999], when the autoregressive order of such a model is very high, the use of non-linear regression likes the kernel estimation procedure can lead to underestimation. An alternative consists of making use of the theory of ANN for non-parametric regression analysis purposes combined with EVT.
White [1981] did solve such regression problems by applying some denseness results of neural network outputs functions but restricted to bounded random variables. Such denseness properties become no longer applicable when dealing with heavy tailed stochastic processes, which are unbounded. The boundedness that requires White's results is not plausible for financial time series as one of the well known stylized facts implies that heavy-tailedness which strongly contradicts any boundedness.

Therefore, in the first chapter an extension of White's denseness results will be given. A new result will be proved in order to derive new denseness and approximation properties, which uses the Sobolev space $L^m(\mu)$ for some $m \geq 1$ and some probability measure and which holds for a certain subclass of square integrable functions. The main idea is to use the Fourier transform in a similar manner as Barron [1993]. Two others new results will also be proved, consisting of an extension of the Bernstein inequality for unbounded random variables from stationary α-mixing processes and a generalization of a result of White and Wooldrige[1990] theorem 3.5, which assumes that the tails of the stationary distribution decrease to 0 faster than exponential function. We require only that they decrease like $b_0 \exp(-b_1 x^\alpha)$ for $x \to \infty$ for some $\alpha > 0$ (not $\alpha > 1$ like in White and Wooldrige). Based on the autoregressive models, one can use the new denseness result to build some ANN estimate of the conditional mean of heavy

[4] EGARCH= Exponential, Generalised Autoregressive and Condiditonaly Heteroskedastic.

tailed stochastic processes. In the same manner, this new approximation property also provides an algorithm for estimating conditional stochastic volatilities. To measure the accuracy of the estimation procedures, the consistency of the ANN estimates of the conditional expectations will be analysed. Within such models, to estimate the VaR or the Expected Shortfall(ES[5]; See Delbaen [1998]), one has to combine ANN functions and EVT (see Embrechts [1997], Mc-Neil [1999-2000]). EVT will be used for estimating the tails and the quantiles of the unexpected returns, which is needed in the final VaR estimation formula. For the assessment of the quantile of the unexpected returns, the use of some results based on Extreme Value Distributions(EVD) and Generalised Pareto Distributions (GPD) will provide the accurate estimates for the quantiles of heavy tailed stochastic processes (See Smith [1987], Embrechts [1997] or Alexander Mc Neil [1999]). The last section of the first chapter deals with numerical simulations using real financial data in the aim to illustrate the accuracy of the VaR methodology via the computation of the daily Value-at-Risk of one share of COMMERZBANK. As explanatory variables, we use daily closing prices of DEUTSCHE Bank, the ones of BASF, SIEMENS and the DAX30 index all traded in the Frankfurt stock exchange.

The purpose of the second chapter is to implement a forecasting ARMA[6]-GARCH algorithm that enables to predict future stock prices of a given security by estimating the conditional expected returns while taking into account the stochastic features of the volatility of financial instruments. Throughout this chapter, after specifying the financial returns model, the first section deals with the estimation of the conditional expected returns and the conditional stochastic volatility by means of ANN output functions and using the new denseness result previously established in the first chapter. The non-parametric ARMA-GARCH algorithm of Mc-Neil and Buehlmann [2000] will be combined with the new denseness results to derive the stochastic volatility estimates. Instead of using the contraction assumptions as implemented in Mc-Neil and Buehlmann [2000], the convergence results of Corradi and White based on Regularized ANN [1995] will be used. Corradi and White have shown that Regularized ANN are capable of learning and approximating (on compacta) elements of certain Sobolev Spaces. The third section is dedicated to the consistency and the convergence of the resulting estimators. In the last section, based on some regularity assumptions imposed on the volatility regression function, a powerful financial forecasting algorithm will be designed. Beside the correction of the constant volatility assumption, the algorithm also enables to make some financial pricing. To illustrate such capability,

[5] ES=Expected Shortfall

[6] ARMA=Autoregressive and Moving Average Processes.

some options pricing formula based on the new non-parametric AR-GARCH and using ANN and EVT under a stochastic volatility framework is implemented. In this subsection, Bootstrapping algorithms are used in order to compare the option pricing methodology of the non-parametric ARMA-GARCH algorithm and the well known Black-Scholes option pricing formula which assumes constant volatility and normality assumptions. One Call option on a SIEMENS share will be computed.

Based on the new approximation and denseness results established in the first and second chapters, one can also derive some straightforward estimates of the conditional quantile of the financial returns using the quantile characterisation of Bassett and Koenker [1978]. The third chapter is making use of the quantile characterisation of Bassett and Koenker while correcting the normality assumptions, the constant volatility and the skewness of financial time series. The third chapter is structured in the following manner. After presenting the existence and convergence results and exhibiting the qualitative features of the nonparametric neural network quantile estimates, follows the analysis of the goodness and the accuracy of the Value-at-Risk methodology based on such a neural network estimates of the conditional quantile of the distributions of the market value of the considered instruments. The goodness and the accuracy of the Value-at-Risk methodology based on such a neural network estimates of the conditional quantile is illustrated throughout the computation of the daily VaR of a holding consisting of one share of DEUTSCHE Bank. As explanatory variables, we use the daily closing prices of BASF, SIEMENS, COMMERZBANK and DAX30 traded on the stock exchange of Frankfurt.

The last chapter is mainly dealing with some theoretical results under a continuous time setting of financial returns. It also attempts to correct the constant volatility assumption usually made when stochastic differential equations and Diffusion Itô processes are used to describe the dynamics of the market value of financial assets. It can be equated as the completion and the continuous time extension of the financial returns model of the three first chapters. Markovian diffusion neural network theory combined with stochastic calculus will be the main tool. Active modern research based on methods of Markovian diffusion theory and diffusion neural networks have shown that, using contrastive Hebbian learning rules (CHL), one can formalize the activation dynamics of diffusion neural networks in order to reproduce the entire multivariate probability distributions of a given financial instrument (see Movellan and Clelland [1993]). The CHL have some appealing features that enable to capture differences between desired and obtained continuous probability distributions. Diffusions Networks are type of recurrent neural network with probabilistic dynamics, as models for learning natural signals that are continuous in time and space. Since the solutions for many decision theoretic problems of interest are naturally formulated using probability distributions, it would be desirable to design flexible neural networks frameworks for approximating probability distributions on continuous path spaces. Instead of using ordinary differential equations for describing the evolution of stock prices or portfolio values, diffusion networks are described by a set of stochastic differential equations. Diffusion neural networks are an extension of recurrent neural networks in which the dynamics are probabilistic. They have been found very useful in stock price prediction (see Mineiro, Movellan and Williams [1997], Kamijo and Tanigawa [1990], Kimoto and Asakawa [1990], Refenes et al [1993]), Movellan [1997]). The main advantages of Diffusion Networks over conventional forecasting methods include simplicity of implementation and good approximation properties (see Warwick et al [1992]). In this chapter, we present some theoretical results illustrating the use of Diffusion Networks for financial prediction. We show that, under the general regularity conditions allowing existence and uniqueness of solutions of stochastic differential equations, under some appropriate settings, one can approximate the transition probabilities and the log-likelihood functions and derive consistent non-biased predicting algorithm of future values of a considered stock.

Combining the ideas of learning probability distributions with symmetric diffusion networks(see Mineiro, Javier Movellan and Williams [1997] or Movellan [1997]) with the Maximum Likelihood estimation algorithm developed Pedersen [1993], which is based on incomplete observations of stochastic processes, one can build a forecasting tool providing consistent and unbiased estimates of the futures values of stock prices. In the first section of the chapter, the definition of the log-likelihood function, the concept of transition probabilities and their density functions will be defined. Up to some regularity conditions, it will be shown that the approximate probability density functions of the transition probabilities converge in law to the underlying ones. The analysis of the qualitative features and the study of the convergence properties, such consistency and asymptotic normality will also be illustrated (see Pedersen [1994], Dachunha and Zmirou [1989]).

Frequently Used Notation

$\Re$	Set of Real Values
$\mathcal{N}$	Set of Integer Values
β	Input weights from input to hidden layers
γ	Input weights from hidden layers to output layers
H	Number of hidden nodes
$f_H(x, \beta, \gamma)$	Neural output function
L_{ind}	Set of linear combinaisons of indicators functions of finite interval
α	Confidence level, either 0,95 or 0,99
T	Risk Horizon: One day, one week or 10 days
VaR_α^T	Value-at-Risk for a confidence α within a risk horizon equal to T
ES_α^T	Expected Shortfall for a confidence α within a risk horizon equal to T
$var(X)$	Variance of the random variable X
Cov	Covariance operator
$Pr(A)$	Probability that event A occurs
$H_{\psi,\mu,\sigma}$	Generalized Extreme Value Distribution
$G_{\psi,\beta}$	Generalized Pareto Distributions
$ANN(\Psi, q_n, \Delta_n)$	Neural output functions with some growth conditions on the weights
S_t	Returns or daily closing price of the considered asset
X_t	Explanatory variable used in the prediction of S_t
r	Dimension of the explanatory variable X_t
$L^2(\mu)$	Set of μ-square integrable functions
x^T	Transpose of x
u	Threshold level used for estimationg tails of heavy tailed distributions
N_u	Number of excesses above the threshold level u
F_u	Excess distibution function
ANN	Artificial Neural Network
EVT	Extreme Value Theory
VaR	Value-at-Risk
$(\Re^N, \mathcal{B}^N)$	The Borel space of $\Re^N$
$\mathcal{M}^N$	Set of measurable functions in $\Re^N$
$\vert\vert \ \vert\vert_m$	$\vert\vert x \vert\vert_m := \int_{\Re^r} \vert\vert x \vert\vert^m \mu(dx)$
$\mathcal{D}^l$	The differential operator $\mathcal{D}$ applied l times.
∇	The gradient operator
$\frac{\partial}{\partial w_i}$	The partial derivative with respect the ANN weight w_i
$\mathcal{N}_\epsilon(\theta_\alpha)$	ϵ neighbourhood of θ_α
$\vert A \vert$	Determinant of the matrix A.

Chapter 1

Value-at-Risk and Expected Shortfall Estimation Using Artificial Neural Networks and Extreme Value Theory

1.1 Introduction

Numerous hard statistical and scientific studies have indicated that the stock market, as well as other financial markets, are, like other complex natural phenomena, to a certain degree predictable by means of newly developing methods and tools. Movements of the stock prices, as well as price movements of other financial instruments, generally present a deterministic trend, on which are superimposed some "noise" signals, in turn composed of truly random and chaotic signals. Deterministic trends can be detected and assessed by some maximum-likelihood methods. Although a truly random signal, often represented by a Brownian motion, is unpredictable, it can be estimated by its mean and standard deviation. The chaotic signal, seemingly random but with deterministic nature, proves predictable to some degree by means of several analysis techniques, among which the Artificial Neural Network (ANN) techniques have proven most effective over the widest range of predictive variables. Artificial Neural Network (ANN) is an important branch of Artificial Intelligence. Motivated in its design by the human nervous system, ANN mimics the human nervous system in its operations. At this extraordinary interface between natural human systems and created electronic ones, ANN is capable of learning by training to generalize from special cases just like human beings can. Beyond the fact that neural network output functions are dense in the huge set of measurable functions, the theory of artificial neural networks also provides a broad range of financial applications more specifically in market risk quantification and portfolio optimisation (see Franke [1998]). Neural networks have been applied to a variety of financial prediction and risk management tasks. In practical applications, focus is frequently on higher-order statistics such as the variance, skewness, and kurtosis of financial returns that serve as the basis for hedging or risk-control strategies. In this chapter, the main focus consists of correcting the classical questionable assumptions of existing Value-at-Risk methodologies. Value at Risk (VaR) has become the standard measure of market risk employed by financial institutions for both internal and regulatory purposes. VaR is defined as the value that a portfolio might lose with a given probability, over a certain time horizon (usually one or ten days). Despite its conceptual simplicity, its measurement is a very challenging statistical problem and none of the methodologies developed so far give satisfactory solutions. For example, the delta normal method is based on a linearization of the portfolio, and thus can perform poorly with portfolios that include large positions in options or instruments with option like payoffs. The Monte Carlo or the Delta Gamma Approach are both based on normality assumptions that also contradict the skewness, the heavy tailedness that logarithmic returns of financial instruments are usually displaying. Interpreting the VaR as the quantile of future portfolio values conditional on current information, this chapter is mainly dealing with a new approach to quantile estimation which does not require any of the questionable assumption invoked by existing methodologies. This chapter is dedicated to the development of efficient methods for computing portfolio VaR where the underlying risk factors can be drawn from heavy tailed and heteroskesdastic distributions. This new methodology overcomes several well-known deficiencies of existing Value at Risk approaches. It enables to avoid potentially disastrous clustering in predicted tail events by accurately estimating the conditional distribution of asset returns by using artificial neural networks and extreme value theory. Autoregressive models will be used; therefore market risk measures have to be computed conditionally to the whole market information up to trading days. Hence the conditional VaR has to be considered. The mathematical schematisation, which enables to take into account, all these considerations can be described by modelling the portfolio returns as an autoregressive and conditionally heteroskedastic financial time series e.g.

$$S_t = m\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right) + \sigma\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right)\mathcal{E}_t \qquad (1.1)$$

Where:

- The autoregressive order τ is a given integer.
- S_t represents the financial return of the portfolio at the t-th trading day.
- X_{t-1} is an $\Re^d$ random variable and can be interpreted as current market information.
- The Predictor function m is the conditional expected return .
- σ represents the conditional volatility of the portfolio changes.

$\mathcal{E}_t$ are iid[1] random variables such that

$$E(\mathcal{E}_t) = 0 \text{ and } Var(\mathcal{E}_t) = 1.$$

The stochastic process S_t is called an AR-ARCH[2] process and is heteroskedastic (time changing volatility). One can also assume that the innovations are heavy tailed. Such an assumption copes with a lot of financial applications due to the fact that errors (or unexpected returns) resulting from financial returns are usually displaying some patterns of heavy tailedness. Its the case when innovations are Cauchy, Pareto, Student, Log-gamma or Frechet distributed. The heavy tailedness imposed on the innovations also corrects the classical unwarrant normality or linearity assumptions usually made for getting a quick and easy computable form of the Value-at-Risk.

For estimating the future market value of a given portfolio, one needs to quantify the expected return m, the volatility σ of the portfolio values and the conditional quantile of the innovations $\mathcal{E}_t$ that can be equated as the unexpected returns (see Engles, EGARCH [1982]). As stated in Franke [1999], when the autoregressive order τ and the dimension d of the fixing space R^d are very high, the use of localized nonlinear regression like the kernel estimation procedure or local polynomials cannot be applied due to the cause of dimensionality, i.e even for large sample sizes, there are too few observations in local neighbourhoods of $\Re^{\tau+d}$ to estimate m and σ reliably by local smoothing . An alternative consists of making use of the theory of ANN for non-parametric regression analysis purposes combined with EVT.

To forecast the VaR or the Expected Shortfall, ANN and EVT (see Embrechts [1997], A.Mc-Neil [1999-2000]) are combined. Based on autoregressive models, White did solve the regression problem by applying some denseness results valid only for bounded random variables. Such denseness properties become no longer applicable when dealing with heavy tailed stochastic processes such as financial returns or portfolio changes. Therefore, later in this chapter the extension of White's denseness results is necessary in order to derive a suitable neural network estimate of the conditional expectation for unbounded random variables. The

[1] iid=independent and identically distributed.

[2] AR-ARCH=Autoregressive-Autoregressive and Conditionally Heteroskedastic.

boundedness that requires White's results is not applicable for unbounded financial time series as one of the well-known stylised facts implies that heavy tailedness which strongly contradicts any boundedness. The extension of H.White's denseness results will be done via some Sobolev Spaces and the use of the Fourier transform algorithm. The metric Sobolev Space $L^m(\mu)$ for some $m \geq 1$ and some probability measure μ are used. The main idea is to use the Fourier transform based on assumptions in a similar manner as Barron [1993].

For the assessment of the quantile of the unexpected returns, we make use of some results based on Extreme Value Theory and providing an estimator for the quantiles of generalized Pareto Distribution (See Smith [1987], Embrechts [1997] or Alexander Mc Neil [1999]). This section is mainly dealing with the estimation of the tail of Generalized Pareto Distribution above some appropriate threshold. The section related to Artificial Neural Network is used for estimating the expected return and the conditional volatility, while the part dealing with EVT will be used for estimating the quantile of the unexpected returns supposed to be heavy tailed and following some generalized Pareto Distributions over some appropriate threshold. The asymptotic properties such as consistency and asymptotic normality of the resulting dynamic Value-at-Risk measure will be analysed. The last section is dedicated to the performance results and the goodness and the level of accuracy via some numerical simulation with real financial data. Throughout the numerical simulation, the daily Value-at-Risk of a one COMMMERZBANK share will be computed. As explanatory variables, the DEUTSCHE Bank daily closing prices, the ones of BASF, SIEMENS and the DAX30 stock index traded in the Frankfurt stock exchange will be used.

1.2 Market Risk Assessment

Market Risk is the risk that a position will not be as profitable as an investor expected because of fluctuations in market prices or rates (e.g. equity prices, interest rates, currency rates or commodity prices). Market Risk can be defined as the uncertainty of the future market values of the portfolio's profits and losses resulting from adverse market movements of the market-risk factors. The market-risk factors help to compute the whole market risk using the Value-at-Risk for aggregating the whole market risk exposure. Although several types of approaches are available for measuring market risk, institutions have increasingly adopted the Value-at-Risk approach for their trading operations.

Given some confidence level $\alpha \in (0,1)$ and a risk horizon T, the Value-at-Risk($VaR(\alpha, T)$) of a portfolio at the confidence level α within the risk horizon $[0, T]$ is defined by the smallest real value l such that, the probability that the portfolio changes(P&L[3]) exceeds l is not larger than $1-\alpha$. Formally

$$VaR(\alpha, T) := \inf\{l \in \Re \;\text{ such that } Pr\,(P\&L \,>\, l\,) \;\leq 1-\alpha\} \tag{1.2}$$

1.2.1 Definition: Value-at-Risk

The definition of the Value-at-Risk in (1.2) can also be viewed quantitatively as an α quantile of the $P\&L$ distribution in terms of generalised inverse of the distribution function $F_{P\&L}$ e.g.

$$\begin{aligned} VaR(\alpha, T): &= \inf\{l \in \Re \;\text{ such that } 1 - F_{P\&L}(l) \leq 1-\alpha\} & (1.3)\\ &= \inf\{l \in \Re \;\text{ such that } F_{P\&L}(l) \geq \alpha\} & (1.4)\end{aligned}$$

where the function $F_{P\&L}$ represents the distribution function of the $P\&L$ distribution. To correct the non-subadditivity of the Value-at-Risk, the Expected Shortfall is also used as a market risk measure.

1.2.2 Definition: Expected Shortfall

The Expected is defined as the expected return given that returns are greater than the Value-at-Risk, i.e

$$ES_T^{\alpha}: \;=\; E\,(P\&L|P\&L \geq VaR(\alpha, T)) \tag{1.5}$$

1.2.2.1 Remarks

Remark
The first important comment regarding the computational aspect of the Value-at-Risk, lies on its different technical implementations that varies closely with respect to the assumptions made on the probability distribution of the portfolio returns. In practice, some linear or quadratic and normality assumptions are usually made on the distribution function driving the portfolio returns, in order to get an easy and direct computable form of the Value-at-Risk. In such cases, the VaR derives from the standard deviation of the portfolio changes, the α quantile Q_α of the standard normal distribution adjusted by the square root of the risk horizon e.g

$$VaR(\alpha, T) := Q_\alpha * \sigma_{portfolio} * \sqrt{T}, \tag{1.6}$$

where

$$\begin{cases} \sigma_{portfolio}: & = & \sqrt{\Pi'\Gamma\Pi},\\ \Pi: & = & \text{Portfolio Weights},\\ \Gamma: & = & \text{Covariance Matrix.}\end{cases}$$

[3] P&L=Profits and Losses.

As stated in Jorion [1997], the greatest advantage of the VaR lies on the fact that the Value-at-Risk summarizes throughout one single quantitative parameter the whole market risk aversion. In practice, in financial corporations that are managing large portfolios like commercial or investment banks and some large credit institutions; the financial reserves or economic capital effectively used as a cushion, is the so-called “Safe Value-at-Risk” computed as K*VaR. Where K represents some safety multiplicative factor, depending on the number of times the VaR was violated during the Back-testing of the Internal Risk Model of the given institution. Note that, the choices of the confidence level α, the risk horizon T and the multiplicative factor K are not uniform and depend mainly on the size of the financial institution where the Value-at-Risk is implemented. Another important remark is about the normality assumption of the portfolio returns. In practice financial portfolio returns usually display some patterns of heteroscedasticity (time changing variance). Therefore the classical normality assumptions do not fully cope with the reality. An alternative consists on modelling the portfolio absolute or logarithmic returns using autoregressive processes as in (1.1) with ARCH innovations. To correct the drawbacks of existing Value-at-Risk methodologies, we estimate the corresponding quantile leading to the Value-at-Risk by using some non-parametric regression methods such as ANN and combine it with EVT. This can be done in the following manner:

1.2.2.2 Proposition

Under the model(1.1), the Conditional Value-at-Risk VaR_α^t is related to the expected return m, the time dependent volatility σ and the α-quantile of the innovations q_α by the following relation:

$$VaR_\alpha^t = m\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right) + q_\alpha * \sigma\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right). \quad (1.7)$$

Proof:
This can be proved by the following reasoning :
$\forall x \in \ \Re$ and $\forall \alpha \in [0,1]$:

$$P_{|\mathcal{F}_{t-1}}\left(S_t \ \leq x\right) = P_{|\mathcal{F}_{t-1}}\left(\mathcal{E}_t \ \leq \frac{x - m\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right)}{\sigma\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right)}\right). \quad (1.8)$$

Using the assumption that the innovations are independent, identically distributed and in particular $\mathcal{E}_t$ is independent of $\mathcal{F}_{t-1}$, we have that:

$$P_{|\mathcal{F}_{t-1}}\left(S_t \ \leq x\right) = P_{\mathcal{E}}\left(\mathcal{E}_t \ \leq \frac{x - m\left(S_{t-1}, S_{t-2}...., S_{t-\tau}, X_{t-1}\right)}{\sigma\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right)}\right). \quad (1.9)$$

Therefore by the definition of VaR_α^t and of the α quantile of the innovations, we have:

$$q_\alpha = \frac{VaR_\alpha^t - m\left(S_{t-1}, S_{t-2}...., S_{t-\tau}, X_{t-1}\right)}{\sigma\left(S_{t-1}, S_{t-2}...., S_{t-\tau}, X_{t-1}\right)}. \tag{1.10}$$

Consequently the Value-at-Risk is given by:

$$VaR_\alpha^t = m\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right) + q_\alpha * \sigma\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right).\diamondsuit$$

To estimate the Value-at-Risk, one can use a method based on the theory of Artificial Neural Networks for estimating the conditional expectation m and the stochastic volatility of the portfolio changes σ (see H.White [1990]). The estimation of the quantile q_α will be done by using Extreme Value Theory, more exactly by the approximation of the tail of probability distribution developed by L.Smith [1987]. This leads to an invertible form of the distribution function of the innovations that help to get easily the estimator of the required quantile with appealing asymptotic properties.
In the same manner, the relationship between the expected shortfall, the volatility, the expected return and the innovations is given by the following formula:

$$ES_\alpha^t = m\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right) + \sigma\left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right) * E\left(\mathcal{E}_t \,|\, \mathcal{E}_t \,>\, q_\alpha\right) \tag{1.11}$$

The expression $E\left(\mathcal{E}_t \,|\, \mathcal{E}_t \,>\, q_\alpha\right)$ written in the following form:

$$E\left(\mathcal{E}_t \,|\, \mathcal{E}_t \,>\, q_\alpha\right) = E\left(\mathcal{E}_t - q_\alpha \,|\, \mathcal{E}_t \,>\, q_\alpha\right) \, + q_\alpha \tag{1.12}$$

becomes a quite instructive expression, as the first term of its right hand side represents the mean excess function of the innovations evaluated at q_α. Some nice results presented in Embrechts[1997] or Alexander Mc.Neil[1999] enable to estimate $E\left(\mathcal{E}_t - q_\alpha \,|\, \mathcal{E}_t \,>\, q_\alpha\right)$ using Maximum Likelihood Estimators for appropriate classes of distributions.

1.3 Non-parametric Regression Analysis Using Artificial Neural Network

Non-parametric artificial neural network estimators are generally Sieve estimators (see: Grenader [1981], White [1981], White-Hornik- Stinchcombe [1984], Chen and Schen [1996] or Shen [1997]). White proved the existence of consistent and asymptotically normal ANN estimators with an initial convergence rate of order $o(\frac{n}{logn}^{-\frac{1}{4}})$ for sigmoid activation function with independent data or with time series having some mixing properties. The convergence rate has been progressively shaped by Barron [1994], Modha and Masry [1996], by Shen and Chen [1996] and by Xiahong Chen and White [1997]. ANN represent one the most powerful tools for non-parametric regression analysis due to their flexibility and high capacity of approximating unknown functions assisted by the increasing computational power of new computers and numerical optimisation software.

An ANN model is a computerized processing method for analysing data based on historical information by mimicking the human brain's ability of classifying patterns or making decisions based on past experiences. A neural network is a system composed of many simple processing elements operating in parallel, whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.

Feed-forward networks with a single hidden layer are statistically consistent estimators of arbitrary square integrable regression functions under certain practically-satisfiable assumptions regarding sampling, target noise, number of hidden units, size of weights, and form of hidden-unit activation function (White [1990]). Such networks can also be trained as statistically consistent estimators of derivatives of regression functions (White and Gallant [1992]) and quantiles of the conditional noise distribution (White [1992a]). Feed-forward ANN with a single hidden layer using threshold or sigmoid activation functions can be equated as universal approximators.

The mathematical description of such a model is generally presented as a triplet (Ψ, H, W) and can be represented as in picture1:

(Insert Page)

where the input variable $x = (x_1, x_2, ..., x_r)^{'} \in \Re^r$ sends signals of intensity γ_{ij} to the hidden layer (processing unit) that provides via β_i and the activation function Ψ the so called artificial neural network output function $f_H(x, w)$ defined by:

$$f_H(x, w) := \beta_0 + \sum_{j=1}^{H} \beta_j \Psi(\tilde{x}^{'}.\gamma_j) \tag{1.13}$$

Where

$$\begin{cases} \bullet \quad H \in \mathcal{N} \text{ is the number of neurons in the hidden layer, i.e the network complexity.} \\ \bullet \quad \tilde{x}^{'} \in \Re^{r+1} \text{denotes the Input Variable } (1, x^{'})^{'} \text{ augmented by a constant.} \\ \bullet \quad w := (\beta_0, \beta_1, ..., \beta_r; \gamma_{ij}), \ \ j = 1, 2, ..., H \text{ and } i = 0, 1, , ..., r \text{ are the Network Weights .} \\ \bullet \quad \Psi \text{ is called the learning or activation function of the Network.} \\ \bullet \quad \gamma_{ji}, \text{ signal intensity from input node } i \text{ to hidden node } j \text{ and } \gamma_j = (\gamma_{j0}, \gamma_{j1}, ..., \gamma_{jr})^{'}. \\ \bullet \quad \beta_j \text{ signal strength from hidden node j to the output.} \end{cases}$$

1.3.1 Denseness Properties of Artificial Neural Network Output Functions

1.3.1.1 Neural Activation Function

A real valued measurable function Ψ is called a Neural Activation Function (or a squashing function in Neural Network parlance) if the following properties are fulfilled:

$$\begin{cases} \lim_{x \to +\infty} \Psi(x) = \Psi(+\infty) \ < +\infty \\ \\ \lim_{x \to -\infty} \Psi(x) = \Psi(-\infty) \ > -\infty \\ \\ \Psi \text{ is monotonically increasing.} \end{cases} \tag{1.14}$$

Later on, we will add some extra requirements such as Lipschitz continuity or l-finiteness.

Examples of Activation Function

The Squasher Tangent Hyperbolic Function:

$$\Psi(\lambda) = \frac{1}{2}\left[1 + \tanh(x)\right] \tag{1.15}$$

Exponential, Antisymmetric Sigmoid:

$$\Psi(u) = \frac{2}{1 + \exp(-u)} - 1 \tag{1.16}$$

1.3.1.2 Definition: Uniform Denseness on Compacta

Let $(X, || \ ||_X)$ be a complete metric space of real valued continuous functions defined on $\Re^n$. A subset S of X is said to be uniformly dense on Compacta in X if and only if:
For all compact subset K of $\Re^n, \forall f \in X, \forall \epsilon > 0, \exists g \in S$ such that

$$\sup_{x \in K} | \, f(x) - g(x) \, | \ < \ \epsilon. \tag{1.17}$$

A sequence of scalar continuous function $(f_n) \subset X$ converges uniformly on Compacta to f if:
For all compact K subset of $\Re^n$,

$$\lim_{n \to +\infty} \left[\sup_{x \in K} |f_n(x) - f(x)| \right] \ = 0. \tag{1.18}$$

The following denseness results have been proved by White,Hornik and Stinchcombe [1989].

1.3.2 Theorem: ANN as Universal Approximators

Given a Lipschitz continuous activation function Ψ, the set of Artificial Neural Network Output Functions with one hidden layer defined by:

$$ANN\,(\Psi) \ := \left\{ \ f \in \mathcal{C}^N \text{ such that } f(x) = \beta_0^f + \sum_{j=1}^{H_f} \beta_j^f \, \Psi(\tilde{x}.\gamma_j^f) \ \right\} \tag{1.19}$$

is uniformly dense on Compacta in $\mathcal{C}^N$, the set of real valued continuous function on $\Re^N$, and dense in the whole set of measurable function $\mathcal{M}^N$ in the following manner:
Given a probability measure μ on the Borel space $(\Re^N, \mathcal{B}^N)$,
$\forall f \in \mathcal{M}^N$, there exists a sequence of neural output functions $(f_n)_{n \in \mathcal{N}}$ such that:

$$\begin{cases} (f_n)_{n \in \mathcal{N}} \ \subset ANN(\Psi) \\ \\ ||f - f_n||_\mu := \inf \{\epsilon > 0 \ \text{ such that } \mu \left(\{x : |f(x) - f_n(x)| > \epsilon\} \right) \leq \epsilon \} \to \ 0. \end{cases} \tag{1.20}$$

White [1981] applied this denseness results and proved the existence of consistent and asymptotically normal estimators for conditional expectation for bounded stochastic processes. To estimate the conditional expectations for heavy tailed distributions, the denseness of $ANN(\Psi)$ in the set of continuous functions with respect to the supremum norm on compacts sets is too restrictive. This assumption is not plausible for financial time series as one of the well known stylized facts implies that heavy tailedness which strongly contradicts any boundedness. White's approximation (1.20) is too weak and technically complicated when dealing with unbounded stochastic processes. Therefore, after presenting White's results, we will prove a new approximation and denseness result which uses a $L^m(\mu)$ metric for some $m \geq 1$ and some probability measure and which holds for a certain subclass of square integrable functions. The main idea is to use the Fourier transform in a similar manner as Barron [1993]. We first prove two auxiliary results, which, together, almost immediately provide the desired result.

1.3.3 Extensions of White's Neural Network Denseness Results

1.3.3.1 Lemma:

Given an activation function Ψ satisfying (1.14). Let $ANN(\Psi)$ be defined by (1.19). Let μ be an absolute continuous probability measure on $\Re^r$ and $m \in \mathcal{N}$. For any $w \in \Re^r,\ b \in \Re$, the function

$$g(x) = cos(w^T x\ +\ b), \quad \forall x \in \Re^r, \tag{1.21}$$

may be arbitrary well approximated by functions in $ANN(\Psi)$ in the following sense:
For any $\epsilon\ >\ 0$, there exists a function $f\ \in\ ANN(\Psi)$ such that

$$\int_{\Re^r} |g(x) - f(x)|^m \, dx \ \leq\ \epsilon. \tag{1.22}$$

(1.22) states that $ANN(\Psi)$ is dense the set $\mathcal{G}$ consisting of all functions

$$g_{(w,b)}(x) := cos(w^T x\ +\ b)$$

with respect to the norm $||\ ||_m$ defined by:

$$||x||_m = \int_{\Re^r} ||x||^m \mu(dx) \tag{1.23}$$

Proof
Let $w = (w_1, w_2, ..., w_r)^T \in \Re^r$. If $w_k = 0$, then $g(x)$ does not depend on x_k, and is essentially a function of $\Re^{r-1}$ only. This means that one can transform a network function f on $\Re^{r-1}$ to one on $\Re^r$, which does not depend on x_k, by augmenting connections of the k^{th} input to all neurons in the hidden layer and setting the weights to 0. Therefore, the approximation problem in $\Re^r$ with $w_k = 0$

is equivalent to the approximation problem in $\Re^{r-1}$. Therefore without loss of generality, one can assume that

$$\forall k = 1, 2, ..., r \quad w_k \neq 0. \tag{1.24}$$

It suffices to consider the case $w_1, ..., w_r > 0$. If an approximating network function for $|w_1|, |w_2|, ..., |w_r|$ is available, then one can derive the same result for arbitrary $w_1, ..., w_d \in \Re^d$ by multiplying the weights of the input x_k with $sign(x_k)$ for $k = 1, 2, .., r$ where

$$sign(x_k) = \begin{cases} 1 \ \text{ if } x_k > 0 \\ \\ -1 \ \text{ if } x_k < 0 \end{cases}$$

i) As a first step, we show that $g(x)$ may be approximated by a linear combinations of indicator functions of finite interval, i.e. by functions in the set:

$$L_{ind} = \{ \ f \quad ; \quad f(x) = \sum_{j=1}^{H} \beta_j 1_{[u_j\,,\,v_j]} \left(w^T x \,+\, b \right) ;$$
$$-\infty < u_j < v_j < \infty, \ \ |v_j - u_j| = \delta, \ \ \delta > 0, \forall j. \tag{1.25}$$

The interval $(-\pi\,,\,\pi)$ is partioned into intervals $[z_{j-1}\,,\,z_j]\,, j = 1, 2, .., N$, with

$$-\pi = z_0 < z_1 < ... < z_N = \pi \ \text{ and } \ |z_j - z_{j-1}| = \frac{2\pi}{N}. \tag{1.26}$$

We set

$$C_0(z) = \sum_{j=1}^{H} cos(z_j) * 1_{[z_{j-1}\,,\,z_j]}(z). \tag{1.27}$$

As the derivative of $cos(z)$ is uniformly bounded by 1, by using the mean-value theorem, one can derive that:

$$|cos(z_j) - cos(z)| \ \leq \ |z_j - z| \ \leq \ |z_j - z_{j-1}| = \frac{2\pi}{N} \quad \text{for } z_{j-1} \leq \ z \ \leq \ z_j \tag{1.28}$$

and therefore,

$$|cos(z) - C_0(z)| \ \leq \ \frac{2\pi}{N} \qquad \forall - \pi \ < \ z \ < \ \pi. \tag{1.29}$$

By the periodicity of $cos(z)$, we have analogously that:

$$\begin{cases} C_k(z) := \sum_{j=1}^{H} cos(z_j) * 1_{[2k\pi + z_{j-1}\,, \ \ 2k\pi + z_j]}(z) \quad \text{for } -\infty < k < \infty \\ \\ |cos(z) - C_k(z)| \ \leq \ \frac{2\pi}{N} \qquad \forall (2k-1)\pi < z \leq \ (2k+1)\pi. \end{cases} \tag{1.30}$$

For any given integer $K \geq 1$, we consider

$$f(x) = \sum_{k=-K}^{k=K} C_k(w^T x \, + \, b) \, \in \, L_{ind} \tag{1.31}$$

and we have with $M = 2K + 1$,

$$|g(z) - f(z)| = |cos(w^T x \, + \, b) - f(x)| \, \leq \, \frac{2\pi}{N} \qquad \forall \, |w^T x \, + \, b| \, \leq M\pi. \tag{1.32}$$

Now we select M large enough such that

$$\mu\left(\left\{x; \; |w^T x \, + \, b| \, > \, M\pi\right\}\right) \leq \frac{\epsilon}{2} \tag{1.33}$$

and we get

$$\begin{aligned} \int_{\Re^r} |g(x) - f(x)|^m \mu(dx) \quad &= \quad \int_{|w^T x \, + \, b| \leq M\pi} |g - f|^m \mu(dx) \, + & (1.34) \\ & \quad\;\; \int_{|w^T x \, + \, b| \, > \, M\pi} |g - f|^m \mu(dx) & (1.35) \\ &\leq \quad \left(\frac{2\pi}{N}\right)^m \, + \, \frac{\epsilon}{2} \, \leq \, \epsilon & (1.36) \end{aligned}$$

if N is chosen large enough.
Here, we have used that:

$$\begin{cases} |g(z)| \, \leq \, 1 \\ f(z) = 0 \;\; \text{on} \;\; \left\{x; |w^T x \, + \, b| \, > \, M\pi\right\}. \end{cases} \tag{1.37}$$

ii) For sigmoid activation functions Ψ satisfying (1.14), we have

$$\Psi(cz) \rightarrow 1_{(0,\infty)}(z) \qquad \text{for } c \rightarrow \infty \; \forall z \neq 0. \tag{1.38}$$

Therefore, we have for arbitrary $-\infty \, < \, u \, < \, v \, < \, \infty, \qquad c \rightarrow \infty$

$$\left|1_{[u,v]}(z) \, - \, \{\Psi(c(z - u)) \, - \, \Psi(c(z - v))\}\right| \rightarrow 0 \;\; \text{for } z \neq u, v. \tag{1.39}$$

By the Lebesgue's theorem of dominated convergence, we get for $c \rightarrow 0$

$$\begin{aligned} \int \Big| 1_{[u,v]}(w^T x \, + \, b) \, - \, \Big\{ & \Psi(c(w^T x \, + \, b - u)) \, - \\ & \Psi(c(w^T x \, + \, b - v))\Big\}\Big|^m \mu(dx) \rightarrow 0, \end{aligned} \tag{1.40}$$

where we use that μ is absolutely continuous and, therefore,

$$\mu\left(\left\{x; \;\; w^T x \, + \, b \, \neq \, u, v\right\}\right) = 1. \tag{1.41}$$

Therefore any function of L_{ind} may be arbitrarily well approximated by a neural network output function in $ANN(\Psi)$.
Together with **i)**, this implies the assertion of the Lemma 1.3.3.1.$\diamondsuit$.

1.3.3.2 Lemma

Let $g \in L^2(\Re^r)$ be a square integrable real-valued function on $\Re^r$ with Fourier transform

$$\tilde{g}(w) = \frac{1}{(2\pi)^r} \int e^{-iw^T x} g(x) dx \; = \; |\tilde{g}(w)| e^{i\Phi(w)}. \tag{1.42}$$

Assume that $|\tilde{g}(w)|$ is bounded and integrable over $\Re^r$ and that $|\tilde{g}(w)|$, $\Phi(w)$ are Lipschitz continuous on any compact subset of $\Re^r$. Let μ be an arbitrary probability measure on $\Re^r$ for which the m^{th} moment exists:

$$\int ||x||^m \mu(dx) \; < \; \infty \text{ for some} \quad m \geq 1. \tag{1.43}$$

Then, g may be arbitrarily well approximated by functions of the form:

$$f(x) = \sum_{j=1}^{N} \beta_j cos(w_j^T x + \Phi_j) \qquad \beta_j, \Phi_j \in \Re, w_j \in \Re^r, \; j = 1, 2, .., N \tag{1.44}$$

in the sense that, for any $\epsilon > 0$, there exists a function f such that:

$$\int_{\Re^r} |g(x) - f(x)|^m \, \mu(dx) \leq \epsilon \tag{1.45}$$

(1.45) represents also a new approximation result stating that the function set consisting of the finite sums of functions g defined by (1.44) is dense in the set of bounded square integrable functions having a Fourier transform defined by (1.42).

Proof:

As g is real-valued, by the Fourier inversion, one can derive that:

$$\int_{\Re^r} e^{iw^T x} \tilde{g}(x) dx = \int_{\Re^r} cos(w^T x + \Phi(w)) |\tilde{g}(w)| dw. \tag{1.46}$$

As $\tilde{g}(w)$ is integrable and $cos(z)$ is bounded, there exists a positive real valued $M > 0$ and a hypercube I_M with side length M such that

$$\left| \int_{I_M^c} cos(w^T x + \Phi(w)) |\tilde{g}(w)| dw \right| \; \leq \; \int_{I_M^c} |\tilde{g}(w)| dw \leq \frac{\epsilon}{2}. \tag{1.47}$$

Let make a partition of each side of I_M in n subintervals of length $\frac{M}{n}$. This corresponds to a partition of I_M into $N = n^r$ small hypercubes $i_1, i_2, ..., i_N$. $\forall j = 1, 2, ..., N, \; \forall w \in i_j, \forall w_j \in i_j$, a straightforward calculation shows that:

$$||w - w_j|| \; \leq \; \sqrt{r} * \frac{M}{n}. \tag{1.48}$$

As $\Phi(w)$, $|\tilde{g}(w)|$ are Lipschitz continuous on I_M with constants, say, L_Φ, L_g, and as the cosine function is uniformly Lipschitz on $\Re$ with constant 1, we have:
$\forall j = 1, 2, ..., N \quad \forall w \in i_j$,

$$\begin{aligned} L(g, \Phi) : &= \left| cos(w_j^T x + \Phi(w_j)) |\tilde{g}(w_j)| \; - \; cos(w^T x + \Phi(w)) |\tilde{g}(w)| \right| && (1.49) \\ &\leq |\tilde{g}(w_j) - \tilde{g}(w)| \; + \\ &\quad \left| cos(w_j^T x + \Phi(w_j)) - cos(w^T x + \Phi(w)) \right| |\tilde{g}(w)| && (1.50) \\ &\leq L_g ||w_j - w|| \; + \; (||w_j - w|| ||x|| \; + \; L_\Phi ||w_j - w||) \, C_g && (1.51) \\ &= (L_g \; + \; C_g . L_\Phi + C_g ||x||) \, ||w_j - w|| && (1.52) \\ &\leq (C_L + C_g ||x||) \sqrt{r} * \frac{M}{n}. && (1.53) \end{aligned}$$

Here C_g is an upper bound for $|\tilde{g}(w)|$ and $C_L = L_g + C_g L_\Phi$.
Using the approximation of integrals by the corresponding Riemann sums and setting

$$\begin{cases} \beta_j = |\tilde{g}(w_j)| \\ \\ \Phi_j = \Phi(w_j) \quad \forall j = 1, 2, ..., N. \end{cases} \qquad (1.54)$$

it derived that

$$\begin{aligned} S : &= \left| \int_{I_M} cos(w^T x + \Phi(w)) |\tilde{g}(w)| dw \; - \; \sum_{j=1}^{N} \beta_j cos(w_j^T x + \Phi_j * \frac{M^r}{N} \right| && (1.55) \\ &\leq \sum_{j=1}^{N} \int_{i_j} \left| cos(w^T x + \Phi(w)) |\tilde{g}(w)| \; - \; \beta_j cos(w_j^T x + \Phi_j) \right| dw && (1.56) \\ &\leq \sum_{j=1}^{N} (C_L + C_g * ||x||) * \sqrt{r} \frac{M}{N} * \int_{i_j} dw && (1.57) \\ &= (C_L + C_g * ||x||) * \sqrt{r} \frac{M^{r+1}}{n}. && (1.58) \end{aligned}$$

Setting

$$f(x) = \sum_{j=1}^{N} \beta_j cos(w_j^T x + \Phi_j) * \frac{M^r}{N}, \qquad (1.59)$$

we derive by combining (1.45) and (1.57) that:

$$||g(x) \ - \ f(x)||_{L^m(\mu)} \ \leq \ \frac{\epsilon}{2} + (C_L + C_g.||x||)\sqrt{d}\frac{M^{d+1}}{n}. \qquad (1.60)$$

Then, as $\int ||x||^m \mu(dx) < \infty$, we can achieve (1.45) by choosing n large enough.$\diamondsuit$.

Both Lemmas together imply that the class of neural networks output functions $ANN(\Psi)$ is dense with respect to $L^m(\mu)$ norm in the class $\mathcal{G}$ of functions defined in the following manner.

1.3.3.3 Definition

Let $\mathcal{G} \subset L^2(\Re^r)$ be the class of real-valued functions of $g(x)$ with Fourier transform

$$\tilde{g}(w) = |\tilde{g}(w)|e^{i\Phi(w)} \qquad (1.61)$$

satisfying

$$\begin{cases} a) \ \tilde{g}(w) \ \text{ is integrable over } \Re^r. \\ \\ b) \ |\tilde{g}(w)|\,, \ \Phi(w) \ \text{ are Lipschitz continuous on any compact subset of } \ \Re^r. \end{cases}$$

$\mathcal{G}$ includes the Schwartz space of infinitely often continuously differentiable functions which decrease rapidly in the sense that all derivatives are converging faster to 0 for $||x|| \to \infty$ than $||x||^{-p}$ for all $p \geq 1$. This follows from the fact that the Fourier transform is a bijection on that space (see Theorem 10.3 of Weidmann, 1976). Barron [1993] has studied a similar, but more restrictive function class in relation to neural networks where he did not only show a universal approximation property but derived also rates for the approximation depending on the size of the network.

1.3.3.4 Theorem: Uniform Approximation Property of ANN in $L^m(\mu)$-sense

Let Ψ be a sigmoid function satisfying(1.14). Let μ be any absolutely continuous probability measure on $\Re^r$ satisfying:

$$\int_{\Re^d} ||x||^m \mu(dx) < \infty \quad \text{for some } m \geq 1. \qquad (1.62)$$

Then, $ANN(\Psi)$ is dense in the function class $\mathcal{G}$ in the $L^m(\mu)$-sense, i.e. $\forall g \in \mathcal{G}, \forall\ \epsilon > 0, \exists f \in ANN(\Psi)$ such that

$$\int_{\Re^d} |f(x) - g(x)|^m \mu(dx) < \mathcal{E}. \tag{1.63}$$

Proof: First, we remark that, for $\forall g \in \mathcal{G}$,

$$|g(x)| = \left| \int_{\Re^r} e^{-iw^T x} \tilde{g}(w) dw \right| \leq \int |\tilde{g}(w)| dw, \tag{1.64}$$

i.e. g is uniformly bounded and, therefore, in $L^m(\mu)$. As functions in $ANN(\Psi)$ are also bounded, they are in $L^m(\mu)$ too.
By Lemma 1.3.3.2, g may be approximated by a function of the form

$$\sum_j \beta_j cos(w_j^T x + \Phi_j).$$

By Lemma 1.3.3.1, each cosine function $cos(w_j^T x + \Phi_j)$ may be approximated by a function of the form

$$\sum_k b_{jk} \Psi(\gamma_{jk}^T x + d_{jk}).$$

Therefore combining both results, $g(x)$ may be approximated by

$$\sum_{j,k} \beta_{jk} \Psi(\gamma_{jk}^T x + d_{jk}) \quad \in \quad ANN(\Psi) \Diamond \tag{1.65}$$

1.4 Extension of White's Results to Unbounded Stochastic Processes: ANN Estimates of Conditional Expectations

In this section we discuss the consistency of neural network estimators for conditional means and volatility. First, we refer White's results for bounded stochastic processes. Then, extend them to unbounded random stochastic processes in order to be able to cover the case of financial time series where boundedness assumptions would be questionable and counterintuitive.
Extending White's results was not too easy and we had to prove also several auxiliary new results, among them a sort of Bernstein inequality for unbounded stochastic processes. To avoid the flow of argument we put these results in an own section at the end of this chapter.

1.4.1 White's ANN Estimates of Conditional Expectations

Let Ψ a l-finite activation function as defined in (1.14) satisfying

$$\int \left|\mathcal{D}^l \Psi(x)\right| dx \ < \ +\infty \tag{1.66}$$

and two increasing and unbounded sequences q_n and Δ_n such that:

$$\begin{cases} 1°)\ (q_n) \subset \mathcal{N}, \\ \\ 2°)\ (\Delta_n) \subset \Re^+, \\ \\ 3°)\ \Delta_n = o(n^{\frac{1}{4}}) \text{ e.g } \lim\limits_{n\to+\infty} \dfrac{\Delta_n}{n^{\frac{1}{4}}} = 0, \\ \\ 4°) q_n {\Delta_n}^2 \log(q_n\Delta_n) = o(n^{\frac{1}{2}}) \end{cases} \tag{1.67}$$

then

$$\cup_{n=1}^{+\infty} ANN(\Psi, q_n, \Delta_n) \text{ is dense in } \mathcal{M}^r. \tag{1.68}$$

where, as above $\mathcal{M}^r$ represents the set of real valued measurable functions defined on $\Re^r$ and $ANN\left(\Psi, q_n, \Delta_n\right)$, also called the connectionist sieve, is defined by:

$$ANN\left(\Psi, q_n, \Delta_n\right) := \left\{ f;\ \text{s.t } f(x) = \beta_0^f + \sum_{j=1}^{q_n} \beta_j^f \psi(\tilde{x}'.\gamma_j^f) \text{ satisfies (1.70)} \right\} \tag{1.69}$$

with (1.70) defined by the following growth conditions:

$$\begin{cases} \sum\limits_{j=0}^{q_n} |\beta_j^f| \le \Delta_n, \\ \\ \sum\limits_{j,i} |\gamma_{ji}^f| \le q_n\Delta_n. \end{cases} \tag{1.70}$$

To introduce the estimation of the conditional autoregression function m of the model (1.1), let $\mathcal{F}_t$ be the σ-algebra generated by

$$\{Z_s \text{ for } \ s \le t\}$$

where

$$Z_s := (S_s, ..., S_{s-\tau+1}, X_s) \tag{1.71}$$

is representing the most recent returns or portfolio changes combined with some latest market information available at the current trading period. Based on these information, one can estimate the expected conditional return

$$m(Z_{t-1}) = \mathcal{E}\left(S_t|\mathcal{F}_{t-1}\right). \tag{1.72}$$

The following theorem states that connectionist sieves provides consistent estimators of expected conditional returns. For sake of illustration, we consider the case where (S_t, Z_{t-1}) are arbitrary random variables satisfying a model like(1.1).

1.4.1.1 Theorem: H.White's Conditional Mean Approximation for Bounded Stochastic Processes

Let $(S_t, Z_{t-1})_{t\in\mathcal{Z}}$ be some bounded random variables with $S_t \in \Re, Z_{t-1} \in \Re^s$ satisfying

$$S_t = m(Z_{t-1}) + \sigma(Z_{t-1})\epsilon_t \tag{1.73}$$

where ϵ_t are independent and identically distributed.
Given Ψ , q_n and Δ_n satisfying (1.14), (1.66), (1.67),and (1.70).
Consider $\hat{\theta}_n$, any optimal solution of the following minimization problem:

$$\min_{\theta\in ANN\,(\Psi,q_n,\Delta_n)} \frac{1}{n}\sum_{t=1}^{n}\ \left[S_t - \theta\left(Z_{t-1}\right)\right]^2 \tag{1.74}$$

a) When $(Z_t)_{t\in\mathcal{Z}}$ is a sequence of independent and identically distributed bounded random variables and

$$q_n{\Delta_n}^4\log(q_n\Delta_n) = o(n) \tag{1.75}$$

then $\hat{\theta}_n$ is a consistent estimator of the expected return $m := E(S_t|\mathcal{F}_{t-1})$.
b) If we assume that $(Z_t)_{t\in\mathcal{Z}}$ is a bounded, stationary and strongly mixing process with mixing coefficients $\alpha(k)$ satisfying:

$$\alpha(k) = \alpha_0{\rho_0}^k \tag{1.76}$$

for some constant α_0 and $0 < \rho_0 < 1$; and let

$$q_n{\Delta_n}^2\log(q_n\Delta_n) = o(n^{\frac{1}{2}}) \tag{1.77}$$

then the conditional expectation can also be estimated consistently by $\hat{\theta}_n$.
c) In both of the precedent cases, up to some regularity conditions, $\hat{\theta}_n$ is generally asymptotically normally distributed if q_n is kept fixed.

White's result becomes no longer applicable when dealing with unbounded stochastic processes. Therefore, later in this chapter we are going to extend such an approximation result to heavy tailed, unbounded stochastic processes in order to derive a neural nerwork estimates for conditional expectation.

Proof

The theorem may be derived from general results in the literature. For later reference, we present the details here.

Before starting the proof, let us recall briefly the concept of mixing processes.

1.4.1.2 Definition: Mixing Processes

A stochastic process $(Z_t)_{t\in\mathcal{Z}}$ is said to be a mixing process when $(Z_t)_{t\in\mathcal{Z}}$ exhibits considerable short-run dependence but displays a form of asymptotic independence, in that events involving elements of $(Z_t)_{t\in\mathcal{Z}}$ separated by increasingly greater time intervals are increasingly closer to independence. There are two common types of mixing processes:

$(Z_t)_{t\in\mathcal{Z}}$ is ϕ-mixing or uniformly mixing if:

$$\lim_{k\to+\infty} \phi(k) = \sup_t \sup_{\left\{A\in\mathcal{F}_1^t,\, B\in\mathcal{F}_{t+k}^{+\infty}\,:\,P(A)\,>0\right\}} [P(B|A)\,-\,P(B)] = 0 \tag{1.78}$$

with

$$\begin{cases} \mathcal{F}_1^t := \sigma(Z_1, Z_2, ..., Z_t) \\ \\ \mathcal{F}_{t+k}^{+\infty} := \sigma(Z_{t+k}, Z_{t+k+1}, ...). \end{cases} \tag{1.79}$$

$(Z_t)_{t\in\mathcal{Z}}$ is α-mixing or strongly mixing if:

$$\lim_{k\to+\infty} \alpha(k) := \sup_t \sup_{\left\{A\in\mathcal{F}_1^t,\, B\in\mathcal{F}_{t+k}^{+\infty}\right\}} \mid P(B\cap A)\,-\,P(A)P(B)\mid =\, 0. \tag{1.80}$$

ϕ-mixing is the stronger assumption and implies α-mixing. Therefore, we mainly consider α-mixing processes.

Now, let us begin with the proof of the theorem.

In fact the proof is essentially based on the theorems 4.1 and 4.2 in White[1990], pages 182,183. To apply theorem 4.1, set:

$$(\Theta, \rho) := \left(L^2(I^r, \mu),\, \rho_2 := ||\;\;||_{L^2}\right). \tag{1.81}$$

Where I^r represents the bounded support of the stochastic process Z_t. We have that (Θ, ρ) is a complete separable metric space (e.g. Kolmogorov and Fomin,

1970,theorem 37.5,problem 37.4).
Define

$$\hat{\Theta}_n := \left\{(\beta_0^f, \beta_1^f, ..., \beta_{q_n}^f, \gamma_1^f, \gamma_1^f, ..., \gamma_{q_n}^f) \text{ satisfying } (1.70)\right\} \subset \Re^{M_n}. \quad (1.82)$$

as the set of parameters defining uniquely the functions of $ANN(\Psi, q_n, \Delta_n)$ with

$$M_n = 1 + q_n + (1 + r)q_n. \quad (1.83)$$

The set of parameter vectors is a nonempty, closed and bounded subset of the finite dimensional space $\Re^{M_n}$ and therefore compact. $\hat{\Theta}_n$ is the image of the compact set of parameters under a continuous mapping, and therefore compact too. $ANN(\Psi, q_n, \Delta_n)$ is a nonempty, compact set and $\cup_{n=1}^{+\infty} ANN(\Psi, q_n, \Delta_n)$ is dense in M^r using the theorem 1.4.1.
Consider

$$Q_n(w, \theta) := \frac{1}{n} \sum_{t=1}^{n} (S_t(w) - \theta(Z_{t-1}(w)))^2. \quad (1.84)$$

As a consequence of Lemma 2.2 of Stinchcombe and White [1989,b], $Q_n(w, \theta)$ is measurable, because for every $w \in \Omega$, $Q_n(w, .)$ is continuous and $\forall \theta$ in the separable metric space Θ, $Q_n(., \theta)$ is measurable. The continuity of $Q_n(w, .)$ implies lower semi continuity. The existence of $\hat{\theta}_n$ follows from theorem 4.1(a) of White[1990]. For sake of simplicity, we write θ for the parameter vector in $\hat{\Theta}_n$ as well as for the corresponding function in $ANN(\Psi, q_n, \Delta_n)$.
Based on the network growth complexity assumptions and the regularity conditions imposed on the activation function we derive also that:

$$\cup_{n=1}^{+\infty} ANN(\Psi, q_n, \Delta_n) \text{ is dense in } L^2(I^r, \mu), .$$

Setting for arbitrary function θ, the function $\bar{Q}$ defined below is well defined:

$$\bar{Q}(\theta) := E\left([S_t - \theta(Z_{t-1})]^2\right), \quad (1.85)$$

using Lemma 4.3, 4.4 and applying 4.2 in White[1990], page 183, we derive that: $\forall \epsilon > 0$,.

$$\mathcal{P}\left(\left\{w \in \Omega \ : \ \sup_{\theta \in ANN(...)} \left|Q_n(w, \theta) - \bar{Q}(\theta)\right| \ > \epsilon\right\}\right) \to 0. \quad (1.86)$$

Using the characterization of the conditional expectation as being the measurable function of $\theta(Z_t)$ that minimizes the mean square error, we get that:

$$\bar{Q}(\theta) - \bar{Q}(m) = E\left(\theta(Z_{t-1}) - m(Z_{t-1})\right)^2 = \rho(\theta, m)^2. \quad (1.87)$$

Therefore with, $B(\theta, \epsilon)$ denoting the ϵ-ball around θ,

$$\inf_{\theta \in B(\theta,\epsilon)^c} \bar{Q}(\theta) - \bar{Q}(m) = \inf_{\theta \in B(\theta,\epsilon)^c} \rho(\theta, m)^2 \geq \epsilon^2 > 0 \tag{1.88}$$

and

$$\bar{Q}(\theta) - \bar{Q}(m) = E\left(\theta(Z_{t-1}) - m(Z_{t-1})\right)^2 = \rho(\theta, m)^2 + \bar{Q}(m) \tag{1.89}$$

is continuous at m, hence using Theorem 4.1(b) of White[1990], we derive that:

$$\rho(\hat{\theta}_n, m) \to 0 \text{ in probability} \tag{1.90}$$

which means that the conditional expectation $m(Z_t)$ can be estimated consistently by $\hat{\theta}_n(Z_{t-1})$.$\diamondsuit$
Therefore, the conditional expectation function

$$m(z) = \mathcal{E}\left(S_t \mid Z_{t-1} = z\right) \tag{1.91}$$

can be estimated consistently by the network output function $\hat{\theta}_n(z)$ defined by the weights $\hat{w}_n := \left(\hat{\beta}^f, \hat{\gamma}^f\right)$.
One has still to discuss how to determine numerically $\hat{\theta}_n(z)$ as the solution of the following constrained and global optimization problem:

$$\min_{w=(\beta,\gamma)\in\hat{\Theta}_n} \frac{1}{n} \sum_{t=1}^{n} \left(S_t - \beta_0^f - \sum_{j=1}^{q_n} \beta_j^f \psi(\tilde{x}_t^{'}.\gamma_j^f)\right)^2 \tag{1.92}$$

with

$$\tilde{x_t} := (1, Z_{t-1}). \tag{1.93}$$

$\hat{\Theta}_n$ is defined by the following constraints:

$$\sum_{j=0}^{q_n} |\beta_j^f| \leq \Delta_n \text{ and } \sum_{i,j} |\gamma_{ij}^f| \leq q_n \Delta_n. \tag{1.94}$$

To find an efficient and accurate estimator of the optimal weights with sufficiently appealing asymptotic properties we use the following stochastic approximation results.
In the following, let M be the dimension of $\hat{w}_n$. We write shorthand

$$Y_t := (S_t, Z_{t-1})' \in \Re^{\mu} \text{ with } \mu = r + 1 \tag{1.95}$$

and we consider an arbitrary loss function $l(Y_t, w)$ instead of the square loss as defined in (1.92). We replace the constraint $\theta \in \hat{\Theta}_n$, with $w \in W$, where W is a compact subset of $\Re^M$. Note that, the domain of minimization no longer depends on the sample size n. The new setting consists of taking a network of given size and study the behavior of $\hat{\theta}_n$ as $n \to +\infty$.

1.4.1.3 Theorem: Stochastic Approximation of Optimal Minima

Assume there exists $l: \ \ R^{\mu} * R^{M} \rightarrow R$ a continuously differentiable function that can be interpreted as a penalty function e.g the one that measures the accuracy of the estimation, and W be the connectionist sieve that can also be equated as a compact subset of R^{M} defined by the network growth complexity inequalities. If there exists an integrable function d, that dominates l e.g:

$$\begin{cases} 1^{\circ}) \ \ |l(z,w)| \ \leq d(z) \quad \forall w \in W, \, z \in \Re^{\mu} \\ \\ 2^{\circ}) \ \ E\left(d(Y_t)\right) \ < \ +\infty, \end{cases} \tag{1.96}$$

then, for each n=1, 2, 3,....there exists an optimal solution $\hat{w}_n$ to the problem

$$\min_{w \in W} \hat{\lambda}_n(w) := \frac{1}{n}\sum_{t=1}^{n} l(Y_t, w) \tag{1.97}$$

and

$$\hat{w}_n \ \rightarrow \ W^* \ \ a.s, \tag{1.98}$$

where

$$\begin{cases} W^* := \{w^* \ \text{ such that } \ \lambda(w) \ \geq \ \lambda(w^*) \ \}\,, \\ \\ \lambda(w) = E\left(l(Y_t, w)\right) \quad \text{is called the expected penalty }, \\ \\ \hat{w}_n \ \rightarrow \ W^* \text{ means that } \inf_{w^* \in W^*} ||\hat{w}_n - w^*|| \rightarrow 0 \quad \text{as } \ n \rightarrow +\infty. \end{cases} \tag{1.99}$$

Proof
The existence of $\hat{w}_n$, is justified by the compactness of the weight subset W and the continuity of the objective function $\hat{\lambda}_n$ which follows from the continuity of l. For independent and identically distributed random variables, it follows from Theorem2 of Jennrich [1969] which is usually interpreted as the uniform law of large numbers that:

$$\sup_{w \in W} \left|\hat{\lambda}_n(w) - \lambda(w)\right| \rightarrow 0 \ a.s.\diamondsuit \tag{1.100}$$

which also proves the consistency of $\hat{\lambda}_n$ toward λ with respect to the supremum norm. To derive the same convergence result for mixing processes, one can also use the result of Stout[1994].
Therefore $\hat{\lambda}_n$ converges uniformly on W towards λ.

Let $\hat{w}_n$, be a sequence of minimizers of $\hat{\lambda}_n$. Because W is compact, there exists a limit point w^0, a subsequence $\hat{w}_{n_k}$ such that:

$$\hat{w}_{n_k} \to w^0 \tag{1.101}$$

Claim: w^0 belongs to W^* and

$$\hat{\lambda}_{n_k}(\hat{w}_{n_k}) \to \lambda(w^0) \text{ as } k \to +\infty. \tag{1.102}$$

The claim can be established in the following manner:
It follows from the triangle inequality that:

$$\left|\hat{\lambda}_{n_k}(\hat{w}_{n_k}) - \lambda(w^0)\right| \leq \left|\hat{\lambda}_{n_k}(\hat{w}_{n_k}) - \lambda(\hat{w}_{n_k})\right| + \left|\lambda(\hat{w}_{n_k}) - \lambda(w^0)\right| < 2\epsilon \tag{1.103}$$

for any positive real number ϵ and sufficiently large natural number n_k, given the the uniform convergence and the continuity already established.
Now for arbitrary $w \in W$

$$\lambda(w^0) - \lambda(w) = \left[\lambda(w^0) - \hat{\lambda}_{n_k}(\hat{w}_{n_k})\right] + \left[\hat{\lambda}_{n_k}(\hat{w}_{n_k}) - \hat{\lambda}_{n_k}(w)\right] + \left[\hat{\lambda}_{n_k}(w) - \lambda(w)\right] \leq 3\epsilon,$$

for any $\epsilon > 0$ and all sufficiently large n_k, because

$$|\lambda(w^0) - \hat{\lambda}_{n_k}(\hat{w}_{n_k})| \leq 2\epsilon \tag{1.104}$$

as just established and $\hat{\lambda}_{n_k}(\hat{w}_{n_k}) - \hat{\lambda}_{n_k}(w) \leq 0$ by the definition of $\hat{w}_{n_k}$, and $\hat{\lambda}_{n_k}(w) - \lambda(w) < \epsilon$ by the uniform convergence.
Because ϵ is arbitrary,

$$\lambda(w_0) \leq \lambda(w) \text{ and } w^0 \in W^*.$$

Now suppose that

$$\inf_{w^* \in W^*} ||\hat{w}_n - w^*|| \not\to 0. \tag{1.105}$$

Then there exists $\epsilon > 0$ and a subsequence $(n_k)_{k \in \mathcal{N}}$ such that

$$||\hat{w}_{n_k} - w^*|| \geq \epsilon \quad \forall n_k \text{ and } w^* \in W^*. \tag{1.106}$$

But $\hat{w}_{n_k}$ has a limit point that, by the preceding argument, must belong to W^*. This is a contradiction to $||\hat{w}_{n_k} - w^*|| \geq \epsilon \ \forall n_k$, so

$$\inf_{w^* \in W^*} ||\hat{w}_n - w^*|| \to 0.\Diamond \tag{1.107}$$

The following result also proved by White and Gallant[1988], provides the asymptotic behaviour of the estimator $\hat{w}_n$. A simpler proof is also given by Franke and Neumann[1998], with some stronger assumptions.

1.4.1.4 Theorem: Asymptotic Properties of $\hat{w}_n$

Let $(\Omega, \mathcal{F}, \mathcal{P})$, (Y_t), W and l as previously defined in Theorem 1.4.1.3, and suppose that

$$\hat{w}_n \to w^* \text{ with probability one,} \tag{1.108}$$

where w^* is an isolated element of W^{int}; the interior of W.
Suppose in addition that for each $z' \in R^\mu$, $l(z, .)$ is twice continuously differentiable, such that

$$E\left([\nabla l(Y_t, w^*)]' \, [\nabla l(Y_t, w^*)] \right) \; < \; +\infty; \tag{1.109}$$

and each element of $\nabla^2 l$ is dominated on W by an integrable function ; and that $A^* := E(\nabla l^2(Y_t, w^*))$ and $B^* := E\left(\nabla l(Y_t, w^*)[\nabla(Y_t, w^*)]'\right)$ are nonsingular, where ∇ and ∇^2 represent the gradient and the Hessian matrices with respect to the weight vector w.
Then

$$\sqrt{n}\,(\hat{w}_n \, - \, w^*) \to \mathcal{N}(0, C^*) \text{ in distribution,} \tag{1.110}$$

where

$$C^* := A^{*-1} B^* A^{*-1}. \tag{1.111}$$

If, in addition each element of $\nabla l \nabla l'$ is dominated on W by an integrable function, then

$$\hat{C}_n \to C^* \text{ almost surely,} \tag{1.112}$$

with

$$\begin{cases} \hat{C}_n := \hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1}, \\ \hat{A}_n := \frac{1}{n} \sum_{t=1}^{n} \nabla^2 l(Y_t, \hat{w}_n), \\ \hat{B}_n := \frac{1}{n} \sum_{t=1}^{n} \nabla l(Y_t, \hat{w}_n) \nabla l(Z_t, \hat{w}_n)'. \end{cases} \tag{1.113}$$

Proof See White[1989].
Although $\hat{w}_n$ has considerable appeal and quite elegant asymptotic properties, when q_n and Δ_n are large, it is computationally demanding to solve the Non-linear Global Optimisation Problem (1.74), which is extremely time consuming and

difficult for practical numerical implementation. Therefore, in a practical framework two alternatives are chosen by fixing the network complexity and looking for numerical approximation of $\hat{w}_n$ which preserve the good asymptotic properties such as consistency and asymptotic normality.
The first alternative focuses, on some generalization of the Nonparametric Back-Propagation Algorithm, while the second lies essentially on purely non-linear global optimisation algorithms such as the Performed Version of Random Optimisation Method of Matyas [1965] developed by Nario Baba [1981] or the method of Simulated Annealing.
To establish the asymptotic properties of the resulting estimators, we recall first the concept of Stochastic Recursive m-Estimators that can be equated as a generalisation of the nonparametric back-propagation method and also a powerful tool for estimating local minima of continuously differentiable functions up to some regularity conditions. In this framework, to correct some eventual deficiencies of the resulting nonparametric estimators that may diverge or get stuck in local minima, we implement the algorithm using many different initial values and select the one providing the more accurate result. This usually yields a consistent estimate and furnishes quite acceptable outputs in a practical point of view, but nothing guarantees that one get close to global minima. There exists also a version of the nonparametric back-propagation estimators due to Kuschner, that provides a nonparametric consistent estimator converging to the global optimum, but this one has a very slow convergence rate.

1.4.1.5 Theorem: Stochastic Recursive m-Estimator

Let (Z_n) be a stochastic process consisting of either independent and identically distributed R^M random variables or a stationary mixing process .
Let m_1 be a continuously differentiable function on $\Re^M * \Re^l$ with values in $\Re^l$ such that:
$\forall w \in \Re^l$:

$$M(w) := E\left(m_1(Z_n, w)\right) \text{ exists.} \tag{1.114}$$

Consider $\eta_n \subset R^+$, a decreasing sequence satisfying:

$$\begin{cases} 1^\circ)\ \sum_{n=0}^{+\infty} \eta_n \ = +\infty, \\ \\ 2^\circ)\ \lim_{n \to +\infty} \sup \left[\dfrac{1}{\eta_n} - \dfrac{1}{\eta_{n-1}} \right] \ = 0, \\ \\ 3^\circ)\ \Sigma_{n=0}^{+\infty} \eta_n^d \ < \ +\infty \ \text{ for some } d > 1. \end{cases} \tag{1.115}$$

The stochastic recursive m-estimator is defined by:

$$\begin{cases} \tilde{w}_n \;:=\; \tilde{w}_{n-1} \;+\; \eta_n m_1\left(Z_n, \tilde{w}_{n-1}\right), \\ \\ \tilde{w}_0 \text{ arbitrarily chosen.} \end{cases} \tag{1.116}$$

Before stating the proposition that provides the asymptotic properties of the stochastic recursive m-estimator, we need to set up some assumptions.

Assumption1

There exists a function $Q \,:\, \; R^l \;\rightarrow\; R$, twice continuously differentiable such that:
$\forall w \in R^l$:

$$\nabla Q(w)^{'}.M(w) \;\leq\; 0. \tag{1.117}$$

Assumption2

There exists $w^* \,\in R^l$ such that:
$\forall \epsilon > 0$ and $\forall n \geq n_\epsilon$:

$$||\tilde{w}_n - w^*|| \leq \epsilon \tag{1.118}$$

Assumption3

Assumption1 holds and Assumption2 is fulfilled for all elements $w^* \in W^*$ defined by:

$$W^* := \{\, w \;\text{ such that }\; \nabla Q(w) = 0\}$$

1.4.1.6 Proposition

Let (Z_n) be the stochastic process consisting of either independent and identically distributed random variables or stationary strongly mixing processes, defined in a complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$.
If Assumption1 Holds then: with probability 1,
either

$$\tilde{w}_n \rightarrow W^* := \{w \;\text{ such that }\; \nabla Q(w) = 0\} \tag{1.119}$$

in the sense that $\inf\limits_{w \in W^*} ||\tilde{w}_n - w|| \rightarrow 0$, or

$$\tilde{w}_n \rightarrow \;+\infty. \tag{1.120}$$

If Assumption2 Holds then: $\quad M(w^*) = 0$.
If Assumption3 Holds then, with probability 1:
either

$$\tilde{w}_n \text{ converges to a local minimum of } \; Q(w) \tag{1.121}$$

or

$$\tilde{w}_n \rightarrow +\infty. \tag{1.122}$$

The proof of this proposition follows from corollary1, theorem2, and corollary2, respectively of Ljung [1977] . For more details, see White [1990].

1.4.1.7 Theorem & Definition: Nonparametric Stochastic Estimators

We now return to our original problem of estimating the autoregressive function m of the model (1.1) nonparametrically using neural networks.
Consider a stochastic process $(S_t, X_t)_{t\in\mathcal{Z}}$ satisfying (1.1), let Z_{t-1} be defined as in (1.71), let (η_n) be a real valued decreasing sequence such that (1.115) is fulfilled, a non parametric Stochastic Estimator under an ANN model with H a fixed number of hidden nodes is the specific stochastic recursive m-estimator $\tilde{w}_n$ defined by:

$$\begin{cases} \tilde{w}_n := \tilde{w}_{n-1} \ + \ \eta_n * m_1\left(S_n, Z_{n-1}, \tilde{w}_{n-1}\right) \\ \\ \tilde{w}_0 \ \text{ is a given arbitrary initial weight.} \end{cases} \tag{1.123}$$

where:

$$m_1\left(S_n, Z_{n-1}, \tilde{w}_{n-1}\right) = \nabla f_H\left(Z_{n-1}, \tilde{w}_{n-1}\right).\left(S_n - f_H(Z_{n-1}, \tilde{w}_{n-1})\right). \tag{1.124}$$

For any network weight vector w, define

$$\begin{aligned} Q(w):\ &= \ E(q_n(w)) && (1.125)\\ &= \ E\left(\frac{1}{2}[S_n - f_H(S_{n-1}, ..., S_{n-\tau}, X_{n-1}, w)]^2\right) && (1.126) \end{aligned}$$

We assume that $E(S_t^2) < \infty$ and $E(X_{t,k}^2) < \infty$ for all components of the exogenous time series X_t. Moreover, the activation function $\Psi(x)$ of the network has a bounded derivative.
If m_1 is satisfying assumption 1, 2, 3 of the proposition 1.4.1.5 and if (1.114), (1.115) and (1.116) hold, then with probability 1:
either

$$\tilde{w}_n \rightarrow \Theta^* := \left\{w \in R^l \,\text{such that}\ E\left(\nabla q_n(w)\right) = 0\right\} \tag{1.127}$$

or

$$\tilde{w}_n \rightarrow +\infty. \tag{1.128}$$

If, in addition, $\exists w^*$ such that

$$J^* := E\left(\nabla q_n(w^*)^{'} \nabla q_n(w^*)\right) \tag{1.129}$$

is positive definite then, with probabiliy 1:
either

$$\tilde{w}_n \text{ converges to a local minimum of } Q(w)$$

or

$$\tilde{w}_n \rightarrow +\infty.$$

Therefore the non-parametric Stochastic Estimator, either diverges or converges to a local minimum of Q almost surely.

Proof:
Setting $x = (y, z)$ and

$$m_1(x, w) := -\nabla q(y, z, w) = \nabla f_H(z, w)(y - f_H(z, w)) \tag{1.130}$$

and using Assumption3, one can derive that m_1 is continuously differentiable. Therefore,

$$M(w) = E\left[-\nabla q(S_t, Z_{t-1}, w)\right]. \tag{1.131}$$

$M(w)$ is finite for any given weight w as, due to the following argument:
$f_H(z, w)$ is uniformly bounded in z, as Ψ is bounded.

$$\frac{\partial}{\partial w_i} f_H(z, w) = \begin{cases} 1 \quad \text{if} \quad w_i = \beta_0, \\ \\ \Psi(\tilde{z}^{'}.\gamma_i) \quad w_i = \beta_h \quad \text{for } h = 1, ..., H, \\ \\ z_j \Psi^{'}(..) \quad \text{if} \quad w_i = \gamma_{hj} \;\; (\text{ where } z_0 = 1). \end{cases} \tag{1.132}$$

Hence $M(w)$ is finite if:
For all coordinates of Z_{t-1}, let say $Z_{t-1,j}$:

$$\begin{cases} E\left(|Z_{t-1,j}\Psi^{'}(..)|)\right) \; < \; \infty \;\; \text{and} \;\; E\left(S_t * |Z_{t-1,j}\Psi^{'}(..))|\right) \; < \; \infty \\ \\ E\left(|Z_{t-1,j}|\right) \; < \; \infty \;\; \text{and} \;\; E\left(S_t * |Z_{t-1,j}|)\right) \; < \; \infty. \end{cases} \tag{1.133}$$

This holds if

$$E\left(|S_t^2|)\right) \; < \; \infty \;\; \text{and} \;\; E\left(|X_{t,k}^2|)\right) \; < \; \infty \;\; \forall \, k \tag{1.134}$$

and $\Psi^{'}$ bounded which is true e.g. for the activation function Ψ defined in (1.115) e.g

$$\Psi^{'}(u) = \frac{2}{(1+e^{-u})(1+e^{u})} \ \leq \ 2. \tag{1.135}$$

$M(w)$ is finite for any given weights w, as, due to the assumptions on Ψ, $f_H(z,w)$ is bounded in z and $\forall i$

$$\left|\frac{\partial}{\partial w_i} f_H(z,w)\right| \leq C_1 + C_2.|z_j| \tag{1.136}$$

for some appropriate constants and an appropriate coordinate z_j of z depending on i. As the second moments of the processes $S_t$and $X_{t,k}$ are finite, we have that, $\forall i$:

$$E\left|\frac{\partial}{\partial w_i} f_H(Z_{t-1},w)\right| \ < \ \infty \quad \text{and} \quad E\left|S_t \frac{\partial}{\partial w_i} f_H(Z_{t-1},w)\right| \ < \ \infty. \tag{1.137}$$

Now, consider

$$Q(w) := \frac{1}{2} E\left[(S_t - f_H(Z_{t-1},w))^2\right]. \tag{1.138}$$

Given Assumption 1 and 2 and applying the localized version of theorem 16.8(ii)of Billingsley[1979], we have:

$$\nabla Q(w) = -E\left[(S_t - f_H(Z_{t-1},w)\nabla f_H(S_t - f_H(Z_{t-1},w)\right]. \tag{1.139}$$

Hence $\nabla Q(w) = -M(w)$, which implies that:

$$\nabla Q(w)^{'}.M(w) = -||M(w)||^2 \leq 0 \ \ \forall w. \tag{1.140}$$

Therefore the condition (1.117) underlying the Theorem 1.4.1.5 holds, hence with probability 1:
either

$$\tilde{w}_n \rightarrow \ W^{*} \tag{1.141}$$

or

$$\tilde{w}_n \rightarrow \ +\infty. \tag{1.142}$$

1.4.1.8 Conclusion

For the t^{th} trading period, the portfolio expected return can be estimated consistently with asymptotic normality by using:
Either the Nonlinear Least Square Estimator $\hat{m}_n^{NLS}$ defined as any optimal solution of the following minimization problem:

$$\min_{(f_G,w)\in ANN(\Psi,H)} \frac{1}{n}\sum_{t=1}^{n}\left[S_t - f_H\left(Z_{t-1},w\right)\right]^2 \tag{1.143}$$

or using the non-parametric Stochastic Estimator e.g

$$f_H(Z_{t-1},\hat{w}_N).$$

1.4.1.9 Consistency of the ANN Estimators for the Conditional Expectation of Unbounded Random Variables

We now return to the original problem of estimating conditional means by neural output functions. Let again $(S_t, Z_{t-1})_{t\in\mathcal{Z}}$ be a stationary stochastic process defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. We asssume that $(\Omega, \mathcal{F}, \mathcal{P})$ is a complete probability space. We want to estimate

$$\theta_0(z) = m(z) = E\left(S_t|Z_{t-1} = z\right), \quad \forall z \in \Re^r, \tag{1.144}$$

by solving the nonlinear least-square problem

$$\min_{\theta\in ANN(\Psi,q_n,\Delta_n)} \frac{1}{n}\sum_{t=1}^{n}\left[S_t - \theta(Z_{t-1})\right]^2 := \min_{\theta\in ANN(\Psi,q_n,\Delta_n)} Q_N(\theta). \tag{1.145}$$

We assume that the activation function Ψ is bounded, continuous and strictly increasing.
Let $\hat{\Theta}_n = ANN(\Psi, q_n, \Delta_n)$, and let Θ be the closure of $\cup_{n=1}^{+\infty}\hat{\Theta}_n$ in the space $L^2(\mu)$ where the probability measure μ on $\Re^r$ denotes the stationary distribution of Z_{t-1}. If μ has finite second moment, then (by Theorem 1.3.3.4) the function class $\mathcal{G}$ in the definition 1.3.3.3 is contained in Θ which, therefore contains functions of interest.
Θ is a subspace of the Hilbert space $L^2(\mu)$; therefore, it is automatically complete and separable (compare, e.g., Weidmann[1976], page33). The subsets $\hat{\Theta}_n$ are compact, as the set of weights of the network functions in $\hat{\Theta}_n$ is closed and bounded, therefore compact in a space $\Re^M$ of appropriate finite dimension $M := 1 + q_n + (1 + r)q_n$, and as the mapping from the weights in $\Re^M$ to the functions in $L^2(\mu)$ is continuous. We use the well-known fact that the continuous image of a compact set is compact(e.g, Theorem 5.1.7 of Ljusternik and Sobolev,[1968]).

Using again the correspondance between network functions and their weights and the special form of the functions in $\hat{\Theta}_n$, we get that

$$\theta_{n_k} \to \theta \in L^2(\mu) \tag{1.146}$$

with $\theta_k, \theta \in \hat{\Theta}$ implies $\theta_k(z) \to \theta(z)$ for all z provided that μ is absolutely continuous. As an immediate consequence, $Q_n(\theta)$ is continuous in θ for any realization $(S_t, Z_{t-1})_{t=1,2,...n}$, and, moreover, it is measurable with respect to $\mathcal{F}$ for given θ. It follows that the conditions of theorem 2.2 of White and Wooldride[1990] are satisfied. As an immediate consequence, there exist measurable functions $\hat{\theta}_n$ such that

$$Q_n(\hat{\theta}_n) = \min_{\theta \in ANN(\Psi, q_n, \Delta_n)} Q_n(\theta). \tag{1.147}$$

for any realization of $(S_t, Z_{t-1})_{t=1,2,...n}$. That result is not surprising as $Q_n(\theta)$ can be interpreted as a rather simple function of the network weights i.e a function on $\Re^r$. Now, we consider

$$\bar{Q}(\theta) = E\left[Q_n\right] = E\left(S_t - \theta(Z_{t-1})\right]^2. \tag{1.148}$$

As, by our assumptions, the functions $\theta \in \hat{\Theta}_n$ are uniformly bounded over $\Re^M$, we get from Lebesgue's theorem of dominated convergence and from the remark above that $\bar{Q}(\theta)$ is also continuous on $\hat{\Theta}_n$. Now, we consider $\theta = \theta_0$.
As $\theta_0(Z_{t-1}) = E\left(S_t | Z_{t-1}\right), \bar{Q}(\theta_0)$ is finite and we have

$$\bar{Q}(\theta) = E\left(S_t - \theta_0(Z_{t-1})\right)^2 \; + \; E\left(\theta_0(Z_{t-1}) - \theta(Z_{t-1})\right)^2 \tag{1.149}$$

for all $\theta \in \Theta$. It follows immediately that $\theta \to \theta_0 \in L^2(\mu)$ implies $\bar{Q}(\theta) \to \bar{Q}(\theta_0)$, i.e continuity of $\bar{Q}$ at θ_0. We assume that $\theta_0 \in \Theta$. Then, Corrolary 2.6 of White and Wooldridge[1990] implies consistency of the neural network estimators $\hat{\theta}_n$ in the following sense

$$\int \left(\hat{\theta}_n(z) - \theta_0(z)\right)^2 d\mu(z) \overset{prob}{\to} 0 \tag{1.150}$$

provided that the following two conditions are satisfied:

$$\sup_{\theta \in \hat{\Theta}_n} \left| Q(\theta) - \bar{Q}(\theta) \right| \overset{prob}{\to} 0 \tag{1.151}$$

and

$$\inf_{\theta \in \mathcal{N}_\epsilon^c(\theta_0)} \bar{Q}(\theta) - \bar{Q}(\theta_0) > 0 \tag{1.152}$$

for arbitrary ϵ-neighbourhood of θ_0 defined by:

$$\mathcal{N}_\epsilon(\theta_0) := \left\{ \theta \in \Theta;\ \int (\theta(z) - \theta_0(z))^2 \, d\mu(z) \ < \ \epsilon^2, \ \forall \epsilon > 0 \right\}$$

The latter condition is an immediate consequence of (1.149) as

$$\bar{Q}(\theta) \ - \ \bar{Q}(\theta_0) = E\left(\theta_0(Z_{t-1}) \ - \ \theta(Z_{t-1})^2 = \int (\theta_0(z) - \theta(z))^2 \, d\mu(z). \quad (1.153)$$

Therefore, we only have to verify (1.151).

For that purpose, we use Theorem 1.6.2 where the crucial assumption (1.277) follows from an application of the Bernstein inequality Theorem 1.6.1.3. We follow basically the arguments of chapter4 in White and Wooldridge[1990] who considered not neural networks but a kind of series expansions as nonparametric regression estimates.

First we remark that by definition of $ANN(\Psi, q_n, \Delta_n)$ we immediately have

$$|\theta(x)| \ \leq \Delta_n \quad \forall x, \ \ \theta \in \Theta_n. \quad (1.154)$$

The existence of an open covering $(U_{n_i})_{i=1,2,\dots,k(d_n)}$, follows from lemma 1.6.2.1. To simplify notation, we choose $\eta = 2 \times \delta_n$ in that lemma and set $C_0 = 2L_1$ such that we get as upper bound for $K(\delta_n)$

$$K(\delta_n) \leq 4 \left(\frac{\Delta_n}{\delta_n}\right)^{q_n(r+2)+1} q_n^{q_n(r+1)}. \quad (1.155)$$

We use the notation

$$S_t(\theta) = g\left(\theta, S_t, Z_{t-1}\right) := \left(S_t \ - \ \theta(Z_{t-1})^2\right. . \quad (1.156)$$

The measurability condition of Theorem 1.6.2 on g is obviously satisfied as it depends continuously on θ, S_t, Z_{t-1}. Mark that all network functions in $\hat{\Theta}_n$ are continous if Ψ is continuous, and we even assume Lipschitz continuity. Condition (1.278) of Theorem 1.6.2 is satisfied as for $\theta, \theta^* \ \in \hat{\Theta}_n$

$$\begin{aligned} |S_t(\theta) - S_t(\theta^*)| \ &= \ \left| [S_t - \theta(Z_{t-1})]^2 \ - \ [S_t - \theta^*(Z_{t-1})]^2 \right| & (1.157) \\ &\leq \ |[S_t - \theta(Z_{t-1})] + \\ &\quad\ [S_t - \theta^*(Z_{t-1})]| \, |\theta^*(Z_{t-1}) - \theta(Z_{t-1})| & (1.158) \\ &\leq \ 2\left(|S_t| + \Delta_n\right) \times |\theta(Z_{t-1}) - \theta^*(Z_{t-1})| & (1.159) \end{aligned}$$

if we choose

$$M_{nt} := 2\left(|S_t| + \Delta_n\right).$$

Then we have

$$\mu_n^2 = E\left(M_{nt}\right)^2 \leq 8\left(E(S_t^2) + \Delta_n^2\right). \tag{1.160}$$

As (S_t, Z_{t-1}) is α-mixing with geometrically decreasing mixing coefficients, the same mixing behaviour is shared by $S_t(\theta)$ due to, e.g., Theorem 3.49 of White[1984]. Now, we assume that the stationary distribution of S_t has exponential decreasing tails, i.e. for some a_o, a_1 and $\alpha > 0$,

$$Pr\left(|S_t| > x\right) \leq a_0 \exp\left\{-a_1 x^{\alpha}\right\} \quad \forall x \geq 0. \tag{1.161}$$

We conclude immediately

$$\begin{aligned}
Pr\left(|S_t(\theta) - E(S_t(\theta))| > x\right) &\leq Pr\left(|S_t(\theta)| >\right. \\
&\quad \left. x - \ E(S_t(\theta))\right) &(1.162)\\
&= Pr\left(\left[S_t - \theta(Z_{t-1})\right]^2 >\right. \\
&\quad \left. x - E(S_t - \theta(Z_{t-1})^2)\right) &(1.163)\\
&\leq Pr\left(\left[S_t - \theta(Z_{t-1})\right]^2 >\right. \\
&\quad \left. x - 2E(S_t^2) - 2\Delta_n^2\right) &(1.164)\\
&\leq Pr\left(|S_t| >\right. \\
&\quad \left. \left\{x - 2E(S_t^2) - 2\Delta_n^2\right\}^{\frac{1}{2}} - \Delta_n\right) &(1.165)\\
&\leq a_0 \exp\left(-f_n(x)\right) &(1.166)
\end{aligned}$$

$\forall x > 3\Delta_n^2 + 2E(S_t^2)$ with

$$f_n(x) = a_1\left(\left\{x - 2E(S_t^2) - 2\Delta_n^2\right\}^{\frac{1}{2}} - \Delta_n\right)^{\alpha}. \tag{1.167}$$

Therefore, the Bernstein inequality of Theorem 1.6.1.3 is applicable to

$$\epsilon_t(n) = S_t(\theta) - E(S_t(\theta)).$$

Choosing $M_n = 8\Delta_n^2$ in that inequality, we get $f_n(M_n) \geq a_1\Delta_n^{\alpha}$ for $\Delta_n^2 \geq E(S_t^2)$, and we have

$$\begin{aligned}
Pr\left(\left|\sum_{t=1}^{n}\left(S_t(\theta) - E(S_t(\theta))\right)\right| > \Delta_n\right) &\leq C_1 \exp\left\{-\frac{C_2}{8}\frac{\Delta_n}{\sqrt{n}\Delta_n^2}\right\} + \\
&\quad n a_0 \exp\left\{-a_1\Delta_n^{\alpha}\right\} &(1.168)
\end{aligned}$$

provided that $nE(8\Delta_n^2) = o(\Delta_n)$. The first term on the right hand side of the last equation coincides with the corresponding term for results of White and Wooldrige[1990] for bounded random variables. Therefore, condition (1.277) of theorem 1.6.2 is satisfied if $\Delta_n \to +\infty$ fast enough if we choose

$$\gamma_n(\epsilon) := C_1 \exp\left\{-\frac{C_2}{8}\frac{\epsilon}{\sqrt{n}\Delta_n^2}\right\} + na_o \exp\left\{-a_1\Delta_n^\alpha\right\}. \tag{1.169}$$

We conclude that the assumptions of Theorem 1.6.2 are satisfied if we additionally assume that the stationary distribution μ of Z_{t-1} also decays exponentially, i.e for some $\beta_0, \beta_1, \tau > 0$

$$Pr\left(||Z_{t-1} > x|\,|\right) \leq \beta_0 \exp\left\{-\beta_1||x||^\tau\right\} \quad \forall x. \tag{1.170}$$

Now, we apply Theorem 1.6.2 for $S_n(\theta) = nQ_n(\theta)$ and $a_n = n$, and we get

$$Pr\left(\sup_{\theta\in\hat{\Theta}_n}\left|Q_n(\theta) - \bar{Q}(\theta)\right| > \epsilon\right) = Pr\left(\sup_{\theta\in\hat{\Theta}_n}\left|S_n(\theta) - E(S_n(\theta)\right| > n\epsilon\right) \xrightarrow{P} 0 \tag{1.171}$$

if for any arbitrary δ_n, ρ_n

$$K(\delta_n)\gamma_n(n\epsilon) \to 0,\;\; K(\delta_n)\frac{n\mu_n}{\epsilon n}\left(1 + \Delta_n\rho_n\right)\delta_n \to 0 \tag{1.172}$$

with

$$K(\delta_n)\frac{n\mu_n}{\epsilon n}\Delta_n \exp\left\{-\frac{\beta_1}{2}\delta_n^2\right\} \to 0. \tag{1.173}$$

To finalize the proof of (1.151) above, we have to show (1.173). There, we replace $K(\delta_n)$ by the upper bound from (1.155). For the first term of (1.173), we need, using the abbreviation $p_n = q_n(r+1)$,

$$\begin{array}{rcl} K(\delta_n)\gamma_n(n\epsilon) & \leq & 4\left(\dfrac{\Delta_n}{\delta_n}\right)^{p_n+q_n+1} q_n^{p_n}\left\{C_1\exp\left\{\dfrac{C_2}{8}\dfrac{\epsilon}{\sqrt{n}\Delta_n^2}\right\} + na_o\exp\left\{-a_1\Delta_n^\alpha\right\}\right\} \\ & \to & 0 \;\text{ for } n \to +\infty. \end{array}$$

For that, it sufficies that

$$\begin{array}{ll} \exp & \left\{(q_n + p_n + 1)log(\dfrac{\Delta_n}{\delta_n}) + p_n log(q_n) - \dfrac{C_2}{8} \times \dfrac{\epsilon\sqrt{n}}{\Delta_n^2}\right\} \to 0 \qquad (1.174) \\ \exp & \left\{(q_n + p_n + 1)\, log(\dfrac{\Delta_n}{\delta_n}) + p_n log(q_n) - a_1\Delta_n^\alpha + log(n)\right\} \to 0. (1.175) \end{array}$$

As p_n is a constant multiple of q_n, and if we assume $log(n) = o(\Delta_n^\alpha)$, these two assertions hold in particular if

$$\begin{aligned} q_n log\left(\frac{\Delta_n q_n}{\delta_n}\right) &= o\left(\frac{\sqrt{n}}{\Delta_n^2}\right) &(1.176)\\ q_n log\left(\frac{\Delta_n q_n}{\delta_n}\right) &= o\left(\Delta_n^\alpha\right). &(1.177)\end{aligned}$$

Using the upper bound for μ_n^2, derived above, the second term of (1.173) is bounded by

$$\left(\frac{\Delta_n}{\delta_n}\right)^{p_n+q_n+1} q_n^{p_n} \frac{\sqrt{8}\left(E(S_t^2)+\Delta_n^2\right)^{\frac{1}{2}}}{\epsilon} C_0\left(1+\Delta_n\rho_n\right)\delta_n. \tag{1.178}$$

Neglecting constants and using that $E(S_t^2)+\Delta_n^2$ behaves like Δ_n^2 for $\to\infty$, and assuming that $\Delta_n\rho_n \to \infty$ for $\to \infty$, that term converges to 0 if

$$\left(\frac{\Delta_n q_n}{\delta_n}\right)^{p_n}\left(\frac{\Delta_n}{\delta_n}\right)^{q_n+1}\Delta_n^2\rho_n\delta_n \to 0. \tag{1.179}$$

Analogously, the last term of (1.173) converges to 0 if

$$\left(\frac{\Delta_n q_n}{\delta_n}\right)^{p_n}\left(\frac{\Delta_n}{\delta_n}\right)^{q_n+1}\Delta_n^2 \exp\left\{-\frac{\beta_1}{2}\rho_n^2\right\} \to 0. \tag{1.180}$$

Now, we choose $\rho_n = n^\delta$ and $\delta_n = n^\gamma \Delta_n q_n$ for some $\rho, \gamma > 0$. As $q_n, \Delta_n \to \infty$, (1.176) implies necessarily $\frac{\sqrt{n}}{\Delta_n^2} \to 0$ i.e

$$\Delta_n = o(n^{\frac{1}{4}}). \tag{1.181}$$

This is the same condition as for bounded random variables(compare Theorem 3.3 of White[1990]). (1.177) and (1.179) hold if

$$-q_n\Delta_n^2\gamma log(n) = o(n^{\frac{1}{2}}) \text{ and } -q_n\Delta_n^2\gamma log(n) = o(\Delta_n^{2+\alpha}).$$

or neglecting the constant γ,

$$q_n\Delta_n^2 log(n) = o(n^{\frac{1}{2}}) \text{ and } q_n\Delta_n^2 log(n) = o(\Delta_n^{2+\alpha})$$

Mark that the second assertion implies the assumption $log(n) = o(\Delta_n^\alpha)$ which we have made above. Also, we have now

$$\delta_n = n^\gamma\Delta_n q_n = o(n^{\frac{1}{2}+\gamma}). \tag{1.182}$$

Together with $\rho_n = n^\rho$, we see immediately that (1.180) is implied by (1.179), and it remains to consider the latter condition. As $\Delta_n \leq q_n \Delta_n$, (1.179) is implied by

$$\frac{(\Delta_n q_n)^{p_n+q_n+3}}{{\delta_n}^{q_n+p_n}} \rho_n = \frac{(\Delta_n q_n)^3 n^\delta}{n^{\gamma(p_n+q_n)}} \to 0 \tag{1.183}$$

as $\Delta_n q_n = o(n^{\frac{1}{2}})$ and $p_n + q_n \to \infty$ for $n \to \infty$.
It remains to discuss the condition $nE_n\left(8\Delta_n^2\right) = o(\Delta_n)$ which we have assumed above. As we have chosen now $\Delta_n = \epsilon n$, we need

$$E_n\left(8\Delta_n^2\right) = o(1). \tag{1.184}$$

But for $x \geq 8\Delta_n^2$, we have

$$\frac{x}{2} \geq 2E\left(S_t^2\right) + 2\Delta_n^2 \tag{1.185}$$

for sufficiently large n and $\frac{1}{2}\sqrt{\frac{x}{2}} \geq \Delta_n$ such that

$$\sqrt{x - 2E\left(S_t^2\right) - 2\Delta_n^2} - \Delta_n \geq \frac{1}{2}\sqrt{\frac{x}{2}} \quad \text{for } x \geq 8\Delta_n^2. \tag{1.186}$$

This implies immediately

$$\begin{aligned} E_n\left(8\Delta_n^2\right) &= \int_{8\Delta_n^2}^{\infty} e^{-a_1\left(\sqrt{x - 2E(S_t^2) - 2\Delta_n^2} - \Delta_n\right)^\alpha} dx & (1.187) \\ &\leq \int_{8\Delta_n^2}^{\infty} e^{-a_1\left(\frac{1}{2}\sqrt{\frac{x}{2}}\right)^\alpha} dx \to 0 & (1.188) \end{aligned}$$

as Δ_n for $n \to \infty$. Therefore, we finally have shown the following result.

1.4.2 Consistency of the ANN Estimators for the Conditional Mean of Unbounded Stochastic processes

1.4.2.1 Theorem

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a complete probability space, and let (S_t, Z_{t-1}) be a stationary stochastic process satisfying an α-mixing condition with exponentially decreasing mixing coefficients, where S_t is real valued and $Z_{t-1} \in \Re^r$. Let the stationary distribution of S_t be absolutely continuous and satisfy

$$P\left(|S_t| > x\right) \leq a_0 \exp\left\{a_1 x^\alpha\right\} \quad \text{for all } x \geq 0 \tag{1.189}$$

for some a_0, a_1 and $\alpha > 0$. Let the stationary distribution μ of Z_t be absolutely continuous and satisfy

$$P\left(|Z_t| > x\right) \leq \beta_0 \exp\left\{\beta_1 x^\tau\right\} \quad \text{for all } x \geq 0 \tag{1.190}$$

for some β_0, β_1 and $\tau > 0$.
Let $m(z)$ denote the best forecast of S_t given $Z_{t-1} = z$:

$$m(z) = E\left(S_t | Z_{t-1} = z\right). \tag{1.191}$$

Let Ψ be bounded in absolute value by 1 and satisfy a Lipschitz condition:

$$|\Psi(x) - \Psi(y)| \leq L.|u - v| \;\; \forall u, v \in \Re. \tag{1.192}$$

Let $\hat{\Theta}_n := ANN(\Psi, q_n, \Delta_n)$ be the usual set of neural network functions of r input variables with q_n neurons in the hidden layer where the sum of absolute values of the weights from hidden to output layer is bounded by Δ_n, and the sum of all absolute wieghts from input to hidden layers is bounded by $q_n\Delta_n$. Let Θ denote the closure of $\cup_{n=1}^{\infty}\hat{\Theta}_n$ in $L^2(\mu)$. Let $\hat{\theta}_n \in \hat{\Theta}_n$ be the network function which provides the best nonlinear least-square fit to the data $(S_1, Z_0), (S_2, Z_1), ..., (S_n, Z_{n-1})$:

$$\hat{\theta}_n = argmin_{\theta \in \hat{\Theta}_n} \frac{1}{n} \Sigma_{t=1}^{n} \left[S_t - \theta(Z_{t-1})\right]^2. \tag{1.193}$$

Assume $m \in \Theta$. Then, $\hat{\theta}_n$ is a consistent estimate of m for $n \to \infty$ in the following $L^2(\mu)$ sense

$$\int \left(m(z) - \hat{\theta}_n(z)\right)^2 d\mu(z) \xrightarrow{p} 0 \tag{1.194}$$

provided that for $n \to \infty$:

$$\begin{aligned} q_n, \Delta_n &\to \infty, & (1.195)\\ \Delta_n &= o(n^{\frac{1}{4}}) \text{ and} & (1.196)\\ q_n \Delta_n^2 log(n) &= o\left(\min(\sqrt{n}, \Delta_n^{2+\alpha})\right). & (1.197) \end{aligned}$$

For bounded random variables S_t, Z_{t-1}, we have from theorem 1.4.2.1 that

$$\int_I \left(m(z) - \hat{\theta}_n(z)\right)^2 \mu(dz) \xrightarrow{p} 0 \tag{1.198}$$

where I denotes the bounded support of Z_{t-1} and m is assumed to be continous, provided that

$$\begin{aligned} \Delta_n = o(n^{\frac{1}{4}}) \text{ and} & & (1.199)\\ q_n \Delta_n^2 log(n) &= o\left(\min(\sqrt{n}, \Delta_n^{2+\alpha})\right). & (1.200) \end{aligned}$$

The difference between $log(q_n\Delta_n)$ and $log(n)$ in (1.197) is of minor impact. The main difference between the bounded and unbounded case is the additional requirement that

$$q_n\Delta_n^2 log(n) = o\left(\Delta_n^{2*\alpha}\right). \tag{1.201}$$

To get an intuition, look at the special case

$$\Delta_n := bn^\beta \text{ for some } 0 < \beta < \frac{1}{4}. \tag{1.202}$$

Then, the $\sqrt{n}$ term on the right hand side of (1.197) is the more severe one if

$$\sqrt{n} = O\left(\Delta_n^{(2+\alpha)}\right) = O\left(n^{(2+\alpha)\beta}\right), \text{ i.e } \beta \geq \frac{1}{2(2+\alpha)}. \tag{1.203}$$

So, if β is close enough to its upper bound $\frac{1}{4}$, i.e if Δ_n grows rather fast, the unboundedness of the random variables has practically no influence on the series of q_n as a function of Δ_n. Of course, the larger α is, which determines the probability of large values of S_t, the smaller β may be to end up in that case. If, on the other hand, $\beta < \frac{1}{2(2+\alpha)}$, then the rate of q_n is determined by

$$q_n\Delta_n^2 log(n) = o\left(\Delta_n^{2+\alpha}\right) \tag{1.204}$$

instead of $o\left(\sqrt{n}\right)$, i.e the number of hidden neurons has to be smaller than for bounded S_t. Here, we have consistency of $\hat{\theta}_n(z)$ if

$$\Delta_n = bn^\beta \text{ for } 0 < \beta < \frac{1}{4}, \, b > 0, \tag{1.205}$$

and either

$$\beta \geq \frac{1}{2(2+\alpha)} \text{ for } q_n = o\left(\frac{n^{\frac{1}{2}-2\beta}}{log(n)}\right) \tag{1.206}$$

or

$$\beta < \frac{1}{2(2+\alpha)} \text{ for } q_n = o\left(\frac{n^{\alpha\beta}}{log(n)}\right) \tag{1.207}$$

We also apply this consistent method of approximating the conditional expected returns by means of neural output functions to forecast the unknown conditional volatilities of the market value of a given financial instrument.

1.4.3 Neural Network Estimate of the Conditional Stochastic Volatility

As in the case of the expected return, one can use Artificial Neural Network for estimating conditional stochastic volatilities. This can be done in the following manner:
Consider the stochastic process that describes the portfolio returns dynamics under the model defined in (1.1).

$$\begin{aligned} \sigma^2(Z_{t-1}) &= \left(\frac{S_t - m(Z_{t-1})}{\mathcal{E}_t}\right)^2 && (1.208)\\ &= E\left(S_t^2|\mathcal{F}_{t-1}\right) - \left[E\left(S_t|\mathcal{F}_{t-1}\right)\right]^2 && (1.209)\\ &= E\left(S_t^2|\mathcal{F}_{t-1}\right) - m^2(Z_{t-1}). && (1.210)\end{aligned}$$

Therefore, one could estimate the stochastic volatility σ^2 by estimating the conditional second moment and, subtracting the squared neural network estimate $f_H(Z_t, \hat{w}_N)$ of the conditional expected return $E\left(S_t^2|\mathcal{F}_{t-1}\right)$.
The result of the previous sections are immediately applicable to $\left(S_t^2, Z_{t-1}\right)$ instead of (S_t, Z_{t-1}). The only additional assumption which have to made is that:

$$E\left(S_t^4\right) < +\infty \qquad (1.211)$$

Then, the volatility estimate

$$\tilde{\sigma}_t^2(z) = f_G(z, \bar{w}_N) - \left[f_G(z, \hat{w}_N)\right]^2 \qquad (1.212)$$

for $\tilde{\sigma}_t^2(z) = var\left(S_n|Z_{n-1} = z\right)$ has the same asymptotic behavior as $f_H(z, \hat{w}_N)$ as an estimate of $m(z) = E\left(S_t|Z_{n-1} = z\right)$. However, that estimate has two slight drawbacks. First, if $G \neq H$, it may happen with small, but positive probability that $\tilde{\sigma}^2(z) < 0$ for some z which is of course not desirable. Moreover, the analogous procedure using kernel estimates has a an additional bias as an estimate of $\sigma^2(z)$ which is caused by the bias in $f_H^2(z, \hat{w}_N)$ as estimate of $m^2(z)$(see Franke, Neumann and Stockis [2001]). Therefore we consider the following alternative as developed in Franke[1999] by treating $\sigma(Z_t)\mathcal{E}_t$ as one special innovation I_t and set:

$$I_t := \sigma(Z_t)\mathcal{E}_t. \qquad (1.213)$$

Hence I_t can be initially estimated by $\hat{I}_t$ defined as follow:

$$\hat{I}_t := S_t - f_H(Z_t, \hat{w}_N) \qquad (1.214)$$

Therefore one can fit this dependence using a new Artificial Neural Network with G hidden nodes by noting that:

$$\sigma^2(Z_t) = E(I_t^2|\mathcal{F}_{t-1}) \tag{1.215}$$

Hence $\sigma^2(Z_t)$ can be estimated by solving the following minimization problem:

$$\min_{(f_G,\bar{w})\in ANN(\Psi,G)} \frac{1}{N}\sum_{t=1}^{N}\left[\hat{I}_t^2 \;-\; f_G\left(Z_t,\bar{w}\right)\right]^2 . \tag{1.216}$$

providing us with $\hat{\sigma}^2(z) = f_G(z,\hat{w}_N)$ as an alternative estimate for $\sigma^2(z)$. $\hat{\sigma}^2(z)$ will always be positive.

Remark:

The theory of the estimate based on $\hat{I}_t$ is technically more demanding as $\hat{I}_t \neq I_t$. A work has been started dealing with such difficulties(see Franke, Stockis and Dimitroff[2002]).

1.5 Quantile Estimation Using Extreme Value Theory

The quantile estimation procedure that is presented throughout this section is making use of EVT and is relying essentially on the papers of Smith [1987] and the one of Mc-Neil [1999] dealing with the approximation of the tail of probability distributions .The initial ideas of this estimation procedure can also be found in Hosking [1987], where the author is presenting some results concerning the estimation of the parameters and quantile for the Generalized Pareto Distributions(GPD). This approach leads to an invertible form of the distribution function of the innovations which help to get easily the estimator of the required quantile with appealing asymptotic properties. The use of EVT and GPD as a tool in financial risk management is also developed in Mc-Neil [1999] or Embrechts [1997]. This approach consists of an appropriate choice of a threshold level u and estimating the distribution function F, by its sample version below the threshold and some GPD over the chosen threshold. For that, the concept of Excess Distribution will be defined and some fundamental results of the theory of extreme value will be recalled. Such results, due to Pickand [1975] and Fischer enable to approximate accurately the Excess Distribution over the threshold level.

1.5.1 Excess Distribution Function Estimation

1.5.1.1 Definition: Excess Distribution

Given an appropriately high threshold u and a strictly white noise $\mathcal{E}_t$ supposed to be heavy tailed with unknown distribution function F, the Excess Distribution Function over the threshold u , is defined by:

$$\begin{aligned} F_u(x) : &= P\left(\mathcal{E} \leq u + x \;|\mathcal{E} > u\right) && (1.217) \\ &= \frac{F(u+x) - F(u)}{1 - F(u)}. && (1.218) \end{aligned}$$

Hence

$$1 - F(x) = [1 - F(u)] * [1 - F_u(x - u)]\,. \qquad (1.219)$$

$F(u)$ is estimated by the sample distribution function evaluated at u e.g

$$\hat{F_n}(u) := \frac{1}{n}\sum_{t=1}^{n} 1_{\{\mathcal{E}_t \;\; \leq u\,\}}. \qquad (1.220)$$

This is equivalent to suppose the existence of N excesses $(Y_i = \mathcal{E}_{t_i} - u\,)$ over the threshold that are independent and identically distributed conditionally to N. The use of Extreme Value Theory leads to the estimation of the distribution function of these excesses and the related mean excess function.

The estimation of $F_u(x - u)$ will be done by using the theorem of Pickands-Balkema-de Hann [1974/1975]. For that one needs to recall the two important classes composed by Generalized Extreme Value and Generalized Pareto Distributions and the theorems of Fischer-Tippett and the one of Pickands-Balkema-de Haan. The two theorems can be considered as the bedrock of Extreme Value Theory.

1.5.1.2 Definition: Generalized Distribution Functions

Generalized Extreme and Pareto Distribution functions play a crucial role in the study of financial market extreme events more specifically in financial market-crashes or extreme loss quantification in insurance mainly during earthquake or hurricane.

The Generalized Extreme Value Distribution $H_{\psi,\mu,\sigma}$ is defined by:

$$H_{\psi,\mu,\sigma}(x) := \begin{cases} \exp\left\{-\left[1 + \psi\dfrac{x-\mu}{\sigma}\right]^{-\frac{1}{\psi}}\right\} & \text{if} \quad \psi \neq 0 \\ \exp\left(-\exp\left[-\dfrac{x-\mu}{\sigma}\right]\right) & \text{if} \;\; \psi = 0 \end{cases} \qquad (1.221)$$

and the Generalized Pareto Distribution $G_{\psi,\beta}$ is given as:

$$G_{\psi,\beta}(x) := \begin{cases} 1 - \left(1 + \dfrac{\psi x}{\beta}\right)^{\frac{-1}{\psi}} & \text{if} \quad \psi \neq 0, \\ \\ 1 - \exp\left(-\frac{x}{\beta}\right) \quad \text{if} \quad \psi = 0. \end{cases} \tag{1.222}$$

The Generalized Pareto Distribution is defined under the following conditions:

$$\begin{cases} 1^\circ) \;\; \beta > 0, \\ \\ 2^\circ) \;\; x \in [0\,,\, -\frac{\beta}{\psi}] \text{ if } \psi < 0, \\ \\ 3^\circ) \;\; x \geq 0 \;\; \text{if} \;\; \psi \geq 0. \end{cases} \tag{1.223}$$

Beyond the important fact that Generalized Distributions help to estimate tails of distributions, they also provide accurate estimation tools that can be used to construct quantile estimation of heavy tailed distributions like the innovations resulting from model (1.1). Before starting the estimation procedure, some fundamental results of EVT need to be introduced.

1.5.2 Fundamental Results of Extreme Value Theory

1.5.2.1 Fischer-Tippett Theorem

let $(\mathcal{E}_t)$ be independent identically distributed random variables with distribution function $F_{\mathcal{E}}$. Let M_n be the random variable defined by:

$$M_n := \max_{1 \leq t \leq n} \mathcal{E}_t. \tag{1.224}$$

If there exist two real valued sequences $a_n > 0$ and $b_n \in \mathcal{R}$ and a distribution function H such that:

$$\frac{M_n - b_n}{a_n} \rightarrow H \quad \text{in distribution} \ , \tag{1.225}$$

then there exist ψ , μ, and σ such that:

$$H = H_{\psi,\mu,\sigma} \quad \text{almost surely.} \tag{1.226}$$

1.5.2.2 Definition: Maximum Domain of Attraction (MDA)

If (1.224), (1.225) and (1.226) hold, we say that $F_{\mathcal{E}}$ belongs to the Maximum Domain of Attraction of $H_{\psi,\mu,\sigma}$.
The Fischer-Tippett Theorem is stating that the distribution function describing the dynamic of extreme events belongs to Maximum Domain of Attraction of a Generalized Extreme Value Distribution.
Gnedenko accomplished an important excursion related to this result in 1943. He proved that The Fischer-Tippett theorem is applicable for heavy tailed distributions functions. More precisely, he shown that heavy tailed distribution functions belong to the Maximum Domain of Attraction of the Frechet Distribution e.g. $H_{\psi,0,1}$ with $\psi > 0$.

1.5.2.3 Theorem of Pickands-Balkema-Gnedenko-de Haan

Under the same condition as the theorem of Fischer-Tippett, given an appropriately high threshold u, there exits a measurable function $\sigma(u)$ such that:

$$F \in (MDA)(H_{\psi,0,1}) \iff \lim_{u \to x_0} \left\{ \sup_{0 \leq x \leq x_0 - u} |F_u(x) - G_{\psi,\beta(u)}(x)| \right\} = 0 \quad (1.227)$$

Where x_0 is defined by:

$$x_0 := \sup \{ x \in \mathcal{R}^r \text{ such that } F(x) < 1 \} \quad (1.228)$$

In other word, it means that:
Once a reasonably high threshold is fixed, the excess distribution F_u can be approximated by a Generalized Pareto Distribution $G_{\hat{\psi},\hat{\beta}(u)}$ (see Embrechts, Resnick and Samorodnisky[1997]). Where $\hat{\psi}$ and $\hat{\beta}(u)$ denotes the corresponding Maximum Likelihood Estimators of ψ and β.

1.5.2.4 Theorem

Let $(\mathcal{E}_t)$ be a heavy tailed strictly white noise with unknown distribution function F. Then given an appropriately high threshold level u, there exits a natural number N_u , a positive real scalar $\hat{\psi}_N$ and a positive measurable function $\hat{\beta}_N(u)$ such that:

$$1 - F_{\mathcal{E}}(x) \simeq \left[1 - \frac{N_u}{n} \right] * \left[1 + \hat{\psi} \, \frac{x-u}{\hat{\beta}(u)} \right]^{\frac{-1}{\hat{\psi}}} . \quad (1.229)$$

Proof:
Using the fact that

$$1 - F_{\mathcal{E}}(x) = [\, 1 - F(u) \,] * [\, 1 - F_u(x-u)] \quad (1.230)$$

and approximating:
$\bullet F(u)$ using the sample distribution function evaluated at u, this means that we can suppose that there exists N_u excesses $(Y_1, Y_2, ..., Y_{N_u})$ over the threshold u.
$\bullet F_u(x-u)$ by $G_{\hat{\psi}_N,\hat{\beta}_N(u)}(x-u)$ using the theorem of Pickand-Balkema-Gnedenko-de Haan .
To construct $G_{\hat{\psi}_N,\hat{\beta}_N(u)}$, one can assume that the excesses are exactly (or even approximately) identically Generalized Pareto distributed and use the fact that they are independent conditionally to N_u.

1.5.3 Quantile Estimation Formula for Heavy Tailed Distributions

Based on the Maximum Likelihood estimators $\hat{\psi}_N$ and $\hat{\beta}_N(u)$ of ψ and $\beta(u)$ fitted with the residual excess sample $\hat{\mathcal{E}}_t$ defined by:

$$\begin{aligned} \hat{\mathcal{E}}_t : &= \frac{S_t - \hat{m}(Z_t)}{\hat{\sigma}(Z_t)} && (1.231) \\ &= \frac{S_t \; - \; f_H(Z_t, \hat{w}_N)}{f_G(Z_t, \bar{w}_N)}, && (1.232) \end{aligned}$$

if the model (1.1) is tenable, the $\hat{\mathcal{E}}_t$ must be iid and:

$$F_{\mathcal{E}}(x) \;\simeq\; \hat{F}_N^u(x) := 1 - \frac{N_u}{N}\left(1 + \frac{\hat{\psi}_N}{\hat{\beta}_N(u)}(x-u)\right)^{\frac{1}{\hat{\psi}_N}}. \qquad (1.233)$$

The unknown heavy tailed distribution function $F_{\mathcal{E}}$, estimated as in (1.229) becomes invertible and

$$x \;\simeq\; \frac{\hat{\beta}_N(u)}{\hat{\psi}_N}\left\{\left[\frac{N}{N_u}(1-\hat{F}_N^u(x))\right]^{\hat{\psi}_N} - 1\right\} + u. \qquad (1.234)$$

Therefore the α quantile q_α of the unexpected returns $\mathcal{E}_t$ can be estimated by $\hat{q}_N^\alpha(u)$ defined by:

$$\hat{q}_N^\alpha(u) := \frac{\hat{\beta}_N(u)}{\hat{\psi}_N}\left\{\left[\frac{N}{N_u}(1-\alpha)\right]^{\hat{\psi}_N} - 1\right\} + u. \qquad (1.235)$$

Under some general conditions, Smith [1987] has proved that $\hat{\sigma}_N$ and $\hat{\psi}_N$ are consistent and asymptotically normal. Hence $\hat{q}_n^\alpha(N,u)$ is consistent and asymptotically normal distributed.

1.5.3.1 VaR Estimation Formula

The Conditional VaR can finally be estimated consistently in the following manner:

$$V\hat{a}R_{\alpha}^{t}(u,N) = f_H(Z_t,\hat{w}_N) \ + \ [\hat{q}_N^{\alpha}(u)] * [f_G(Z_t,\bar{w}_N)] \tag{1.236}$$

where

$$\begin{cases} f_H(Z_t,\hat{w}_N) = \hat{\beta}_0^N + \sum\limits_{j=1}^{H} \hat{\beta}_j^N \psi(\tilde{x}_t.\hat{\gamma}_j^N), \\ \\ f_G(Z_t,\bar{w}_N) = \hat{\nu}_0^N + \sum\limits_{j=1}^{G} \hat{\nu}_j^N \psi(\tilde{x}_t.\hat{\lambda}_j^N), \\ \\ \hat{q}_N^{\alpha}(u) := \dfrac{\hat{\beta}_N(u)}{\hat{\psi}_N} \left\{ \left[\dfrac{N}{N_u}(1-\alpha) \right]^{\hat{\psi}_N} - 1 \right\} + u \end{cases} \tag{1.237}$$

with

$$\begin{cases} \hat{w}_N = \left(\hat{\sigma}_0^N, \hat{\sigma}_1^N, ..., \hat{\sigma}_H^N, \hat{\gamma}_j^N, \ j = 1, 2...H\right), \\ \\ \bar{w}_N = \left(\hat{\nu}_0^N, \hat{\nu}_1^N, ..., \hat{\nu}_G^N, \hat{\lambda}_j^N, \ j = 1, 2...G\right). \end{cases} \tag{1.238}$$

1.5.4 Expected Shortfall Estimation

1.5.4.1 Proposition

Under the same previous assumptions of the model (1.1), the Expected Shortfall ES_{α}^{t} is given by the following expression:

$$ES_{\alpha}^{t} = m(Z_t) + \sigma(Z_t) * q_{\alpha} * \left[\frac{1}{1-\psi} + \frac{\beta - \psi u}{(1-\psi) * q_{\alpha}} \right]$$

Proof:

$$\begin{aligned} ES_{\alpha}^{t} : &= E_{t-1}\left(S_t | S_t > VaR_{\alpha}^{t}\right) \\ &= E_{t-1}\left(m(..) + \sigma(..)\mathcal{E}_t | m(..) + \sigma(..)\mathcal{E}_t > m(..) + \sigma(..) * q_{\alpha}\right) \\ &= m(..) + \sigma(..)E\left(\mathcal{E}_t | \mathcal{E}_t > q_{\alpha}\right). \end{aligned} \tag{1.239}$$

Therefore estimating ES^t_α requires the valuation of the unconditional expected shortfall of the innovations.
Given the assumption that $\mathcal{E}_t - u|\mathcal{E}_t > u$ follows a GPD, it follows that:

$$\mathcal{E}_t - q_\alpha|\mathcal{E}_t > q_\alpha = \left[(\mathcal{E}_t - u) + (q_\alpha - u)\,|\mathcal{E}_t - u > q_\alpha - u\right]. \tag{1.240}$$

The right hand side of the previous equation follows a Generalized Pareto Distributed with parameter ψ and $\sigma + \psi(q_\alpha - u)$.
Using the fact that, if a random variable $\mathcal{E}$ follows a GPD GPD(ψ, β), then:

$$E\left(\mathcal{E}|\mathcal{E} > x\right) = \frac{x + \beta}{1 - \psi}, \tag{1.241}$$

one has:

$$E\left(\mathcal{E}_t|\mathcal{E}_t > q_\alpha\right) = q_\alpha \left[\frac{1}{1 - \psi} + \frac{\beta - \psi u}{1 - \psi q_\alpha}\right]. \tag{1.242}$$

Therefore the Expected Shortfall of the innovation is given by:

$$ES^t_\alpha = m(Z_t) + \beta(Z_t) * q_\alpha * \left[\frac{1}{1 - \psi} + \frac{\beta - \psi u}{(1 - \psi) * q_\alpha}\right]. \tag{1.243}$$

For more details about (1.241) and (1.242), we refer to the paper of Mc-Neil [2000], page 11, formula (14).

1.5.5 Expected Shortfall Estimation Formula

Finally the Expected Shortfall can be estimated by:

$$\begin{aligned} \hat{ES}^t_\alpha &:= f_H(Z_t, \hat{w}_N) + \\ & f_G(Z_t, \bar{w}_N) * \hat{q}^\alpha_N(u) \left[\frac{1}{1 - \hat{\psi}} + \frac{\hat{\beta} - \hat{\psi} u}{(1 - \hat{\psi})} * (\hat{q}^\alpha_N(u))\right]. \end{aligned} \tag{1.244}$$

1.6 Some technical Results

1.6.1 A Bernstein Inequality for Unbounded Stochastic Processes

In this section we prove some auxiliary results needed to prove the main theorem 1.4.2.1. The following result is a Bernstein inequality for unbounded random variables from stationary α-mixing processes. It generalizes a result of White and Wooldrige[1990] theorem 3.5, which assumes that the tails of the stationary distribution decrease to 0 faster than exponential function. We only require that they decrease like $b_0 \exp(-b_1 x^\alpha)$ for $x \to \infty$ for some $\alpha > 0$ (not $\alpha > 1$ like in White and Wooldrige). The second new result reprensents a variation of the theorem of White and Wooldridge[1990].

1.6.1.1 Theorem: Generalization of the Bernstein Inequality to Unbounded Random Variables

Let $(\mathcal{E}_t)_{-\infty<t<+\infty}$, be a stationary stochastic process with zero mean, $E(\mathcal{E}_t) = 0$, and satisfying an α-mixing condition with exponentially decreasing mixing coefficients. Suppose

$$Pr\left(|\mathcal{E}_t| > x\right) \le b_0 \exp(-b_1 x^\alpha) \quad \text{for all} \;\; x \tag{1.245}$$

for some b_0, b_1 and $\alpha > 0$. Then, there exist some constant d_1, d_2 such that for all sufficiently large N, C and $\delta > 0$

$$Pr\left(\left|\Sigma_{t=1}^N \mathcal{E}_t\right| > CN^{\frac{1}{2}+\delta}\right) \le d_1 \exp\left(-d_2 N^{\frac{\delta\alpha}{(1+\alpha)}}\right). \tag{1.246}$$

The constant d_1, d_2 are not depending on N.

Proof

We truncate $\mathcal{E}_t$ at some bound $M_N > 0$ that will be specified later, and set

$$\begin{aligned} \bar{\mathcal{E}}_{t,N} &= \mathcal{E}_t - \min\left(\mathcal{E}_t, M_N\right), \;\; \tilde{\mathcal{E}}_{t,N} = \max\left(\mathcal{E}_t, -M_N\right), & (1.247)\\ \mathcal{E}_{t,N} &= \max\left(\min\left[\mathcal{E}_t, M_N\right)\right], -M_N) = \mathcal{E}_t - \bar{\mathcal{E}}_{t,N} - \tilde{\mathcal{E}}_{t,N}. & (1.248) \end{aligned}$$

a) The $\mathcal{E}_{t,N}$ are bounded by M_N in absolute value. As functions of finitely many observations from a stationary mixing process, they also form a stationary process with the same type of mixing behaviour (compare, e.g., theorem 3.4.a of White, [1984). Therefore the $\mathcal{E}_{t,N}$ represent a stationary process with exponentially decreasing α-mixing coefficients too. If we center them around 0, we may apply Bosq´s [1975] Bernstein inequality for bounded mixing time series in the version of theorem 3.3 of White and Wooldrige [1990], and we get

$$Pr\left(\left|\Sigma_{t=1}^N \left(\mathcal{E}_{t,N} - E(\mathcal{E}_{t,N})\right)\right| > \Delta\right) \le C_1 \exp\left(-C_2 \frac{\Delta}{\sqrt{N} M_N}\right) \tag{1.249}$$

for all $\Delta > 0$ with constants C_1, C_2 not depending on N.

b) By definiton, $\bar{\mathcal{E}}_{t,N} \ge 0$, and $\bar{\mathcal{E}}_{t,N} > 0$ iff[4] $\mathcal{E}_t > M_N$.

[4] iff=if and only if.

Therefore, for all $\Delta > 0$,

$$\begin{aligned} Pr\left(|\Sigma_{t=1}^{N}\bar{\mathcal{E}}_{t,N}| > \Delta\right) &\leq Pr\left(\Sigma_{t=1}^{N}\bar{\mathcal{E}}_{t,N} > 0\right) && (1.250)\\ &\leq Pr\left(\bar{\mathcal{E}}_{t,N} > 0 \text{ at least for one } t = 1, 2, .., N\right) && (1.251)\\ &\leq \Sigma_{t=1}^{N} P\left(\bar{\mathcal{E}}_{t,N} > 0\right) = N\bar{P}_N && (1.252)\end{aligned}$$

as the $\bar{\mathcal{E}}_{t,N}$ are identically distrbuted, where

$$\bar{P}_N = Pr\left(\bar{\mathcal{E}}_{1,N} > 0\right) = Pr\left(\mathcal{E}_1 > M_N\right) \leq b_0 \exp\left(-b_1 M_N^{\alpha}\right). \qquad (1.253)$$

Together, we have

$$Pr\left(\left|\Sigma_{t=1}^{N}\bar{\mathcal{E}}_{t,N}\right| > \Delta\right) \leq b_0 N \exp\left(-b_1 M_N^{\alpha}\right). \qquad (1.254)$$

Analogously, it can proved that

$$Pr\left(\left|\Sigma_{t=1}^{N}\tilde{\mathcal{E}}_{t,N}\right| > \Delta\right) \leq N\tilde{p}_N \leq b_0 N \exp\left(-b_1 M_N^{\alpha}\right) \qquad (1.255)$$

where

$$\tilde{p}_N := Pr\left(|\tilde{\mathcal{E}}_{1,N}| > 0\right) = Pr\left(\tilde{\mathcal{E}}_1 < -M_N\right). \qquad (1.256)$$

c) As $E(\mathcal{E}_t) = 0$, we have

$$E(\mathcal{E}_{t,N}) = -E(\bar{\mathcal{E}}_{t,N}) - E(\tilde{\mathcal{E}}_{t,N}) = -E(\bar{\mathcal{E}}_{1,N}) - E(\tilde{\mathcal{E}}_{1,N})$$

by stationarity. As $\mathcal{E}_{t,N} \geq 0$, we have

$$\begin{aligned} E(\bar{\mathcal{E}}_{t,N}) &= \int_0^{+\infty} P(\bar{\mathcal{E}}_{t,N} > x)dx = \int_0^{+\infty} Pr(\mathcal{E}_t - M_N > x)dx && (1.257)\\ &= \int_{M_N}^{+\infty} Pr(\mathcal{E}_t > x)dx \leq b_0 \int_{M_N}^{+\infty} \exp(-b_1 x^{\alpha})dx && (1.258)\\ &= o\left(\exp(-b_1 M_N^{\beta})\right) && (1.259)\end{aligned}$$

for all $0 < \beta < \alpha$, where the latter relation follows easily from de l'Hospital's rule. A similar argument applies to $E(\tilde{\mathcal{E}}_{t,N})$, and we get

$$\left|\Sigma_{t=1}^{N} E(\mathcal{E}_{t,N})\right| = N\left|E(\mathcal{E}_{1,N})\right| = o\left(N \exp\left[-b_1 M_N^{\beta}\right]\right). \qquad (1.260)$$

d) Now we choose $\Delta = CN^{\frac{1}{2}+\delta}$ and $M_N = N^\gamma$ for some $\delta > 0$ with

$$\gamma = \frac{\delta}{1+\alpha} < \delta. \tag{1.261}$$

From (1.260), we have that $\left|\Sigma_{t=1}^N E\left(\mathcal{E}_{t,N}\right)\right|$ decreases faster to 0 than the multiple of $\exp\left(-b_1 M_N^\beta\right)$ for all $0 < \beta < \alpha$. Therefore, it is negligeable compared to Δ, and we get from (1.249) for suitable constants C_1, C_2 (not necessarily the same as in (1.249):

$$P\left(\left|\Sigma_{t=1}^N \mathcal{E}_{t,N}\right| > CN^{\frac{1}{2}+\delta}\right) \le C_1 \exp\left(-C_2 N^{\delta-\gamma}\right). \tag{1.262}$$

By (1.254) and (1.255), the large deviations of $\Sigma_{t=1}^N \bar{\mathcal{E}}_{t,N}$ and $\Sigma_{t=1}^N \tilde{\mathcal{E}}_{t,N}$ have probabilities of order $\exp\left(-b_1 N^{\gamma\alpha}\right)$ which is of the same order as (1.262) as by our choice of γ, we have

$$\delta - \gamma = \delta - \frac{\delta}{1+\alpha} = \delta\frac{\alpha}{1+\alpha} = \gamma\alpha. \tag{1.263}$$

Here, we use that

$$N\exp\left(-bN^\beta\right) = O\left(exp\left[-b'N^\beta\right]\right) \quad \forall\beta > 0, 0 < b' < b. \tag{1.264}$$

Therefore, (1.254), (1.255) and (1.262) together imply

$$\begin{aligned} P\left(\left|\Sigma_{t=1}^N \mathcal{E}_t\right| > CN^{\frac{1}{2}+\delta}\right) &\le d_1 exp\left(-d_2 N^{\delta-\gamma}\right) && (1.265)\\ &\le d_1 exp\left(-d_2 N^{\frac{\delta\gamma}{(1+\alpha)}}\right) && (1.266)\end{aligned}$$

for appropriately chosen constants d_1, d_2 depending on b_0, b_1 but not on N.♢

1.6.1.2 Corollary

Under the condition of the theorem 1.6.1.1, we have for $\Delta_N, M_N \to +\infty$

$$Pr\left(\left|\Sigma_{t=1}^N \mathcal{E}_t\right| > \Delta_N\right) \le C_1 \exp\left(-C_2 \frac{\Delta_N}{\sqrt{N} M_N}\right) + b_0 N \exp\left(-b_1 M_N^\alpha\right) \tag{1.267}$$

for some constant C_1, C_2 independent of M_N provided that

$$N\exp\left(-b_1 M_N^\beta\right) = o(\Delta_N) \quad \text{for some } 0 < \beta < \alpha. \tag{1.268}$$

Proof:
The result follows from the proof of the theorem, relation (1.249), (1.254) and

(1.255) with $\Delta = \Delta_N$ where we take into account that either $\bar{\mathcal{E}}_{t,N} = 0$ or $\tilde{\mathcal{E}}_{t,N} = 0$. The last assumption guaranties that $\left|\Sigma_{t=1}^N E(\mathcal{E}_{t,N})\right|$ is negligeable compared to Δ_N. Mark that by theorem 3.3 of White and Wooldridge[1990], C_1, C_2 do not depend on N even for $M_N \to +\infty$ $\diamondsuit$.

For the intended application, we need a more general version of the corollary (1.6.1.2), which, however, is proved exactly along the same lines of arguments.

1.6.1.3 Theorem

For each $N = 1, 2, ...$, let $\{\mathcal{E}_t(N)\}$ be a stationary stochastic process with zero mean, $E(\mathcal{E}_t(N)) = 0$, satisfying an α-mixing condition with exponential decreasing mixing coefficients. Suppose for all $N = 1, 2, ...$, that

$$Pr(\mathcal{E}_t(N) > x) \leq b_0 \exp(-f_N(x)) \quad \forall x \geq M_N \tag{1.269}$$

for some sequence $M_N \to +\infty$ and functions $f_N(x) \geq 0, x \geq M_N$ which are increasing and $f_N(x) \to +\infty$ for $x \to +\infty$.
Then, there are some constant C_1, C_2 not depending on N, such that for all large enough N

$$Pr\left(\left|\Sigma_{t=1}^N \mathcal{E}_t(N)\right| > \Delta_N\right) \leq C_1 \exp\left(-C2\frac{\Delta_N}{\sqrt{N}M_N}\right) + Nb_0 \exp(f_N(M_N)) \tag{1.270}$$

where $\Delta_N \to +\infty$ such that $NE_N(M_N) = o(\Delta_N)$ for $N \to +\infty$ where

$$E_N(v) = \int_v^{+\infty} \exp(-f_N(u))\, du \tag{1.271}$$

Proof:
The proof follows exactly as the proof of the Theorem (1.6.1.1) and we use the notation of that proof. The crucial Theorem 3.3 of White and Wooldrige[1990] holds also for a sequence of bounded stochastic processes. Therefore, the right hand side of (1.249), remains unchanged. The analogous results to (1.254) and (1.255) follow exactly as in part b) of the proof of Theorem (1.6.1.1), using the more general tail condition(1.269). Finally, as in part c) of that proof

$$E\left(\bar{\mathcal{E}}_{t,N}(N)\right) \leq b_0 \int_{M_N}^{+\infty} \exp(-f_N(u))\, du = b_0 E_N(M_N), \tag{1.272}$$

and, therefore, our last assumption guaranties that $\Sigma_{t=1}^N \mathcal{E}_{t,N}(N)$ is negligeable compared to $\Delta_N \diamondsuit$.

1.6.2 Theorem: Variation of a Theorem by White and Wooldridge

Let $(S_t, Z_{t-1})_{-\infty<t<+\infty}$, be a stationary stochastic process, $S_t \in \Re, Z_{t-1} \in \Re^d$. Let μ denote the stationary distribution of Z_{t-1}.
Suppose

$$Pr\left(||Z_{t-1}|| > x\right) \leq \beta_0 \exp\left(-\beta_1 ||x||^\tau\right) \quad \forall x \tag{1.273}$$

for some β_0, β_1 and $\tau > 0$.
Let Θ_n be a compact set of continuous functions in $L^2(\mu)$ satisfying for some $\Delta_n > 0$

$$|\theta(x)| \leq \ \Delta_n \quad \forall x, \ \forall \theta \in \Theta_n. \tag{1.274}$$

Assume further that for all δ_n there exist open subsets O_{n_i} for $i = 1, 2, ..., K(\delta_n)$, of Θ_n and $\theta^*_{in} \in O_{n_i}$

$$\Theta_n = O_{n_1} \cup O_{n_2} \cup ... \cup O_{n_{K(\delta_n)}}$$

and such that for some constant C_0 and all ρ we have

$$\sup_{||x||\leq\rho} |\theta(x) - \theta^*_{in}(x)| \leq C_0\left(1 + \Delta_n \rho\right)\delta_n \quad \forall \theta \in O_{n_i}. \tag{1.275}$$

Let $g: \ \Theta_n \times \Re^{r+1}$ be a measurable function, and denote

$$S_n(\theta) = \sum_{t=1}^{n} g(\theta, S_t, Z_{t-1}). \tag{1.276}$$

Assume that there are functions $\gamma_n(\epsilon)$ such that

$$Pr\left(|S_n(\theta) - E(S_n(\theta))| \geq \epsilon\right) \leq \gamma_n(\epsilon) \quad \forall \theta \in \Theta_n, \ \epsilon > 0, \tag{1.277}$$

and random variables M_{nt} with $\mu_n^2 := E(M_{nt}^2) < \infty$ such that:
$\forall \ \theta; \theta^* \in \Theta_n$,

$$|g(\theta, S_t, Z_{t-1}) - g(\theta^*, S_t, Z_{t-1})| \ \leq \ M_{nt} \left|\theta(Z_{t-1}) - \theta^*(Z_{t-1})\right|. \tag{1.278}$$

Then, for all $\epsilon, \rho > 0$ and all n sufficiently large,

$$\begin{aligned} Pr\left(\sup_{\theta\in\Theta_n} |S_n(\theta) - E(S_n(\theta))| > \epsilon\right) \ &\leq \ k(\delta_n)\gamma_n(\epsilon) \\ &+ \ k(\delta_n)\frac{4n\mu_n}{\epsilon} C_0(1 + \Delta_n\rho)\delta_n \\ &+ \ k(\delta_n)\frac{4n\mu_n}{\epsilon}\sqrt{2\beta_0}\Delta_n exp\left(-\frac{\beta_1}{2}{\delta_n}^2\right). \end{aligned}$$

If for some sequences a_n, δ_n we have for $n \to \infty$

$$\begin{cases} K(\delta_n)\gamma_n(\epsilon a_n) \to 0, \\ K(\delta_n)\frac{n\mu_n}{\epsilon a_n}(1 + \Delta_n \delta_n) \to 0, \\ K(\delta_n)\frac{n\mu_n}{\epsilon a_n}\Delta_n exp\left(-\frac{\beta_1}{2}\delta^2\right) \to 0 \end{cases}$$

then

$$Pr\left(\sup_{\theta \in \Theta_n} |S_n(\theta) - E(S_n(\theta))| > \epsilon a_n\right) \to 0 \text{ for } n \to \infty \quad \forall \epsilon > 0. \tag{1.279}$$

Proof:
For $\theta, \theta^* \in \Theta_n$, we use the abbreviations

$$G_t = g(\theta, S_t, Z_{t-1}) \tag{1.280}$$

$$G_t^* = g(\theta^*, S_t, Z_{t-1}). \tag{1.281}$$

For any ϵ we have

$$Pr\left(\sup_{\theta \in \Theta_n} |S_n(\theta) - E(S_n(\theta))| > \epsilon\right) \leq Pr\left(\max_{1 \leq i \leq K(\delta_n)} \sup_{\theta \in O_{n_i}} P\,|S_n(\theta) - E(S_n(\theta))| \geq \epsilon\right).$$

As

$$Pr\left(\sup_{\theta \in O_{n_i}} |S_n(\theta) - E(S_n(\theta))| \geq \epsilon \text{ for some } i \leq K(\delta_n)\right) \leq$$
$$\sum_{i=1}^{K(\delta_n)} P\left(\sup_{\theta \in O_{n_i}} |S_n(\theta) - E(S_n(\theta))| \geq \epsilon\right) \tag{1.282}$$

we have

$$|S_n(\theta) - E(S_n(\theta))| = \left|\sum_{i=1}^{n}(G_t - E(G_t))\right| \leq \tag{1.283}$$
$$\sum_{i=1}^{n}|G_t - G_t^* - E(G_t - G_t^*)| +$$
$$\left|\sum_{i=1}^{n}(G_t^* - E(G_t^*)).\right| \tag{1.284}$$

The second term does not depend on θ such that for i fixed, $\theta^* = \theta^*_{in}$

$$\begin{aligned} Pr\left(\sup_{\theta\in O_{n_i}} |S_n(\theta) - E(S_n(\theta))| > \epsilon\right) &\leq \qquad (1.285) \\ & Pr \quad \left(\left|\sup_{\theta\in O_{n_i}} \sum_{t=1}^{n} |G_t - G^*_t - E(G_t - G^*_t)\right| > \epsilon\right) + \\ & Pr \quad \left(\left|\sum_{t=1}^{n} (G^*_t - E(G^*_t)\right| > \epsilon\right). \qquad (1.286) \end{aligned}$$

Using Markov's inequality, we have that the first term on the right hand side of (1.286) is bounded by

$$\begin{aligned} \frac{1}{\epsilon} E\left(\sup_{\theta\in O_{n_i}} \sum_{i=1}^{n} |G_t - G^*_t - E(G_t - G^*_t)|\right) &\leq \frac{1}{\epsilon}\sum_{t=1}^{n} E\left(\sup_{\theta\in O_{n_i}} |G_t - G^*_t - E(G_t - G^*_t)|\right) \\ &= \frac{n}{\epsilon} E\left(\sup_{\theta\in O_{n_i}} |G_t - G^*_t - E(G_t - G^*_t)|\right) \\ &= \frac{2n}{\epsilon} E\left(\sup_{\theta\in O_{n_i}} |G_t - G^*_t|\right) \end{aligned}$$

where we have used the stationary of (S_t, Z_{t-1}) for the second line. By assumption(1.277) and the Cauchy-Schwarz inequality we finally get

$$\begin{aligned} Pr \quad &\left(\sup_{\theta\in O_{n_i}} \sum_{t=1}^{n)} |G_t - G^*_t - E(G_t - G^*_t)| > \epsilon\right) \leq \\ &\frac{2n}{\epsilon} E\left(M_{nt} \sup_{\theta\in O_{n_i}} |\theta(Z_{t-1} - \theta^*(Z_{t-1}|\right) \qquad (1.287) \\ \leq \quad &\frac{2n}{\epsilon} E\left(M^2_{nt}\right)^{1/2} \left(E \sup_{\theta\in O_{n_i}} |\theta(Z_{t-1} - \theta^*(Z_{t-1}|^2\right)^{1/2}. \qquad (1.288) \end{aligned}$$

Let $\rho > 0$. Using the boundeness of all $\theta \in \Theta_n$ and a truncation argument

$$\begin{aligned} E\left(\sup_{\theta\in O_{n_i}} |\theta(Z_{t-1}) - \theta^*(Z_{t-1})|^2\right) &\leq \sup_{\theta\in O_{n_i}} \sup_{||x||\leq\delta} |\theta(x) - \theta^*(x)|^2 + 2\Delta^2_n Pr\left(||z_{T-1}|| \geq \rho\right) \\ &\leq C^2_0 (1 + \Delta^2\rho)^2 {\delta_n}^2 + 2\Delta^2_n \beta_0 \exp(-\beta_1\rho^2) \end{aligned}$$

by assumptions. Putting(1.286) and (1.289) together and using assumption (1.277) we get

$$\begin{aligned} Pr\left(\sup_{\theta\in O_{n_i}} |S_n(\theta) - E(S_n(\theta))| \geq \epsilon\right) &\leq \frac{4n}{\epsilon}\mu_n \left(C_0(1 + \Delta_n\rho)\delta_n + 2\sqrt{2\beta_0}\Delta_n \exp(-\frac{\beta_1}{2}\rho^2)\right) \\ &+ \gamma(\epsilon). \end{aligned}$$

As the right-hand side does not depend on i, we finally get from (1.277)

$$Pr \left(\sup_{\theta \in O_{n_i}} |S_n(\theta) - E(S_n(\theta))| \geq \epsilon \right) \leq K(\delta_n)\gamma_n(\epsilon)$$

$$+ \quad K(\delta_n)\frac{4n\mu_n}{\epsilon}\left(C_0(1 + 1 + \Delta_n\rho)\delta_n + 2\sqrt{2\beta_0}\Delta_n \exp(-\frac{\beta_1}{2}\rho^2)\right) + \gamma(\epsilon)\diamondsuit$$

The following lemma guaranties that the set $\Theta_n = ANN(\Psi, q_n, \Delta_n)$ of neural network functions satisfies the compactness assumptions of theorem 1.6.2.
It is a variation of lemma 4.3 of White[1990] which provides an upper bound for the metric entropy of Θ_n with respect to the supremum norm over a compact set.

1.6.2.1 Lemma

Let Ψ be bounded in absolute value by 1 and satisfy a Lipschitz condition i.e

$$|\Psi(u) - \Psi(v)| \leq L|u - v| \quad \forall u, v \in \Re. \tag{1.289}$$

Consider $\Theta_n = ANN(\Psi, q_n, \Delta_n)$ as a subset of $L^2(\mu)$ for some probability measure μ on $\Re^r$. Then, there exists for all $\eta > 0$ open subsets $O_i, i = 1, 2, ..., k(\eta)$, of Θ covering Θ, i.e.

$$\Theta = O_1 \cup O_2 \cup ... \cup O_{k(\eta)},$$

and there are $\theta_i^* \in O_i$ such that for all $\rho \geq 1$ and with $L_1 = max(L, 1)$

$$\sup_{||X|| \leq \rho} |\theta(x) - \theta_i^*(x)| \leq L_1(1 + \Delta\rho).\eta \qquad \forall \theta \in O_i. \tag{1.290}$$

Moreover, we have

$$K(\eta) \leq 4\left(\frac{2\Delta}{\eta}\right)^{q(r+2)+1} q^{q(r+1)} \tag{1.291}$$

Proof: Let

$$V = \left\{ v \in \Re^{r+1}; \sum_{i=0}^{q} |v_i| \leq \Delta \right\}$$

and

$$W = \left\{ w \in \Re^{q(r+1)}; \sum_{k=1}^{q}\sum_{i=0}^{r} |w_{ki}| \leq \Delta \right\}$$

be the set of weight vectors of network functions in Θ and let $V \times W$ the network parameter set corresponding to functions in Θ. For $\eta > 0$, let V_η be an η-net for V with respect to the l_1-norm, i.e a subset

$$V_\eta = \left\{ v_1^*, v_2^*, ..., v_q^* \right\} \subset V$$

such that for any $v \in V$ there is a $v_i^* \in V_\eta$ with

$$||v - v_i^*||_1 = \sum_{k=0}^{q} |v_k - v_{ik}^*| < \eta. \tag{1.292}$$

Let

$$W_\eta = \{w_1^*, w_2^*, ..., w_M^*\}$$

be a corresponding defined η-net for W_η, and, the $V_\eta \times W_\eta$ is an η-net for $V \times W$. Consider $\theta \in \Theta$ with weight vectors u, w. There are $v^* \in V_\eta$, $w^* \in W_\eta$ such that

$$\sum_{k=0}^{q} |v_k - v_k^*| < \eta \quad , \quad \sum_{k=0}^{q} \sum_{i=0}^{r} |w_{ki} - w_{ki}^*| < \eta. \tag{1.293}$$

Let $\theta^* \in \Theta$ be the network function with weights v^*, w^*. Then,

$$\begin{aligned}
|\theta(x) - \theta^*(x)| &\leq |v_0 - v_0^*| + \sum_{k=1}^{q} |v_k - v_k^*| + \\
&\qquad \sum_{k=1}^{q} |v_k^*||\Psi(\tilde{x}^T w_k) - \Psi(\tilde{x}^T w_k^*) && (1.294)\\
&\leq \eta + \Delta \sum_{k=1}^{q} |\Psi(\tilde{x}^T w_k) - \Psi(\tilde{x}^T w_k^*)| && (1.295)\\
&\leq \eta + \Delta L \sum_{k=1}^{q} |\tilde{x}.(w_k - w_k^*)| && (1.296)\\
&\leq \eta + \Delta L \sum_{k=1}^{q} \left(\sum_{i=1}^{r} |x_i||\tilde{x}|w_{k0} - w_{k0}^*| \right) && (1.297)\\
&\leq \eta + \Delta L \eta \rho = (1 + L\Delta\rho)\,\rho \leq L_1(1 + \Delta\rho)\eta && (1.298)
\end{aligned}$$

for all $x \in \Re^r$ with $||x|| \leq \rho$.
Let

$$\left\{ \left(v^*(j), w_1^*(j), w_1^*(2), ..., w_q^*(j) \right), \quad j = 1, 2, ..., k(\eta) \right\} = V_\eta \times W_\eta$$

be an enumeration of the weight vectors in the η-net $V_\eta \times W_\eta$. Let $\theta_1^*, \theta_2^*, ..., \theta_{k(\eta)}^*$ be the network function with weights in $V_\eta \times W_\eta$, and let

$$O_j = \left\{ \theta \in \Theta; \sum_{i=0}^{q} |v_i - v_i^*(j)| < \eta \quad , \quad \sum_{k=1}^{q} \sum_{i=0}^{r} |w_{ki} - w_{ki}^*(j)| < \eta. \right\} \tag{1.299}$$

As $V_\eta \times W_\eta$ is an η-net for the set of weight vectors $V \times W$ is an η-net of the functions in Θ, we have

$$O_1 \cup O_2 \cup ... \cup O_{K(\eta)} = \Theta,$$

and we have just shown that

$$\sup_{||x|| \leq \rho} |\theta(x) - \theta_i^*(x)| \leq L_1(1 + \Delta\rho)\eta \quad \forall \theta \in O_i. \tag{1.300}$$

Now $K(\eta)$ is the number of elements in $V_\eta \times W_\eta$. For this, we can use the upper bound derived in the proof 4.3 of White[1990] and get

$$K(\eta) \leq 4\left(\frac{2\Delta}{\eta}\right)^{q(r+2)+1} q^{q(r+1)} \quad \diamondsuit \tag{1.301}$$

1.7 Financial Applications

Throughout this section, the simulations have been done with real financial data and illustrate the goodness and accuracy of the proposed Value-at-Risk methodology via the computation of the daily VaR of a one COMMERZBANK share. As explanatory variables, the daily closing prices of DEUTSCHE Bank, the ones of BASF, SIEMENS and the DAX30 (all traded on the Frankfurt stock exchange) have been used. The back testing results are quite successful. Therefore follows the conclusion that ANN and EVT represent extremely powerful tools to the special task of daily market risk measurement without the need on making any of the questionnable assumptions underlying current Value-at-Risk methodologies. At the closing day of each considered period, the method is applied using the 255 previous closing prices and setting the threshold level u of the innovation as the 90^{th} sample percentile of the fitted residual $\hat{\mathcal{E}}$. τ is equal to 5 which means that we use the 5 previous closing price to forecast the future market value. The Value-at-Risk estimation is then back tested by comparing the estimates with the actual losses observed on the next day. The goodness of the estimation procedure is then measured by computing the number of violation throughout the back testing. Only 3 violations have been observed for a period of 577 trading days.

Chapter 2

Financial Forecasting via Non-parametric AR-GARCH and Artificial Neural Networks

2.1 Introduction

Forecasting financial stock prices, predicting daily returns or modelling stochastic volatilities have been a very active area of research in recent years. Several financial, statistical and econometrical theories attempting to explain the features, the patterns and the dynamics of stock prices have been largely elaborated by traders, academic and market makers. Due to the huge and complex sets of random technical indicators that are driving the dynamics of stock prices, modelling or forecasting financial markets behaviours still remain a very difficult task. From the point of view of market makers or traders, the returns distribution through a day is a very important statistic not solely for the information contents it might carry but also because it might help him to anticipate the market trends and execute orders at some better prices. For hedging against risk, efficient portfolio management via accurate forecast of conditional returns and reliable volatility estimates are crucial for adopting optimal trading strategies with increasing margins. The stylised facts (non-linearity, skewness, fat tails, volatility clustering, leverages effects, co-movements in volatility) of the existing financial returns models, from the Integrated Autoregressive Moving Average models (ARIMA) to the General Autoregressive and Conditionally Heteroskedastic (GARCH) and others stochastic volatility models including those by Bera and Higgins [1995], Bollerslev, Chou and Kroner [1992], Engle and Nelson [1994], Ghysels, Harvey and Renault [1996] provide serious reasons for thinking about elaborating alternative forecasting statistical methods. Non-parametric AR-GARCH, combined with ANN enable to correct some of these stylised facts. To overcome the non-linearity, the skewness or the heteroscedasticity that financial time series are usually displaying, one can use of the theory of ANN, and take profit of the universal approximating power and denseness properties of neural output functions. Neural output functions represent a powerful tool for estimating conditional expected returns and the stochastic volatilities of financial securities while using a fully non-parametric model. ANN can be equated as a black box consisting of some computing systems containing many simple non-linear processing units or nodes interconnected by synaptic links. ANN is a well-tested method for financial analysis on the stock market (see Franke [1999], White [1989,1990], Jingtao and Poh [1998], Chapman [1994]). During the last decades, ANN have been actively used for financial stock trading: forecasting stock prices (see Freisleben: Stock Market Prediction with back-propagation networks [1992]), trading patterns recognitions (see Tanigawa: Stock Price Pattern matching system, [1992]), rating of corporate bonds (see Dutta and Shekar: Bond Rating, [1990]) or hedging and trading derivative products (see Hutchinson, Poggion: A non-parametric approach to pricing and hedging derivative securities via learning networks [1994]). The research fund for ANN applications from financial institutions is the second largest (see Trippie and Turban [1996]: Neural Network in Finance and Investing). For example, the American defence department invests $400 millions in a six-year project, and Japan has $250 millions ten-year-neural-computing project (see The Economist, April 1995: More in a Cockroach's brain than your computers dreams).

The purpose of this chapter is to implement a forecasting algorithm that enables to predict future stock prices of a given security by estimating the conditional expected returns while taking into account the stochastic feature of the conditional volatility of the considered financial instrument. Under the same setting, the associated market risk exposure will also be computed via the calculation of the conditional VaR without making any normality assumptions and also without referring to any linear dependence of the portfolio values with respect the underlying risk elements. After the description of the financial returns model, the first section consists on estimating the conditional expected returns by means of neural output functions as in the previous chapter. The non-parametric ARMA-GARCH algorithm of Mc-Neil and Buehlmann [2000] will implemented in order to derive the corresponding estimates of the time dependent conditional volatilities. For matter of consistency, some smoothing regularity or contraction properties can be imposed on the volatility regression function. Instead of using the contraction assumptions as implemented in Mc-Neil and Buehlmann[2000], we use the convergence results of Corradi and White dealing with Regularized ANN [1995]. Corradi and White have shown that Regularized ANN are capable of learning and approximating (on compacta) elements of certain Sobolev space at a non-parametric rate that optimally exploits the smoothness properties of the unknown mapping. If the unknown mapping has an order of differentiability equal to m, the mean square error of the estimation procedure reaches zero at the rate of $n^{\frac{-2m}{2m+1}}$. Therefore such regularity assumptions enable to build consistent volatility estimates.

The market settings that will be imposed on the unexpected returns that are underlying the financial returns model, justify the existence of some additive volatility noise and lead to the estimation of the stochastic volatility as a regression function of the squared conditional returns centred with the ANN estimates of the conditional expected returns. To build sufficiently accurate estimates of the volatility regression function, on can use of the standard GARCH volatility estimation procedure (see Tim Bollerslev [1992]) to provide the starting volatility parametric estimates used for the initialisation of the non-parametric ARMA-GARCH algorithm. In the second section, we recall the standard ARMA-GARCH forecasting algorithm that will provide the starting estimates. The third section is dealing with the consistency of the resulting estimators. In this section, based on some regularity assumptions imposed on the volatility regression function, we combine the use of Regularized ANN and apply Luka's theorem (see Corradi and White [1995]) to derive the convergence rate of our estimation procedure. We also provide some aggregate market risk analysis by estimating the VaR under the assumptions that the unexpected returns are heavy tailed and heteroskedastic and follow some GPD above a specific threshold. Beside this aggregate market risk analysis, we also provide some options pricing formula based on non-parametric AR-GARCH and ANN. In this subsection, it will be shown how one can use Bootstrapping Algorithms for comparing the ANN based option pricing methodologies and the well known Black-Scholes option pricing formula. The goodness of the ANN based daily VaR methodology will be illustrated via the computation of the daily VaR of a position on the DAX30 stock index traded on the Frankfurt stock exchange. As explanatory variables, the Deutsche Bank daily closing prices, the ones of Commerzbank , BASF and SIEMENS will be used.

2.1.1 Security Price Model

$$\begin{cases} S_t = \ \mu_t \ + \ \sigma_t \mathcal{E}_t \\ \\ \mu_t := \mu\left(S_{t-1}, ..., S_{t-\tau}, X_{t-1}\right) \\ \\ \sigma_t^2 := \sigma\left(S_{t-1} - \mu_{t-1}, ..., S_{t-p} - \mu_{t-p}, \sigma_{t-1}^2, \sigma_{t-2}^2, ..., \sigma_{t-q}^2\right) \end{cases} \tag{2.1}$$

where

$$\left\{ \begin{array}{l} \mu_t = E\left(S_t|\mathcal{F}_{t-1}\right) \\ \sigma_t^2 = Var\left(S_t|\mathcal{F}_{t-1}\right) \end{array} \right. \tag{2.2}$$

and

- S_t:= Financial return[1]of the t^{th} trading period
- X_{t-1}:= Available market information for the considered trading period
- $\mathcal{F}_t$:= Set of all market information up to the t^{th} trading period.

The model (2.1)-(2.2) is known as a non-parametric AR-GARCH, where the conditional expected return μ is modelled as a nonparametric autoregressive process of order τ whose dynamic is governed by an unspecified non-linear functional of some lagged returns of the stock, while the conditional stochastic volatility regression function σ is assumed to be sufficiently smooth to enable the application of Luka's theorem (see Corradi and White [1995] about Regularized ANN). Instead of assuming such regularity conditions, one can also impose (see Mc-Neil and Buehlmann [2000]) some contraction properties on the volatility regression function. under the following market assumptions:
The unexpected returns $(\mathcal{E}_t)_{t\in\mathcal{Z}}$ that are driving the market randomness are drawn from a time series consisting of iid random variables verifying:

$$\left\{ \begin{array}{l} 1^\circ)\ \ E(\mathcal{E}_t) = 0, \\ \\ 2^\circ)\ \ Var(\mathcal{E}_t) = 1, \\ \\ 3^\circ)\ \ E(\mathcal{E}_t^4)\ <\ \infty, \\ \\ 4^\circ)\ \ \forall t\ \in\ \mathcal{Z},\ \mathcal{E}_t\ \text{ is independent to }\ \mathcal{F}_{t-1}. \end{array} \right. \tag{2.3}$$

In the aim to predict the most accurately the future market value S_t of a given holding at the t^{th} trading period, the theory of ANN as used in the previous and described in Franke [1999] and suggested in White [1990], Hornick [1989] or Tan and Poh [1998] for estimating consistently and non parametrically the regression function μ and the conditional volatility function σ will be used.
After training the resulting networks e.g. estimating the regression function μ , follows the forecast of the associated conditional stochastic volatility. The forecasting will be done using the conditional centred lagged returns $S_{t-1} - \mu_{t-1}$

[1] $S_t = log\left(\dfrac{Price_t}{Price_{t-1}}\right)$ or $S_t = \left(\dfrac{Price_t - Price_{t-1}}{Price_{t-1}}\right)$

$S_{t-2}-\mu_{t-2},...,S_{t-p}-\mu_{t-p}$ and the unobservable conditional volatilities $\sigma^2_{t-1},\sigma^2_{t-2}$ ..., andσ^2_{t-q}. The Nonparametric ARMA-GARCH algorithm, due to Mc-Neil and Buehlmann [2000], that will be described in the following sections, overcomes the difficulties resulting from the non observability of $(\mu_t)_{t\in\mathcal{Z}}$ and $\left(\sigma^2_t\right)_{t\in\mathcal{Z}}$ and helps to estimate consistently and non-parametrically the two regression functions μ and σ.

First, one can start estimating the conditional mean in the line of the previous chapter. This step provides consistent estimates of the regression function μ e.g $(\hat{\mu}_t)_{t=1,2,...n}$. After getting sufficiently accurate estimates for the conditional expected returns, on can then use the centred returns $S_t - \hat{\mu}_t$, to implement the nonparametric GARCH estimation procedure (see Mc-Neil and Buehlmann [2000]) and derive the corresponding stochastic volatility estimates.

Due to the importance of the accuracy of the starting estimates in the algorithm, to initialise it, we choose the best estimators between some optimal neural output functions and the one provided by some standard linear and parametric GARCH (see T.Bollersev and R.Baillie [1992]) or the GARCH predictor of John Knight and Stephen.E.Satchell [1998].

Under the market settings (2.1), (2.2) and (2.3), the stochastic process $(V_t)_{t\in\mathcal{Z}}$ defined by:

$$V_t := \sigma^2\left(S_{t-1}-\mu_{t-1},...,S_{t-p}-\mu_{t-p},\sigma^2_{t-1},\sigma^2_{t-2},...,\sigma^2_{t-q}\right)\times\left[\mathcal{E}^2_t\ -\ 1\right] \qquad (2.4)$$

can be equated as the random noise driving the uncertainty of the volatility.This statement can be proved by the following proposition:

2.1.1.1 Proposition

Under $(2.1)\,,(2.2)$ and (2.3), the process $(V_t)_{t\in\mathcal{Z}}$ can be equated as a martingale difference and:

$$\begin{cases} E\left(V_t\right)=0, \\ \\ Cov\left(V_t,V_s\right)=0 \quad \forall(t,s) \ \text{ such that } \ t\ <\ s. \end{cases} \qquad (2.5)$$

Proof

$\forall t\in\ \mathcal{Z}$,

$$E\left(V_t\right)=E\left[E\left(V_t|\mathcal{F}_{t-1}\right)\right]. \qquad (2.6)$$

Since $\sigma\left(S_{t-1}-\mu_{t-1},...,S_{t-p}-\mu_{t-p},\sigma^2_{t-1},\sigma^2_{t-2},...,\sigma^2_{t-q}\right)$ is $\mathcal{F}_{t-1}$ measurable, it can be derived that:

$$E\left(V_t|\mathcal{F}_{t-1}\right)=\sigma^2\left(S_{t-1}-\mu_{t-1},...,S_{t-p}-\mu_{t-p},\sigma^2_{t-1},\sigma^2_{t-2},...,\sigma^2_{t-q}\right)\times E\left(\mathcal{E}^2_t\ -\ 1\right) \qquad (2.7)$$

Therefore using (2.3), the right hand side of (2.7) is equal to zero.
To see that the $(V_t)_{t\in\mathcal{Z}}$ are uncorrelated, we use the fact that:
$\forall(s,t)\in\ \mathcal{Z}^2$ with $t\ <\ s$,

$$Cov\,(V_t, V_s) = Cov\,[Cov\,(V_t, V_s|\mathcal{F}_{s-1})] \tag{2.8}$$

and

$$Cov\,(V_t, V_s|\mathcal{F}_{s-1}) \;=\; E\left(\sigma_t^2\sigma_s^2 \times (\mathcal{E}_t^2\ -1)(\mathcal{E}_s^2\ -1)|\mathcal{F}_{s-1}\right) \tag{2.9}$$

where

$$\left\{\begin{array}{l} \sigma_s^2 := \sigma\left(S_{s-1}-\mu_{s-1},...,S_{s-p}-\mu_{s-p},\sigma_{s-1}^2,\sigma_{s-2}^2,...,\sigma_{s-q}^2\right) \\ \\ \sigma_t^2 := \sigma\left(S_{t-1}-\mu_{t-1},...,S_{t-p}-\mu_{t-p},\sigma_{t-1}^2,\sigma_{t-2}^2,...,\sigma_{t-q}^2\right). \end{array}\right. \tag{2.10}$$

Since

$$\left.\begin{array}{l} \sigma_s^2 \text{ is } \mathcal{F}_{s-1}\text{meas} \\ \\ \sigma_t^2(\mathcal{E}_t^2\ -1) \text{ is } \mathcal{F}_t\text{meas}\ \text{ with } \mathcal{F}_t\subset\mathcal{F}_{s-1} \end{array}\right\} \Rightarrow \sigma_s^2\sigma_t^2(\mathcal{E}_t^2\ -1)\ \text{ is } \mathcal{F}_{s-1}\text{meas.} \tag{2.11}$$

The abbreviation σ_s^2 is $\mathcal{F}_{s-1}$meas just denotes that σ_s^2 is $\mathcal{F}_{s-1}$ measurable.
Therefore (2.9) becomes:

$$Cov\,(V_t, V_s|\mathcal{F}_{s-1}) = \sigma_s\sigma_t(\mathcal{E}_t^2\ -1)\times E\left(\mathcal{E}_s^2\ -1)|\mathcal{F}_{s-1}\right) = 0. \tag{2.12}$$

Hence the process $(V_t)_{t\in\mathcal{Z}}$ can effectively be seen as a real volatility noise.
Therefore, after estimating the conditional return regression function μ, the volatility regression function can be estimated in the following manner:
From(2.1), we see that the squared centered conditional returns

$$[\,S_t-\mu\,(S_{t-1},...,S_{t-p_1},\mu_{t-1},\ ,\mu_{t-q_1})\,]^2 \tag{2.13}$$

drives the conditional stochastic volatility up to the additional volatility noise V_t e.g:

$$\begin{aligned} [\,S_t\ -\mu_t]^2 &= \sigma_t^2\times\mathcal{E}_t^2 &(2.14)\\ &= \sigma_t^2\left(\mathcal{E}_t^2\ -\ 1\right)\ +\ \sigma_t^2 &(2.15)\\ &= \sigma^2\left(S_{t-1}-\mu_{t-1},...,S_{t-p}-\mu_{t-p},\sigma_{t-1}^2,...,\sigma_{t-q}^2\right)\ +\ V_t &(2.16) \end{aligned}$$

This suggest to regress the squared centered conditional returns $\left([\,S_t-\mu_t]^2\right)_{t=2,3,...n}$ against $(S_{t-1}-\mu_{t-1},..,S_{t-p}-\mu_{t-p})_{t=2,...n}$ and the unobservable volatilities σ_{t-1}^2,

$\sigma^2_{t-2}, ..., \sigma^2_{t-q}$ for estimating the conditional stochastic volatility function σ.
During the estimation process of the regression function μ the neural activation function that will be used throughout the whole training steps, is assumed to be continuously lipchits, l-finite and having all the universal approximating and denseness properties. This means

$$\begin{cases} 1^\circ)\Psi \quad \text{is Lipschitz} \\ \\ 2^\circ)\,|\Psi(x)| \leq 1, \\ \\ 3^\circ)\Psi \text{ is monotonically increasing and l-finite,} \end{cases} \tag{2.17}$$

for example

$$\Psi(x) = \frac{1 \ - \ \exp(-x)}{1 \ + \ \exp(-x)}. \tag{2.18}$$

The non-parametric ARMA-GARCH algorithm, due to Mc-Neil and Buehlmann [2000] that will be described throughout the coming sections overcomes the difficulties resulting from the non observability of $\sigma^2_t, \sigma^2_{t-1}, ..., \sigma^2_{t-q}$ and helps to estimate consistently and non-parametrically the corresponding volatility function.
In every step of the conditional expected returns estimation procedure, the two approaches of White [1990] and Mc-Neil [2000] will be combined in order to update the approximate regression function $\hat{\mu}_m$ of the conditional expected return. This will be done by using the optimal solution of the following minimisation problem:

$$\min_{\theta \in ANN\,(\Psi,H)} \frac{1}{N} \sum_{t=2}^{N} \mathcal{L}\left(S_t, \theta\left(Z_{t-1}, \hat{\mu}_{t-1,m-1}, \hat{\mu}_{t-2,m-1}, ..., \hat{\mu}_{t-\tau,m-1}\right)\right) \tag{2.19}$$

for some arbitrary loss function $\mathcal{L}$. The neural output function θ is defined by

$$\theta(x) := \beta_0 \ + \sum_{j=1}^{H} \beta_j \Psi\left(\gamma_{0,j} \ + \ \sum_i \gamma_{ij} x_i\right). \tag{2.20}$$

and the initial estimate $\hat{\mu}_{t,0}$ of the regression function $\hat{\mu}$ is given exactly in the line of the previous chapter e.g.

$$\min_{\theta \in ANN\,(\Psi,H)} \frac{1}{N} \sum_{t=2}^{N} \left[S_t - \theta\left(Z_{t-1}\right)\right]^2. \tag{2.21}$$

In fact, (2.19), (2.20) state that, at the t^{th} trading period, the most recent daily returns $S_{t-1}, S_{t-2} ...,$ and $S_{t-\tau}$ combined with some important trading information X_{t-1} and the corresponding conditional market expectation $\hat{\mu}_{t,m-1}, \hat{\mu}_{t-1,m-1}, ...,$ $\hat{\mu}_{t-\tau,m-1}$ are used as ANN inputs for predicting the target value S_t.

2.1.2 Mc-Neil and Buehlmann Nonparametric ARMA-GARCH Algorithm

The algorithm can be subdivided into five basic steps that can be described in the following manner:

Parameter Settings

- Specify the n-forecasting sample consisting of some historical daily returns and market performances of the stock e.g specify $(S_t)_{t=1,2,...,n}$ and $(X_t)_{t=1,2,...,n}$.
- Choose M and K, the maximum number of iteration and a final smoothing coefficient.

Initialization

- Provide some initial neural network estimates $\hat{\mu}_{t,0}$ of the conditional expected returns $\hat{\mu}_t$ and some initial parametric GARCH estimates $\hat{\sigma}_{t,0}$ of the conditional volatilities σ_t and set m=1 (iteration counter).

Estimation Updating Phase

- **Conditional Expected Returns re-estimation**
  Regress S_t against $Z_{t-1}, \hat{\mu}_{t-1,m-1}, \ \ , \hat{\mu}_{t-p_1,m-1}$. By means of $\hat{\theta}^{\mu}_{m}$ the neural output function defined as the optimal solution of following minimization problem.

$$\min_{\beta,\gamma} \frac{1}{n} \sum_{t=2}^{n} \left[S_t - \theta\left(Z_{t-1}, \hat{\mu}_{t-1,m-1}, \ \ , \hat{\mu}_{t-p_1,m-1}\right)\right]^2, \tag{2.22}$$

  and derive the new and updated estimates $\hat{\mu}_{t,m}$ of the conditional returns as follow:
  $\forall t = 2, 3, ...,$

$$\hat{\mu}_{t,m} : \ = \ \hat{\mu}^{\mu}_{m}\left(Z_{t-1}, \hat{\mu}_{t-1,m-1}, \ \ , \mu_{t-q_1,m-1}\right) = \hat{\beta}_0 \ + \tag{2.23}$$

$$\sum_{j=1}^{H} \hat{\beta}^{\mu}_{j} \, \Psi \left(\hat{\gamma}^{\mu}_{0j} \ + \ \hat{\gamma}^{\mu}_{Xj} X_{t-1} \ + \ \sum_{i=1}^{p_1} \hat{\gamma}^{\mu}_{ij} S_{t-i} \sum_{l=1}^{q_1} \hat{\gamma}^{\mu}_{lj} \hat{\mu}_{t-l,m-1} \right) \tag{2.24}$$

- **Updating the Conditional Stochastic Volatility Estimates**
  Regress $(S_t \ - \ \hat{\mu}_{t,m})^2$ against $(S_{t-1} - \hat{\mu}_{t-1,m})$, $(S_{t-2} - \hat{\mu}_{t-2,m})$, ..., $(S_{t-p} - \hat{\mu}_{t-p,m})$ and $\hat{\sigma}^2_{t-1,m-1}, \hat{\sigma}^2_{t-2,m-1}, ..., \hat{\sigma}^2_{t-q,m-1}$.
  This consists on solving the following optimisation problem:

$$\min_{(\beta,\gamma) \text{ s.t } \theta(\beta,\gamma) \in ANN\,(\Psi, q_N, \Delta_N)} \frac{1}{N} \sum_{t=2}^{N} \left[(S_t \ - \ \hat{\mu}_{t,m})^2 \ - \ \theta\left(a(S, t, \hat{\mu}_m, \hat{\sigma}^2_{.,m-1})\right) \right]^2 \tag{2.25}$$

where $a(S, t, \hat{\mu}_m, \hat{\sigma}^2_{.,m-1})$ represents the neural network input defined by

$$a(S, t, \hat{\mu}_{.,m}, \hat{\sigma}^2_{.,m-1}) = \left(S_{t-1} - \hat{\mu}_{t-1,m}, .., S_{t-p} - \hat{\mu}_{t-p,m}, \hat{\sigma}^2_{t-1,m-1}, .., \hat{\sigma}^2_{t-q,m-1}\right) \quad (2.26)$$

This provide the updated forecasted conditional stochastic volatility $\hat{\sigma}^2_{t,m}$ defined by:
$\forall t = 2, 3, ...,$

$$\begin{aligned} \hat{\sigma}^2_{t,m} &= \hat{\theta}^\sigma_m \left(S_{t-1} - \hat{\mu}_{t-1,m}, .., S_{t-p} - \hat{\mu}_{t-p,m}, \hat{\sigma}^2_{t-1,m-1}, .., \hat{\sigma}^2_{t-q,m-1}\right) \quad (2.27) \\ &= \hat{\beta}^\sigma_0 \ + \\ &\quad \sum_{j=1}^{H^\sigma} \hat{\beta}^\sigma_j \psi \left(\hat{\gamma}^\sigma_0 \ + \ \sum_{i=1}^{p} \hat{\gamma}^\sigma_{ij}(S_{t-i} - \hat{\mu}_{t-i,m}) \ + \ \sum_{l=1}^{q} \hat{\gamma}^\sigma_{lj} \hat{\sigma}^2_{t-l,m-1} \right) \quad (2.28) \end{aligned}$$

where $\hat{\theta}^\sigma_m$ represents the neural output function defined by any optimal solution of the optimisation problem (2.22).

- Set m=m+1, and chek if m=M, otherwise update the estimates.

Final Averaging Step

The algorithm terminates by averaging over the K final estimates $\left(\hat{\sigma}^2_{t,m}\right)_{M-K+1 \leq m \leq M}$ e.g.

$$\hat{\sigma}^2_{t,*} := \frac{1}{K} \sum_{m=M-K+1}^{M} \hat{\sigma}^2_{t,m} \quad (2.29)$$

and regressing $(S_t - \hat{\mu}_{t,M})^2$ against $(S_{t-1} - \hat{\mu}_{t-1,*})$ $(S_{t-2} - \hat{\mu}_{t-2,*}), ..., (S_{t-p} - \hat{\mu}_{t-p,*})$ and $\hat{\sigma}^2_{t-1,*}, \hat{\sigma}^2_{t-2,*}, ..., \hat{\sigma}^2_{t-q,*}$.
This final averaging step helps to increase the efficiency of the algorithm (see Mc-Neil and Buehlmann [2000].)
Due the importance of the qualitative properties of the starting initial estimates, in this following section, some basic results about the classical ARMA-GARCH forecasting methodology that will be used to initialise the stochastic volatility estimate of the algorithm are needed.

2.2 Conditional Stochastic Volatility Estimates of GARCH Models

In order to implement the nonparametric estimation algorithm that has been announced in the previous sections, one needs to recall the classical GARCH methodology due to Bollerslev and Baillie [1992] that will provide the starting initial estimator of the algorithm. For a stochastic volatility forecast of one or two days ahead, a non-linear recursive formula and the characteristic function of the corresponding Mean Square Error of the volatility estimates will be derived.

2.2.1 Classical ARMA-GARCH Predicting Methods

2.2.1.1 Definition: Autoregressive and Moving Average Processes

A stochastic process $(y_t)_{t \in \mathcal{Z}}$ is said to be a linear autoregressive and moving average process of order (p,q) if:

$$y_t = \sum_{k=1}^{p} \phi_k \, y_{t-k} \ + \sum_{k=1}^{q} \theta_k \mathcal{E}_{t-k} + \ \mathcal{E}_t \tag{2.30}$$

with

$$\begin{cases} \mathcal{E}_t \quad \text{independent to} \quad \mathcal{F}_{t-1} \\ \\ E(\mathcal{E}_t | \mathcal{F}_{t-1}) = 0 \quad \forall t. \end{cases} \tag{2.31}$$

2.2.1.2 Mean Square Error for The s-step-ahead Predictor in ARMA Models

Before providing the recursive formula leading to the MSE[2] for the s-step-predictor, let recall first the matrix representation of ARMA models, that helps to handle more easily the following computation in a more compact framework.The matrix form of (2.30) is given by:

$$\mathbf{Y}_t = \begin{pmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \\ \mathcal{E}_t \\ \mathcal{E}_{t-1} \\ \vdots \\ \mathcal{E}_{t-q+1} \end{pmatrix} = \underbrace{\begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p & \theta_1 & \dots & \theta_{q-1} & \theta_q \\ 1 & 0 & \dots & & 0 & 0 & \dots & 0 & 0 \\ . & . & . & . & . & . & . & . & \\ 0 & . & \dots & 1 & 0 & 0 & \dots & . & 0 \\ 0 & . & \dots & . & 0 & 0 & \dots & . & 0 \\ 0 & . & \dots & . & 0 & 1 & \dots & 0 & 0 \\ . & . & . & . & . & . & . & . & \\ 0 & . & \dots & . & 0 & . & \dots & 1 & 0 \end{pmatrix}}_{\Phi} \underbrace{\begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \\ \mathcal{E}_{t-1} \\ \mathcal{E}_{t-2} \\ \vdots \\ \mathcal{E}_{t-q} \end{pmatrix}}_{Y_{t-1}} + \begin{pmatrix} \mathcal{E}_t \\ 0 \\ . \\ 0 \\ \mathcal{E}_t \\ 0 \\ . \\ 0 \end{pmatrix} \tag{2.32}$$

This means,

$$Y_t = \Phi . Y_{t-1} \ + \ (e_1 \ + e_{p+l}) \times \mathcal{E}_t, \tag{2.33}$$

where e_i denotes the i^{th} unit vector. At the $(t+s)^{th}$ trading period, the corresponding asset price or the expected market value of the portfolio y_{t+s} can be forecasted using $E_t(y_{t+s})$ defined in the following manner.

[2] MSE=Mean Square Error

2.2.1.3 Proposition

Using an ARMA model, the portfolio returns or the expected market values of a given instrument are forecasted as follow:

$$E_t(y_{t+s}) = \sum_{i=0}^{p-1} \tau_{i,s} y_{t-i} \ + \ \sum_{i=0}^{q-1} \lambda_{i,s} \mathcal{E}_{t-i} \tag{2.34}$$

with

$$\begin{cases} \tau_{i,s} = \ e_1^{'} \Phi^s e_{i+1} \quad i = 0, ..., p-1, \\ \\ \lambda_{i,s} = \ e_1^{'} \Phi^s e_{p+i+1} \quad i = 0, ..., q-1. \end{cases} \tag{2.35}$$

Proof: See T.Bollerslev and R.Baillie[1992]
Therefore the s-step error $e_{t,s}$ defined by

$$e_{t,s} := y_{t+s} \ - \ E_t(y_{t+s}) = \sum_{i=1}^{s} \Psi_{s-i} \mathcal{E}_{t+i} \tag{2.36}$$

with

$$\Psi_i := \ e_1^{'} \Phi^i \left(e_1 \ + \ e_{p+1}\right) \quad i = 0, ..., p-1. \tag{2.37}$$

Hence the conditional Mean Square Error $E_t(e_{t,s}^2)$ is given by:

$$E_t(e_{t,s}^2) := Var_t(y_{t+s}) = \sum_{i=1}^{s} {\Psi_{s-i}}^2 E_t(\sigma_{t+i}^2) \tag{2.38}$$

2.2.2 Mean Square Error for The s-step-ahead Predictor in GARCH Models

2.2.2.1 Definition: LINEAR ARMA-GARCH

A stochastic process $(y_t)_{t \in \mathcal{Z}}$ is said to be a linear $ARMA(p_1, q_1) - GARCH(p, q)$ if:

$$\begin{cases} y_t = \sum_{k=1}^{p_1} \phi_k \, y_{t-k} \ + \sum_{k=1}^{q_1} \theta_{t-k} \mathcal{E}_{t-k} + \ \mathcal{E}_t \\ \\ \sigma_t^2 := Var\left(\mathcal{E}_t | \mathcal{F}_{t-1}\right)) = \alpha_0 \ + \ \sum_{k=1}^{p} \alpha_k \, \sigma_{t-k}^2 \ + \ \sum_{k=1}^{q} \beta_k \mathcal{E}_{t-k}^2 \end{cases} \tag{2.39}$$

with

$$\begin{cases} \mathcal{E}_t \quad \text{independent to } \ \mathcal{F}_{t-1} \\ \\ E(\mathcal{E}_t|\mathcal{F}_{t-1}) = 0 \quad \forall t. \end{cases} \tag{2.40}$$

Based on the ARMA representation of GARCH models, the squared innovation $\mathcal{E}_t^2$ issued from a linear GARCH(p,q) can conveniently be rewritten as:

$$\mathcal{E}_t^2 = \omega \ + \sum_{i=1}^{\max(p,q)} (\alpha_i \ + \ \beta_i)\, \mathcal{E}_{t-i}^2 \ - \ \sum_{i=1}^{p} \beta_i \nu_{t-i} \ + \ \nu_t \tag{2.41}$$

where $(\nu_t)_{t \in \mathcal{Z}}$ are the serially uncorrelated random variables defined by.

$$\nu_t := \mathcal{E}_t^2 \ - \ \sigma_t^2. \tag{2.42}$$

Therefore setting $m = \max(p, q)$:

$$\underbrace{\begin{pmatrix} \alpha_1 \ +\beta_1 & \alpha_2 \ +\beta_2 & \dots & \alpha_{m-1} \ +\beta_{m-1} & \alpha_m \ +\beta_m & -\beta_1 & \dots & \beta_{q-1} & \beta_q \\ 1 & 0 & \dots & & 0 & 0 & \dots & 0 & 0 \\ . & . & . & . & . & . & . & . & \\ 0 & . & \dots & 1 & 0 & 0 & \dots & . & 0 \\ 0 & . & \dots & . & 0 & 0 & \dots & . & 0 \\ 0 & . & \dots & . & 0 & 1 & \dots & 0 & 0 \\ . & . & . & . & . & . & . & . & \\ 0 & . & \dots & . & 0 & . & \dots & 1 & 0 \end{pmatrix}}_{\Gamma} \tag{2.43}$$

we derive the compact version of (2.41)-(2.42):

$$V_t^2 = we_1 \ + \ \Gamma V_{t-1}^2 \ + \ (e_1 + e_{m+1})\, \nu_t \tag{2.44}$$

where

$$V_t^2 := \begin{pmatrix} \mathcal{E}_t^2 \\ \mathcal{E}_{t-1}^2 \\ . \\ . \\ \mathcal{E}_{t-m+1} \\ \nu_t \\ . \\ \nu_{t-q+1} \end{pmatrix} \tag{2.45}$$

Therefore, computing (2.44) s-steps more implies that:

$$V_{t+s}^2 = \sum_{i=0}^{s-1} \Gamma^i \left((e_1 + e_{m+1})\nu_{t+s-i} + we_1\right) \ + \ \Gamma^s V_t^2. \tag{2.46}$$

2.2.2.2 Proposition

Under the previous setup, the minimum MSE s-step-ahead predictor for the conditional variance from the GARCH(p,q) model is given by:

$$\begin{aligned} E_t(\mathcal{E}^2_{t+s}) : &= E\left(\sigma^2_{t+s}\,|\mathcal{F}_t\right) & (2.47)\\ &= \omega_s \;+\; \sum_{i=0}^{q-1}\delta_{i,s}\sigma^2_{t-i} \;+\; \sum_{i=0}^{m-1}\rho_{i,s}\mathcal{E}^2_{t-i} & (2.48)\end{aligned}$$

where:

$$\begin{cases} \omega_s := {e_1}'\left(\sum\limits_{i=1}^{s-1}\Gamma^i\right)e_1\omega, \\ \delta_{i,s} :=\; -{e_1}'\Gamma^s e_{m+i+1} \qquad i=0,1,...,p-1, \\ \rho_{i,s} :=\; -{e_1}'\Gamma^s\left(e_{i+1} \;+\; e_{m+i+1}\right) \qquad i=0,1,...,p-1, \\ \rho_{i,s} :=\; -{e_1}'\Gamma^s e_{i+1} \qquad i=0,1,...,m-1. \end{cases} \tag{2.49}$$

2.2.3 Mean Square Error for The s-step-ahead Predictor in ARMA-GARCH Models

Combining (2.36) and (2.46) provide the total mean square error in the s-step prediction of the ARMA-GARCH model.

$$Var(y_{t+s}|\mathcal{F}_t) = \sum_{i=1}^{s}\Psi^2_{s-i}\omega_i \;+\; \sum_{i=1}^{s}\Psi^2_{s-i}\left(\sum_{i=1}^{p-1}\delta_{j,i}\sigma^2_{t-j} \;+\; \sum_{i=1}^{m-1}\rho_{j,i}\mathcal{E}^2_{t-j}\right) \tag{2.50}$$

Now, we come into the section dealing with the consistency of the resulting nonparametric neural network estimates.

2.3 Consistency

To study the consistency of the neural network estimates of the stochastic volatility $\hat{\sigma}^2_{t,m}$, we consider the estimated squared centered returns $(S_t \;-\; \hat{\mu}_{t,m})^2$, use the volatility noise V_t defined in (2.4) and apply some convergences results of Regularized ANN(see Corradi and White[1995]).
Based on (2.16), (2.4) can be seen as a regression problem for estimating σ^2, where $(S_t \;-\; \hat{\mu}_{t,m})^2$ represents the output variable, $a_{t,m}$ defined by

$$a_{t,m} := \left(S_{t-1} \;-\; \hat{\mu}_{t-1,m}, .., S_{t-p} \;-\; \hat{\mu}_{t-p,m}, \hat{\sigma}^2_{t-1,m-1}, .., \hat{\sigma}^2_{t-q,m-1}\right) \tag{2.51}$$

the input or explanatory variable and V_t the additional noise. The resulting regression problem is given as follow:

$$o_{t,m} := \sigma^2(a_{t,m}) \; + \; V_t. \tag{2.52}$$

We make use of the concept of Kernel Hilbert spaces and the general result on convergence rate for Regularized ANN: Luka's Theorem, see Corradi and Halbert White[1995]:
The Regularized solution β_n is defined as the minimizer with respect to $\beta \in L^2(\Re^r)$ of:

$$\min_{\beta \in L^2(\Re^r)} \frac{1}{n-1} \sum_{t=2}^{n} \left[o_{t,m} - \mathcal{K}\beta\left(a_{t,m}\right)\right]^2 \; + \; \alpha_n ||\beta|| \tag{2.53}$$

where:

- α_n is a scalar regularization factor such that:

  $$\alpha_n \rightarrow 0 \;\; \text{as} \; n \rightarrow +\infty,$$

- $||\beta||^2 = \int_0^1 \beta^2(x)dx$

- $\mathcal{K}$ is defined as an operator on the space of integrable functions. For example if K represents the Green function, $\mathcal{K}$ can be defined by

  $$\mathcal{K}(\beta(x)) := \int K(x,y)\beta(y)dy \;\; \forall \beta. \tag{2.54}$$

  Based on predefined unit activation function Ψ, an explicit solution of the previous minimization problem is given by Wahba[1977] as:

  $$\beta_n(.) = \eta(.) \;\; \left(Q_n \; + \; \alpha_n.nI\right)^{-1} o_{t,m} \tag{2.55}$$

  where the output vector $o_{t,m}$ and the input based vector $\eta_{t,m}$ are given by:

  $$\begin{aligned} o_{t,m} &= (o_2, o_3..., o_n) \\ \eta &= \left(\eta_{a_{1,m}}, \eta_{a_{2,m}}, ..., \eta_{a_{n,m}}\right) \end{aligned}$$

  with

  $$\left\{ \begin{array}{l} \eta_{a_{i,m}} := \Psi(a_{i,m}, \gamma) \\ \Psi(x,\gamma) := \left\{ \begin{array}{l} x(1-\gamma) \;\; if \;\; 0 \leq \gamma \leq x \leq 1 \\ \gamma(1-x) \;\; if \;\; 0 \leq x \leq \gamma \leq 1 \end{array} \right. \\ Q_n(i,j) = \left(\eta_{a_{i,m}}, \eta_{a_{j,m}}\right) \\ \mathcal{K}\beta_n(.) = Q(.)\left(Q_n \; + \; \alpha_n.I\right)^{-1} o_{t,n} \end{array} \right. \tag{2.56}$$

where

$$Q(.) = \left[Q_{a_{1,m}}(.), ..., Q_{a_{n,m}}(.) \right] \tag{2.57}$$

and

$$Q_{x_j} = \int_0^1 \Psi(x_j, s)\Psi(s, .)ds \tag{2.58}$$

The term $\alpha_n.nI$ in (2.55) is used for increasing the convergence speed in case α_n do not approach 0 very fast.
This approach has connections to a procedure known in the statistics literature as Adaptive Ridge Regression see Judge et al.1985,ch22. When:

$$\begin{cases} \beta(\gamma) := \beta \\ \\ \Psi(x, \gamma) := x \ \forall \gamma \in [0, 1], \end{cases} \tag{2.59}$$

then the optimal solution of the previous minimization problem is the adaptive ridge estimator $\tilde{\beta}_n$ defined by:

$$\tilde{\beta}_n = \left(\sum_1^n a_{i,m}{}^2 - \alpha_n.n \right)^{-1} \sum_1^n a_{i,m} o_{i,m} \tag{2.60}$$

To study the asymptotic behavior of the Regularized Solution, we impose the following assumptions, which rely on the paper of Corradi and White[1995] and some theory of reproducing kernel Hilbert spaces.
Assumption1:
The volatility regression function σ^2 belongs to the Sobolev H^s for some $s \in [0, 1]$, where H^s is a reproducing Kernel Hilbert space(see Definition.A16, page 1242, Neural Computation 7,1225-1244[1995],MIT) and

$$\sigma^2 = \mathcal{K}\beta_0(x) \tag{2.61}$$

with

$$\beta_0 \in \mathcal{N}(\mathcal{K})^\perp \subset L^2 \tag{2.62}$$

where $\mathcal{N}(\mathcal{K})^\perp$ represents the orthogonal complement of $\mathcal{N}(\mathcal{K})$ defined by

$$\mathcal{N}(\mathcal{K}) = \left\{ \beta \in L^2 \text{ such that } \mathcal{K}(\beta) = 0 \right\}. \tag{2.63}$$

Assumption2:
The volatility noise V_t are assumed to be iid with zero mean and finite variance.

This assumption is a natural extension of the fact that the V_t are uncorrelated random variable having a zero conditional mean and unit finite conditional variance as illustrated in (2.4) and (2.5).

Assumption3

Let $Q(.,.)$ be the reproducing kernel (RK) of H^s, for some $s \geq 1$.The eigenvalues of the associated operator $Q(.,.)$ satisfy:

$$a_1 j^{-2p} \leq \lambda_j \leq a_2 j^{-2p} \tag{2.64}$$

for some constants $0 < a_1 \leq a_2 < \infty$ and $p > \frac{1}{2}$.

This assumption requires that the eigenvalues of $Q(.,.)$, say λ_j declines to zero as $j \rightarrow +\infty$. Therefore such an assumption imposes rectriction on the choice of activation function.

Assumption4

Let p be as in Assumption3, there exist a constant $0 < \nu < 1 - \frac{1}{4p}$ depending also on the input vector $a_{t,m}$ and a sequence $k_n \rightarrow 0$ such that:

$\forall f, g \in H^s, s \geq 1$ we have:

$$\left| \int_0^1 fgdF - \frac{1}{N} \sum_{i=n}^{N} f(a_{i,m})(g(a_{i,m}) \right| \leq k_n ||f||_\nu ||g||_\nu . \tag{2.65}$$

For the definition of $||f||_\nu$ and $||g||_\nu$, see Definition:A16 in Valentina Corradi and Halbert Whites[1995].

This assumption specifies some goodness requirements of the input data.

Assumption5

There exist two sequences α_n and k_n such that: If $s \geq \max(\nu, \mu)$,$\mu < 2 - \nu - \frac{1}{4p}$ then

$$k_n * \alpha_n^{-\frac{\nu}{2} - \frac{\mu}{2} - \frac{1}{4p}} \rightarrow 0 \tag{2.66}$$

2.3.1 Theorem: Luka's Theorem[1988]

Under the 5 previous assumptions, the volatility regression function $\tilde{\sigma}$, can be consistently estimated in the following manner:

-If $s \geq \mu + 2$, then α_n is optimal, in the sense of guaranteeing that the squared bias and the variance of the estimate of the volatility regression function σ^2 have the same order of magnitude, if and only if:

$$\alpha_n \sim [\frac{1}{n}]^{\frac{2p}{(4p+2p\mu+1)}}. \tag{2.67}$$

With this choice of α_n, it follows that:

$$E||\beta_n - \mathcal{K}^+ \sigma^2||^2_{H^\mu} = E||\mathcal{K}\beta_n - \sigma^2||^2_{H^\mu} \sim [\frac{1}{n}]^{\frac{4p}{(4p+2p\mu+1)}} \tag{2.68}$$

-If $\mu < s \leq \mu + 2$, then α_n is optimal if and only if:

$$\alpha_n \sim [\frac{1}{n}]^{\frac{2p}{(2ps+1)}}. \tag{2.69}$$

and with this choice of α_n:

$$E||\beta_n - \mathcal{K}^+\sigma^2||^2_{H^\mu} = E||\mathcal{K}\beta_n - \sigma^2||^2_{H^\mu} \sim [\frac{1}{n}]^{\frac{2p(s-\mu)}{2ps+1}} \tag{2.70}$$

Proof: See Valentina Corradi and Halbert White[1995]$\diamondsuit$

Hence, based on (2.70), we derive that $\mathcal{K}\beta_n$ represent consistent ANN estimates of the conditional stochastic volatility σ^2.

We remark that this result holds for bounded random variables only due to the considering only βs which should be normally be defined on compact set. But our extension of consistency results for neural network estimators for unbounded stochastic processes in chapter 1 and 3, however suggest that theorem 2.3.1 can also be extended to a more general setting which applies to financial applications.

2.4 Financial Applications

2.4.1 Financial Valuation on a Risk Adjusted Basis

2.4.1.1 Forecasting Stock Price Conditional Expected Returns

Optimal forecasts must have the features to minimize the Mean Square Error, therefore accordingly to (2.24) and (2.28), at the t^{th} trading day, the daily expected return μ_t can be forecasted using:

$$\begin{aligned} \hat{\mu}_{t,m} : &= \hat{\mu}^{\mu}_{m}\left(S_{t-1}, ..., S_{t-p_1}, \hat{\mu}_{t-1,m-1}, \ \ , \mu_{t-q_1,m-1}\right) & (2.71) \\ &= \hat{\beta}_0 \ + \ \sum_{j=1}^{H} \hat{\beta}^{\mu}_{j}\, \Psi \left(\hat{\gamma}^{\mu}_{0} \ + \ \sum_{i=1}^{p_1} \hat{\gamma}^{\mu}_{i} S_{t-i} \ + \ \sum_{l=1}^{q_1} \hat{\gamma}^{\mu}_{l} \hat{\mu}_{t-l,m-1} \right). & (2.72) \end{aligned}$$

where $\hat{\mu}_{t,0}$ represents some initial estimates of the conditional expected returns. To determine $\hat{\mu}_{t,0}$, unwarrant classical normality assumptions are usually initially imposed on the unexpected return generating process $(\mathcal{E}_t)_{t\in\mathcal{Z}}$ and finding $\hat{\mu}_{t,0}$ by means of some maximum likelihood estimation procedures under some linear models like the ARMA ones.

2.4.2 Value-at-Risk Quantification

Defined as the conditional quantile of the daily returns, accordingly to the model (2.1) and (2.2), the daily value-at-risk Var^t_α defined by:

$$P(S_t \ \leq VaR^t_\alpha | \mathcal{F}_{t-1}) = \alpha \tag{2.73}$$

is given as follow:

$$VaR^t_\alpha := \mu_t \ + \ \sigma_t \times q_\alpha. \tag{2.74}$$

Therefore, at the t^{th} trading period, the maximum amount of $P\&L$ that might occur for the given holding can be estimated by:

$$\hat{Var}^t_\alpha := \ \hat{\mu}_{t,m_{opt}} \ + \ \hat{\sigma}_{t,m_{opt}} \times q_\alpha. \tag{2.75}$$

where q_α represents the quantile of the unexpected returns $(\mathcal{E}_t)_{t\in\mathcal{Z}}$.
To overcome the difficulties resulting from the non observability of $(\mathcal{E}_t)_{t\in\mathcal{Z}}$, one can replace $(\mathcal{E}_t)_{t\in\mathcal{Z}}$ by the fitted residuals defined by:

$$\hat{\mathcal{E}}_t := \frac{S_t \ - \ \hat{\mu}_{t,m_{opt}}}{\hat{\sigma}_{t,m_{opt}}} \tag{2.76}$$

and use if necessary the same machinery based on ANN, EVT and GPD for estimating the conditional expected return, the conditional stochastic volatiliy and the quantile of the heavy tailed distribution. This can be done exactly in the same manner as the results of the previous chapter. In such framework, q_α is estimated by:

$$\hat{q}^\alpha_n(N,u) := \frac{\hat{\sigma}_N(u)}{\hat{\psi}_N}\left\{\left[\frac{n}{N}(1-\alpha)\right]^{\hat{\psi}_N} - 1\right\} + u. \tag{2.77}$$

Therefore the daily Value-at-Risk is estimated by:

$$\hat{VaR}^t_\alpha := \ \hat{\mu}_{t,m_{opt}} + \left[\frac{\hat{\sigma}_N(u)}{\hat{\psi}_N}\left\{\left[\frac{n}{N}(1-\alpha)\right]^{\hat{\psi}_N} - 1\right\} + u\right] \times \ \hat{\sigma}_{t,m_{opt}} \tag{2.78}$$

2.4.3 Applications in Option Pricing

This section outlines the uses of non-parametric ARMA-GARCH in option pricing. Similarly to the two factor volatility model of Hull-White[1987], an Autoregressive-Sieve Bootstrapping or Monte Carlo Simulation method can also be combined with the non-parametric ARMA-GARCH algorithm for pricing a European Call option whose underlying security has a price given by an ARMA-GARCH process.

2.4.3.1 Bootstrapping Fitted Unexpected Returns For Pricing European Style Options.

Similarly to the Historical Simulation approach, one can estimate the empirical distribution of the unexpected returns $(\mathcal{E}_t)_{t\in\mathcal{Z}}$ using the Bootstrap methodology. The method was initially proposed by Efron [1979], as a non-parametric randomisation technique that draws from the observed distribution of the data to model the distribution of a statistic of interest. The AR-Sieve Bootstrapping method that will be used throughout this section can be fundamentally equated as a hybrid between the original Sieve estimation procedures of Grenander [1981] and the classical bootstrap method (see Efron [1979], Freedman [1984], Bose [1988], Franke and Kreiss [1992], Buehlmann [1999]).

As the sample size tends to infinity, AR-Sieve Bootstraps provide correct non-parametric model-specification (see P.Buehlmann, page 4). Therefore, AR-Sieve Bootstraps are robust against model-misspecification.

Under our market settings, the AR-Sieve Bootstrapping method is carried out by considering the fitted residuals $\hat{\mathcal{E}}_t$ defined by:

$$\hat{\mathcal{E}}_t := \frac{S_t \ - \ \hat{\mu}_{t,m_{opt}}}{\hat{\sigma}_{t,m_{opt}}} \tag{2.79}$$

and define $\hat{F}_{\mathcal{E}}$ e.g the empirical distribution function of the innovation $(\mathcal{E}_t)_{t\in\mathcal{Z}}$ by:

$$\hat{F}_{\mathcal{E}}(x) := \frac{1}{N-p}\sum_{t=p+1}^{N} 1_{\left[\hat{\mathcal{E}}_t \ - \ \bar{\mathcal{E}}_t \ \leq x\right]} \tag{2.80}$$

where

$$\bar{\mathcal{E}}_t := \frac{1}{N-p}\sum_{t=p+1}^{N} \hat{\mathcal{E}}_t. \tag{2.81}$$

Now, we consider the AR-Sieve Bootstrap model defined by:

$$S^*_{t+1} := \hat{\mu}_{t,m_{opt}} \ + \ \hat{\sigma}_{t,m_{opt}} \times \mathcal{E}^*_t \tag{2.82}$$

where

$$\mathcal{E}^*_t \ \text{ are iid and drawn from } \ \hat{F}_{\mathcal{E}}. \tag{2.83}$$

To construct some artificial market scenarios, we consider a large bootstrap sample

$$\left[\left(\mathcal{E}^*_{t,p}\right)\right]_{t=1,2,\ldots,N;p=1,2,\ldots,P=10.000} \tag{2.84}$$

generated from $\hat{F}_{\mathcal{E}}$ or the fitted residuals $\hat{\mathcal{E}}_t$.
At each trading period, P bootstrap artificial market scenarios representing some admissible market values of the underlying stock prices are given by:
$\forall t = 1, 2, ...N,$ and $\forall p = 1, 2, ..., P = 10.000$

$$S^*_{t+1,p} = \hat{\mu}_{t,m_{opt}} + \hat{\sigma}_{t,m_{opt}} \times \mathcal{E}^*_{t,p}. \tag{2.85}$$

Assuming that the option expires at the T^{th} trading period with a strike price K and a current price S_0 under a risk free interest rate equal to r, we derive the theoretical market value of a European Call Option by discounting the expectation of the option's payoff e.g:

$$\hat{C}(S_0, T, K, r) := \exp(-rT) \left(\frac{1}{P} \sum_{p=1}^{P} \max\{S^*_{T,p} - K, 0\} \right) = \tag{2.86}$$

$$\exp(-rT) \left(\frac{1}{P} \sum_{p=1}^{P} \max\{\hat{\mu}_{T,m_{opt}} + \hat{\sigma}_{T,m_{opt}} \times \mathcal{E}^*_{t,p} - K, 0\} \right). \tag{2.87}$$

Using the Put-Call Parity, one can derive similar results for pricing European Put options.

2.4.3.2 Combining Monte Carlo Simulation and Non-parametric ARMA-GARCH For Pricing European Options

In this subsection, after specifying some underlying models of the unexpected returns$(\mathcal{E})_{t\in\mathcal{Z}}$, we combine the use of Non-parametric ARMA-GARCH and some independent series of Monte Carlo Simulated values of the unexpected returns for pricing European Style Options. We implement the three most frequently used models for approximating underlying stock price innovations.

- I:$(\mathcal{E}_{t\in\mathcal{Z}})$ iid Student distributed (for capturing the heavy tailedness)

- II:$(\mathcal{E}_{t\in\mathcal{Z}})$ iid Generalized Pareto distributed (for extreme market events).

In each case, independent series consisting of Monte Carlo Simulated values drawn from the underlying models help to generate large set of simulated terminal prices $(S_{T,i})_{i=1,2,...,N}$ given by:

$$S_{T,i} := \hat{\mu}_T + \hat{\sigma}_T \mathcal{E}_{T,i}. \tag{2.88}$$

Therefore the discounted expectation of the option payoff defined by:

$$\hat{C}(S_0, T, K, r) := \exp(-rT]) \left(\frac{1}{N} \sum_{i=1}^{N} \max\{S_{T,i} - K, 0\} \right) \tag{2.89}$$

can be used for estimating the theoretical market value of the European call option $C(S_0, T, K, r)$.
To test the goodness and efficiency of these option-pricing methodologies, we recall the famous Black-Scholes pricing formula and compare it with the non-parametric ARMA-GARCH based methods while assuming that at the initial trading period, we have a known initial unconditional volatility of the underlying security denoted by $\hat{\sigma}_0^2$.

2.4.3.3 Recall: Black-Scholes Option Pricing Formula.

If the underlying asset price of a given option is modelled by geometric Brownian motions e.g.

$$dS(t) = S(t)\ [\mu dt\ +\ \sigma dW(t)]\,, \tag{2.90}$$

then it is log-normally distributed and:

$$\ln(S_t)\ \sim\ \mathcal{N}\left(\ln(S_0)\ +\ \left[\mu\ -\ \frac{\sigma^2}{2}t\right]\ ,\ \sigma t\right). \tag{2.91}$$

Using the Itô Formula, the price $f(S,t)$ of a European call option is given by the following stochastic differential equation:

$$df(S,t) = \left(\frac{\partial f}{\partial S}\mu S\ +\ \frac{\partial f}{\partial t}\ +\ \frac{1}{2}\frac{\partial^2 f}{\partial S^2}\sigma^2 S^2\right)dt\ +\ \frac{\partial f}{\partial S}\sigma S dW. \tag{2.92}$$

Therefore, under the Black-Scholes market settings (see Hull-White[1997]), one can build a risk-less portfolio by shorting one share of option and longing $\frac{\partial f}{\partial S}$ of the underlying stock. Due to the absence of arbitrage opportunities underlying the Black-Scholes model, such portfolio must provide the same expected return as a risk-free bond. This leads to the Black-Scholes-Merton partial differential equation

$$\frac{\partial f}{\partial t}\ +\ rS\frac{\partial f}{\partial S}\ +\ \frac{1}{2}\sigma^2 S^2\frac{\partial^2 f}{\partial S^2}\ = rf \tag{2.93}$$

subject to the initial boundary conditions

$$f\ =\ \max(S-K,0) \text{ For a European Call Option },\tag{2.94}$$
$$f\ =\ \max(K-S,0) \text{ For a European Put Option }.\tag{2.95}$$

The solutions of these partial differential equations, known as the Black-Scholes option pricing formula provide at time zero, the theoretical market value of a

European Call option on a non-dividend paying stock and the corresponding price for a European Put Option e.g:

$$C(S_0, T, K, r, \mu, \sigma) = S_0\mathcal{N}(d_1) \ - \ K\exp(-rT)\mathcal{N}(d_2) \tag{2.96}$$

and

$$P(S_0, T, K, r, \mu, \sigma) = K\exp(-rT)\mathcal{N}(-d_2) \ - \ S_0\mathcal{N}(-d_1) \tag{2.97}$$

where

$$d_1 : \ = \ \frac{\ln(S_0) \ - \ \ln(K) \ + \ (r \ + \ \frac{\sigma^2}{2})T}{\sigma\sqrt{T}} \tag{2.98}$$

$$\tag{2.99}$$

$$d_2 : \ = \ \frac{\ln(S_0) \ - \ \ln(K) \ + \ (r \ - \ \frac{\sigma^2}{2})T}{\sigma\sqrt{T}} = \ d_1 \ - \sigma\sqrt{T} \tag{2.100}$$

2.4.3.4 ANN; EVT and ARMA-GARCH versus Black Scholes

We illustrate the goodness and the accuracy of the ANN; EVT and ARMA-GARCH based daily Value-at-Risk methodology via the computation of the daily VaR of a holding consisting of one European Call Option on one share of SIEMENS as the underlying asset. As explanatory variables, we use the daily closing prices of BASF, DAX30 and COMMERZBANK traded on the stock exchange of Frankfurt. We have assumed a flat term structure. The simulation results fit with the current market practices and expectation. The correction made by the ANN; EVT and ARMA-GARCH based discounted payoff is illustrated by some additional payoff above the black scholes payoff. The ANN; EVT and ARMA-GARCH based discounted payoff is slightly greater than the market value of the considered Black Scholes European Call. The difference between the two payoffs can be equated as the added value provided by the ANN; EVT and ARMA-GARCH based approach.

Chapter 3

Market Risk Controlling Based on Artificial Neural Networks

Neural Networks is now a vibrant and mature subject. It began over 50 years ago, based on very simple models of the real neurons of the brain. It had great acceptance initially but was oversold as being able to solve all problems of information processing up to consciousness itself. This hubris caused a reduction of support for the subject, but more recent work has led to deeper foundations as well as increasingly powerful ability to simulate large networks of neurons. The areas of industrial applications of neural networks are now very broad, they being important components in power distribution and control systems in numerous chemical and engineering plants, as well as leading to efficient pattern recognisers such as the IRIS scan system recently launched on the NY Stock Exchange. They also play an important role as predictors in the financial markets. At the same time the increased understanding of the powers of neural networks has allowed an ever deeper understanding of the nature of the processing in various parts of the brain. There is now a strong move to broaden this to the global brain and to its greatest subtlety, that of consciousness. There are now numerous groups dedicated to understanding this last great bastion of the scientific unknown; the neural correlates of consciousness are being carefully tracked down and the underlying neural mechanisms being exposed.

The major criticism of many market risk measurement models is the need of normality settings. Cleary, for some assets such as options and short-term securities (bonds), normality assumptions are highly questionable. For example, the most an investor can lose if he or she buys a call option on equity is the call premium; however, the investor's potential upside returns are unlimited. In a statistical sense, the returns on call options are nonnormal since they exhibit positive skew. To overcome the normality assumptions, consider $(S_t)_{t\in\mathcal{Z}}$ the stochastic process driving the returns (Profits and Losses) of a given financial portfolio. Beside the use of the conditional volatility as a market risk measure, the Value-at-Risk is nowadays widely adopted for aggregating market risk exposures, estimating the riskyness of trading strategies or determining the economic capital for fulfilling financial regulators risk capital standards (see Bank of International Settlement, [1996]).

There are various ways of defining the Value-at-Risk, there exist also different technical approaches for its implementation, but the crucial quantity is always the conditional quantile of the Profits and Losses ($P\&L$) within a certain confidence level over some liquidation or holding period. For estimating easily the daily Value-at-Risk or forecasting the future market values of financial assets, some unwarranted normality assumptions are usually made (see Variance Covariance; Delta-Gamma, Monte Carlo Simulation). The normality settings of financial assets hide a lot of drawbacks due to the fact that financial returns usually display some patterns of skewsness or heavy tailedness. Consequently the normality assumptions become no longer appropriate for describing the dynamics of the marked-to-market values of financial stocks. An alternative consists on making use of the theory of ANN by applying the new denseness results established in chapter one that extends White's neural network denseness results to heavy tailed, unbounded stochastic processes. Based on the new approximation results establisched in the first chapter, one can derive the conditional quantile of the stochastic process of the financial returns of a given position. For that, we do combine, Bassett and Koenker [1978] conditional quantile estimation algorithm and our new ANN denseness results. This enables to estimate the VaR without the need to estimate the conditional means or stochastic volatility or assuming any of the questionable hypothesis except that financial returns are either independent identically distributed or are mixing stochastic processes.

Given $\alpha \in [0\,,\,1]$, generally chosen in $[0.95\,,\,0.99]$, the daily α-conditional Value-at-Risk VaR_α^t, is defined as the α-conditional quantile of the financial return S_t given trading information and financial fixings of the portfolio up to time $t-1$ e.g.:

$$P\left(S_t \leq VaR_\alpha^t | S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right) = \alpha. \tag{3.1}$$

Using the new approximation properties of neural output functions, their flexible learning capacities, combined with the characterization of conditional quantiles due to Bassett and Koenker [1978], one can derive a consistent neural network estimator for the conditional daily VaR VaR_α^t by solving the following optimisation problem:

$$\min_{\theta \in ANN(\Psi, q_n, \Delta_n)} \frac{1}{N} \sum_{t=1}^{N} \mathcal{L}\left(S_t, \theta(Z_{t-1})\right) \tag{3.2}$$

where

$$Z_{t-1} := (S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}) \in \Re^r.$$
$$\theta(z) := \beta_0 + \sum_{j=1}^{H} \beta_j \Psi(\tilde{z}.\gamma_j). \tag{3.3}$$

with

$$\tilde{z} := (1, z_1, z_2, ..., z_r)^T. \tag{3.4}$$

and

$$\mathcal{L}\left(S_t, \theta(Z_{t-1})\right) := |\, S_t - \theta(Z_{t-1})| \left[\alpha 1_{[0,+\infty[}\left(S_t - \theta(Z_{t-1})\right) + (1-\alpha) 1_{]-\infty,0]}\left(S_t - \theta(Z_{t-1})\right)\right]$$

Let $\hat{\theta}_n$ denote any optimal solution of (3.2). Up to some regularity conditions on Ψ, q_n and Δ_n where, as in the first chapter q_n and Δ_n determine the connectionist sieve controlling the complexicity of the network for increasing sample size n. $\hat{\theta}_n$ can be used for estimating consistently and non-parametrically the daily Value-at-Risk VaR_{α}^{t}.
The first section of the chapter is dealing with general results on the existence and consistency of the neural network estimators $\hat{\theta}_n$. The second section is specifying the underlying assumptions that enable the estimation of the conditional quantile of the returns using ANN and leading to the daily VaR estimates. We illustrate the goodness and the accuracy of the VaR methodology via the computation of the daily VaR of a holding consisting of one share of DEUTSCHE Bank. As explanatory variables, we use the daily closing prices of BASF, DAX30 and COMMERZBANK traded on the stock exchange of Frankfurt.

3.1 Consistent and Nonparametric Conditional Quantile Estimation Using ANN

Before the statement of the main theorem leading to the consistent neural network estimator of the daily VaR, we refer to the basic existence and consistency result of White [1990] which we already applied in chapter 1.

Theorem: Existence and Consistency defined as Solutions of a Minimization Problem

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a complete probability space, and $(\Theta, ||\ \ ||_\Theta)$ a separable normed space. For n=1, 2, 3,..., consider $\Theta_n, \subset\ \Theta$ and $Q_n\ : \Omega \times \Theta \to \bar{\Re}$ such that:
1°) $(\Theta_n)_{n\in N}$ is an increasing sequence of compact subsets of Θ such that

$$\cup_{n=1}^{+\infty} \Theta_n \text{ is dense in } \Theta \tag{3.5}$$

2°) $Q_n(\omega, \theta)$ is measurable in ω for any θ and continuous in θ for any ω. Then, there exists measurable mappings $\hat{\theta}_n : \Omega \to \Theta_n$ such that:

$$Q_n(w, \hat{\theta}_n) := \min_{\theta \in \Theta_n} Q_n(w, \theta). \tag{3.6}$$

If additionally, there exists a continuous function $\bar{Q} : \Theta \to \bar{\Re}$ such that for some $\theta_0 \in \Theta$:
3°)

$$\sup_{\theta \in \hat{\Theta}_n} \left| Q_n(\omega, \theta) - \bar{Q}(\theta_0) \right| \xrightarrow{p} 0 \quad \text{for } n \to \infty, \tag{3.7}$$

4°)

$$\inf_{\theta \in \Theta \, ; \, ||\theta - \theta_0|| \geq \epsilon} \left\{ \bar{Q}(\theta) - \bar{Q}(\theta_0) \right\} > 0 \;\; \text{for all} \epsilon > 0, \tag{3.8}$$

then $\hat{\theta}_n$ is a consistent estimator of θ_0, i.e.

$$||\hat{\theta}_n - \theta_0||_\Theta \xrightarrow{p} 0 \quad \text{for } n \to \infty. \tag{3.9}$$

Proof:
The theorem is a direct consequence of Theorem 2.2 and corollary 2.6 of White and Wooldridge [1990]. We only have to strengthened some of the assumptions a bit, e.g. assuming continuity instead of lower semi continuity of Q_n or a normed space instead of a metric space, to simplify formulations .◇

3.2 Theorem: Consistent and Nonparametric Estimator for the daily Value-at-Risk

3.2.1 Consistent Neural Network Conditional Quantile Estimator

Let q_n , Δ_n and Ψ be chosen as in (1.14) and (1.67) , and $(S_t)_{t \in \mathcal{Z}}$ be the stochastic process describing the dynamics of the financial returns of a given portfolio. Let X_t represent some exogenous information on the market, and set as before

$$Z_{t-1} := \left(S_{t-1}, S_{t-2}, ..., S_{t-\tau}, X_{t-1}\right). \tag{3.10}$$

In contract to chapter1, we do not assume that an AR-ARCH-model like (1.1), but allow (S_t, Z_{t-1}) to be rather arbitrary stationary time series. Our goal is to estimate the α-quantile of S_t given $Z_{t-1} = z$.

We consider neural networks with inputs Z_{t-1}, i.e. the corresponding output functions are defined on $\Re^l$, where $l = \tau + dim(X_t)$. We fit the neural networks to the data by minimizing

$$Q_n(\theta) := \frac{1}{n}\sum_{t=1}^{n} \mathcal{L}\left(S_t, \theta(Z_{t-1})\right) \tag{3.11}$$

where

$$\mathcal{L}\left(S_t, \theta(Z_{t-1})\right) = \left|\, S_t - \theta(Z_{t-1})\right| \Big[\alpha 1_{[0,+\infty[}\left(S_t - \theta(Z_{t-1})\right) \,+\, (1-\alpha)1_{]-\infty,0]}\left(S_t - \theta(Z_{t-1})\right)\Big] .$$

The resulting network function is called $\hat{\theta}_n$:

$$Q_n(\hat{\theta}_n) := \min_{\theta \in ANN(\ldots)} Q_n(\theta). \tag{3.12}$$

We prove that the conditional α-quantile of S_t given $Z_{t-1} = z$ which we call $\theta_\alpha(z)$, can be estimated nonparametrically and consistently using the neural output function $\hat{\theta}_n(z)$ under appropriate assumptions on the growth of the network complexicity.

Our arguments follow closely those in section 1.4.2 where we discussed the estimation of conditional means. We, therefore, assume the assumptions of theorem 1.4.2.1 are fulfilled. Write again

$$\hat{\Theta}_n = ANN(\Psi, q_n, \Delta_n), \tag{3.13}$$

and Θ be the closure of $\cup_{n=1}^{+\infty}\hat{\Theta}_n$ in $L^2(\mu)$, where μ is the stationary distribution of Z_{t-1}. We remark that the summands

$$\mathcal{L}\left(S_t, \theta(Z_{t-1})\right) = \alpha\left(S_t - \theta\right)^+ + (1-\alpha)\left(S_t - \theta\right)^- ,$$

where u^+, u^- defined as the positive and negative part of $u \in \Re$, are also continuous in θ. Moreover

$$\bar{Q} = E\left(Q_n\right) = E\Big[\alpha(S_t - \theta(Z_{t-1}))^+ \,+\, (1-\alpha)(S_t - \theta(Z_{t-1}))^-\Big] \tag{3.14}$$

is continuous on $\hat{\Theta}_n$ by the same arguments as in section 1.4.2. Now we consider $\theta = \theta_\alpha$. Let

$$q(s, y) = \alpha(S - y)^+ \,+\, (1-\alpha)(S - y)^-. \tag{3.15}$$

Then, $\theta_\alpha(z)$ minimizes by definition $E\left(q(S_t, \theta_\alpha)|Z_{t-1} = z\right)$. Therefore, for all z,

$$\begin{aligned} E\left(q(S_t, \theta_\alpha)|Z_{t-1} = z\right) &\leq E\left(q(S_t, 0)|Z_{t-1} = z\right) && (3.16)\\ &= E\Big(\alpha S_t^+ \,+\, (1-\alpha)S_t^- |Z_{t-1} = z\Big) && (3.17)\\ &= E\left(|S_t| \;|Z_{t-1} = z\right). && (3.18) \end{aligned}$$

The right-hand side is integrable with respect to μ, giving as result just $E(|S_t|) < \infty$, and, therefore

$$\bar{Q}(\theta_\alpha) = E\left\{E(q(S_t, \theta_\alpha(Z_{t-1})|Z_{t-1}))\right\} \leq E(|S_t|) < \infty. \tag{3.19}$$

Moreover, Lemma 3.2.1.2 stated below, implies that

$$\begin{aligned} \left|\bar{Q}(\theta) - \bar{Q}(\theta_\alpha)\right| &\leq E\left|q(S_t, \theta(Z_{t-1})) - q(S_t, \theta_\alpha(Z_{t-1})\right| & (3.20) \\ &\leq E\left|\theta(Z_{t-1}) - \theta_\alpha(Z_{t-1})\right| & (3.21) \\ &\leq \left\{\int (\theta(z) - \theta_\alpha(z))^2 \, d\mu(z)\right\}^{\frac{1}{2}} & (3.22) \end{aligned}$$

where we use Jensen's inequality for the last line. Therefore, $\bar{Q}$ is continuous w.r.t $L^2(\mu)$-norm in θ_α. We assume $\theta_\alpha \in \Theta$. Then, Theorem 3.1 implies that there exist $\hat{\theta}_n \in \hat{\Theta}_n$ such that

$$Q_n(\hat{\theta}_n) = \min_{\theta \in \hat{\Theta}_n} Q_n(\theta) \tag{3.23}$$

which estimate θ_α consistently, i.e

$$\int \left(\hat{\theta}_n(z) - \theta_\alpha(z)\right)^2 d\mu(z) \xrightarrow{p} 0 \tag{3.24}$$

provided that

$$\begin{aligned} &\sup_{\theta \in \hat{\Theta}_n} \left|Q_n(\theta) \; - \; \bar{Q}(\theta)\right| \xrightarrow{p} 0 & (3.25) \\ &\inf_{\theta \in \mathcal{N}_\epsilon^c(\theta_\alpha)} \; \bar{Q}(\theta) \; - \; \bar{Q}(\theta_\alpha) > 0 & (3.26) \end{aligned}$$

for arbitrary ϵ-neighbourhoods $\mathcal{N}_\epsilon(\theta_\alpha)$ of θ_α. For the latter condition, consider θ with $||\theta - \theta_\alpha|| \geq \epsilon$. Then, we have for arbitrary $\delta > 0$,

$$\begin{aligned} \bar{Q}(\theta) \;\; - & \;\; \bar{Q}(\theta_\alpha) = E\left(q(S_t, \theta(Z_{t-1}) - q(S_t, \theta_\alpha(Z_{t-1}))\right) = \\ & \int E\left\{q(S_t, \theta(Z_{t-1}) - q(S_t, \theta_\alpha(Z_{t-1})) \mid Z_{t-1} = z\right\} d\mu(z) \\ = & \int E\left\{..|\; Z_{t-1} = z\right\} 1_{(\delta,\infty)}(|\theta(z) - \theta_\alpha(z)|) d\mu(z) + \\ & \int E\left\{..|\; Z_{t-1} = z\right\} 1_{[0,\delta]}(|\theta(z) - \theta_\alpha(z)|) d\mu(z) \\ \geq & \int E\left\{..|\; Z_{t-1} = z\right\} 1_{(\delta,\infty)}(|\theta(z) - \theta_\alpha(z)|) d\mu(z) - \end{aligned}$$

$$
\begin{aligned}
& \int |\theta(z) - \theta_\alpha(z)|.1_{[0,\delta]}(|\theta(z) - \theta_\alpha(z)|)d\mu(z) \\
\geq & \int C_\delta(z, \theta_\alpha(z))\frac{1}{2}|\theta(z) - \theta_\alpha(z)|^2 1_{(\delta,\infty)}(|\theta(z) - \theta_\alpha(z)|)d\mu(z) - \\
& \int |\theta(z) - \theta_\alpha(z)| 1_{[0,\delta]}(|\theta(z) - \theta_\alpha(z)|)d\mu(z)
\end{aligned}
$$

where we have used Lemma 3.2.1.2 for the first inequality and Lemma 3.2.1.3. for the second one. $C_\delta(z, \theta_\alpha)$ denotes the lower bound for the conditional density $f_s(y|z)$ of S_t given $Z_{t-1} = z$ on the interval $\theta_\alpha - \delta \leq y \leq \theta_\alpha + \delta$.
Now for $\delta \searrow 0$,

$$u^2 1_{(\delta,\infty)}(u) \nearrow u^2 \quad \forall u \geq 0 \tag{3.27}$$

$$u 1_{[0,\delta]}(u) \searrow 0 \quad \forall u \geq 0 \tag{3.28}$$

and, using the continuity of $f_s(y|z)$ in a neighbourdhood of θ_α for all z

$$C_\delta(z, \theta_\alpha(z)) = \inf_{|y-\theta_\alpha(z)|\leq\delta} f_s(y|z) \ \nearrow f_s(\theta_\alpha(z)|z) \tag{3.29}$$

as soon as δ is small enough. By the monotone convergence theorem we may exchange the limit operator $\lim_{\searrow 0}$ and the integration $\int ..d\mu(z)$ and get

$$\bar{Q}(\theta) \ - \ \bar{Q}(\theta_\alpha) \ \geq \ \frac{1}{2}\int f_s(\theta_\alpha(z)|z).|\theta(z) - \theta_\alpha(z)|^2 d\mu(z) \tag{3.30}$$

$$\geq \ \frac{C}{2}||\theta - \theta_\alpha||^2 \ \geq \ \frac{C}{2}\epsilon^2 > 0. \tag{3.31}$$

We have to assume that for some $C, \delta, f_s(y|z)$ is continuous in
$y \in (\theta_\alpha(z) - \delta \, , \, \theta_\alpha(z) + \delta)$ and $f_s(\theta_\alpha(z)|z) \geq C$ for almost all (w.r.t measure μ) z.
So, we have proven (3.26), and it remains to prove (3.25).
For that purpose, we follow the proof of theorem 1.4.2.1. First, we consider the same open covering $U_{n_i}, \ i = 1, 2, ..., K(\delta_n)$, of the compact set $\hat{\Theta}_n$, and we again remark

$$K(\delta_n) \leq 4\left(\frac{\Delta_n}{\delta_n}\right)^{q_n(r+2)+1} q_n^{q_n(r+1)} \tag{3.32}$$

by lemma 1.6.2.1. Now, we use the notation

$$S_t(\theta) = g(\theta, S_t, Z_{t-1}) = \alpha(S_t - \theta(Z_{t-1}))^+ + (1-\alpha)(S_t - \theta(Z_{t-1}))^-. \tag{3.33}$$

By continuity of y, the measurability condition of Theorem 1.6.2 is satifsfied. Condition of Theorem 1.6.2 is satisfied with $M_{n,t} \equiv 1$ as for $\theta, \theta^* \in \hat{\Theta}_n$

$$|S_t(\theta) - S_t(\theta^*)| \leq |\theta(Z_{t-1}) - \theta^*(Z_{t-1})| \tag{3.34}$$

by lemma 3.2.1.2. The α-mixing property of (S_t, Z_{t-1}) is again inherited by $S_t(\theta)$. Assuming an exponentially decreasing tail of the law of S_t, i.e.

$$\begin{aligned}
P\left(|S_t(\theta) - E(S_t(\theta))| > x\right) &\leq P\left(|S_t(\theta)| > x - E(S_t(\theta))\right) \\
&= P\left(\alpha(S_t - \theta(Z_{t-1})^+ + (1-\alpha)(S_t - \theta(Z_{t-1})^- > x - E(S_t(\theta))\right) \\
&\leq P\left(|S_t - \theta(Z_{t-1})| > x - E(|S_t - \theta(Z_{t-1})|)\right) \\
&\leq P\left(|S_t| > x - E(|S_t|) - E(|\theta(Z_{t-1})|) - |\theta(Z_{t-1})|)\right) \\
&\leq P\left(|S_t| > x - E(|S_t|) - 2\Delta_n\right) \\
&\leq a_0 \exp\left\{-f_n(x)\right\}
\end{aligned}$$

for all $x > E(|S_t|) + 2\Delta_n$ with

$$f_n(x) = a_1(x - |S_t|) - 2\Delta_n)^\beta. \tag{3.35}$$

Therefore, the Bernstein inequality of Theorem 1.6.1.3 is applicable to $\epsilon_t(n) = S_t(n) - E(S_t(n))$. Choosing $M_n = 4\Delta_n$ in that inequality, we have

$$f_n(M_n) \geq a_1 \Delta_n^\beta \quad \text{for } \Delta_n \geq E(|S_t|), \tag{3.36}$$

and, then

$$P\left(\left|\sum_{t=1}^n |S_t| - E(|S_t|)\right| > \Delta_n\right) \leq C_1 \exp\left\{-\frac{C_2}{2}\frac{\Delta_n}{\sqrt{n}\Delta_n}\right\} + na_0 \exp\{a_1\Delta_n^\beta\} \tag{3.37}$$

provided that $nE_n(4\Delta_n) = o(\Delta_n)$. We postpone discussion of that property, and we remark that, then, condition (1.277) of Theorem 1.6.2 is satisfied if $\Delta_n \to \infty$ fast enough for the choice

$$\gamma_n(\epsilon) = C_1 \exp\{-\frac{C_2}{4}\frac{\epsilon}{\sqrt{n}\Delta_n}\} + na_0 \exp\{-a_1\Delta_n^\beta\}. \tag{3.38}$$

We conclude that the assumptions of Theorem 1.6.2 are satisfied if we additionally assume that the stationary distribution of Z_{t-1} aslo decays exponentially, i.e. for some $\beta_0, \beta_1, \tau > 0$

$$P\left(||Z_{t-1}|| > x\right) \leq \beta_0 \exp\{-\beta_1||x||^2\} \quad \text{for all } x. \tag{3.39}$$

Now, we can apply Theorem 1.6.2. for some $S_n(\theta) = nQ_n(\theta)$ and $a_n = n$, and get

$$P\left(\sup_{\theta\in\hat{\Theta}_n}\left|Q_n(\theta) - \bar{Q}(\theta)\right| > \epsilon\right) = P\left(\sup_{\theta\in\hat{\Theta}_n}|S_n(\theta) - E(S_n(\theta)| > \epsilon_n\right) \xrightarrow{p} 0 \tag{3.40}$$

if for yet arbitrary δ_n, ρ_n

$$K(\delta_n)\gamma_n(\epsilon n){\rightarrow}0, \quad K(\delta_n)\frac{n}{\epsilon n}(1+\Delta_n\rho_n)\delta_n{\rightarrow}0, \tag{3.41}$$

$$K(\delta_n)\frac{n}{\epsilon n}\Delta_n \exp\{-\frac{\beta_1}{2}\rho_n^2\}{\rightarrow}0. \tag{3.42}$$

To show the previous equation(3.42), we replace $K(\delta_n)$ by the upper bound given above. For, the first term, we need, using $p_n \equiv q_n(r+1)$,

$$K(\delta_n)\gamma_n(\epsilon n) \leq 4\left(\frac{\Delta_n}{\delta_n}\right)^{q_n+p_n+1} *$$

$$q_n^{p_n}\left\{C_1 \exp\left(-\frac{C_2}{4}\frac{\epsilon\sqrt{n}}{\Delta_n}\right) + na_0 \exp\left(-a_1\Delta_n^\beta\right)\right\} \rightarrow 0 \quad \text{for } n \rightarrow \infty. \tag{3.43}$$

For that it sufficies

$$\exp\left\{(p_n+q_n+1)log(\frac{\Delta_n}{\delta_n}) + p_n log(q_n) - \frac{C_2}{4}\frac{\epsilon\sqrt{n}}{\Delta_n}\right\}{\rightarrow}0, \tag{3.44}$$

$$\exp\left\{(p_n+q_n+1)log(\frac{\Delta_n}{\delta_n}) + p_n log(q_n) - a_1\Delta_n^\beta + log(n)\right\}{\rightarrow}0. \tag{3.45}$$

Assuming $log(n) = o(\Delta_n^\beta)$, then (3.44) and (3.45) hold if

$$q_n log\left(\Delta_n q_n \delta_n\right) = o\left(\frac{\sqrt{n}}{\Delta_n}\right), \tag{3.46}$$

$$q_n log\left(\Delta_n q_n \delta_n\right) = o\left(\Delta_n^\beta\right). \tag{3.47}$$

The left term of (3.42) is bounded by

$$\left(\frac{\Delta_n}{\delta_n}\right)^{q_n+p_n+1} q_n^{p_n}\left(1+\Delta_n\rho_n\right)\delta_n. \tag{3.48}$$

Assuming that $\Delta_n\rho_n{\rightarrow}\infty$, that term converges to 0 if

$$\left(\frac{\Delta_n q_n}{\delta_n}\right)^{q_n}\left(\frac{\Delta_n}{\delta_n}\right)^{q_n+1}\Delta_n\delta_n\rho_n{\rightarrow}0. \tag{3.49}$$

Therefore, the left term of (3.42) converges to 0 if

$$\left(\frac{\Delta_n q_n}{\delta_n}\right)^{p_n}\left(\frac{\Delta_n}{\delta_n}\right)^{q_n+1}\Delta_n \exp\left\{-\frac{\beta_1}{2}\rho_n^2\right\}{\rightarrow}0. \tag{3.50}$$

Now we choose $\rho_n = n^\rho, \delta_n = n^\gamma\Delta_n q_n$ for some $\rho, \gamma > 0$.
As $q_n, \Delta_n{\rightarrow}\infty$ (3.46) implies necessarily

$$\Delta_n = o(n^{\frac{1}{2}}). \tag{3.51}$$

(3.46), (3.47) hold if, neglecting the constant factor γ,

$$\Delta_n q_n log(n) = o(n^{\frac{1}{2}}) \text{ and } \Delta_n q_n log(n) = o\left(\Delta_n^{(1+\beta)}\right). \tag{3.52}$$

The second assertion implies the assumption $log(n) = o\left(\Delta_n^{\beta}\right)$ made above. Also, we have now

$$\delta_n = \Delta_n q_n^{\gamma} = o\left(n^{\frac{1}{2}+\gamma}\right). \tag{3.53}$$

Therefore, together with $\rho_n = n^{\rho}$, (3.50) is implied by (3.49). The latter condition is implied by, using $\Delta_n \leq q_n \Delta_n$,

$$\frac{(\Delta_n q_n)^{p_n+q_n+2}}{{\delta_n}^{p_n+q_n}} n^{\rho} = \frac{(q_n \Delta_n)^2 n^{\rho}}{n^{\gamma(p_n+q_n)}} \rightarrow 0, \tag{3.54}$$

as

$$q_n \Delta_n = o\left(n^{\frac{1}{2}}\right) \text{ and } p_n + q_n \rightarrow \infty \text{ for } n \rightarrow \infty.$$

It remains to discuss the condition $nE_n(4\Delta_n) = o(\Delta_n)$ which we have assumed above. We have chosen $\Delta_n = \epsilon n$ and, therefore, need $E_n(4\Delta_n) = o(1)$. Let n be large enough such that $\Delta_n \geq E(|S_t|)$, then for $x \geq 4\Delta_n$, we have

$$x - E(|S_t|) - 2\Delta_n \geq \frac{x}{4}, \tag{3.55}$$

and therefore

$$\begin{aligned} E_n(4\Delta_n) &= \int_{4\Delta_n}^{\infty} e^{-a_1}(x - E(|S_t|) - 2\Delta_n)^{\beta} dx && (3.56)\\ &\leq \int_{4\Delta_n}^{\infty} e^{-a_1(\frac{x}{4})^{\beta}} dx \quad \rightarrow 0 \text{ as } \Delta_n \rightarrow \infty. && (3.57)\end{aligned}$$

Therefore, we have finally proven the following result.

3.2.1.1 Theorem

Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, and let (S_t, Z_{t-1}) be a stationary stochastic process satisfying an α-mixing condition with exponentially decreasing mixing coefficients, where S_t is real valued and $Z_{t-1} \in \Re^r$. Let the stationary distribution of S_t be absolutely continous and satisfy

$$P(|S_t| > x) \leq \beta_0 \exp\{-\beta_1 x^{\tau}\} \quad \text{for } x \geq 0 \tag{3.58}$$

for some $\beta_0, \beta_1, \tau > 0$. Let q_α denote the conditional α-quantile of S_t given $Z_{t-1} = z$, i.e,

$$P\left(S_t \leq q_\alpha | Z_{t-1} = z\right) = \alpha \quad \text{for } \mu - \text{almost all } z. \tag{3.59}$$

Let $f_s\left(x|z\right)$ denote the conditional density function of S_t given $Z_{t-1} = z$, and assume that there are some $C, \delta > 0$ such that for μ-almost all z, $f_s\left(x|z\right)$ is continuous in $x \in (q_\alpha(z) - \delta, q_\alpha(z) + \delta)$ and

$$f_s\left(q_\alpha(z)|z\right) \geq C > 0. \tag{3.60}$$

Let Ψ be bounded in absolute value by 1 and satisfy the Lipschitz condition:

$$|\Psi(u) - \Psi(v)| \leq L.\, |u - v| \quad \text{for } u, v \in \Re. \tag{3.61}$$

Let $\hat{\Theta} = ANN\left(\Psi, q_n, \Delta_n\right)$ be the usual set of neural network output functions of r variables with q_n neurons in the hidden layer where the sum of absolute values of the weights from the hidden layer to the output layer is bounded by Δ_n, and the sum of all absolute values of the weights from the from input to hidden layer is bounded by $q_n\Delta_n$. Let Θ denote the closure of $\hat{\Theta}$ in $L^2(\mu)$. Let $\hat{\theta}_n(z) \in \hat{\Theta}$ be the neural network estimate given by

$$\hat{\theta}_n = argmin_{\theta\in\hat{\Theta}} \frac{1}{n} \sum_{t=1}^{n} \left\{\alpha\left(S_t - \theta(Z_{t-1})\right)^+ + (1-\alpha)\left(S_t - \theta(Z_{t-1})\right)^-\right\}. \tag{3.62}$$

Assume $q_\alpha \in \hat{\Theta}$. Then, $\hat{\theta}_n$ is a consistent estimate of q_α for $n \to \infty$ in $L^2(\mu)$-sense

$$\int \left(q_\alpha(z) - \hat{\theta}_n(z)\right)^2 d\mu \quad \xrightarrow{p} 0 \tag{3.63}$$

provided that for $n \to \infty$:

$$\begin{cases} q_n, \Delta_n \to \infty \\ \\ \Delta_n = o(n^{\frac{1}{2}}) \quad \text{and} \\ \\ q_n \Delta_n log(n) = o\left(\min\left(\sqrt{n}, \Delta_n^{(1+\beta)}\right)\right). \end{cases}$$

White[1992] has proven a similar result for bounded random variables. In that case, the growth conditions are

$$\begin{cases} q_n, \Delta_n \to \infty \\ \Delta_n = o(n^{\frac{1}{2}}) \quad \text{and} \\ q_n \Delta_n log(q_n \Delta_n q_n) = o\left(n^{\frac{1}{2}}\right), \quad \text{i.e.} \end{cases} \tag{3.64}$$

the unboundedness essentially introduces the additional condition

$$q_n \Delta_n log(n) = o\left(\Delta_n^{(1+\beta)}\right). \tag{3.65}$$

compare also the discussion after Theorem 1.4.2.1: Our growth conditions are e.g., satisfied if

$$\Delta_n = bn^{\gamma} \quad \text{for } 0 < \gamma < \frac{1}{2}\,, \, b > 0, \tag{3.66}$$

and either

$$\gamma \geq \frac{1}{2(1+\beta)}, \quad q_n = o\left(\frac{n^{\frac{1}{2}-\gamma}}{log(n)}\right), \tag{3.67}$$

or

$$\gamma < \frac{1}{2(1+\beta)}, \quad q_n = o\left(\frac{n^{\beta\gamma}}{log(n)}\right). \tag{3.68}$$

To guarantee the uniqueness property (3.26) for $\theta_\alpha(\equiv q_\alpha)$ White [1992] directly assumed that:
For all small $\epsilon > 0, \exists \delta_\epsilon > 0$ such that $E\left|\theta(Z_{t-1}) - \theta_\alpha(Z_{t-1})\right| > \epsilon$ implies

$$p\left(\left\{\frac{\theta(Z_{t-1}) + \theta_\alpha(Z_{t-1})}{2}\right\} \leq S_t < \theta_\alpha(Z_{t-1}) | \theta(Z_{t-1}) < \theta_\alpha(Z_{t-1}))\right) \;\; > \;\; \delta_\epsilon$$

$$p\left(\theta(Z_{t-1}) \leq S_t < \left\{\frac{\theta(Z_{t-1}) + \theta_\alpha(Z_{t-1})}{2}\right\} | \theta(Z_{t-1}) \geq \theta_\alpha(Z_{t-1})\right) \;\; > \;\; \delta_\epsilon$$

(compare Assumption A.3 of White[1992]). We have replaced this technical assumption which exclude certain degenerate non-uniqueness of the quantiles by the somewhat stronger, but more easily verified assumptions on the conditional density $f_s(x|z)$ in a neibourghood of the quantile $q_\alpha(z)$. Our type of assumptions are rather standard in the context of studying quantile estimators. For the AR-ARCH model (1.1), we have, e.g.,

$$f_s(x|z) = f_\epsilon\left(\frac{s - m(z)}{\sigma(z)}\right) \tag{3.69}$$

where f_ϵ denotes the density of the innovations ϵ_t. Therefore, the assumption on $f_s(x|z)$ is satisfied if $f_\epsilon(u)$ is continuous in a neighbourdhood of the α-quantile Q_α of the innovation distribution and if $f_\epsilon(Q_\alpha) > 0$.
We conclude this section by adding two technical Lemmas used above.

3.2.1.2 Lemma

For real-valued s, y let

$$\begin{aligned} q(s,y) &= |s-y| \left\{ \alpha 1_{[0,\infty)}(s-y) + (1-\alpha) 1_{(-\infty,0)}(s-y) \right\} && (3.70) \\ &= \alpha(s-y)^{+}(1-\alpha)(s-y)^{-}. && (3.71) \end{aligned}$$

Then, for all $y,\ \bar{y},\ s$

$$|q(s,y) - q(s,\bar{y})| \leq \max(\alpha, 1-\alpha).|y-\bar{y}| \leq |y-\bar{y}| \qquad (3.72)$$

Proof:
For $y \leq \bar{y} \leq s$, we have

$$\begin{aligned} \alpha(y-\bar{y}) - (\bar{y}-s) &= q(s,y) - q(s,\bar{y}) \\ &= (s-y) - (1-\alpha)(\bar{y}-y) && (3.73) \end{aligned}$$

which implies, as $(\bar{y}-s), (s-y) \geq 0$,

$$\begin{aligned} \alpha(y-\bar{y}) &\geq q(s,y) - q(s,\bar{y}) \\ &\geq -(1-\alpha)(\bar{y}-y), && (3.74) \end{aligned}$$

and, therefore, $|q(s,y) - q(s,\bar{y})|$ is bounded from above by at least one of the terms $\alpha(\bar{y}-y)$ and $(1-\alpha)(\bar{y}-y)$. For other choices

$$y \leq \bar{y} < s \quad \text{and} \quad s < y \leq \bar{y} \qquad (3.75)$$

we have

$$q(s,y) - q(s,\bar{y}) = \alpha(\bar{y}-y) \quad \text{resp} \quad q(s,y) - q(s,\bar{y}) = (1-\alpha)(y-\bar{y}) \qquad (3.76)$$

which immediately implies the assertion (3.73).$\diamondsuit$

3.2.1.3 Lemma

Let $q(s,y)$ be defined as in the Lemma 3.2.1.3. Let S be a real random variable with density $f(x)$. Let η be the α-quantile of S, i.e. $\alpha = P(S \leq \eta)$.
a)

$$E\left(q(S,y) - E\left(q(S,\eta)\right) = \begin{cases} E\left(S-y\right) 1_{[y,\eta]}(S) & \forall y \leq \eta \\ \\ E\left(y-S\right) 1_{[\eta,y]}(S) & \forall y \geq \eta \end{cases}$$

b) Let $|y-\eta| \geq \delta > 0$. Then, for any lower bounded C of $f(x)$ on $[\eta-\delta, \eta+\delta]$,

$$E\left(q(S,y)\right) - E\left(q(S,\eta)\right) \geq C.\frac{\delta^2}{2}. \tag{3.77}$$

Proof:
a) If $y \leq \eta$, we have

$$\begin{aligned} q(s,y) - q(s,\eta) &= \alpha(\eta - y)1_{[\eta,\infty)}(S) + (1-\alpha)(y-\eta)1_{(-\infty,y)}(S) + \\ &\quad (S - \alpha y - (1-\alpha)\eta)\, 1_{[y,\eta]}(S) \\ &= \alpha(y-\eta)1_{[\eta,\infty)}(S) - (1-\alpha)(\eta-y)1_{(-\infty,\eta)}(S) + \\ &\quad (S - \alpha y - (1-\alpha)\eta + (1-\alpha)(\eta - y))\, 1_{[y,\eta]}(S). \end{aligned}$$

The expectation of the first term is 0, as

$$E\left(1_{[\eta,\infty)}(S)\right) = P\left(S > \eta\right) = 1 - \alpha \tag{3.78}$$

and

$$E\left(1_{(-\infty,\eta)}(S)\right) = P\left(S < \eta\right) = \alpha. \tag{3.79}$$

The second term is just $(S-y)1_{[y,\eta]}(S)$, and the assertion follows for $y \leq \eta$.
The statement for $\eta \geq y$ follows completely anagously.
b) If $y \leq \eta$, we have even $y \leq \eta - \delta$ by assumption.
Using a)

$$\begin{aligned} E\left(q(S,y) - E\left(q(S,\eta)\right.\right. &= \int_y^\eta (u-y)f(u)du \qquad (3.80) \\ &\geq \int_{\eta-\delta}^\eta (u-y)f(u)du \qquad (3.81) \\ &\geq C\int_{\eta-\delta}^\eta (u-y)du \qquad (3.82) \\ &= C\delta\left(\eta - y - \frac{\delta}{2}\right) \geq C.\frac{\delta^2}{2} \qquad (3.83) \end{aligned}$$

where for the first inequality we use that the integrand is nonnegative, for the second one that

$$f(u) \geq C > 0 \quad \text{for} \quad |u-\eta| \leq \delta \tag{3.84}$$

and for the third one $\eta - y \geq \delta$. The case $y \geq \eta$ dealt with analogously.$\Diamond$

3.3 Financial Application

The goodness and the accuracy of this Value-at-Risk methodology based solely on the neural network estimates of the conditional quantile is illustrated throughout the computation of the daily VaR of a holding consisting of one share of DEUTSCHE Bank. As explanatory variables, we use the daily closing prices of BASF, SIEMENS, COMMERZBANK and DAX30 traded on the stock exchange of Frankfurt. Beside the fact, the simulation takes relatively longer time before delivering the ANN VaR estimate, the proposed method provides better back testing results compared to the historical simulation VaR approach or the ANN-EVT-AR-GARCH based VaR.
After a good fitting of the historical daily closing prices on one period, a VaR validation and back testing period consisting of 255 trading days is used. The ANN based VaR displays only 4 exceeds against 6 for the historical simulation based VaR and 4 for the ANN-EVT-AR-GARCH based VaR implemented in the first chapter.

Chapter 4

Financial Predictions using Diffusion Networks

4.1 Introduction

Markovian Diffusion Theory combined with Stochastic Calculus has always been a fundamental concept in the analysis of the daily returns, closing prices or market performances of financial instruments. Active modern research based on methods of Markovian Diffusion Theory and Diffusion Neural Networks have shown that, using Contrastive Hebbian Learning rules (CHL), one can formalize the activation dynamics of diffusion neural networks in the aim to reproduce the entire multivariate probability distributions of a given financial portfolio (see Movellan and Clelland [1993]) up to an acceptable level of accuracy. The Contrastive Hebbian Learning rules (CHL) have some appealing features that enable to capture differences between desired and obtained continuous probability distributions.
Diffusions Networks are type of recurrent neural network with probabilistic dynamics, as models for learning natural signals that are continuous in time and space. Since the solutions for many decision theoretic problems of interest are naturally formulated using probability distributions, it would be desirable to design flexible neural networks frameworks for approximating probability distributions on continuous path spaces. Instead of using ordinary differential equations for describing the evolution of stock prices or portfolio values, diffusion networks are described by a set of stochastic differential equations. Diffusion Neural Networks are an extension of recurrent neural networks in which the dynamics are probabilistic. They have been found very useful in stock price prediction (see Mineiro, Movellan and Williams [1997], Kamijo and Tanigawa [1990], Kimoto and Asakawa [1990], Refenes et al [1993]), Movellan [1997]).
The main advantages of Diffusion Networks over conventional forecasting meth-

ods include simplicity of implementation and good approximation properties (see Warwick et al [1992]). The notion of state plays a vital role in the mathematical formulation of the dynamical system used for describing the changes of the values of financial instruments. The state of such a dynamical system is formally defined as a set of quantities that summarizes all the information about the past behaviour of the system that is needed to uniquely describe its future behaviour. Such approach usually called Contingency can be seen as a class of functions mapping the space of inputs onto the space of possible probability distributions of the outputs. In this chapter, some theoretical results illustrating the use of Diffusion Networks for financial prediction will be the main focus. It will be shown that, under some general regularity conditions allowing existence and uniqueness of solutions of stochastic differential equations, one can approximate the transition probabilities and the log-likelihood functions and derive consistent non-biased predicting algorithm of future values of a considered financial instrument.

The ideas of learning probability distributions with Symmetric Diffusion Networks of Mineiro, Movellan and Williams [1997] and Movellan [1997] combined with the Maximum Likelihood estimation and forecasting algorithm developed in Pedersen [1993], which is based on incomplete observations of stochastic processes, will be combined in order to build consistent estimates of the future market values of a given position. In the first section, the log-likelihood function is will be defined; the concept of transition probabilities will be specified. The first section ends with some general results dealing with the existence and uniqueness of solutions of stochastic differential equations (see Oksendal[1995]). The second section of mainly dedicated to the approximation of the log-likelihood function. Up to some regularity conditions, it will be shown that the approximate probability density functions of the transition probabilities converge in law to the underlying ones. The analysis of the qualitative features and the study of the convergence properties, such consistency and asymptotic normality will also be illustrated (see Pedersen [1994], Dacunha-Castelle and Zmirou [1989]).

Security Price Model in a Continuous Time Setting

$$\begin{cases} dX(t) = b\,(t, X(t), \theta)\, dt \,+\, \sigma\,(t, X(t), \theta)\, dW(t) \\ \\ X(0) = x_0 \end{cases} \tag{4.1}$$

Where:
W denotes a n-dimensional Brownian Motion.

The drift terms $b = (b_1, b_2, ..., b_n)$ and the dispersion matrix, i.e. the volatility matrix σ, are modelled as some neural output functions given by:
For a given synaptic weights $\theta = (\beta, \gamma)$ $\forall i, j = 1, 2, .., n,$, $\forall t \in \Re^+$, $\forall x \in \Re$

$$b_i\,(t, x, \theta) = \beta_i^0(t) + \sum_{j=1}^{q_i^N} \beta_i^j(t)\psi_i(\tilde{x}^T.\gamma_i^j). \tag{4.2}$$

and

$$\sigma_{ij}\,(t, x, \theta) = \beta_{ij}^0(t) + \sum_{k=1}^{q_{ij}^N} \beta_{ij}^l(k)\psi_{ij}(\tilde{x}.\gamma_{ij}^k). \tag{4.3}$$

verifying the following regularity conditions:
Market Coefficients Regularity Conditions
$\forall N \in \mathcal{N}, \forall\ L = ij$ or $l = i$ for $i, j = 1, 2, ..., n,$ there exist a sufficiently large scalar q_L^N and some positive real numbers Δ_L^N, such that:

$$\begin{cases} 1^\circ)\ (q_L^N) \subset \mathcal{N}, \\ \\ 2^\circ)\ (\Delta_L^N) \subset \mathcal{R}^+, \\ \\ 3^\circ)\ \Delta_L^N = o(N^{\frac{1}{4}}) \\ \\ 4^\circ) q_L^N \left(\Delta_L^N\right)^4 \log(q_{ij}^N \Delta_L^N)) = o(N^{\frac{1}{4}}) \end{cases} \tag{4.4}$$

and for any trading interval $[0,\ T]$, $\forall i, j = 1, 2, ...$, $\beta_l(t)$ and $\gamma_l(t)$ are continuous and uniformly bounded functions of the variable t such that:

$$\begin{cases} \sum_{h=0}^{q_L^N} ||\beta_h||_\infty \leq \Delta_N^L, \\ \\ \sum_{h,e=1}^{q_L^N} ||\gamma_{he}||_\infty \leq q_L^N \Delta_L^N. \end{cases} \tag{4.5}$$

where

$$||\beta_h||_\infty = \max_{t\in[0,\ T]} ||\beta_h(t)|| \tag{4.6}$$

The neural activation functions ψ_i and ψ_{ij} are assumed to be monotonically increasing, continuously Lipschitz, bounded, l-finite, twice continously differentiable and verifying:

$$\begin{cases} \lim\limits_{x \to +\infty} \Psi(x) < +\infty, \\ \\ \lim\limits_{x \to -\infty} \Psi(x) > -\infty. \end{cases} \tag{4.7}$$

In fact the drift term and the dispersion matrix are some neural output functions as defined in the previous chapters, with the suitable universal approximating growth conditions. The regularity conditions (4.4), (4.5) and (4.7) justify the financial returns model (4.1) because of the fact that connectionist sieve as described in White [1989] and extended to unbounded random variables in the first chapter do have some universal approximation properties. They are in some sense, dense in the huge set of continuous functions (denseness on compacta) or dense in some rich subset of square integrable functions (denseness with respect to L^m norm). Consequently the neural output functions as chosen in the model (4.1) can be used to approximate accurately, consistently and non-parametrically the mean rate of returns of the portfolio that can be equated as the drift term and the corresponding volatility. We remark that the difference to the set up of the previous chapters is the fact that we consider a multivariate input X_t instead of a univariate one, which complicates notation, but under appropriate regularity conditions; share the same properties with the univariate case. Moreover, we allow for a dynamic feature by letting the weights from hidden to output layer to be stochastic in a continuous time setting.

Question:

Given some historical trading performances of a portfolio, how to train the network to match some desired trading strategies . In others words, given

$$0 = t_0 < \ t_1 \ < \ t_2 \ < \ ... \ < t_k = T$$

some historical trading days and the corresponding financial returns or daily closing $P\&L$ of the portfolio

$$\left(X(t_0)\, X(t_1)\ ...\, X(t_k)\right),$$

how to train the network e.g how to choose and update the time depending synaptic weights θ such that the resulting optimal weights maximizes the log-likelihood function for θ defined by:

$$l_K(\theta) = \sum_{i=1}^{K} \log p\left(t_{i-1}, X(t_{i-1}), t_i, X(t_i); \theta\right) \tag{4.8}$$

where

$$p\left(t_{i-1}, X(t_{i-1}), t_i, X(t_i); \theta\right) \tag{4.9}$$

represents the probability that the network generates a continuous trading path realizing the market value $X(t_{i-1})$ at time t_{i-1} and $X(t_i)$ at the ${t_i}^{th}$ corresponding trading period. $p\,(t_{i-1}, X(t_{i-1}), t_i, X(t_i); \theta)$ are called the transition probabilities. When they are explicitly known, Billingsley [1961]; Dacunha-Castelle and Florence-Zmirou [1986] have shown that the maximum likelihood estimators $\hat{\theta}_K$ which maximizes the likelihood function $l_k(\theta)$ defined in (4.8) are in many cases consistent and asymptotically normally distributed (see Pedersen [1994]). Unfortunately transition densities are usually unknown.
Before starting the heavy artillery leading to a consistent maximum likelihood approximation of the desired synaptic weights with other appealing properties (such as asymptotic normality and consistency), we need to recall first, some fundamental results dealing with the existence and uniqueness of solutions of stochastic differential equations (see Karatzas and Shreves [1991]).

4.2 Existence and Uniqueness for Stochastic Differential Equations

4.2.1 Definitions

4.2.1.1 Strong Solutions of Stochastic Differential Equations

A strong solution of a stochastic differential equation as the one of the considered model, on the given probability space $(\Omega, \mathcal{F}, \mathcal{P})$ with respect to the Brownian Motion W and the initial condition $X(0) = x_0$, is a process
$X = \{\, X_t\,,\ 0 \,\leq\, t \,<\, +\infty \,\}$ with continuous path and verifying:
1°) $\mathcal{P}\,\{\, X(0) = x_0 \,\} = 1$,

2°) $\forall i = 1, 2, .., n\,,\ \forall j;\ \ , 0 \,\leq\, t \,<\, +\infty\,;$

$$Pr\left(\left\{\int_0^t \{|b_i(s, X_s)| + \sigma_{ij}(s, X_s)\}\, ds \,<\, \infty\right\}\right) \,=\, 1, \tag{4.10}$$

3°) $\forall\ 0 \,\leq\, t \,<\, +\infty$, X_t is given almost surely by:

$$X_t \,=\, X_0 \,+\, \int_0^t b(s, X_s)ds \,+\, \sigma(s, X_s)dW(s) \tag{4.11}$$

This strong solution can be viewed as the output of a dynamical system described by the pair of coefficients (b, σ), whose inputs are the initial value x_0, and the Brownian motion W. According to the model (4.1), such output is given as a neural output function of the time depending synaptic weights (β, γ) , adjusted by the time depending bias β_0.

4.2.1.2 Weak Solutions of Stochastic Differential Equations

A weak solution of the stochastic differential equation (4.1) is a triple (X, W), $(\Omega, \mathcal{F}, \mathcal{P})$, and a filtration $\{\mathcal{F}_t\}$ such that:
The process X_t is adapted[1] to $\{\mathcal{F}_t\}$ and 2°) and 3°) of the definition of strong solutions are verified.

4.2.1.3 Theorem: Karatzas and Shreve[1999] or Yatanabe[1971]

Suppose that the coefficients $b(t, X_t)$ and σ satisfy the global Lipschitz and linear growth conditions e.g:
for every $0 \leq t < +\infty$, $(x, y) \in \Re^n \times \Re^n$ and some positive constant K
Assumption1 Global Lipschitz Conditions

$$||b(t,x) - b(t,y)|| + ||\sigma(t,y) - \sigma(t,y)|| \ \leq \ K \times ||x - y||,$$

Assumption2 Linear Growth Bound

$$||b(t,x)||^2 \ + \ ||\sigma(t,x)||^2 \ \leq K \times (1 + ||x||^2)$$

Assumption3 Measureability
Both $b(t,x)$ and $\sigma(t,y)$ are progressively measurable functions.
If A1-A3 are fulfilled, then there exists a uniquely defined stochastic process, which is a strong solution of the stochastic differential equation (4.1).
Consequently we have also existence and uniqueness of a weak solution. In fact the regularity conditions can be weaker to get only the existence and a uniqueness of the weak solution that we need in the coming sections for a proper characterisation of the transition probabilities.
Beyond the three previous conditions, we add a fourth one such that, the martingale problem problem for, $b = (b_1, b_2, ..., b_n)$ and $a = \sigma\sigma^T$ is well posed (see Oksendal).
As stated in Pedersen [1994], to ensure that the stochastic differential equation (4.1), has a weak solution, it is sufficient to require that for $\forall\theta$, the martingale problem for the drift term $b = (b_1, b_2, ..., b_n)$ and $a = \sigma\sigma^T$ is well posed (see Rogers and Williams [1987] or Strook and Varadhan [1979]).
Assumption.4 Coercivess

$$a(t,x) = \sigma\sigma^T \quad \text{is positive definite.}$$

[1] X_t is adapted to $\mathcal{F}_t$ if $\forall t$, X_t is $\mathcal{F}_t$ measurable

4.3 Approximates Log-likelihood Function

This section represents the core of this chapter. It will shown that, under some general conditions, one can accurately estimate the future financial returns of the portfolio (X_t), by generating some artificial market scenarios leading to some consistent estimate to the future financial returns. It will be shown that, the expected returns corresponding to the generated artificial market scenarios converge to the true underlying ones (convergence in L^1). The regularity conditions imposed on the market coefficients are sufficient to provide an explicit formula of the continuous density of the stochastic process driving the approximate returns. It will be shown that the approximate probability density functions of the transition probabilities converge in law to underlying ones. The second part of this section is dealing with the qualitative features and the study of some convergence properties, such consistency and asymptotic normality of the resulting estimators (see Pedersen [1994], Dacunha-Castelle and Zmirou [1989]). The appealing characteristics of the approximate financial returns and the approximate log-likelihood estimators for the synaptic weights justify the choice of neural output functions for drawing the portfolio market values dynamics.

4.3.1 Transition Probabilities

4.3.1.1 Definition

For a given synaptic weight θ and a pair (x, y) consisting of two admissible market values, the corresponding transition probabilities denoted by

$$(P_{s,x,t,y;\theta})_{(0 \leq s \leq t < +\infty; x, y; \theta)}$$

represents the probability that the network generates a continuous trading path realizing the market values y at the trading time t while providing the value x at the previous trading step s. In order to approximate the transition probabilities (usually unknown), one can make use of the following theorem that provides an alternative way for describing the randomness of the dynamic of the returns. In others words, it helps to replace the initial Brownian motion W, by a new one that enables to derive more easily the probability distribution of the returns.

4.3.1.2 Theorem

Under the assumptions A1-A4, one has:
1°) The process $\left(W_t^\theta\right)_{t\geq 0}$ defined by:

$$\begin{cases} W_t^\theta = \int_0^t \sigma(s, X_s; \theta)^{-1} d\left[X_s - x_o - \int_0^s b(u, X_u; \theta)du\right] \\ \\ t \geq 0. \end{cases} \tag{4.12}$$

is a d-dimensional Brownian Motion.
2°) Any solution of (4.1) is also solution of the following stochastic differential equation

$$\begin{cases} dX_t = b(t, X_t; \theta)dt \, + \, \sigma(t, X_t; \theta)dW_t^\theta \\ \\ X_0 = x_0 \end{cases} \tag{4.13}$$

3°) Using the Brownian Motion $\left(W_s^\theta\right)_{s\geq 0}$, $\forall t \geq 0$ and $\forall\, x_0 \,\in\, \Re^d$,one has:

$$X_t = x_0 \,+\, \int_0^t b(s, X_s; \theta)ds \,+\, \int_0^t \sigma(u, X_u; \theta)^{-1} d[W_s^\theta]_u$$

where $[W_s^\theta]_u$ is given by:

$$[W_s^\theta]_u := \int_s^u \sigma(u, X_t; \theta)^{-1} d\left[X_t \,-\, x \,-\, \int_s^t b(v, X_v; \theta)dv\right] \tag{4.14}$$

Proof: See Friedman [1975]; Strook and Varadhan[1979].♢

4.3.1.3 Corollary: Forecasted Market Values

Using the third part of the previous theorem; if at a given time s , the portfolio or the stock prices achieved a market value equal to x, then the future market values are given by:
$\forall\ \theta$, $\forall x \in \Re^d$,$\forall s \geq 0$ such that: $X_s = x$ then,

$$\begin{cases} X_t = x + \int_s^t b(u, X_u; \theta)du \,+\, \int_s^t \sigma(u, X_u; \theta)d[W_s^\theta]_u \\ \\ t \geq s. \end{cases} \tag{4.15}$$

Due to the existence and the uniqueness of a weak solution X_t of (4.1) verifying $X_s = x$, there exist a unique probability measure induced by X_t,. This probability measure, denoted by $P_{\theta;s,x}$, is defined on the sets of the Borel-sigma algebra by: $\forall A \,\in\, \mathcal{B}(\Re^d)$

$$P_{\theta;s,x}(A) := Pr\left(\{X_t \,\in\, A\}\right).$$

For any Borel set A, $P_{\theta;s,x}\left\{X_t \in A\right\}$, represents the probability that the diffusion network generates a trading strategy realizing at a future time t, a portfolio market value belonging to the set A, provided that it equals x at the previous trading time s.
Before starting the approximation procedure leading to the approximate log-likelihood function or providing the approximate transition probabilities, we need to recall some results due to Kloeden and Platen [1991] and dealing with the approximation of solutions of stochastic differential equations. This results can also be equated as the forecasting tool helping to approximate the future market values of the portfolio or the future stock prices.

4.3.2 Proposition: Approximated Financial Returns

Based on the discrete version of (4.15), one can define the approximate financial returns of the given portfolio as follow:
For $N \geq 1$, $\forall\,(s,t) \in [0\ \ t_n]^2$; $\forall\,(x,y) \in \Re^d \times \Re^d$, let $\left(Y_l^N\right)_{l\in[s\ ,\ t]}$ be the stochastic process defined by:

$$\begin{cases} Y_s^N &= & x \\ \\ Y_{\tau_k}^N &= & Y_{\tau_{k-1}}^N + \frac{t-s}{N} b(\tau_{k-1}, Y_{\tau_{k-1}}^N;\theta) + \\ & & a(\tau_{k-1}, Y_{\tau_{k-1}}^N;\theta)^{\frac{1}{2}}.\left([W_s^\theta]_{\tau_k} - [W^\theta s]_{\tau_{k-1}}\right) \\ \\ \tau_k &= & s + k \times \frac{t-s}{N} \end{cases} \tag{4.16}$$

Under the Assumptions A1-A4, one has:

$$Y_{\tau_N}^N = Y_t^N \rightarrow\ X_t \quad \text{in}\ \ L^1(P_{\theta,s,x}) \tag{4.17}$$

Proof: see Kloeden and Platen[1991].$\diamondsuit$

4.3.3 Approximate Likelihood Function

In the cases where the transition probabilities are explicitly known or the volatility term is not depending on θ as in the case of constant volatility like the Black-Scholes world, Liptser and Shiryayev [1977] have shown that, in a continuous time setting, the likelihood function $\mathcal{L}_K(\theta)$ can be defined as follow:
Observing continuously the realizations of the stochastic process X_t driving the portfolio market values or the joint stock prices over some trading interval $[0\ ,\ t_n]$; the continuous likelihood function for the synaptic weights θ is given by:

$$\mathcal{L}_{t_n}^c(\theta) : \ = \ \int_0^{t_n} [b(s, X_s;\theta)]^T a(s, X_s)^{-1} dX_s \tag{4.18}$$

$$- \quad \frac{1}{2}\int_0^{t_n} [b(s,X_s;\theta)]^T a(s,X_s)^{-1}[b(s,X_s;\theta)]ds \qquad (4.19)$$

where

$$a(s,X_s) = \sigma(s,X_s)\sigma(s,X_s)^T.$$

Therefore using the Euler approximation (see Kloeden and Platen [1991]) of Riemann and stochastic integrals and discretizing the interval $[0\ ,\ t_n]$ into

$$0 = t_0\ <\ t_1\ <\ t_2\ <,...\ < t_{n-1}\ <\ t_n$$

leads to the discrete version of the likelihood function defined by:

$$\begin{aligned}\tilde{\mathcal{L}}_n^d(\theta):\ &= \sum_{i=1}^n b(t_{i-1},X_{t_{i-1}})^T\left(a(t_{i-1},X_{t_{i-1}})\right)^{-1}(X_{t_i}-X_{t_{i-1}}) \\ &- \frac{1}{2}\sum_{i=1}^n\left[b(t_{i-1},X_{t_{i-1}};\theta)\right]^T\left(a(t_{i-1},X_{t_{i-1}})\right)^{-1}\left[b(t_{i-1},X_{t_{i-1}};\theta)\right](t_i-t_{i-1}).\end{aligned}$$

Unfortunately, unless the sampling step

$$\Delta := max_{1\leq\ i\ \leq n}\,|t_i - t_{i-1}|$$

is constant or sufficiently small, the Discrete Maximum Likelihood Estimators of θ maximizing $\tilde{\mathcal{L}}_n^d$ are usually strongly biased or inconsistent(see Florens-Zimiou[1989]). Seeking for better alternatives, Pedersen[1994,1999] proposes a sequence of Approximate Likelihood Functions $\left(\mathcal{L}_n^N(\theta)\right)_{N\geq 1}$ replacing $\tilde{\mathcal{L}}_n^d(\theta)$. The basic ideas underlying Pedersen's likelihood approximation approach consists on approximating the transition probabilities $p(s,x,t,y;\theta)$ of the stochastic process X by a sequence of continuous transition densities $p^N(s,x,t,y;\theta)$ consisting of approximating processes that converge to $p(s,x,t,y;\theta)$ in $L^1_{P_{\theta;s,x}}$.

4.3.4 Approximate Transition Probabilities

4.3.4.1 Definition:

<u>**For $N=1$:**</u> $\forall\ (s,t)\in[0\ \ t_n]^2\ ,\ \forall\,(x,y)\ \in\ \Re^d\times\Re^d;$

$$\begin{aligned}p^1(s,x,t,y;\theta)\ &=\ \frac{1}{\sqrt{2\pi(t-s)}^d}\,|\sigma(s,x,\theta)|^{-1} * \qquad (4.20)\\ &\exp\left\{-\frac{1}{2(t-s)} * [g(x,y,\theta)]^T\,|a(s,x,\theta)|^{-1}\,[g(x,y,\theta)]\right\} \qquad (4.21)\end{aligned}$$

where

$$g(x, y, \theta) := y - x - (t - s)b(s, x; \theta)$$

and $|a|$ denotes the determinant of a.
<u>For $N \geq 2$</u>,

$$\begin{aligned} p^N(s, x, t, y; \theta) &= E_{P_{\theta,s,x}} \left(p^{(1)}(\tau_{N-1}, Y^{(N)}_{\tau_{N-1},t,y;\theta}) \right) && (4.22) \\ & && (4.23) \\ &= \int_{\Re^{d \times N-1}} \prod_{k=1}^{N} p^1(\tau_{k-1}, \zeta_{k-1}, \tau_k, \zeta_k) d\zeta_1 d\zeta_2 ... d\zeta_{N-1} && (4.24) \end{aligned}$$

where the sequence $\tau_{k=0,1,2,...}$ is defined by:

$$\tau_k = s + k \times \frac{t - s}{N} \qquad (4.25)$$

and

$$\begin{cases} \zeta_0 = x \\ \zeta_N = y \end{cases} \qquad (4.26)$$

Now we have all the tools, to prove the result stating that the density of the approximates financial returns $Y^N_{\tau_k}$ can be effectively used as the approximate transition densities.

4.3.4.2 Theorem

For some fixed trading periods $0 \leq s < t$, and some admissible market value $x \in \Re^d$, $\forall N \in \mathcal{N}$ the distribution of the approximate return Y^N_t under the probability measure $P_{\theta,s,x}$ has a density $p^N(s, x, t, y; \theta)$ with respect to the d-dimensional Lebesgue measure λ^d, and :
For $N = 1$: $\forall (s, t) \in [0 \ , \ t_n]^2$; $\forall (x, y) \in \Re^d \times \Re^d$,

$$p^1(s, x, t, y; \theta) = \frac{1}{\sqrt{2\pi(t - s)}^d} |a(s, x, \theta)|^{-\frac{1}{2}} \times$$
$$\exp \left\{ -\frac{1}{2(t - s)} \times [y - x - (t - s)b(s, x; \theta)]^T |a(s, x, \theta)|^{-1} [y - x - (t - s)b(s, x; \theta)] \right\}$$

For $N \geq 2$,

$$\begin{aligned} p^N(s, x, t, y; \theta) &= \int_{\Re^{d \times N-1}} \prod_{k=1}^{N} p^1(\tau_{k-1}, \zeta_{k-1}, \tau_k, \zeta_k) d\zeta_1 d\zeta_2 ... d\zeta_{N-1} && (4.27) \\ &= E_{P_{\theta,s,x}} \left(p^1(\tau_{N-1}, Y^{(N)}_{\tau_{N-1},t,y;\theta}) \right) && (4.28) \end{aligned}$$

with $\zeta_0 = x \ \zeta_N = y$

Proof:

Based on the equality (4.16) of the proposition 4.3.2, we derive that:
$\forall 0 \leq \ s \ \leq \ t$ with $X_0 = x$

$$Y_t^1 = x + (t-s) \times b(s,x;\theta) + \sigma(s,x;\theta)\left([W_s^\theta]_t \ - \ [W_s^\theta]_s\right). \tag{4.29}$$

Therefore, Y_t^1, can be seen as an invertible affine transformation of the multinormal random variable $\left([W_s^\theta]_t \ - \ [W_s^\theta]_s\right)$. Using the regularity condition of positive definitness imposed on the diffusion matrix $a(s,x,\theta) = [\sigma(s,x;\theta)] \times [\sigma(s,x;\theta)]^T$ we have that:

$$Y_t^1 \ \sim \ \mathcal{N}_d\left(x + (t-s)b(s,x;\theta),\, (t-s)a(s,x;\theta)\right). \tag{4.30}$$

Therefore, under the probability measure $P_{\theta;s,x}$ the approximate return Y_t^1 , has a density given by:

$$\left.\begin{array}{l} p^1(s,x,t,y;\theta) = \dfrac{1}{\sqrt{2\pi(t-s)}^d}\,|a(s,x,\theta)|^{-\frac{1}{2}} \times \\ \exp\left\{-\dfrac{1}{2(t-s)} \times [y - x - (t-s)b(s,x;\theta)]^T |a(s,x,\theta)|^{-1}\,[y - x - (t-s)b(s,x;\theta)]\,.\right\} \end{array}\right\} \tag{4.31}$$

Based on the Markov Property of $\left\{Y_{\tau_k}^N\right\}_{k=0}^N$ under the probability measure $P_{\theta;s,x}$, the multivariate distribution Y^N defined by:

$$\left(Y_{\tau_1}^N, Y_{\tau_2}^N, ..., Y_{\tau_N}^N = Y_t^N\right) \tag{4.32}$$

is absolutely continuous with respect to the Nd-dimensional Lebesgue measure. Consequently, the corresponding Radon Nykodim derivative provides its density e.g

$$\frac{dY^N}{d\lambda^{Nd}}(y_1, y_2, ..., y_N) = \prod_{k=1}^N p^1\left(\tau_{k-1}, y_{k-1}, \tau_k, y_k; \theta\right) \tag{4.33}$$

Hence, its N^{th} component which is equal to Y_t^N, is absolutely continous with respect the d-dimensinal Lebesgue measure, and has the following density:

$$P^N(s,t,y,x;\theta) = \int_{\Re^{d(N-1)}} \prod_{k=1}^N p^1\left(\tau_{k-1}, \zeta_{k-1}, \tau_k, \zeta_k; \theta\right) d\zeta_1 ... \zeta_{N-1} \tag{4.34}$$

where ζ_0 and ζ_N are given as in (4.26).
In fact (see Pedersen[1993]), the previous equality, can be equated as the Chapman-Kolmogorov equations for the Markov chain $\left\{Y_{\tau_k}^N\right\}_{k=0} N$ under the probability measure $P_{\theta;s,x}$.Therefore,

$$E_{P_{\theta,s,x}}\left(p^1(\tau_{N-1}, Y_{\tau_{N-1},t,y;\theta}^N)\right) = P^N(s,x,t,y;\theta) \tag{4.35}$$

4.3.5 Approximate log-Likelihood Functions for the Synaptic Weights

Having the approximate transition probabilities, we derive the approximate log-likelihood function $\mathcal{L}_n^N(\theta)$ in the following manner:
For N=1; $\mathcal{L}_n^1(\theta)$ is given by:

$$\begin{aligned}\mathcal{L}_n^1(\theta) \;=\; & -\tfrac{nd}{2}\log(2\pi) - \tfrac{d}{2}\sum_{i=1}^{n}\log(t_i - t_{i-1}) - \frac{1}{2}\sum_{i=1}^{n}\log\left(|a(t_{i-1}, X_{t_{i-1}};\theta)|\right) - \\ & \tfrac{1}{2}\sum_{i=1}^{n}(X_{t_i} - X_{t_{i-1}})^T a(t_{i-1}, X_{t_{i-1}};\theta)^{-1}(X_{t_i} - X_{t_{i-1}}) + \\ & \sum_{i=1}^{n} b(t_{i-1}, X_{t_{i-1}};\theta)^T a(t_{i-1}, X_{t_{i-1}};\theta)^{-1}(X_{t_i} - X_{t_{i-1}}) - \\ & \tfrac{1}{2}\sum_{i=1}^{n}(t_i - t_{i-1}) \times b(t_{i-1}, X_{t_{i-1}};\theta)^T a(t_{i-1}, X_{t_{i-1}};\theta)^{-1} b(t_{i-1}, X_{t_{i-1}};\theta)\end{aligned} \tag{4.36}$$

For $N \geq 2$; considering any n-tuples $\bar{N} = (N_1, N_2, ..., N_n)$ sufficiently large and dividing each of the trading subinterval $[t_{i-1}\ \ t_i]$ into a subdivision consisting of N_i parts (non necesssarily equidistant), we obtain the approximate log-likelihood functions $\mathcal{L}_n^{\bar{N}}$ defined by:

$$\mathcal{L}_n^{\bar{N}}(\theta) = \sum_{i=1}^{n}\log\left[p^{N_i}\left(t_{i-1}, X(t_{i-1}), t_i, X(t_i);\theta\right)\right]. \tag{4.37}$$

One important feature about the approximate log-likelihood function $\mathcal{L}_n^1(\theta)$ is the fact that, when the volatility term is not explicitely depending on θ, $\mathcal{L}_n^1(\theta)$ can be equated as the generalisation of the discrete version $\mathcal{L}_n^1(\theta)$ of the underlying log-likelihood function defined in (4.20) because of the fact that:

$$\mathcal{L}_n^1(\theta) = \text{ Constant } + \mathcal{L}_n^d(\theta) \tag{4.38}$$

4.4 Consistency and Asymptotic Normality of the Maximum Likelihood Estimators of the Synaptic Weights

After choosing the n-tuple $\bar{N} = (N_1, N_2, ..., N_n)$, finding the corresponding maximum likelihood estimator consists on maximizing the approximate log-likelihood $\mathcal{L}_n^{\bar{N}}$ e.g:

$$\hat{\theta}_n^{\bar{N}} := \max_{\theta}\mathcal{L}_n^{\bar{N}}(\theta) := \sum_{i=1}^{n}\log\left[p^{N_i}\left(t_{i-1}, X(t_{i-1}), t_i, X(t_i);\theta\right)\right]. \tag{4.39}$$

Before studying the qualitative features of such sequence of estimators, we present first the limiting properties of the transition probabilities in order to derive consequently the convergence (in probability) of the approximate log-likelihood functions $\mathcal{L}_n^N(\theta)$ toward the true underlying one $\mathcal{L}_n(\theta)$. In fact, this convergence, will be used for establishing that the approximate maximum log-likelihood estimator $\hat{\theta}_n^{\bar{N}}$ converges in probability toward the true maximum log-likelihood estimator $\hat{\theta}_n$. Therefore, the usual appealing properties such as consistency or asymptotic normality of the classical maximum likelihood estimator will induce to $\hat{\theta}_n^{\bar{N}}$ the same appealing properties.

4.4.1 Theorem: Limit of the transition densities

Under the regularity conditions and the existence/uniqueness assumptions imposed on the market coefficients b and σ , the approximate transition probabilites $p^N(s,x,t,y;\theta)$ converge toward the true underlying ones e.g.

$$p^N(s,x,t,y;\theta) \;\rightarrow\; p(s,x,t,y;\theta) \quad \text{in} \;\; L^1(\lambda^d) \tag{4.40}$$

Proof: The proof of this theorem will be structured into three main steps:
First Step

In this first step, we are showing that, for a vanishing drift term (b=0), the family $q_{\theta;s,x}$ of probabilitiy measures induced by the weak solutions of the corresponding stochastic differential equations e.g

$$\begin{cases} dX(t) = \sigma\left(t, X(t), \theta\right) dW(t) \\ \\ X(s) = x \end{cases} \tag{4.41}$$

are equivalent to the probability measures $p_{\theta;s,x}$ corresponding to the non vanishing drift terms verifying the regularity conditions.

The discrete version corresponding for the vanishing drift term is given by

$$X_{\tau_k} = X_{\tau_{k-1}} \; + a^{\frac{1}{2}} \left([\tilde{W}_s^{\theta}]_{t_k} \; - \; [\tilde{W}_s^{\theta}]_{t_{k-1}} \right) \tag{4.42}$$

When the drift term is assumed to be identically zero, all the regularity conditions and the assumptions of existence and uniqueness are trivially fulfilled, this implies the existence of a family of probability measures $q_{\theta;s,x}$ induced by the weak solutions of the corresponding equations. Using the same arguments as in the proof of the theorem (4.3.4.2), one can derive the Radon Nykodim derivative of $q_{\theta;s,x}$ with respect to the d-dimensional Lebesgue measure in the following manner:

$$\frac{dq_{\theta;s,x}.X^{(N)}}{d\lambda^{Nd}}(x_1, x_2, ..., x_n) = \prod_{k=1}^{N} \phi_d \left(x_k, x_{k-1}, \frac{t-s}{N} a(\theta) \right) \tag{4.43}$$

where:

$$\begin{cases} N \geq 0, \\ \\ X_{t_k} = X_{t_{k-1}} \; + \; a^{\frac{1}{2}} \left([\tilde{W}^{\theta}_s]_{t_k} \; - \; [\tilde{W}^{\theta}_s]_{t_{k-1}} \right) \\ \\ [\tilde{W}^{\theta}_s]_t \; := \; a^{-\frac{1}{2}} (X_t - x) \quad \text{for} \quad t \geq s \end{cases} \tag{4.44}$$

For the same reasons, for a non vanishing drift term fulfilling the regularity conditions, one has:

$$\frac{dp_{\theta;s,x}.X^{(N)}}{d\lambda^{Nd}}(x_1, x_2, .., x_n) = \prod_{k=1}^{N} \phi_d \left(u(x_k, x_{k-1}, t, s, \theta, N) \right) \tag{4.45}$$

where

$$u(x_k, x_{k-1}, t, s, \theta, N) := \left(x_k, x_{k-1} + \frac{t-s}{N} b(\tau_{k-1}, x_{k-1}), \frac{t-s}{N} a(\theta) \right)$$

and ϕ_d denotes the density function of the d-dimensional normal distribution. From (4.43), (4.44) and (4.45), we derive that $\frac{dp_{\theta;s,x}}{dq_{\theta;s,x}}$ is equal to:

$$\exp \left(\sum_{k=1}^{N} v_k(\theta)^T a(\theta)^{-1} (x_k - x_{k-1}) - \frac{(t-s)}{2N} \sum_{k=1}^{N} b(\tau_{k-1}, x_{k-1}; \theta)^T a(\theta)^{-1} v_k(\theta) \right) \tag{4.46}$$

where

$$v_k(\theta) := b(\tau_{k-1}, x_{k-1}; \theta).$$

Therefore $q_{\theta;s,x}|_{\mathcal{F}_t} \;\sim\; p_{\theta;s,x}|_{\mathcal{F}_t}$ and

$$S := \frac{dp_{\theta;s,x}}{dq_{\theta;s,x}} | \mathcal{F}_t \tag{4.47}$$

is given by:

$$S = \exp \left(\int_s^t b(u, X_u; \theta)^T a(\theta)^{-1} dX_u \; - \; \frac{1}{2} \int_s^t b(u, X_u; \theta)^T a(\theta)^{-1} b(u, X_u; \theta) du \right) \tag{4.48}$$

Second Step

This step consists on proving that:
$\forall N = 1, 2, ...$, the process defined by:

$$S_N := \frac{dp_{\theta;s,x}.Y^{(N)}}{dq_{\theta;s,x}.X^{(N)}} (X^{(N)}) \tag{4.49}$$

converges in L^1 toward S defined in (4.47) and

$$\begin{cases} 1°) \ E(S_N) \ \rightarrow \ E(S) \\ \\ 2°) \ S_N \ \rightarrow \ S \ \text{ in probability.} \end{cases} \tag{4.50}$$

This first claim of (4.50) derives from the use of the regularity conditions and the application of some results in Jacob&Shiryaev[1987], chapter4; Revuz&Yor[1991]. For more details, see Pedersen[1993].
Furthermore,

$$E_{q_{\theta;s,x}}(S_N) = E_{q_{\theta;s,x}}(S) = 1, \tag{4.51}$$

therefore the second claim of (4.50) also holds.
Finally, to prove the convergence of S_N we apply Lemma 1 and Lemma2 in Pedersen[1993].
From Lemma2 we derive that:

$$E_{q_{\theta;s,x}}(S_N|X_t)\phi(.;x,(t-s)a(\theta)) \rightarrow E_{q_{\theta;s,x}}(S|X_t)\phi(.;x,(t-s)a(\theta)) \text{ in } L^1.(4.52)$$

Final Step
The computation of the Radon Nykodim derivatives of $p_{\theta;s,x}.X_t$ with respect to the d-dimensional Lebesgue measure gives:

$$\begin{aligned} p(s,x,t,y;\theta) &= \frac{dp_{\theta,s,x}.X_t}{d\lambda^{Nd}}(y) \\ &= E_{q_{\theta;s,x}}\left(\frac{dp_{\theta;s,x}}{dq_{\theta;s,x}}|_{\mathcal{F}_t}|X_t = y\right) \times \frac{dq_{\theta;s,x}.X_t}{d\lambda^d}(y) \\ &= E_{q_{\theta;s,x}}\left(S|X_t = y\right)\phi_d(y;x,(t-s)a(\theta)) \end{aligned} \tag{4.53}$$

and

$$\begin{aligned} E_{q_{\theta;s,x}}\left(S_N|X_t = y\right) &= E_{q_{\theta;s,x}}\left(\frac{dp_{\theta;s,x}.Y^{(N)}}{dq_{\theta;s,x}.X^{(N)}}\left(X_{\tau_1},...,X_{\tau_{N-1}},y\right)|X_t = y\right) \\ &= \int_{R^{d(n-1)}} \frac{dp_{\theta;s,x}.Y^{(N)}}{dq_{\theta;s,x}.X^{(N)}}\left(\xi_1,...,\xi_{N-1},y\right) \times \\ &\quad \frac{dq_{\theta;s,x}.X^{(N)}}{d\lambda^{Nd}}(\xi_1,...,\xi_{N-1},y).\left(\frac{dq_{\theta;s,x}.X_t}{d\lambda^d}(y)\right)^{-1} d\xi_1...d\xi_{N-1} \\ &= \phi_d(y;x,(t-s)a(\theta))^{-1}p^{(N)}(s,x,t,y;\theta). \end{aligned} \tag{4.54}$$

Therefore, using the convergence established in (4.50) and the equality (4.51), we derive that:

$$p^{(N)}(s,x,t,y;\theta) = E_{q_{\theta;s,x}}(S_N|X_t = y) \times \phi_d(y;x,(t-s)a(\theta)) \tag{4.55}$$

$$\to E_{q_{\theta;s,x}}(S|X_t = y)\phi(.;x,(t-s)a(\theta)) = p(s,x,t,y;\theta) \tag{4.56}$$

One important consequence of the fact that $p^{(N)}(s,x,t,y;\theta) \to p(s,x,t,y;\theta)$, that can also be used for justifying the use of the approximate log-likelihood function for determining the optimal weights is given by the following result:

4.4.1.1 Proposition:

Since $p^{(N)}(s,x,t,.;\theta) \to p(s,x,t,.;\theta)$ then:
$\forall\ 0 \ \le s \ < \ t \ x \in \mathcal{R}^d$ and θ

$$\mathcal{L}_n^N \to \mathcal{L}_n \quad \text{in probability under } P_{\theta_0} \tag{4.57}$$

where θ_0 denotes the true synaptic weights.
Proof: Pedersen[1993] $\diamondsuit$.
Therefore the neural network based approximate transition probabilities converge in law toward the true underlying ones.

Conclusion

In the present work, we investigated how to correct the questionable normality, linear and quadratic assumptions underlying existing Value-at-Risk methodologies. In order to take also into account the skewness, the heavy tailedness and the stochastic feature of the volatility of the market values of financial instruments, the constant volatility hypothesis widely used by existing Value-at-Risk appproches has also been investigated and corrected and the tails of the financial returns distributions have been handled via Generalized Pareto or Extreme Value Distributions. Artificial Neural Networks have been combined by Extreme Value Theory in order to build consistent and nonparametric Value-at-Risk measures without the need to make any of the questionable assumption specified above. For that, either autoregressive models (AR-GARCH) have been used or the direct characterization of conditional quantiles due to Bassett, Koenker [1978] and Smith [1987]. In order to build consistent and nonparametric Value-at-Risk estimates, we have proved some new results extending White Artificial Neural Network denseness results to unbounded random variables and provide a generalisation of the Bernstein inequality, which is needed to establish the consistency of our new Value-at-Risk estimates. For an accurate estimation of the quantile of the unexpected returns, Generalized Pareto and Extreme Value Distributions have been used. The new Artificial Neural Networks denseness results enable to build consistent, asymptotically normal and nonparametric estimates of conditional means and stochastic volatilities. The denseness results uses the Sobolev metric space $L^m(\mu)$ for some $m \geq 1$ and some probability measure μ and which holds for a certain subclass of square integrable functions. The Fourier transform, the new extension of the Bernstein inequality for unbounded random variables from stationary α-mixing processes combined with the new generalization of a result of White and Wooldrige[1990] have been the main tool to establich the extension of White's neural network denseness results. To illustrate the goodness and level of accuracy of the new denseness results, we were able to demonstrate the applicability of the new Value-at-Risk approaches by means of three examples with real financial data mainly from the banking sector traded on the Frankfort Stock Exchange.

Bibliography

Ango, Buehlmann and Doukhan: Weak dependence beyond mixing and asymptotics for nonparametric regression.Annals of Statistics(2001).

Balkema, A. A. and L. de Haan: Residual lifetime at great age, Annals of Probability, 2, 792-804 (1974)

Barnett, E. Berndt, H. White, eds., Dynamic Econometric Modelling. New York: Cambridge University Press, 3-26 (1988).

Barron: Universal Approximation Bounds for Superpositions of Sigmoid Function.IEEE Trans.Inform.Theory 39, 930-945, (1993).

Bassi, Embrechts, Kafetzaki: Risk management and quantile estimation In: A Practical Guide to Heavy Tails, eds. R.J. Adler et al., Boston, Birkhaeuser, pp. 111-130,(1998).

Bassett G.S and Roger Koenker: Regression Quantiles. Econometria, 46:33-50, (1998).

Bassett G.S and Quinshui Zhao: Conditional Quantiles Estimation and Inference for ARCH models. Econometric-Therory, 12:793-813, (1996).

Bates and White: "A Unified Theory of Consistent Estimation for Parametric Models," Econometric Theory, 1, 151-178 (1985).

Bates and White: "Determination of Estimators with Minimum Asymptotic Covariance Matrices," Econometric Theory, 9, 633-648 (1993).

Baxt and White: "Bootstrapping Confidence Intervals for Clinical Input Variable Effects in a Network Trained to Identify the Presence of Acute Myocardial Infarction," Neural Computation, 7, 624-638 (1995).

Bera and Higgins: A Test for Conditional Heteroskedasticity in Time Series Models, Journal of Time Series(1995).

Bera and Higgins: ARCH Models, Properties, Estimation and Testing, Journal of Economic Surveys (Vol. 7 No. 4 pp305-362) (1993).

Bera and Higgins: ARCH and Bilinearity as Competing Models for Nonlinear Dependence, Forthcoming J of Business and Economic Statistics (1996).

Bera and Higgins and Lee: Interaction Between Autocorrelation and Conditional Heteroscedasticity: A Random-Coefficient Approach, J of Business and Economic Statistics(1992).

Bollerslev : Generalized Autoregressive Conditional Heteroskedasticity, J of Econometrics(1986).

Bollerslev : A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return, RES(1987).

Bollerslev : Modelling the Coherence in Short Run Nominal Exchange Rates: A Multivariate Generalized ARCH Model, Review of Economics and Statistics(1990).

Bollerslev and Baille: Prediction in Dynamic Models with Time-Dependent Conditional Variances. Econometrica, 50: 91-114. Carlstein, A. (1992).

Bollerslev, Baillie and H. Mikkelsen: Fractionally integrated generalized autoregressive conditional heteroskedasticity, Working Paper No. 168, Department of Finance, Northwestern University(1993).

Bollerslev and Chou and Kroner: ARCH Modeling in Finance, J of Econometrics (1992).
Bollerslev and Engle and Nelson: ARCH Models, Chapter 49 of the Handbook of Econometrics(1994).

Bollerslev, T., Chou, R.Y., Kroner, K.F: ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence, Journal of Econometrics, 52, 5-59(1992).

Bollerslev, T. and J. Wooldridge: Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances, Econometric Reviews 11, 143-172(1992).

Bosq: Inegalites de Bernstein pour les processus Fortement Mlangeantes non Ncessairement Stationaires. Compte Rendu Hebdomadaire des seances de l'Academie des Sciences Paris, Ser.A, 281, 1095-98(1975).

Boyd and White: "Estimating Data Dispersion Using Neural Networks," Proceedings of the 1994 IEEE Congress On Computational Intelligence, forthcoming.

Buehlmann, Delbaen, Embrechts and Shiryaev, A.: On Esscher Transforms in Discrete Finance Models ASTIN Bulletin 28,171-186,(1998).

Buehlmann: Sieve bootstrap with variable length Markov chains for stationary categorical time series. To appear in Journal of the American Statistical Association(2001).

Buehlmann: Bootstraps for time series. To appear in Statistical Science(2001).

Buehlmann: Bootstrapping time series. Bulletin of the International Statistical Institute, 52nd session. Proceedings, Tome LVIII, Book1, 201-204(1999).

Buehlmann: Confidence regions for trends in time series: a Simultaneous Approach with a Sieve Bootstrap. Tech. Rep. 447. UC Berkeley(1996).

Buehlmann, and Mc Neil: Nonparametric GARCH models(1999).

Buehlmann and Mc Neil: Nonparametric GARCH Models,ETHZ,(1999).

Chapman and A. Wellings and A. Burns: Integrated Program Proof and Worst-Case Timing Analysis of SPARK Ada, Proceedings of the Workshop on Language, Compiler and Tool Support for Real-Time Systems, June (1994).

Chen and White: "Laws of Large Numbers for Hilbert Space-Valued Mixingales With Applications," Econometric Theory, 12, 284-304 (1996).

Chen and White: "Central Limit and Functional Central Limit Theorems for Hilbert Space-Valued Dependent Processes," Econometric Theory 14, 260-284 (1998).

Chen and White: "Nonparametric Learning With Feedback," Journal of Economic Theory, 82, 190-222 (1998).

Chen and White: "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," IEEE Transactions on Information Theory, 45, 682-691 (1999).

Chu, Stinchcombe and White: "Monitoring Structural Change," Econometrica, 64, 1045-1066 (1996).

Corradi and White: " Regularized Neural Networks: Some Convergence Rate Results," Neural Computation, 7, 1201-1220 (1995).

Corradi and White: "Specification Tests for the Variance of a Diffusion," Journal of Time Series Analysis, 20, 253-270 (1999).

Corradi, Swanson, and White: "Testing for Stationarity-Ergodicity and for Comovement Between Nonlinear Discrete Time Markov Processes." Journal of Econometrics, forthcoming.

Cox and White: "Unanticipated Money, Output and Prices in the Small Economy," Economic Letters, 1, 23-27 (1978).

Davidson, MacKinnon and White: "Tests for Model Specification in the Presence of Alternative Hypotheses: Some Further Results," Journal of Econometrics, 21, 53-70(1983).

Delbaen Freddy,Philippe Artzner, Jean-Marc Eber and David Heath: Coherent Measures of Risk, Math. Finance 9 , no. 3, 203-228(1999).

Domowitz and White: "Misspecified Models with Dependent Observations," Journal of Econometrics, 20, 35-50, (1982).

Duffie: Dynamic Asset Pricing Theory, Princetin University Press(1992)

Efron, B: Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1-26(1979).

Embrechts, Mikosch: Mathematical Models in Finance,(2000).

Embrechts: Actuarial versus financial pricing of insurance. Risk Finance 1, no. 4, 17-26,(2000).

Embrechts: Extreme Value Theory: Potential and Limitations as an Integrated Risk Management Tool Derivatives Use, Trading and Regulation 6, 449-456, (2000).

Embrechts,Walk: Recursive estimation of distributional fix-points Journal of Applied Probability 37, 73-87,(2000).

Embrechts, Haan, Huang: Modelling multivariate extremes Extremes and Integrated Risk Management (Ed. P. Embrechts) RISK Books, 59-67,(2000).

Embrechts, Mc Neil, Straumann: Correlation: Pitfalls and alternatives A short, non-technical article, RISK Magazine, May,69-71,(1999).

Embrechts, Resnick, Samorodnitsky: Living on the Edge RISK, January 1998, 96-100. Also published in: Hedging with Trees:Advances in Pricing and Risk Managing Derivatives, M. Broadie and P. Glasserman (eds.), Risk Books, New York, pp. 239-243,(1998).

Embrechts, Klueppelberg, Mikosch: Modelling Extremal Events for Insurance and Finance,(1997).

Embrechts, Resnick, Samorodnitsky: Extreme value theory as a risk management tool North American Actuarial Journal 3, 30-41,(1999).

Embrechts, Mc Neil, Straumann: Correlation: pitfalls and alternatives. RISK, May 1999: pages 69-71,(1999).

Embrechts, Mc Neil and Straumann: Correlation and dependency in risk management: properties and pitfalls. In Risk management: value at risk and beyond, edited by Dempster M and Moffatt HK, published by Cambridge University Press (yet to appear),(2000).

Embrechts, Samorodnitsky: Ruin problem, operational risk and how fast stochastic processes(1997).

Embrechts, et al.: An Academic Response to Basel II. Financial Markets Group, London School of Economics,(2001).

Embrechts, Mc Neil and Straumann.: Correlation and dependency in risk management: properties and pitfalls ,(2001).

Embrechts, Hoeing, Juri: Using Copulae to bound the Value-at-Risk for functions of dependent risk,(2001).

Embrechts, Chavez-Demoulin: Smooth Extremal Models in Finance and Insurance,(2001).

Embrechts,Lindskog and Mc Neil: Modelling Dependence with Copulas and Applications to Risk Management(2001).

Embrechts, Frey, Furrer: Stochastic Processes in Insurance and Finance In: Handbook of Statistics, vol. 19 'Stochastic Processes: Theory and Methods', Elsevier Science, Amsterdam, pp. 365-412,(2001).

Engle, R.F:Autoregressive conditional heteroskedasticity with estimates of the united kingdom inflation, Econometria 50(4), pp 987-1008,[1982]

Engle, R.F:Autoregressive conditional heteroskedasticity with estimates of the united kingdom inflation, Econometria 50(4), pp 987-1008,[1982]

Engle, R.F; Granger, C.W.J and Kraft: Combining Competing Forecasts of inflation using a Bivariate ARCH model, Journal of Economics Dynamics and Control 8, pp 151-165,[1984]

Engle, R.F; Lilien, D,M and Robins, R,P: Estimating time varying risk premia in the term structure: the ARCH-m-model, Combining Competing Forecasts of inflation using a Bivariate ARCH model, Econometria 55(2), pp 391-407,[1987]

Engle: Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, Econometrica (1982).

Engle: Volatility: Statistical Models for Financial Data, WP Notes UCSD (1991).

Engle: Statistical Models for Financial Volatility [low tech summary of types of ARCH models], FAJ, 49(1), pp72-78 (2 copies) (1993).

Engle and Bollerslev : Modelling the Persistence of Conditional Variances, Econometric Reviews (incl. Comments from Diebold, Geweke, Pantula, Zin, and Hendry's "An Excursion into Conditional Varianceland") (1986).

Engle and Gonzalez-Rivera: Semiparametric ARCH Models, J of Business and Economic Statistics (2 copies) (1991).

Engle and Hendry and Trumble: Small Sample Properties of ARCH Estimators and Tests, Canadian Journal of Economics (1985).
Engle and Ito and Lin: Meteor Showers or Heat Waves? Heteroskedastic Intra-Daily Volatility in the Foreign Exchange Market, Econometrica (1990).

Engle and Lilien and Robins: Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model, Econometrica(1987).

Engle and Mustafa: Implied ARCH Models from Option Prices (incl. ARCH and Options), JET (2 copies), (1992).

Engle and Ng: Measuring and Testing the Impact of News on Volatility, WP UCSD, (1991) .
Engle and Rothschild: Editors Introduction to Statistical Models for Financial Volatility, J of Econometrics (1992).

Franke, Schlinder and Siedow: Finanzinnovationenen (Grundlagen und Praxis der Optionpreisbestimmung), Report 7 in WirtschaftsMathematik, University of Kaiserslautern (1996).

Franke, Wolfang and Kreiss: Nonparametric Estimation in a Stochastic Volatility Model,Report 37 in WirtschaftsMathematik, University of Kaiserslautern (1997).

Franke and Neunmann: Bootstrapping Neural Networks,Report 38 in Wirtschafts-Mathematik, University of Kaiserslautern (1998).

Franke : Nonlinear and Nonparametric Methods for Analysing Financial Time Series, Report 44 in WirtschaftsMathematik, University of Kaiserslautern (1998).

Franke and Kreiss: Bootstrap Autoregressive Order Selection Report 46 in WirtschaftsMathematik, University of Kaiserslautern (1999).

Franke and Klein: Optimal Portfolio Management Using Neural Networks, Report 49 in WirtschaftsMathematik, University of Kaiserslautern (1999).

Franke: Portfolio Management and Market Risk Quantification Using Neural Networks, Report 58 in WirtschaftsMathematik, University of Kaiserslautern (1999).

Freisleben, B: The neural composer: A network for musical applications. In Proceedings of the 1992 International Conference on Artificial Neural Networks (Vol. 2, pp. 1663-1666). Amsterdam: Elsevier(1992) .

Freisleben, B: Thilo Kielmann: Automatic Parallelization of Divide-and-Conquer Algorithms. CONPAR pp 849-850 (1992).

Freisleben, B: Stock Market Prediction with Backpropagation Networks. IEA/AIE pp 451-460,(1992)

Freisleben, B and Hans-Henning Koch, Oliver E. Theel: Providing Low Cost Read Access to Replicated Data with Multi-Level Voting. INDC : pp 357-376, (1992).

Frey and Mc Neil A: Modelling Dependent Defaults,ETHZ,(2000).

Gallant and White: A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models. Oxford: Basil Blackwell (1988).

Gallant and White: "On Learning the Derivatives of an Unknown Mapping with Multilayer Feedforward Networks," Neural Networks, 5, 129-138 (1992).

Gallant and White: "There Exists a Neural Network That Does Not Make Avoidable Mistakes," Proceedings of the Second Annual IEEE Conference on Neural Networks,I:657-664 (1988).

Ghysels, E., A. Harvey and E. Renault : Stochastic Volatility, in Maddala, G.S. and C.R. Rao (eds.): Handbook of Statistics, Vol. 14, Elsevier Science B.V(1996).

Gnedenko , B: Sur la distribution limite du terme maximum d'une serie aleatoire, Annals of Mathematics, 44, 423-453.[17] (1943).

Goldbaum, Sample, White and Weinreb: "Interpretation of Automated Perimetry for Glaucoma by Neural Networks," Investigative Ophthamology and Visual Science, 35, 3362-3373 (1994).

Granger, White and Kamstra: "Interval Forecasting: An Analysis Based Upon ARCH-Quantile Estimators," Journal of Econometrics, 40, 87-96 (1989).

Grenader and M. Rosenblatt: Statistical analysis of stationary time series. Wiley. New York(1957).

Hong and White: "Consistent Specification Testing Via Nonparametric Series Regression," Econometrica, 63, 1133-1160 (1995).

Hornik, Stinchcombe and White: "Multilayer Feedforward Networks are Universal Approximators," Neural Networks, 2, 359-366 (1989).

Hornik, Stinchcombe and White: "Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks," Neural Networks, 3, 551-560 (1990).

Hornik, Stinchcombe, White and Auer: "Degree of Approximation Results for Feedforward Networks Approximating Unknown Mappings and Their Derivatives," Neural Computation, 6, 1262-1274 (1994).

Hosking, J.R.M. and Wallis, J.R: Parameter and quantile estimation for the generalised Pareto distribution. Technometrics 29, 339-349 (1987)

Hull, J., and A. White: The Pricing of Options on Assets with Stochastic Volatilities, Journal of Finance, 52, 281-300(1987).

Ikeda and Watanabe: Stochastic Differential Equation and Diffusion Processes, North Holland, New York(1981).

James Chu and White: "A Direct Test For Changing Trends," Journal of Business and Economic Statistics, 10, 289-299 (1992).

Jingtao Yao, Chew Lim Tan and Hean-Lee Poh: Neural Networks for Technical Analysis: a Study on KLCI, International Journal of Theoretical and Applied Finance(Quarterly), Vol. 2, No.2, pp221-241 (1999).

Jingtao Yao, Nicholas Teng, Hean-Lee Poh, Chew Lim Tan: Forecasting and Analysis of Marketing Data Using Neural Networks, Journal of Information Science and Engineering (Quarterly), Vol. 14, No.4, pp523-545(1998).

Jingtao Yao, Hean-Lee Poh,Teo Jasic: Neural Networks for the Analysis and Forecasting of Advertising and Promotion Impact, International Journal of Intelligent Systems in Accounting, Finance and Management (Quarterly), Vol. 7, No. 4, pp253-268(1998).

Jingtao Yao, Hean-Lee Poh, Teo Jasic: Foreign Exchange Rates Forecasting with Neural Networks, ICONIP'96 (International Conference on Neural Information Processing), Hong Kong, Sept. 24-27, pp754-759(1996).
Jorion: Value-at-Risk:The new Benchmark for Controlling Market Risk(1997).

Karatzas and Shreve: Brownian Motion and Stochastic Calculus,Springer(1991)

Komolgorov and S.V. Fomin, Introductory Real Analysis. Dover, New York, (1970).

Korn : Optimal Portfolio, World Scientific, Singapore(1997).

Korn and C.Klueppelberg: Optimal Porfolios with Bounded Capital-at-Risk, Johannes Gutenberg Universtaet Mainz(1998).

Kuan and White: "Artificial Neural Networks: An Econometric Perspective," Econometric Reviews, 13, 1-92 (1994).

Kuan, Hornik and White: "A Convergence Result for Learning in Recurrent Neural Networks," Neural Computation, 6, 420-440 (1994).

Kuan and White: "Adaptive Learning with Nonlinear Dynamics Driven by Dependent Processes," Econometrica, 62, 1087-1114 (1994).

Lee, White and Granger: "Testing for Neglected Nonlinearity in Time-Series Models: A Comparison of Neural Network Methods and Standard Tests," Journal of Econometrics, 56, 269-290 (1992).

MacDonald and White: "Some Large Sample Tests for Nonnormality in the Linear Regression Model," Journal of the American Statistical Association, 75, 16-27 (1980).

MacKinnon and White: "Some Modified Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," Journal of Econometrics, 29,305-325 (1985).

Matyas, J: Random optimization. Automation and Remote Control, 26, 244-251 (1965) .

Mc Neil: Calculating quantile risk measures for financial time series using extreme value theory,ethz (1998).

Mc Neil and Frey: Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. Journal of Empirical Finance, 7: 271-300,(2000).

Mc Neil and Saladin T: Developing scenarios for future extreme losses using the POT method. In Extremes and Integrated Risk Management, edited by Embrechts PME, published by RISK books, London,(2000).

Mc Neil: Reading the Riskometer. In Extremes and Integrated Risk Management, edited by Embrechts PME,London,(2000).

Mc Neil: Extreme value theory for risk managers. Internal Modelling and CAD II published by RISK Books, 93-113,(1999).

Mc Neil: On Extremes and Crashes. RISK, January 1998: page 99.

Mc Neil: Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bulletin, 27: 117-137.

Mc Neil and Saladin: The peaks over thresholds method for estimating high quantiles of loss distributions. Proceedings of 28th International ASTIN Colloquium.

Merton: Continuous-Time Finance, Basis Blackwell, Cambridge MA(1990).

Messer and White: "A Note on Computing the Heteroskedasticity Consistent Covariance Matrix Using Instrumental Variable Techniques," Oxford Bulletin of Economics and Statistics, 46, 181-184 (1984).

Movellan Javier R, Paul Mineiro, R.J. Williams: Modeling Path Distributions Using Partially Observable Diffusion Networks: A Monte-Carlo Approach. Technical Report USCD, (1997)

Movellan Javier R, McCelland: Learning Continous probability distributions with symmetric diffusion networks. Cognitive Science, 17, 463-496. 8 (1993)

Nelson, D: ARCH models as diffusion approximations. Journal of Econometrics 45, 7-38(1990).

Norio Baba: Global optimization of functions by the random optimization method. Int. J. Control 30, 1061-1065, (1979).

Norio Baba: Convergence of a random optimization method for constrained optimization problems. J. Optimization Theory Appl. 33, 451-461, (1981).

Norio Baba and Akira Morimoto: Three approaches for solving the stochastic multiobjective programming problem. Stochastic optimization. Numerical methods and technical applications, Proc. GAMM/IFIP-Workshop, Neubiberg/Ger. 1990, Lect. Notes Econ. Math. Syst. 379, 93-109, (1992).

Norio Baba and Akira Morimoto: Stochastic approximation method for solving the stochastic multiobjective programming problem. Int. J. Syst. Sci. 24, No.4, 789-796, (1993).

Oksendal: Stochastic Differential Equations, Springer-Verlag,(1995)

Olson, Shefrin and White: ”Optimal Investment in Schooling When Incomes Are Risky,” Journal of Political economy, 87, 522-539 (1979).

Ormoneit and White: ”An Efficient Algorithm to Compute Maximum Entropy Densities,” Econometric Reviews, 18, 127-141 (1999).

Pedersen, A.R: Spurious results in therapeutic drug monitoring research. Research Report No. 2002-1, Department of Biostatistics, University of Aarhus.(2002)

Pedersen, A.R: Likelihood inference by Monte Carlo methods for incompletely discretely observed diffusion processes. Research Report No. 2001-1, Department of Biostatistics, University of Aarhus.(2001)

Pedersen, A.R., Petersen, S.O., Vinther, F.P: A stochastic diffusion model for estimating trace gas emissions with static chambers. Research Report No. 2000-2, Department of Biostatistics, University of Aarhus (2000).

Pedersen, A.R : Measuring the nitrous oxide emission rate from the soil surface by means of the Cox, Ingersoll and Ross process. Technical Report No. 11, Biometry Research Unit, Danish Institute of Agricultural Sciences (1998).

Pedersen, A.R : Statistical Analysis of Gaussian Diffusion Processes Based on Incomplete Discrete Observations. Research Reports No. 297, Department of Theoretical Statistics, University of Aarhus (1994).

Pedersen, A.R : Quasi-likelihood Inference for Discretely Observed Diffusion Processes. Research Reports No. 295, Department of Theoretical Statistics, University of Aarhus (1994).

Pedersen, A.R : Uniform Residuals for Discretely Observed Diffusion Processes. Research Reports No. 292, Department of Theoretical Statistics, University of Aarhus (1994).

Pedersen, A.R : Maximum Likelihood Estimation Based on Incomplete Observations for a Class of Discrete Time Stochastic Processes by Means of the Kalman Filter. Research Reports No. 272, Department of Theoretical Statistics, University of Aarhus (1993).

Pickands, J.: Statistical inference using extreme order statistics, Annalsof Statistics, 3, pp 119-131.[35] (1975).

Plutowski, Sakata and White: "Cross-Validation Estimates Integrated Mean Squared Error," in J. Cowan, G. Tesauro, and J. Alspector, eds., Advances in Neural Information Processing Systems 6. San Francisco: Morgan Kaufmann, 391-398 (1994).

Plutowski and White: "Selecting Exemplars for Training Feedforward Networks From Clean Data," IEEE Transactions on Neural Networks, 4, 305-318 (1993).

Plutowski and Cottrell and White: "Experience with Selecting Examplars From Clean Data," Neural Networks, 9, 273-294 (1996).

Refenes and White: "Neural Networks and Financial Economics," International Journal of Forecasting, 17 (1998).

Plutowski, Cottrell and White: "Learning Mackey-Glass from 25 Examples, Plus or Minus 2," in J. Cowan, G. Tesauro, and J. Alspector, eds., Advances in Neural Information Processing Systems 6. San Francisco: Morgan Kaufmann, 1135-1142 (1994).

Sakata and White: "High Breakdown Point Conditional Dispersion Estimation with Application to $S\&P$ 500 Daily Returns Volatility," Econometrica 66, 529-568 (1998).

Sin and White: "Information Criteria for Selecting Possibly Misspecified Parametric Models," Journal of Econometrics, 71, 207-225 (1996).

Smith, R.L.,Estimating tails of probability distributions. A. Statist. 15, 11741207 (1987)

Stinchcombe and White: "Universal Approximation Using Feedforward Networks with Non-Sigmoid Hidden Layer Activation Functions," Proceedings of the International Joint Conference on Neural Networks, I: 612-617 (1989).

Stinchcombe and White: "Approximating and Learning Unknown Mappings Using Multilayer Feedforward Networks with Bounded Weights," in Proceedings of the International Joint Conference on Neural Networks, III: 7-16 (1990).

Stinchcombe and White: "Some Measurability Results for Extrema of Random Functions over Random Sets," Review of Economic Studies, (1992).

Stinchcombe and White: "Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative," Econometric Theory, 14, 295-324(1998).

Stinchcombe and White: "Using Feedforward Networks to Distinguish Multivariate Populations," Proceedings of the International Joint Conference on Neural Networks, (1992).

Stout : A Wide Area Computation System. Technical report, School of Computer Science, Carnegie Mellon University, PhD thesis. Available as Technical Report CMU-CS94-230 (1994).

Sullivan, Timmermann, and White: "Data Snooping, Technical Trading Rule Performance, and the Bootstrap," Journal of Finance, 54, 1647-1692 (1999).

Swanson and White: "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks," Review of Economics and Statistics, 79, 540-550 (1997).

Swanson and White: "Forecasting Economic Time Series Using Adaptive Versus Nonadaptive and Linear Versus Nonlinear Econometric Models," International Journal of Forecasting, 13, 439-461 (1997).

Swanson and White: "A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks," Journal of Business and Economic Statistics, 13, 265-276 (1995).

Weidmann: Lineare Operatoren in Hilbertraeumen.Teubner Stuttgart, (1976).

White: Asymptotic Theory For Econometricians. New York: Academic Press (1984).

White: Estimation, Inference, and Specification Analysis. New York: Cambridge University Press (1994).

White: "Model Specification: Annals," Journal of Econometrics, 20 (1982).
White: Artificial Neural Networks: Approximation and Learning Theory. Oxford: Basil Blackwell (1992).

White: Advances in Econometric Theory: The Selected Works of Halbert White, Cheltenham: Edward Elgar (1998).

White: "Using Least Squares to Approximate Unknown Regression Functions," International Economic Review, 21, 149-170 (1980).

White: "Nonlinear Regression on Cross-Section Data," Econometrica, 48, 721-746 (1980).

White: "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, 48, 817-838 (1980).

White: "Consequences and Detection of Misspecified Nonlinear Regression Models," Journal of the American Statistical Association, 76, 419-433 (1981).

White and Olson: "Conditional Distribution of Earnings, Wages and Hours for Blacks and Whites," Journal of Econometrics, 17, 263-285 (1981).

White: "Maximum Likelihood Estimation of Misspecified Models," Econometrica, 50, 1-25 (1982).

White: "Regularity Conditions for Cox's Test of Non-nested Hypotheses," Journal of Econometrics, 19, 301-318 (1982).

White and Domowitz: "Nonlinear Regression with Dependent Observations," Econometrica, 52, 143-162 (1984).

White: "Maximum Likelihood Estimation of Misspecified Dynamic Models," in T.K. Dijkstra, ed., Misspecification Analysis. New York: Springer-Verlag, 1-19 (1984).

White: "Instrumental Variables Analogs of Generalized Least Squares Estimators," Advances in Statistical Analysis and Statistical Computing, 1, 173-227 (1986).

White: "Specification Testing in Dynamic Models," in Truman Bewley, ed., Advances in Econometrics. New York: Cambridge University Press (1987). Also appears in French as "Test de Specification dans les Modeles Dynamiques," Annales de l'INSEE, 59/60, 125-181 (1985).

White: "Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns," Proceedings of the Second Annual IEEE Conference on Neural Networks,II:451-458. (1988).

White: Asymptotic Theory for Econometricians, Academic Press, Orlando, Florida.(1984)

White: "The Encompassing Principle for Non-Nested Dynamic Model Specification," American Statistical AssociationProceedings of the Business and Economics Statistics Section, 101-109 (1988).

White: "A Consistent Model Selection Procedure Based on m-Testing," in C.W.J. Granger, ed., Modelling Economic Series: Readings in Econometric Methodology. Oxford: Oxford University Press, 369-403 (1989).

White: "Some Asymptotic Results for Learning in Single Hidden Layer Feedforward Network Models," Journal of the American Statistical Association, 84, 1003-1013 (1989).

White: "Learning in Artificial Neural Networks: A Statistical Perspective," Neural Computation, 1, 425-464 (1989).

White: "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings," Neural Networks, 3, 535-549 (1990).

White and Wooldridge: "Some Results for Sieve Estimation with Dependent Observations," in W. Barnett, J. Powell and G. Tauchen, eds., Nonparametric and Semi-Parametric Methods in Econometrics and Statistics. New York: Cambridge University Press, 459-493 (1991).

White and Stinchcombe: "Adaptive Efficient Weighted Least Squares with Dependent Observations," in W. Stahel and S. Weisberg, eds., Directions in Robust Statistics and Diagnostics, IMA Volumes in Mathematics and Its Applications. New York: Springer-Verlag, 337-364 (1991).

White: "Nonparametric Estimation of Conditional Quantiles Using Neural Networks," in Proceedings of the Symposium on the Interface. New York: Springer-Verlag, 190-199(1992).

White: "Parametric Statistical Estimation Using Artificial Neural Networks: A Condensed Discussion," in V. Cherkassky ed., From Statistics to Neural Networks: Theory and Pattern Recognition Applications. NATO-ASI Series F. New York: Springer-Verlag, 127-146 (1994).

White: "Parametric Statistical Estimation Using Artifical Neural Networks," in P. Smolensky, M.C. Mozer and D.E. Rumelhart, eds., Mathematical Perspectives on Neural Networks. Hilldale, NJ: L. Erlbaum Associates, 719-775 (1996).

White and Hong: "M-Testing Using Finite and Infinite Dimensional Parameter Estimators," in R. Engle and H. White, eds., Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger. Oxford: Oxford University Press, 326-345 (1999).

White: "Comment on The Unification of the Asymptotic Theory of Nonlinear Models'," Econometric Reviews, 1, 201-205 (1982).

White: "Comment on Tests of Specification in Econometrics' by Paul A. Ruud," Econometric Reviews, 3, (1985).

White: "Misspecification, Tests for," Encyclopedia of the Statistical Sciences, v. 5. New York: Wiley, 552-555 (1985).

White: "Least Squares," The New Palgrave. London: MacMillian (1987).

White: "Some Asymptotic Results for Back-Propagation," Proceedings of the First Annual IEEE Conference on Neural Networks, III:261-266 (1987).

White: "White Tests of Misspecification," Encyclopedia of the Statistical Sciences, v. 9. New York: Wiley 594-596 (1988).

White: "An Additional Hidden Unit Test for Neglected Nonlinerarity in Multilayer Feedforward Networks," Proceedings of The International Joint Conference on Neural Networks, II:451-455. (1989).

White: "Neural Network Learning and Statistics," AI Expert, 4, 48-52 (1989).

White: "Comment on Basic Structure of the Asymptotic Theory in Dynamic Nonlinear Econometric Models. II. Asymptotic Normality," Econometric Reviews, 10, 345-348 (1991).

White and Gallant: A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models. Oxford: Basil Blackwell (1988).

White and Stinchombe: Multilayer Feedforward Networks are Universal Approximators, Neural Networks, 2, 359-366, (1984).

Wooldridge and White: "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes," Econometric Theory, 4, 210-230 (1988).

Wooldridge and White: Some Results on Sieve Estimation with dependents Observations. In W.Banett, Powell and G.Taucher(eds) Nonparametric and Semiparametric methods in econometrics and statistics.New York: Cambridge University Press(1990)

Yukich, Stinchcombe and White: "Sup Norm Approximation Bounds for Networks Through Probabilistic Methods," IEEE Transactions on Information Theory, 41,1021-1027 (1995).