Eur. Phys. J. B **20**, 481–492 (2001)

THE EUROPEAN
PHYSICAL JOURNAL B
EDP Sciences

Value-at-risk prediction using context modeling

K. Denecker[1,a], S. Van Assche[1], J. Crombez[2], R. Vander Vennet[2], and I. Lemahieu[1]

[1] Department of Electronics and Information Systems, Ghent University Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
[2] Department of Financial Economics, Ghent University Sint-Pietersplein 5, 9000 Ghent, Belgium

Received 2 September 2000 and Received in final form 12 October 2000

Abstract. In financial market risk measurement, Value-at-Risk (VaR) techniques have proven to be a very useful and popular tool. Unfortunately, most VaR estimation models suffer from major drawbacks: the log-normal (Gaussian) modeling of the returns does not take into account the observed fat tail distribution and the non-stationarity of the financial instruments severely limits the efficiency of the VaR predictions. In this paper, we present a new approach to VaR estimation which is based on ideas from the field of information theory and lossless data compression. More specifically, the technique of context modeling is applied to estimate the VaR by conditioning the probability density function on the present context. Tree-structured vector quantization is applied to partition the multi-dimensional state space of both macroeconomic and microeconomic priors into an increasing but limited number of context classes. Each class can be interpreted as a state of aggregation with its own statistical and dynamic behavior, or as a random walk with its own drift and step size. Results on the US S&P500 index, obtained using several evaluation methods, show the strong potential of this approach and prove that it can be applied successfully for, amongst other useful applications, VaR and volatility prediction. The October 1997 crash is indicated in time.

PACS. 02.50.-r Probability theory, stochastic processes, and statistics – 89.70.+c Information science

1 Introduction

Faced with volatile financial markets, both banks and non-financial companies are investing considerable resources in risk management systems. As a result, risk management is increasingly becoming a quantitative discipline. According to international standards elaborated by multinational organizations, most notably the Bank for International Settlements, banks and other financial intermediaries have to maintain capital against a number of potential risks, of which counterparty risk, market risk and interest rate risk are the most important ones. Most countries and financial supervisors have translated these guidelines into their financial legislation and their regulatory practice. The ultimate goal is to guarantee a sufficient degree of financial stability, in view of the potential contagion effects of situations of financial distress in parts of the financial sector and their negative spill-overs to the real sector.

The approach adopted for the calculation of capital adequacy standards has traditionally been rule-based. In such a framework, the types of risks are identified and quantified within each institution according to established methods of computation, and a predetermined level of capital has to be allocated. Increasingly, however, it has become clear that this framework may induce regulatory arbitrage whereby innovative financial contracts are used to migrate certain risks to the risk category with the lowest capital adequacy requirements [1,2]. Moreover, regulators and supervisors are confronted with a rapidly changing competitive financial environment in which both the organization of financial intermediaries (*e.g.*, local commercial banks *versus* internationally diversified financial conglomerates) and the types of risk (*e.g.*, operational risk *versus* market risk) are shifting. In this setting, international regulators and supervisors are gradually moving from a purely rule-based approach of capital adequacy to a more market-based approach in which eligible banks are allowed to use good-practice internal risk management systems to calculate the optimal level of capital coverage.

In the area of market risk, value-at-risk (VaR) models are widely used by financial institutions and non-financial companies [3]. Market risk are the losses arising from adverse movements in market prices (*e.g.* equity prices) or market rates (*e.g.* interest and exchange rates). Value-at-risk is a summary statistical measure of possible portfolio losses under normal market conditions. Losses greater than the VaR are suffered only with a pre-specified probability, assuming a specific distribution of the relevant market variables. The intuitive appeal of VaR estimates arises from the fact that it provides a consistent measure of risk across different positions and risk factors, taking into account the correlation structure between the risk factors. Since the VaR methodology yields the maximum amount that can be lost with a particular confidence level over a specific time period, the VaR forecast can be used to

[a] e-mail: denecker@elis.rug.ac.be

determine capital requirements at the firm level. The accuracy of the VaR estimates under different methodologies is of crucial importance since there is a cost associated with holding both too low and too high levels of capital.

This paper presents some results of using context modeling, a state-of-the-art statistical data compression technique, for increasing the accuracy of the VaR forecast. Data compression is the science that aims at finding the shortest equivalent representation of a given data stream. Both in text and image compression, context models are statistical models that have shown to be very efficient [4,5]. Instead of estimating one comprehensive probability distribution function for the whole text or image, it builds multiple distributions in parallel based upon the value of the context. The probability associated with each new data sample is then determined from the distribution of the samples corresponding with the same context class. Typically, in text compression, the context is the combination of a limited number of nearby characters.

The use of context modeling is intuitively appealing. By defining a set of priors that are theoretically or empirically found to be informative in forecasting future market movements, different context classes can be defined. The set of priors should reflect the different forms in which market risk can occur. Examples are changes in interest rates, exchange rates and business cycle conditions. Every context is designed to describe a combination of priors and implies a possible state of the world, or, in physical terms, a state of aggregation. Past market data is used to define a relevant set of contexts. New observations of the priors automatically lead to the identification of a specific context class and a specific return distribution for that class. A simple example is that when the yield curve flattened in the past period of observation, and this is the only prior, the forecasted VaR will depend on the past observations of returns in the cases where the yield curve also flattened in the preceding period.

In this paper, context modeling is applied to capture the dynamics of the market risk associated with movements in the US stock market. A past window of several thousand daily stock market return observations is used as a data-training frame to form the contexts and delimit the distributions. Once the contexts are defined, the present state of the world can be identified and the VaR forecast can be estimated from the accompanying distribution that is derived from the training period. The analysis is performed on a daily basis. The choice of a forecast horizon is somewhat arbitrary, but a daily frequency is a reasonable choice because it can be assumed that the rebalancing of equity portfolios by market participants follows a similar pattern [6]. Moreover, financial supervisors also require banks to calculate VaRs on a daily basis since the high degree of liquidity of the US stock and derivatives markets allow investors to close risky positions rapidly.

A second part of the paper deals with the evaluation of the accuracy of VaR estimates. Often, the supervisory authorities require that the estimated VaR produced by the internal risk management system of banks is multiplied by a factor to determine the minimum required capital [7]. The standard method prescribed for banks is to count the number of exceptional observations given the VaR forecasts over a horizon of 250 trading days. Other methods have been developed, including the minimization of complicated loss functions. This paper also evaluates the forecasting capabilities of the VaRs obtained through context modeling. An important question in finance is whether forecasting models are able to predict periods of higher and lower volatility. In the case of VaR, periods of higher volatility indicate that the required capital should be increased.

Section 2 outlines the basic concepts of information theory and context modeling. It also shows the analogy between the goals of financial modeling and data compression. In Section 3, the tree-structured vector quantization algorithm, which is needed for partitioning the space of priors, is described. Section 4 deals with the different evaluation criteria used in this research and Section 5 describes the data and reports the test results. Section 6 concludes.

2 Context modeling basics

2.1 Risk analysis and data compression

The series of daily returns $\{y_t\}$ of a financial instrument can be regarded as a realization of an underlying stochastic process $\{Y_t\}$. Precise knowledge of this process is of fundamental importance to predict future evolution and to quantify future risk. Unfortunately, because only one realization of the process is known, properties such as stationarity (or quasi-stationarity) and ergodicity need to be assumed in order to allow significant predictions.

The purpose of risk analysis of financial instruments is to determine the maximal amount of money that can be lost under a certain specified probability p. This amount is usually called the "Value-at-Risk". One way to achieve the goal is to efficiently estimate the underlying probability density function (pdf) of the return Y_{t+h} on day $t + h$ based on all financial and other information accessible on day t. Typically, the probability p will be 1% or 5% and the horizon h will be 1, 5 or 25 days, corresponding with a day, a week and a month respectively.

This kind of statistical prediction, where a complete pdf is estimated rather than the most likely or expected value, is exactly the same goal of data compression. The state-of-the-art techniques in data compression are statistical by nature. Based on an environment called the "context", a pdf of the upcoming new symbol is constructed and used to drive an entropy coder, which actually generates the compressed bitstream [4,5]. In the case of text compression, this context can be the combination of the previous two characters, while in the case of image compression, it can be a combination of the rounded difference and sum of the upper and left pixel.

Though the fundamental goal in risk analysis and data compression may be identical, there are significant differences too. In data compression, millions or even billions of data samples (characters for text compression of pixels for image compression) are available, while in financial modeling the number of daily samples is limited to thousands. Therefore, the risk of data snooping, negligible in data or image compression, is of fundamental importance in financial modeling [8]. Another difference is that in data compression, the overall correctness of the entire pdf is important, while for risk analysis only the tail distribution is taken into account. Also, the only predictive factors in text compression are the neighboring text characters, while in financial modeling, there exists a multitude of factors, both microeconomic and macroeconomic, which may have predictive power. Finally, in data compression, in order to be useful, the compression algorithms must satisfy certain speed and memory consumption criteria, which are almost non-existent for financial modeling.

2.2 Information theory

2.2.1 Entropy of a random variable

The "entropy" H of a discrete random variable Y is defined as $H(Y) = -\sum_{y\in\mathcal{Y}} \Pr[y] \log \Pr[y]$, where $\mathcal{Y}$ is the set of all possible values Y can take, and $\Pr[y]$ is the probability that Y takes the value y [9]. The entropy cannot be negative and is always smaller than or equal to $\log(l(\mathcal{Y}))$, where $l(\cdot)$ indicates the number of elements of a set. It is a measure of the randomness or unpredictability of the random variable. It is also a lower bound for the achievable expected length per symbol when some type of entropy coding is applied.

The joint entropy of two random variables Y and Z is defined as $H(Y,Z) = -\sum_{y,z} \Pr[y,z] \log \Pr[y,z]$. Furthermore, the conditional entropy $H(Y|Z)$ is defined as $\sum_z \Pr[z] H(Y|Z=z)$. It can be shown that $H(Y|Z) \leq H(Y)$, with equality if and only if Y and Z are independent. This property is often referred to as "conditioning reduces entropy": the randomness or unpredictability can only decrease if information about other random variables is used.

2.2.2 Entropy rate of a stochastic process

For a stochastic process $\{Y_t\}$, the "entropy rate" is defined as $H(\mathcal{Y}) = \lim_{t\to\infty} H(Y_1, Y_2, \ldots, Y_t)/t$, when the limit exists. This definition is based upon the observation that, for independently and identically distributed random variables $\{Y_t\}$, the joint entropy $H(Y_1, Y_2, \ldots, Y_t)$ grows linearly with t. The entropy rate is a measure of the average amount of uncertainty about each random variable Y_t, when all $\{Y_t\}$ are considered simultaneously.

A related quantity for the entropy rate is defined as $H'(\mathcal{Y}) = \lim_{t\to\infty} H(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1)$, when the limit exists. It can be shown that, for a stationary stochastic process, the limits for both $H(\mathcal{Y})$ and $H'(\mathcal{Y})$ exist and are equal.

2.2.3 Entropy in physics and other fields

This probabilistic notion of entropy is also known as Shannon's "source entropy" and it was defined and used successfully in the fields of information theory, communication theory, and coding theory.

The actual roots of entropy lie in the field of thermodynamics through the notion of "thermodynamical entropy". This concept was later elaborated in statistical mechanics, which connected the macroscopic property of "physical entropy" and the number of microscopic states of a system through Boltzmann's formula $S = k \ln \Omega$. The relationship between information theory and thermodynamics has been discussed extensively by Brillouin [10] and Jaynes [11].

Later on, Shannon's probabilistic notion of entropy was imported by Kolmogorov into the field of dynamical systems where the "metric" or "Kolmogorov entropy" is defined [12].

Kolmogorov, Solomonoff and Chaitin independently further elaborated this concept to the field of logic and the theory of algorithms by defining the "algorithmic" or "descriptional" entropy (also known as the Kolmogorov complexity). Algebra uses the notion of "galois entropy".

All notions of entropy are similar in that they all aim at quantifying the amount of randomness, unpredictability or incompressibility of the system under investigation. Though they are all defined in different fields, some kind of numerical equivalence can be shown.

2.3 Context modeling

The goal of both data compression and financial modeling is to estimate, given only one data sample series, a pdf that allows to predict the upcoming values. The efficiency of the modeling can be quantified by the achieved entropy (or compression rate). If a good probability model is applied, then the entropy will be lower. The fact that conditioning reduces entropy is the fundamental principle of context modeling: conditioning the random variable on other random variables, which are not independent, can be an efficient way to achieve a reduction in entropy. The other random variables are called the "priors" and a specific combination of priors is called a "context". Usually, the contexts are grouped into "context classes" to avoid the (almost) continuous nature of the context space.

Of course, in lossless data compression, the context class must be known to both encoder and decoder, so only priors from the past may be used. Moreover, those priors are limited to the values of the already encoded characters or pixels. In financial modeling, only the first condition remains: it is obvious that no priors from the future can be used. However, among the priors, not only the past values of Y but also other microeconomic and macroeconomic values Z may be used. Hence, the key idea of context modeling is to substitute the probabilities $\Pr[y_{t+h}]$ by, typically, the probabilities $\Pr[y_{t+h}|c_t]$. The context class c_t is derived from the prior vector $\boldsymbol{z}_t$ through the context mapping function C. Hence, the context mapping function maps a context $\boldsymbol{z} \in \mathcal{Z}$, which is a vector of priors, onto a context class $c \subset \mathcal{Z}$. The set of all context classes is denoted as $\mathcal{C}$.

2.4 Practical implementation

In practice, only one data sample series $\{y_t\}$ of a particular asset or index $\{Y_t\}$ is given and a multitude of dependent priors $\{Z_t^i\}$ are available for building the context classes. Assumptions such as stationarity and ergodicity are made to estimate the probabilities of the underlying model. Specifically in financial modeling, the way the contexts are constructed and adapted is of great importance. A context model can be regarded as a collection of a number of separate probability models without contexts running in parallel, where one probability model is associated with every context class.

2.4.1 Probability model without contexts

In the case of non-parametric probability models, observed counts of samples are used to estimate the probabilities. For every value $y^0 \in \mathcal{Y}$, the probability $\Pr[y_{t+h} = y^0]$ is approximated by $n_t(y^0)/\sum_{y\in\mathcal{Y}} n_t(y)$, where $n_t(y)$ represents the number of times the value y has occurred in the time-interval $[0, t]$. Therefore, a practical implementation will count the occurrences of every symbol $y \in \mathcal{Y}$ and use these to estimate the probabilities. Initially, theses counts are initialized to zero and after a sufficient number of samples has been parsed, the array of counts will reflect the true pdf. This approach is often called the “historical” approach.

If the probability model is parametric, a class of distributions is assumed and only the parameters discerning these distributions are estimated. Very often, a lognormal distribution is presupposed and the mean μ and the variance σ^2 are estimated from the samples.

The cumulative density function is constructed from the derived pdf and used to predict the VaR. Often, the samples will be weighted by a time-varying factor so that older samples have less importance.

2.4.2 Probability model with contexts

If context modeling is used, instead of one pdf, multiple pdf’s are estimated in parallel, and, based upon the value of $\boldsymbol{z}_t$, each event is associated with one of these pdf’s.

In the case of non-parametric modeling, for each value $c \in \mathcal{C}$, the probabilities $\Pr[y_{t+h} = y^0|c]$ are approximated by $n_{t,h}(c; y^0)/\sum_{y\in\mathcal{Y}} n_{t,h}(c; y)$, where $n_{t,h}(c; y)$ is defined as the number of times in the interval $[0, t]$ where a context $\boldsymbol{z}_k$ at time k, belonging to class c, was followed by a sample y at time $k + h$. In the case of parametric modeling, for each value $c \in \mathcal{C}$, parameters are estimated based on the samples corresponding with that particular context class.

2.5 Limitations

The application of probability models to real-life data samples suffers from severe shortcomings. First of all, for some types of financial data, the assumed stationarity does not always hold. Based on Timmermann [13], who explores the relationship between volatility clustering and regime switches in time-series models, it can be argued that part of the non-stationarity may be caused by volatility clustering. In the finance literature, conditional volatility models and change-point models, among others, have been used to remedy this shortcoming. It has become standard practice to model asset returns as a mixture of distributions and to assume that they are conditionally normal [8].

A first step to solve this intricate problem is to transform the price series $\{Y_t\}$ into a set of equivalent values with approximately time-invarying support. For simplifying the calculation of consecutive price differences, usually the “continuously compounded returns” (also called “log returns”) $R_t = \log(Y_t/Y_{t-1})$ are used. However, statistical analysis has shown that this series still is non-stationary. Therefore, each referenced data sample associated with a context is multiplied by a weight $w(\delta_t)$, which is a monotonically decreasing function of the time difference δ_t of the referenced sample and the current time.

The time difference δ_t can be measured in an “absolute” way or in a “relative” way. If measured in an absolute way, the arithmetic difference between the two time indices is used. If measured in a relative way, the samples within the corresponding context class are sorted by time index and the difference in order index is taken. For example, if the referenced sample happened 10 days ago, but it was the previous sample within that particular context class, then δ_t takes the value of 10 in the case of absolute weighting and 1 in the case of relative weighting. Typically, the weighting function $w(\delta_t) = \lambda^{\delta_t}$ with $0 < \lambda \leq 1$ is used.

Moreover, in our application, the context-dependent distributions are conditioned on a parameter which is itself random and which is modeled by the state of the priors defining the context. Consequently, rather than identifying whether the stock return series are stationary in the mean or in the variance, the states of the world (context classes) in which the expected return and the volatility can reasonably be assumed to be constant are generated endogenously.

As such, removing the non-stationarity is achieved by introducing adaptivity into the context model in multiple ways: by using an alternative representation $\{R_t\}$, by weighting the data samples according to their age, by separating the samples into distinct context classes, and, by introducing new context classes which are to be trained with recent data.

Another severe shortcoming of the model is that, since the model is trained on previous samples, it is only able to recognize situations that have already happened once before. This aspect is twofold: firstly, highly unlikely situations will be considered as impossible, so they will not be predicted and secondly, if such a highly unlikely event has occurred and it is used for training, it will be regarded as a typical situation. Especially in the case of risk analysis, this puts heavy constraints on the efficiency of the VaR estimation. For this reason, we have omitted the 1987 crash in most of our experiments.

3 Tree-structured vector quantization

The partitioning of the context space imposes some additional training problems. Since the whole set of available data samples is to be divided over a number of context classes, less samples are available for each context class. But to be statistically significant, the occupation of every context class should be high enough. This is even more so for risk analysis, because then, the focus of the modeling is on extreme value analysis, which is described by the less populated tails of the distribution. The problem of having context classes with a level of occupation that is too small, is often referred to as the "context dilution" problem. While thousands or even millions of context classes can be applied successfully in image compression [14], only about tens or maybe hundreds are to be used in financial modeling.

On the other hand, the dimensionality of the space of priors tends to be high. A simple context mapping function, such as the value of the previous character in text compression, cannot be used: a simple division of each prior into a limited number of distinct intervals gives rise to an exponentially growing number of context classes. This "curse of dimensionality" is a problem that calls for an intelligent partitioning algorithm of the space of the priors.

While processing the first few samples, the model has absolutely no statistically significant information for making predictions. Therefore, a training phase processing a first part of the samples is started. During this first phase, no predictions are made and initial statistics are gathered exclusively for training. After this phase, the model enters the evaluation phase, where training is combined with accurately predicting and evaluating the VaR. During this second phase, a VaR prediction is made for each sample and all data up to the previous day are used for training.

3.1 Context tree partitioning

Typically, if the model is trained using daily samples covering a period of about 30 years, between 2000 and 8000 samples are available. To provide statistically significant tails of the pdf, at least about 100 to 200 samples are needed for each context class. In total, at most about 10 to 40 context classes are to be created. If about 10 priors were used, even a simple division of each prior into two intervals would give rise to more than 1000 different context classes.

This problem is solved in two steps. Firstly, the prior space is partitioned into context classes $c \in \mathcal{C}$ and each context $\boldsymbol{z}$ is mapped onto one context class $C(\boldsymbol{z})$ based on a minimum distance criterion. This context mapping function C is a type of vector quantization [15]. Secondly, the context classes are organized into a growing tree structure, which can change on a daily basis. If new context classes are created in such a way that they take into account the corresponding returns, the advantage is that, after sufficient training, the structure of the classes may reveal "hidden" information about the predictability of the returns.

The processing of an individual sample consists of two steps. Firstly, its context is determined and mapped onto a context class, and the risk for the future sample is estimated using the corresponding pdf. Secondly, the information contained in the co-occurrence of the context and the sample is fed back into the probability model.

3.1.1 Context mapping and VaR estimation

Let $\mathcal{Z}$ be the prior space. For each context class $c \in \mathcal{C}$, a center of mass $\bar{\boldsymbol{z}}_c = \sum_{j:\boldsymbol{z}_j \in c} \boldsymbol{z}_j / l(c)$ can be determined. To make predictions about the future return R_{t+h}, the context $\boldsymbol{z}_t \in \mathcal{Z}$ is first determined. Of course, only information available at time t can be used, not only for making predictions, but also during the training stage.

The context $\boldsymbol{z}_t$ is then mapped onto the context class c for which $||\bar{\boldsymbol{z}}_c - \boldsymbol{z}_t||$ is minimal. The pdf corresponding to that context class is used to estimate the VaR. The pdf can either be parametric or non-parametric.

3.1.2 Observation feedback

At time $t + h$, the combined observation of the return r_{t+h} with the context $\boldsymbol{z}_t$ is the sort of information the context model is trained with, so this observation must be entered back into the model. For this purpose, the context $\boldsymbol{z}_t$ is added to the associated context class c and a new center of mass $\bar{\boldsymbol{z}}_c$ is calculated. The pdf corresponding to that context class is adapted. In parametric modeling, new parameters are calculated for the enlarged set of contexts. In historical modeling, the observation is added to the list of observations. This implies that the state in the prior space corresponding to a particular context class is not constant in time. After incorporating the observation into the model, the model checks if the context tree structure needs being adapted, which is achieved by splitting nodes.

3.2 Splitting algorithm

In the beginning, the context tree consists of a single root node and all samples are mapped onto the same context class. When a specified splitting condition (the "maturity criterion") is met, a context node splits into a number of child nodes (typically two). Usually a node is split whenever a certain level of occupation (*i.e.* a specified number n_m of associated samples) is reached. The old node becomes a parent node and its associated samples are distributed over the two child nodes. After a parent node has split, it is no longer functional. Children nodes can be created from a given parent node in a few distinct ways: "random node creation", "fast min-max node creation" and "full min-max node creation".

3.2.1 Random node creation

Associated with a parent node c, there is a center of mass $\bar{\boldsymbol{z}}$ and a list of associated context samples $\{\boldsymbol{z}_j\}$. In the case of "random node creation", the two child nodes are created by adding and subtracting a randomly generated small disturbance vector $\boldsymbol{\epsilon} \in \mathcal{Z}$ to the parent center of mass. Two new initial attractors, $\bar{\boldsymbol{z}} \pm \boldsymbol{\epsilon}$, have thus been created and each of the samples $\{\boldsymbol{z}_j\}$ is classified into the child node with the closest center of mass. Since the Euclidean distance measure is applied, the different dimensions of the prior space need to be normalized. After distributing the samples over the child nodes, the initial attractor of each child node is replaced by the effective center of mass, which can now easily be determined.

3.2.2 Fast min-max node creation

Using the random node creation splitting technique, the values of the returns r_{t+h} are not taken into account and vector quantization is performed only in the prior domain. However, the combination of the observed returns together with the observed contexts of a specific context class, might also carry useful information. Therefore, in the case of "fast min-max node creation", the observed returns are also taken into account. Moreover, in this case the final context tree might reveal information about the significance of the distinct priors. From all contexts $\{\boldsymbol{z}_j\}$ belonging to samples of the parent context class, the ones with the extreme corresponding returns r_{t+h} are determined and used to create two child nodes. Let r^+ and r^- be the maximal and minimal return respectively and let $\boldsymbol{z}^+$ and $\boldsymbol{z}^-$ be the corresponding contexts. These are then used as the two initial attractors of the two child nodes. As in the case of random node creation, each context is classified into one of the child nodes depending on the smallest distance criterion. After classification, each initial attractor is replaced by a new center of mass and the parent node is no longer used.

3.2.3 Full min-max node creation

The above technique has more potential than the random technique because the information about the returns is fed back into the quantization process. Unfortunately, it is very sensitive to outliers and it assumes a certain degree of monotonicity. These problems can be avoided by using every associated return to classify the contexts. In the case of "full min-max node creation", a threshold return $\hat{r}$ is defined as $(r^+ + r^-)/2$. Each context $\boldsymbol{z}_j$ originally corresponding with the parent node is classified into one of the child nodes, depending on whether the corresponding return $r_j > \hat{r}$ or $r_j < \hat{r}$. After classification, a center of mass corresponding with each child node is calculated.

3.3 Additional improvements

Two additional improvements to the growing context tree algorithm are suggested in this paper. One is the "reverse model restart" which aims at decreasing the consequences of the non-stationarity of the data by enlarging the effect of the most recent data on the growing of the context tree. The other improvement is the "feedback mechanism" which aims at reducing training time and removing repetitive over- or underestimation.

3.3.1 Reverse model restart

Normally, the context tree is built starting from the first samples and adopts itself to the most recent events. However, since the initial node splits have initiated the main branches of the tree, the most important decisions with respect to the structure of the tree are based upon the oldest samples. Therefore, the modeling might improve if more recent events are used first. This goal is achieved if "reverse model restart" is periodically applied with period t_r. After every period, the order of the samples is reversed and the context tree is completely rebuilt. Most recent samples decide on the initial branches and the oldest samples are used for the fine-tuning. It is clear that the order of the processing of the samples is a tradeoff because ideally, the most recent samples should be used for both initial training of the model and for fine-tuning.

3.3.2 Feedback mechanism

If the prediction efficiency is entered back into the model, the dynamics of the training can be changed dramatically. Also, consistent misprediction due to changing statistical behavior can be intercepted and avoided. The "feedback mechanism" adds an artificial prior to the list of economic priors. This additional prior can be regarded as a binary flag which indicates whether the previous sample exceeded its prediction or not. Of course, this variable too is normalized before it is incorporated into the context space.

3.4 Discussion: non-linear modeling

The predictability of the presented modeling technique differs from the one encountered abundantly in non-linear science in multiple ways.

Common non-linear models use a system of non-linear differential equations that comprises a few parameters and a few variables. The time variable is continuous by nature but is usually discretized to allow numerical solutions. The input of the real world consists of parameters and boundary values (usually the present state of the system). The solution to the system is deterministic in theory but chaotic in practice. The system of differential equations itself is time independent and explicitly describes the dynamics.

The proposed context modeling approach, which is a successful technique from the field of data compression, is much more generic since more types of behavior can be modeled. The main difference compared to the conventional model lies in its stochastic approach: multiple outcomes are possible and the probability of each of these outcomes is estimated on statistical grounds. The time variable is discrete and the real world input is much greater since all the information of the system is obtained by training. Only a few assumptions about stationarity and continuity of the probability density function are made. The signal is described as a mixing of multiple stationary sources. The parameters of the model are optimized by an exhaustive search. The model is time dependent and describes the dynamics in an implicit way. Since so many types of behavior can be modeled and so few assumptions are made, the system needs large amounts of data in order to make adequate predictions. The training is similar to Markov modeling, but the approach differs because the model does not estimate state transitions but rather uses external information (the priors) to construct the states.

4 Evaluation techniques

Though evaluating VaR estimates is difficult because most tests have limited power, recently some improved methods have been proposed [7]. In this paper, we basically use three types of evaluation measures: the average hit ratio, a χ^2-distance criterion with respect to the binomial distribution and a cost and a loss function.

Firstly, the binary random process $\{X(t)\}$ is defined as 1 if $\{Y(t)\}$ is smaller than the predicted VaR, and 0 otherwise. It can be interpreted as an indicator whether the loss exceeds the absolute value of the VaR, or similarly, as an "exception" flag. If the statistical model captures all deviations from the ideal and perfectly matches the observed data, then for every t, $\{X(t)\}$ is a random variable which takes the value 1 with probability p, and 0 with probability $1-p$.

Secondly, the entire evaluation period, covering m samples, is divided into q non-overlapping windows of l samples each. For every window i, the random variable T_i is defined as

$$T_i = \sum_{k=0}^{l-1} X(il+k), \tag{1}$$

and, if perfect modeling is achieved, its expected value equals pl. Moreover, the set of random variables $\{T_i\}$ is distributed independently and identically and for every i, the random variable T_i obeys the binomial distribution

$$P[T_i = j] = \begin{pmatrix} l \\ j \end{pmatrix} p^j (1-p)^{l-j}. \tag{2}$$

These values are used to construct three sets of evaluation criteria: (1) "min-mean-max" statistics, (2) the χ^2 statistic and (3) cost and loss functions.

Min-mean-max statistics

The first set of criteria involves the observed values for T_i. The observed minimum $m^- = \min_i\{T_i\}$, the observed mean $\bar{m} = \sum_i T_i/q$, and the observed maximum $m^+ = \max_i\{T_i\}$ are three interesting test statistics. Their distributions are given by:

$$\Pr[m^- < j] = 1 - \left(\sum_{k=j}^{l} \Pr[T = k] \right)^q, \tag{3}$$

$$\Pr[q\bar{m} = j] = \begin{pmatrix} ql \\ j \end{pmatrix} p^j (1-p)^{ql-j}, \tag{4}$$

$$\Pr[m^+ \geq j] = 1 - \left(\sum_{k=0}^{j-1} \Pr[T = k] \right)^q. \tag{5}$$

Ideally, if the heteroskedasticity is intercepted by the modeling, the mean should equal pl and the maximum should not be too large. For low p, the observed minimum is a useless statistic.

χ^2 statistic

The hypothesis that the observed variable T_i obeys the binomial distribution, as given by equation 2, can be tested using Pearson's χ^2 statistic [16,17]. Since T can take values in the interval $[0, l]$, the χ^2 test statistic is given by

$$\chi^2 = \sum_{k=0}^{l} \frac{(n(k) - e(k))^2}{e(k)}, \tag{6}$$

where $n(k)$ and $e(k)$ are the observed and expected number of windows where $T_i = k$ respectively, according to equation (2). The statistic has $\nu = l$ degrees of freedom.

Cost and Loss functions

Previous criteria merely use counts of events where the VaR was exceeded and have no quantitative power. An artificial 1% VaR defined as $+\infty$ on every first day and $-\infty$ for every other 99 days would achieve great score, but does not meet the requirement of a financially useful VaR. Based on the idea of regulatory loss functions [7], both a loss and a cost function are used as an evaluation criterion. The loss and the cost functions are based on exceptional and regular observations respectively.

The loss L is defined as

$$L = \sum_{t=1}^{m} H(\mathrm{VaR}_t - r_t)(\mathrm{VaR}_t - r_t)^2, \tag{7}$$

where $H(x)$ is the Heaviside function, defined as 1 if $x > 0$, 1/2 if $x = 0$, and 0 if $x < 0$. The loss function by Lopez is similar, but adds the number of exceptions, so $L_{\mathrm{Lopez}} = L + \sum_t X(t) = L + q\bar{m}$. Both loss functions are based only on the "exceptional" observations where the loss exceeds the VaR, *i.e.* where $r_t < \mathrm{VaR}_t$. The loss function is a measure of the loss involved in underestimating the risk capital.

The cost function C, is defined as

$$C = \sum_{t=1}^{m} H(r_t - \mathrm{VaR}_t)(r_t - \mathrm{VaR}_t), \qquad (8)$$

so the quadratic form is replaced by a linear form and it only takes into account the "regular" observations, *i.e.* where $r_t > \mathrm{VaR}_t$. It expresses the cost involved in overestimating the risk capital. Neither the cost nor loss function can be interpreted as a standalone criterion. They must be evaluated together, and in combination with the previously introduced criteria. Also, a financial institution might decide to assign different weights to cost and loss functions.

5 Experimental results

The proposed statistical model was implemented in C++ and tested on both Microsoft Windows NT and Linux platforms. Depending on the algorithm options and the choice of the parameters, a typical run on about 8000 samples takes between 10 and 120 seconds. In the current implementation, about 2 megabytes of memory are needed.

5.1 Financial data

The data compression technique is applied to the daily return series of the Standard & Poor's 500 US stock index from October 1969 until December 1999. Since this index contains the largest stocks, and thus represents a major part of the total market capitalization on the New York Stock Exchange, it can be assumed to capture the associated market risk.

A total of 8089 samples were available for training and evaluation. The first 200 samples were not used for training because of initialization conditions (*e.g.* for obtaining useful values of long-term priors). Sample 201 to 2000 (covering the period October 7, 1969 until August 30, 1976) were used for training only. Samples 2001 to sample 8089 (covering the period August 31, 1977 until December 31, 1999) were used for both training and validation. Of course, only data from the past is used to make predictions. Not all priors were available from the beginning of the training period; they are substituted by zero in those cases. Hence, the total number of evaluation samples is $m = 6089$. If windows of width $l = 100$ are used, a total of $q = 60$ windows is available for evaluation.

To avoid the training problems that may arise from highly unlikely events, a volatile period covering 100 returns around the 1987 crash was omitted from the data. In the last paragraph, some numerical results are presented obtained by taking the crash into account.

Different priors are used to construct the contexts. The choice of the priors is based on theoretical models and empirical findings reported in asset pricing research. A number of influential asset pricing studies have concluded that stock returns are driven both by fundamental and technical factors [18].

First there is evidence of persistence in daily returns, particularly in the short run, and mean reversion over the medium term [8]. We use four technical variables to capture these effects and define them as the momentum priors. The first three variables are intended to reflect the short-run dynamics and include the one-day, one-week and one-month past returns. The fourth technical factor is the degree of expected volatility at a given date, measured as the dispersion of the stock market returns over the past 100 trading days. This conditioning variable is calculated as the ratio of the difference between the maximum level of the index and the minimum level of the index in the 100-day window relative to the minimum level of the index.

A second set of priors that is assumed to contain information about future returns are macroeconomic factors. These variables have been widely used in multi-factor models and were found to have predictive power [19,20].

The first macroeconomic variable is the daily change in the US yield curve. This "term spread factor" is measured as the difference between the long-term riskless interest rate (benchmark US 10-year government bond) and the riskless short term interest rate (3-month US treasury bill rate). Harvey finds that the slope of the term structure contains information about future economic growth [21]. Campbell finds a direct link between the term structure of interest rates and excess returns on financial markets [22]. As a consequence, changes in the yield curve influence expected stock returns, although the direction and the exact magnitude of this effect depends on the source of the change in the term structure, *i.e.* whether the change was caused by variations in the short or the long-term interest rate.

The second factor is the "default spread", which is intended to capture the pervasive influence of the economy-wide default risk on financial markets. Theoretically, an increase in the expected distress risk of corporations should increase the required return on equities. We measure the default spread as the difference between a corporate bond return series (the US benchmark BAA corporate bond yield) and a riskless interest rate (the US benchmark 10-year government bond yield)[1]. As in the calculation of the term spread, we use the daily change of the default risk variable as a prior. This procedure ensures that the relevant information is known to investors at the date of the VaR measurement.

Finally, the third fundamental variable is the "dividend yield" [22,8]. We compute the changes in the daily dividend yield series to capture the investors' expectations about the dividend payoff in the US stock market. Theoretically an increase in the dividend yield should reflect improved earnings.

[1] Since the US government has a AAA-rating, this difference effectively captures expected default risk.

Table 1. Summary of investigated priors and parameters. The lower-case letters indicate real or integer values; the capitals indicate a limited number of choices.

prior	symbol
1-day return	z_1
5-day cumulative return	z_2
25-day cumulative return	z_3
100-day volatility	z_4
differential term structure	z_5
differential default spread	z_6
differential dividend yield	z_7
model parameter	
maturity occupation level	n_m
number of child nodes	n_c
node creation algorithm	A
weighting type	T_w
weighting factor	λ
model type	T_m
reverse model restart interval	t_r
feedback mechanism flag	F

5.2 Model parameters

As indicated in previous sections, the tree-structured context model uses a lot of parameters, for which an optimal combination must be empirically derived.

Table 1 gives an overview of the priors available for training and the parameters used. The "maturity occupation level" describes the maturity criterion: a context class node splits whenever the number of associated samples exceeds this level. The "number of child nodes" indicates how many new nodes are created when a node has reached the maturity level. The "node creation algorithm" can be any of the three algorithms described in Section 3.2 (random, fast min-max and full min-max). The "weighting type" can be either absolute or relative, as described in Section 2.5. The "weighting factor" corresponds to the base λ of the weighting function $w(\delta_t) = \lambda^{\delta_t}$. The "model type" can be either parametric (Gaussian) or non-parametric (historical). The "reverse model restart interval" indicates the period after which the model is completely rebuilt, by training using the observed samples in reversed order; a value of ∞ indicates that this never happens. Finally, the "feedback mechanism flag" describes whether the feedback mechanism is applied, see Section 3.3.

5.3 Results

A global optimization of all parameters for every combination of $h \in \{1, 5, 25\}$, $p \in \{0.01, 0.05\}$, and for every combination of priors is not achievable in acceptable time using an exhaustive search algorithm. Therefore, in a first stage, a limited set of parameter combinations was derived using trial and error. This set is summarized in Table 2 and is used for an exhaustive search in the second stage. During the parameter optimization stage, all seven priors are included for building the contexts. The table already shows that using more than 2 child nodes in the splitting stage produced no significant improvement, that relative weighting consistently outperforms absolute weighting, and that only high weights are interesting compared to the RiskMetrics approach.

Table 2. Exhaustive parameter optimization space.

parameter	values
n_m	100, 200, 300, 500, 1000
n_c	2
A	random, fast min-max, full min-max
T_w	relative
λ	1, 0.9995, 0.999, 0.995, 0.99
T_m	Gaussian, historical
t_r	100, 200, 500, 1000, ∞
F	yes, no

A fundamental problem in interpreting the numerical results is the joint evaluation of the five numerical criteria $\bar{m}$, m^+, χ^2, L and C. For the first three criteria, confidence intervals can be numerically derived based on the assumption that the results are modeled correctly. The two-sided 92% confidence for $\bar{m}$ is given by $[0.783, 1.22]$ and $[4.52, 5.47]$ for the case where $p = 0.01$ and $p = 0.05$ respectively. Furthermore, $\Pr[m^+ \leq 5] = 96.84\%$ and $\Pr[m^+ \leq 14] = 91.58\%$ for $p = 0.01$ and $p = 0.05$ respectively. Finally, the one-sided 95% confidence interval for χ^2 is given by $[0, 124.34]$. The other two criteria, L and C, should both be as low as possible. All criteria influence each other, so they should be evaluated simultaneously. This discussion questions the fundamental goal of the VaR.

Optimizing a set of parameters in this way is very sensitive to data snooping, since it is not clear how robust the optimal combination of parameters will be for modeling the values of other financial instruments or other periods.

5.3.1 Parameter optimization

For each combination of p and h, Figure 1 plots the loss L *versus* mean $\bar{m}$ using the parameter combinations from Table 2, except that only historical modeling is used. Different marks are used for each splitting algorithm and the simulated RiskMetrics result is also shown on each plot. For each of the six cases, a combination of parameters is available producing acceptable results, *i.e.* $\bar{m}$ is close to the expected value and L is relatively low. On average, better results are obtained for a 1% VaR than for a 5% VaR, so historical context modeling is better for solving the fat tail problem. Also, better results are obtained for shorter horizons; this is mainly due to the fact that most priors represent short-term dynamics so they do not carry long-term information. Comparing context modeling to RiskMetrics, significantly better results are obtained for

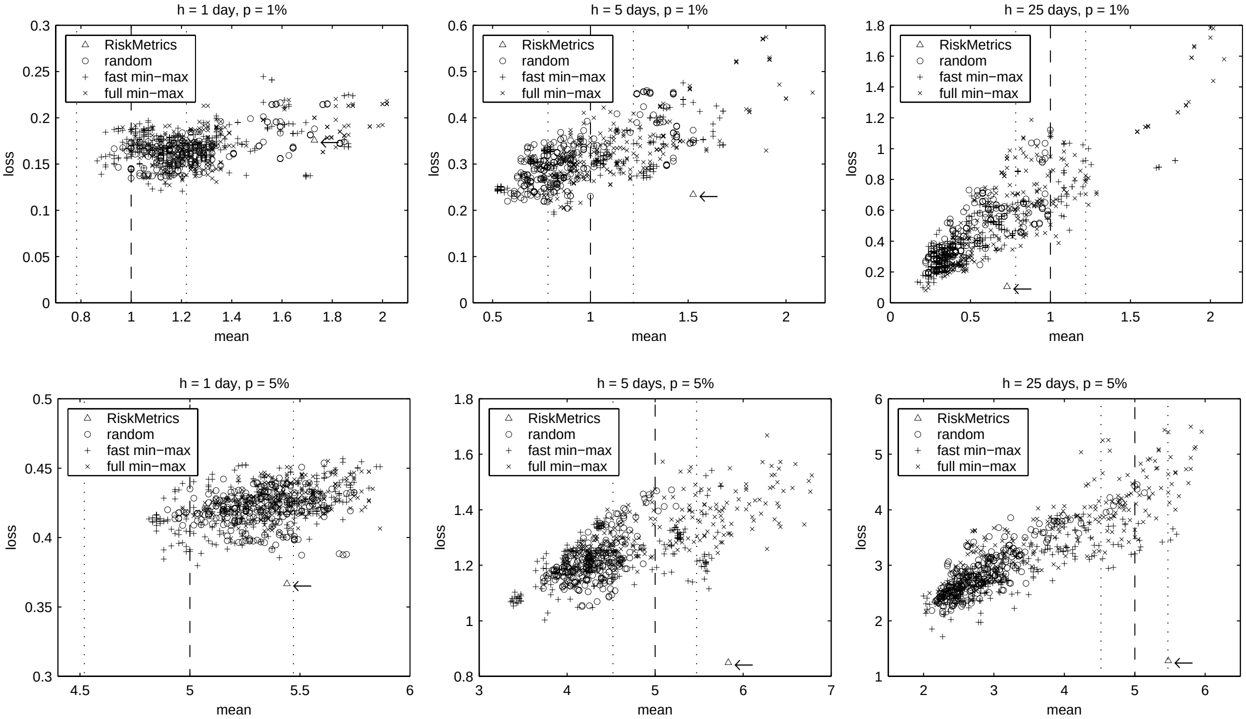

Fig. 1. For each combination of $h \in \{1, 5, 25\}$ and $p \in \{0.01, 0.05\}$, a plot shows the loss $L(\times 10^{-3})$ *versus* the mean $\bar{m}$ for every parameter combination (only historical modeling). The dotted lines represent the 92% confidence interval for $\bar{m}$. The RiskMetrics-based approach is marked with a "←". All priors are included.

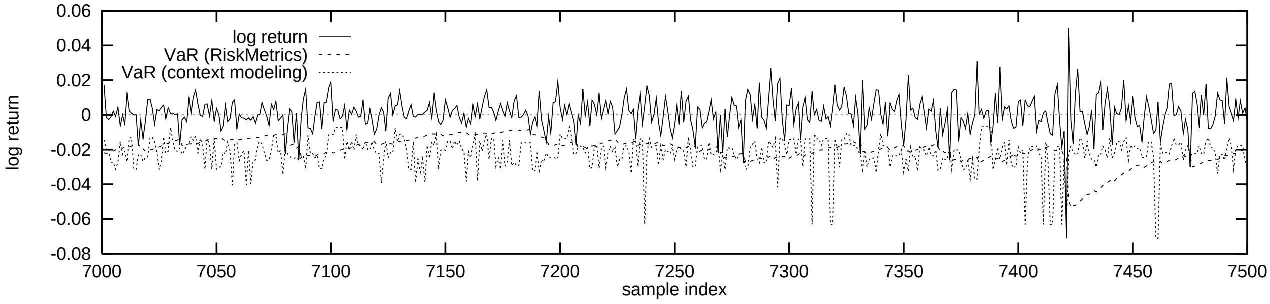

Fig. 2. A typical course of the log return, the RiskMetrics based VaR and the VaR based on context modeling for the case $p = 0.01$ and $h = 1$. Note the difference between the two VaRs, especially before and after sample 7421 (October 1997 crash).

the $(h = 1, p = 0.01)$ and the $(h = 5, p = 0.01)$ case, whereas significantly worse results are obtained for the $(h = 25, p = 0.05)$ case. However, remember that the primary goal of this research was to improve the modeling of extreme events on a short-time horizon. Surprisingly, the fast min-max splitting algorithm performs always optimal or close-to-optimal. In some cases, the full min-max algorithm achieves slightly better results. Probably, the full min-max approach has more potential but adapts slower to the presented data. Especially for long horizons, the differences between the splitting algorithms become larger.

Table 3 presents numerical results using the optimal parameter combination for each of the six cases. The best $\bar{m}$ for each case is printed in boldface. If no context modeling is applied, weights based on the RiskMetrics method are used, *i.e.*, $\lambda = 0.94$ if $h = 1$, $\lambda = 0.95$ if $h = 5$ and $\lambda = 0.97$ if $h = 25$. If optimal parameters are used for every case, historical context modeling achieves the best results with respect to the mean $\bar{m}$. However, the maximum m^+ and χ^2 statistic are also often higher. The greatest improvements are to be expected for short horizons and low probabilities. This is because the priors reflect short-term behavior and because the non-parametric approach is a good solution for the fat tail problem. For a 1% VaR, Gaussian context modeling or historical modeling without contexts does not improve the results compared to the RiskMetrics based approach, but the real improvement lies in the simultaneous application of both context and historical modeling.

Figure 2 shows the typical behavior of the log return and two VaR estimates for sample 7000 to 7500 for the case $p = 0.01$ and $h = 1$ and using the optimal parameters from the previous table. The expected number T of returns exceeding the VaR is 5. The plot shows a big

Table 3. Numerical results on S&P500 (without 1987 crash) using Gaussian and historical modeling and optimal parameters. Key: "CM" = context modeling, "NC" = no context modeling (using RiskMetrics weights), "C" = context modeling, "T_m" = model type, "G" = Gaussian, "H" = historical. For every combination of h and p, the best $\bar{m}$ results are marked in boldface.

CM	T_m	$\bar{m}$	m^+	χ^2	$L(\times 10^{-3})$	C
Case: $h = 1$, $p = 0.01$						
NC	G	1.73	5	52.1	0.18	1.96
NC	H	2.59	5	222.6	0.23	1.93
C	G	1.53	6	95.81	0.20	1.97
C	H	**1.08**	6	42.82	0.12	2.37
Case: $h = 1$, $p = 0.05$						
NC	G	5.44	9	20.9	0.37	1.41
NC	H	6.37	10	40.5	0.38	1.40
C	G	4.64	13	83.85	0.38	1.43
C	H	**5.03**	14	115.84	0.38	1.42
Case: $h = 5$, $p = 0.01$						
NC	G	1.53	5	54.7	0.23	4.34
NC	H	2.80	6	384.0	0.44	4.10
C	G	1.44	13	571×10^6	0.40	4.78
C	H	**1.00**	8	2.32×10^3	0.23	5.62
Case: $h = 5$, $p = 0.05$						
NC	G	5.83	13	31.9	0.85	3.11
NC	H	6.51	11	57.9	0.92	3.09
C	G	5.71	21	1.04×10^6	1.40	3.29
C	H	**4.92**	16	956.02	1.13	3.52
Case: $h = 25$, $p = 0.01$						
NC	G	0.73	5	55.9	0.11	8.03
NC	H	2.83	9	393×10^3	0.45	7.32
C	G	1.22	13	2.29×10^9	0.78	10.14
C	H	**0.98**	12	39.3×10^6	0.54	11.17
Case: $h = 25$, $p = 0.05$						
NC	G	5.47	17	22.0×10^3	1.28	5.74
NC	H	7.15	19	47.8×10^3	1.34	5.64
C	G	5.02	26	7.42×10^9	3.92	7.18
C	H	**5.00**	19	261×10^3	3.51	7.46

Table 4. Optimal parameters and their sensitivity for the case $p = 0.01$ and $h = 1$.

parameter	value	sensitivity
n_m	200	$+$
n_c	2	$-$
A	fast min-max	$+$
T_w	relative	$+$
λ	0.99	$+$
T_m	historical	$+$
t_r	200	$-$
F	no	$-$

Table 5. Optimal priors if the number of priors n_p is limited.

n_p	z_1	z_2	z_3	z_4	z_5	z_6	z_7	$\bar{m}$	$L(\times 10^{-3})$
0	-	-	-	-	-	-	-	1.08	0.16
1	-	-	-	X	-	-	-	1.98	0.16
2	-	-	-	-	-	X	X	1.00	0.12
3	-	-	X	-	X	X	-	1.00	0.14
4	-	X	-	-	X	X	X	1.02	0.14
5	X	X	-	-	X	X	X	1.00	0.14
6	X	X	X	X	X	-	X	1.07	0.15
7	X	X	X	X	X	X	X	1.08	0.13

qualitative difference between the two VaR estimates. The RiskMetrics based VaR achieves a bad number of VaR excess returns (17 times) and is characterized by its slow decay in periods of low volatility, its consequent misses of extreme negative returns and its sudden raise immediately after those extreme situations. The context modeling approach achieves a better number of VaR excess returns (8 times), a better non-stationarity reduction with respect to the VaR excess, but also a very irregular behavior, caused by the constant change of context class. This VaR course is counterintuitive to the notion of slowly varying risk and might be interpreted as a sign of bad modeling. It is an inherent consequence of the modeling approach, though it could be improved if more data were available for training. The October 1997 crash is not predicted at all by the RiskMetrics approach: the VaR slowly decays before the crash and rises immediately after it. The context modeling VaR on the other hand repeatedly predicts more and more returns of high risk as time continues towards the crash. Immediately after the crash, a low VaR is predicted, indicating the danger of high loss is over. This indicates that the context model senses an upcoming period of higher risk and falls back to safe behavior shortly after it.

Table 4 gives the optimal parameters and the sensitivity to that parameter for the ($h = 1, p = 0.01$) case. It is important to note that because of the context modeling, higher weights can be used. Also, the fast min-max performs best, though the difference with full min-max and random splitting is small.

5.3.2 Importance of priors

During the parameter optimization stage, all priors were available to build the contexts. However, not all of them are equally important so every possible combination of seven or less priors is investigated. For the ($h = 1, p = 0.01$) case, Table 5 presents which priors produce the best results if only a limited number of priors n_p were to be used. The differential dividend yield, the differential default spread, the differential term structure and the 5-day cumulative return show to have the most predictive power. However, as the number of priors n_p increases, not always the same priors are selected. This indicates that there is a lot of mutual information between the priors, but this is difficult to quantify and analyze.

Table 6. Numerical results on S&P500 (1987 crash included) using Gaussian and historical modeling for the ($h = 1, p = 0.01$) case. The parameters were not optimized but carefully chosen based on previous experiments. Key: see Table 3.

CM	T_m	$\bar{m}$	m^+	χ^2	$L(\times 10^{-3})$	C
NC	G	2	5	120.9	0.75	2.00
NC	H	2.62	5	231.4	0.76	2.13
C	G	2.03	9	24.3×10^3	0.92	1.97
C	H	**1.43**	8	2.29×10^3	0.73	2.51

5.3.3 Parameter robustness

Many parameters are used in the model and they need to be optimized using only one data series. The presented optimal results are sensitive to the problem of data snooping. The question remains whether parameter values, optimized from the past, will remain good parameters in the future.

To investigate this problem, we performed a limited experiment by independently optimizing the parameters on two separate time intervals: the first 6000 samples and the last 2000 samples. Though the data series clearly show to be non-stationary when comparing these periods, the results of the experiment show that the optimal values are almost identical and that only the splitting algorithm differs. This is an indication that the parameter optimization procedure is reasonably robust.

5.3.4 The 1987 crash test

Some numerical results for the ($h = 1, p = 0.01$) case including the 1987 crash data is shown in Table 6. The parameters for the context modeling were not optimized but chosen based on previous experiments; if no context modeling is applied, $\lambda = 0.94$.

When comparing the historical context modeling with the classical approach, we see an improvement in the mean $\bar{m}$ and the loss L, but the maximum m^+, the χ^2 and the cost C deteriorate. Several of these measures, especially m^+ and χ^2, are non-linear and their values depend mainly on the extreme values. The extremal behavior is mainly concentrated in the period around the 1987 crash. The averaging criterium $\bar{m}$, which is improved by context modeling, does not suffer from this aspect.

6 Conclusion

This paper presents some results of applying context modeling, a state-of-the-art technique in data compression, to the field of financial modeling and risk analysis. The goals of both data compression and financial modeling are shown to be similar, but because of the limited number of data samples and the large presence of useful priors, some adaptations must be added to the modeling. The partitioning of the state space of priors into separate context classes is achieved by a growing tree-structured vector quantization algorithm. An optimal combination of parameters is exhaustively searched for the S&P500 US stock index, covering more than 30 years of data, but omitting the 1987 crash. Multiple evaluation criteria are used for this purpose. Though the approach is very universal, the task of VaR prediction was used to show one possible application. The results show that, for low probability VaRs and short horizons, significantly better predictions are obtained using historical context modeling compared to the RiskMetrics approach. The strength of the approach lies in the combination of introducing contexts and non-parametric modeling. In contrast with the RiskMetrics approach, the October 1997 crash was anticipated in time, and the model recovered from the crash much faster.

We are grateful to Wilfried Philips from Ghent University (Belgium), for his encouraging ideas, and Jeroen Van Overloop, for introducing this interesting subject.

References

1. Bank for International Settlements, *Capital requirements and bank behaviour: The impact of the Basel Accord*, Working Paper No. 1 (1999).
2. Bank for International Settlements, *A new capital adequacy framework*, No. 50 (1999).
3. K. Dowd, *Beyond Value at Risk: The New Science of Risk Management* (J. Wiley & Sons, 1998).
4. K. Sayood, *Introduction to Data Compression* (Morgan Kaufmann, San Francisco, USA, 1996).
5. T.C. Bell, J.G. Cleary, I.H. Witten, *Text Compression*, Advanced Reference Series Computer Science (Prentice Hall, Englewood Cliffs, New Jersey, 1990).
6. K. Simons, New England Econ. Rev., Sept./Oct., 3 (1996).
7. J.A. Lopez, Federal Bank of San Francisco Econ. Rev. 3–17 (1999).
8. J.Y. Campbell, A.W. Lo, A.C. MacKinlay, *The Econometrics of Financial Markets* (Princeton, New Jersey, 1997).
9. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
10. L. Brillouin, *Science and information theory* (Academic Press, New York, 1962).
11. E.T. Jaynes, *Papers on Probability, Statistics and Statistical Physics* (Reidel, Dordrecht, NL, 1982).
12. A. Katok, B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems* (Cambridge University Press, New York, NY, USA, 1995).
13. A. Timmermann, J. Econometrics **96**, (1)75 (May 2000).
14. K. Denecker *et al.* J. Electronic Imaging **8**, (4)404 (Oct. 1999).
15. A. Gersho, R.M. Gray, *Vector Quantization and Signal Compression* (Kluwer, 1992).
16. M. Fisz, *Probability Theory and Mathematical Statistics* (Wiley, New York, 1963).
17. R. Shiavi, *Introduction to applied statistical signal analysis.* 2nd edn. (Academic Press, San Diego, USA, 1999).
18. *The internationalization of equity markets*, edited by J.A. Frankel (The University of Chicago Press, 1994).
19. N.-F. Chen, R. Roll, S.A. Ross, J. Business **59**, 383 (1986).
20. W.E. Ferson, C.R. Harvey, J. Political Econ. **99**, 385 (Apr. 1991).
21. C.R. Harvey, J. Fin. Econ. **22**, 305 (Dec. 1988).
22. J.Y. Campbell, J. Fin. Econ. **18**, 373 (1987).