Construction of multi-step ahead predictions in a normal BVAR(p) model using Monte Carlo sampling

Jan Šindelář

***Abstract*— In Bayesian normal vector AR model (BVAR) of data evolution in a discrete time we are trying to predict the distribution of data up to horizon $t+h$. Since the analytical solution of such a prediction is difficult due to the high dimensionality of the problem, we are forced to search for approximative solutions. We propose a solution using Monte Carlo sampling from parameter distribution and later reconstruction of the predictive distribution of data.**

I. INTRODUCTION

Modelling dynamic systems is a building stone for system optimization and control. In the case of modelling time series, the probably most systematically used tool is the Box-Jenkins methodology [1], where an ARMA process is associated with the time series and fitted to the data. In this paper we restrict ourselves to a Gaussian AR process of order p. The process can be fitted to the data using classical statistics [2], where the parameters are considered unknown and replaced by their estimates or from a Bayesian viewpoint, where parameters are taken as random variables [3], [4].

Since the goal of our project is modelling financial time series and optimal portfolio selection [5], we need to construct multi-step ahead prediction of the VAR process associated with the time series to use for optimization by Dynamic Programming. Such predictions have been already proposed under the assumptions of classical statistics [6] or when mean values of estimated parameters were taken as true values [7]. In this paper we propose an approximate solution of h-step prediction of evolution of the stochastic process related to the data in a full distribution form. Such a solution hasn't been given to the best knowledge of the author. Such a solution is needed if we need to find a mean-variance optimization approximation of Dynamic Programming – then we only need the first two moments of the distribution or when the optimizing agent has a non-flat (convex or concave) utility function [8]. To obtain the distribution we draw Monte Carlo samples from the estimated parameter distribution, we let the AR process evolve with the sampled parameters and reconstruct the predictive distribution by simple numerical integration.

In the next section of this article a detailed description of the modelling method is given. In the first subsection we give a brief introduction into the notation used throughout the text. Subsection (B) describes model choice and a few reasons for choosing such a model. Such a choice have to be made based on expertise, even though we try to minimize such influence on the model choice, by trying to choose the model objectively based on the observed data entries. An example of such automatization is the model structure estimation presented in (C). In (D) we describe the Bayesian parameter estimation. In (E) we try to compensate for ignoring parameter evolution of the model by the use of exponential forgetting. Section (F) is the main part of this article. It is the section where the Monte Carlo approximation of prediction is presented.

This work was supported by contract 2C06001 of Ministry of Education, Youth and Sports of the Czech Republic

J. Šindelář is with the Department of Adaptive Systems, Institute of Information Theory and Automation CAS, Pod Vodárenskou věží 4, Prague 182 08, Czech Republic e-mail:sindys@volny.cz

II. MODELLING

A. Basic concepts and notation

We try to model the time series of a market price y_t of some commodity futures contract, where t is from a discrete finite set of times – an index set $\mathcal{T} = \{1, \dots, T\}$. We suppose that the price evolution is also influenced by other observable data and we collect all these data channels into a vector of data $\mathbf{d}_t = (d_{1;t} = y_t, d_{2;t}, \dots, d_{k;t})'$, where the apostrophe stands for transposition. To such data, we try to assign a discrete-time stochastic process $\mathbf{D}_t$ – an *adpated stochastic process* meaning that now the $\mathbf{D}_t$ are *random vectors*, defined on a probability space $(\Omega, \mathcal{F}, \mu)$, where the space is equipped with a filtration $\mathcal{F}_t$[1] (collection of σ-algebras) and for each time $s \geq t$ the random vector $\mathbf{D}_t$ is measurable $\mathcal{F}_s$ and $\mathbf{D}_t = \mathbf{d}_t$ – the realization of the random vector $\mathbf{D}_t$ is known to the observer from time t on. Capital letters are used for random variables and matrices and small letters for realizations and values.

We suppose, the joint probability distribution of $\mathbf{D}_1, \dots, \mathbf{D}_T$ is absolutely continuous with respect to the underlying Lebesgue measure λ^{Tk}, so that there exists a joint probability density $f(\mathbf{d}_1, \dots, \mathbf{d}_T)$ specifying the distribution. For a more detailed discussion on this issue see Chapter 6 in [10].

Remark 1: In the following text more densities of a different functional form can be denoted by f if they differ in arguments (either in number or type). This is a concept similar to that of function overloading often used in computer programming and leads to a less complicated notation.

Except for random vectors $\mathbf{D}_t$, there are other random variables $\boldsymbol{\theta}$ defined on $(\Omega, \mathcal{F}, \mu)$, called *parameters*, some of which can describe the relations between $\mathbf{D}_1, \dots, \mathbf{D}_T$ by a parameteric model - one of such will be introduced in the next section. These variables are not measurable $\mathcal{F}_t$ $\forall t \in$

[1] See [9] page 51 for details

$\mathcal{T}$, but are measurable $\mathcal{F}$. We again suppose, there exists a joint density of data and parameters determining their joint distribution.

B. Model choice

There are various ways how to choose an appropriate *parametric* model to describe the behavior of stochastic process $\mathbf{D}_1, \ldots, \mathbf{D}_T$ (see for example [11],[2]). We will use a special type of ARMA processes - multivariate AR processes reduced to the first p time-lags (AR processes were first systematically studied by Box and Jenkins in [1] and are also presented in great detail in [12]). The reason for choosing such processes is their relative richness and also computational ease, when it comes to their parameter estimation. We start by splitting the joint probability density mentioned earlier into factors similar to each other, but shifted in time with the use of basic theorems of probability theory [10], [13]

$$\begin{aligned} f(\mathbf{d}_1, \ldots, \mathbf{d}_T, \boldsymbol{\theta}) = \\ f(\mathbf{d}_T, \boldsymbol{\theta}_T | \boldsymbol{\theta}_{T-1}, \ldots, \boldsymbol{\theta}_p, \mathcal{F}_{T-1}) \\ f(\mathbf{d}_{T-1}, \boldsymbol{\theta}_{T-1} | \boldsymbol{\theta}_{T-2}, \ldots, \boldsymbol{\theta}_p, \mathcal{F}_{T-2}) \cdots \\ f(\mathbf{d}_{p+1}, \boldsymbol{\theta} | \boldsymbol{\theta}_p, \mathcal{F}_p) \end{aligned} \tag{1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_T, \ldots, \boldsymbol{\theta}_p)$ and we start modelling at time $p+1$, when the first p data values are available.

In a general case, the set of parameters $\boldsymbol{\theta}$ can be very large and they can evolve over time similarly to the measured data. In such a case, we would also need to associate a stochastic process with the parameter evolution. In such a situation, we would index the parameters by $t \in \mathcal{T}$ and we would model them similarly to the data. We would then have to describe the causal dependence of the parameters in terms of probability density $f(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \ldots, \boldsymbol{\theta}_p, \mathcal{F}_{t-1}) \ \forall t \in \mathcal{T}$. Instead we choose a smaller set of parameters $\boldsymbol{\theta}$, which we believe evolve slowly over time, so that we can account for their evolution by applying an exponential forgetting in their estimation [11],[14]. In this approximation we have

$$f(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \ldots, \boldsymbol{\theta}_p, \mathcal{F}_{t-1}) = f(\boldsymbol{\theta} | \mathcal{F}_{t-1}) \quad \forall t \in \mathcal{T} \tag{2}$$

We now come to the modelling of individual factors in (1). We split the factors again to get

$$f(\mathbf{d}_t, \boldsymbol{\theta} | \mathcal{F}_{t-1}) = f(\mathbf{d}_t | \boldsymbol{\theta}, \mathcal{F}_{t-1}) f(\boldsymbol{\theta} | \mathcal{F}_{t-1}) \tag{3}$$

where the first factor on the right-hand side is the actual parameteric model we will choose now conditioned on the past and the parameters. The second factor in (3) is the posterior probability density we will estimate from past data.

We choose the probability density of data conditioned on the parameters to be multivariate Gaussian of the form

$$\begin{aligned} f(\mathbf{d}_t | \boldsymbol{\theta}, \mathcal{F}_{t-1}) = \\ \frac{1}{[2\pi\,|\mathbf{R}|]^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left[\mathbf{R}^{-1} \begin{bmatrix} \mathbf{I} \\ -\mathbf{A} \end{bmatrix}' \begin{bmatrix} \mathbf{d}_t \\ \boldsymbol{\phi}_{t-1} \end{bmatrix} \begin{bmatrix} \mathbf{d}_t \\ \boldsymbol{\phi}_{t-1} \end{bmatrix}' \begin{bmatrix} \mathbf{I} \\ -\mathbf{A} \end{bmatrix} \right] \right\} \end{aligned} \tag{4}$$

where $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{R}\}$ are matrices of parameters (in large capitals because of their matrix nature despite in this context, these are realizations, not random matrices), $\boldsymbol{\phi}_t = [\mathbf{d}_{t-1}\ \mathbf{d}_{t-2} \cdots \mathbf{d}_{t-p}]'$. Parameter $\mathbf{R}$ stands for a covariance matrix of the model, $\mathbf{A}$ are parameters of the autoregression relating past observations of data to present or future observations (more detailed description will follow). There are several reasons for choosing such a model:

- Bayesian estimation of parameters of such a model is feasible, since it is from the so called *exponential family* of models [11], [14] and reduces the possibly difficult assimilation of data to a simple algebraic operation.
- The model contains as a subset the kind of models used by modern Financial Mathematics in the case that Efficient Market Hypothesis holds. These models are usually of the continuous-time type – they have to be discretized to obtain values at times $t \in \mathcal{T}$. For financial models of such kind see [9], [15], [16], [8] and others.

Since the probability density in (4) represents a conditional density of random vector $\mathbf{D}_t$, we can use the rules of probability theory [10] and decompose the random vector into conditional mean value and innovation

$$\begin{aligned} \mathbf{D}_t = \mathbb{E}\left[\mathbf{D}_t | \boldsymbol{\theta}, \mathcal{F}_{t-1}\right] + \boldsymbol{\varepsilon}_t = \mathbf{A}\boldsymbol{\Phi}_{t-1} + \boldsymbol{\Sigma}\,\mathbf{e}_t = \\ = \mathbf{A}_1\mathbf{D}_{t-1} + \mathbf{A}_2\mathbf{D}_{t-2} + \cdots + \mathbf{A}_p\mathbf{D}_{t-p} + \mathbf{c} + \boldsymbol{\Sigma}\,\mathbf{e}_t \end{aligned} \tag{5}$$

where $\mathbf{e}_t$ is a noise vector having a normal distribution with mean value $\mathbf{0}$ and covariance matrix $\mathbf{I}$ and $\mathbb{E}$ stands for mathematical expectation. The matrices $\mathbf{A}_1, \ldots, \mathbf{A}_p$, vector $\mathbf{c}$ and matrix $\boldsymbol{\Sigma}$, representing square root of the covariance $\mathbf{R}$ being now the parameters of the model. If the original noise is uncorrelated, the matrix $\boldsymbol{\Sigma}$ is diagonal.

Remark 2: Note we have changed the realizations of data variables on the right-hand side of (5) to their random variable representation. Since at time $t-1$ or greater the σ-algebra $\mathcal{F}_{t-1}$ is available and $\boldsymbol{\Phi}_{t-1}$ is measurable $\mathcal{F}_{t-1}$, the relationship (5) therefore holds unchanged. The new feature of (5) is that it holds also for future times, when the data in the condition are unknown. This feature will allow us to construct predictions of data evolution into the future.

For further computation, previous equation can be embedded into a wider scheme to be recursive and have the Markov property – see [15] page 49

$$\begin{aligned} \underbrace{\begin{bmatrix} \mathbf{D}_t \\ \mathbf{D}_{t-1} \\ \mathbf{D}_{t-2} \\ \vdots \\ \mathbf{D}_{t-p} \\ 1 \end{bmatrix}}_{\boldsymbol{\Phi}_t} = \underbrace{\begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_{p-1} & \mathbf{A}_p & \mathbf{c} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots \vdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & 1 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \mathbf{D}_{t-1} \\ \mathbf{D}_{t-2} \\ \mathbf{D}_{t-3} \\ \vdots \\ \mathbf{D}_{t-1-p} \\ 1 \end{bmatrix}}_{\boldsymbol{\Phi}_{t-1}} + \\ + \underbrace{\begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \ldots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}}_{\boldsymbol{\Sigma}} \underbrace{\begin{bmatrix} \mathbf{e}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}}_{\mathbf{e}_t} \end{aligned} \tag{6}$$

where we have used same notation for the embedded as for some of the original quantities, which should cause no

confusion, as we will speak of the extended quantities only from now on.

C. Structure estimation

Although we could use a full dimensional model of a predefined maximal time-lag and predefined number of data channels k such a practice can lead to large models with unnecessary AR component or even unnecessary channel, bringing additional inaccuracy to the model. For that reason Kárný and Kulhavý [17] have proposed a systematic way of Bayesian testing of most suitable model hypothesis through maximum aposteriori likelihood estimation.

D. Parameter estimation

To estimate parameters from past data means to evaluate the second factor on the right-hand side in (3) at each time. We will perform the estimation in a Bayesian manner [4],[3] and use the Bayes theorem. First we have to note that the new σ-algebra $\mathcal{F}_t$ contains only information about the world that doesn't influence the process and the information making $\mathbf{D}_t$ measurable $\mathcal{F}_t$. Now we can use Bayes theorem to write

$$f(\boldsymbol{\theta}|\mathcal{F}_t) = \frac{f(\mathbf{d}_t|\boldsymbol{\theta},\mathcal{F}_{t-1})f(\boldsymbol{\theta}|\mathcal{F}_{t-1})}{\int_{\mathcal{R}} f(\mathbf{d}_t|\boldsymbol{\theta},\mathcal{F}_{t-1})f(\boldsymbol{\theta}|\mathcal{F}_{t-1})d\boldsymbol{\theta}} \tag{7}$$

where $\mathcal{R}$ is the range of the parameters. Usually $\mathcal{R} = \mathbb{R}^n$, where n is the number of parameters and $\mathbb{R}$ is the set of real numbers. We can use this procedure at every time step to update the posterior probability density function (pdf), since we have already chosen the first factor in numerator on right hand-side in (4). At the first estimation step we need the initial condition - the *prior* pdf. At the first estimation step we need the initial condition - the *prior* pdf. Since we don't know anything about the time series, before the first data are obtained, we should choose a *non-informative* or *Jeffrey's* prior pdf. Such a choice is a special case of a wider class of prior distributions defined by a *conjugate* prior density [4]. Since the model chosen is from the exponential family of models a conjugate prior pdf can be chosen in a closed form. In [17] it is shown, such a pdf is of Gauss-Inverse-Wishart (Inverse-Gamma in one-dimensional case) type

$$GiW(\mathbf{V},\nu) \propto \tag{8}$$
$$|\mathbf{R}|^{-\frac{\nu+kp+1-k}{2}} \exp\left\{-\frac{1}{2}\left[\mathbf{R}^{-1}\begin{bmatrix}\mathbf{I}\\-\mathbf{A}\end{bmatrix}'\mathbf{V}\begin{bmatrix}\mathbf{I}\\-\mathbf{A}\end{bmatrix}\right]\right\}$$

where $\mathbf{V}$ is a positive definite *extended information matrix* and ν is a positive number of degrees of freedom. These parameters have to be chosen before the estimation starts. Such a choice can have a considerable impact on estimation of parameters $\mathbf{A}, \mathbf{R}$, but since in time series analysis $\mathcal{T}$ is usually large, the prior pdf choice can be treated with a little less care.

As described in [4], conjugate priors are chosen so that they are *self-reproducing* when estimation (7) is performed - the probability density function retains form (8) with $\mathbf{V}$, ν replaced by $\mathbf{V}_t$, ν_t respectively. The estimation step comes down to simple algebraic operation on these parameters, written recursively

$$\begin{aligned}\mathbf{V}_t &= \mathbf{V}_{t-1} + \begin{bmatrix}\mathbf{d}_t\\ \boldsymbol{\phi}_{t-1}\end{bmatrix}\begin{bmatrix}\mathbf{d}_t & \boldsymbol{\phi}_{t-1}\end{bmatrix} \\ \nu_t &= \nu_{t-1}+1\end{aligned} \tag{9}$$

where

$$\mathbf{V}_0 = \mathbf{V} \qquad \nu_0 = \nu \tag{10}$$

Remark 3: For computational reasons, the model can be decomposed into individual one dimensional regression models as follows. Because $\mathbf{R}$ is a regular positive definite and symmetric covariance matrix, it can be Cholesky-decomposed [18] and we obtain

$$\mathbf{R} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}' = \mathbf{L}\mathbf{D}\mathbf{L}' \tag{11}$$

$\mathbf{L}$ is a lower triangular matrix with units on the diagonal and its inverse $\mathbf{B}$ is also a lower triangular matrix with units on the diagonal. By multiplying the model in (6) by $\mathbf{B}$ and transferring additional terms from left-hand to right-hand side we obtain

$$\boldsymbol{\Phi}_t = [\mathbf{I}-\mathbf{B}]\,\boldsymbol{\Phi}_t + \mathbf{B}\mathbf{A}\boldsymbol{\Phi}_{t-1} + \mathbf{D}^{\frac{1}{2}}\mathbf{e}_t \tag{12}$$

where the square-root of $\mathbf{D}$ is well defined, since all the diagonal elements of $\mathbf{D}$ are strictly positive. The channels of such a model are no longer correlated and the parameters of the model can be estimated for k univariate models instead. Then by a backward transformation, the model can be brought back to its original form. When such a transformation is carried through, the prior information on the parameters is also transformed. Anyhow, in time series analysis, such a change shouldn't be important for the reasons mentioned.

E. Exponential forgetting

In case of slow parameter evolution, we can use the exponential forgetting technique, described in [11], page 46. Before the parameter estimation step a forgetting step is added, accounting for parameter evolution. This step replaces the parameter time-update step [11]. A forgetting factor $\kappa \in (0,1)$ is chosen and previous parameter probability density is flattened

$$f(\boldsymbol{\theta}_t = \boldsymbol{\theta}|\mathcal{F}_{t-1}) = f(\boldsymbol{\theta}_{t-1} = \boldsymbol{\theta}|\mathcal{F}_{t-1})^{\lambda} \tag{13}$$

For GiW model, inclusion of the forgetting causes a change of (9) to

$$\begin{aligned}\mathbf{V}_t &= \lambda\mathbf{V}_{t-1} + \begin{bmatrix}\mathbf{d}_t\\ \boldsymbol{\phi}_{t-1}\end{bmatrix}\begin{bmatrix}\mathbf{d}_t & \boldsymbol{\phi}_{t-1}\end{bmatrix} \\ \nu_t &= \kappa\nu_{t-1}+1\end{aligned} \tag{14}$$

The choice of optimal forgetting factor and also the structure of the forgetting are difficult tasks. These tasks can be left for consideration of an expert, but attempts were made to choose this factor systematically [19].

Remark 4: Exponential forgetting influences the model structure choice discussed in subsection (B). To the best knowledge of the author, no satisfactory feasible algorithm of structure estimation has been proposed yet for $\kappa < 1$. Therefore, the practice used is to estimate model structure using $\kappa = 1$.

F. Prediction using Monte Carlo sampling from parameter distribution

Let's now suppose we know the model parameters perfectly – e.g. their estimated joint probability density is a delta function $f(\boldsymbol{\theta}|\mathcal{F}_t) = \delta(\boldsymbol{\theta} = [\mathbf{A}, \boldsymbol{\Sigma}])$ where $\mathbf{A}, \boldsymbol{\Sigma}$ are now matrices of numbers, not random variables. We now want to construct the *prediction* of the stochastic process $\boldsymbol{\Phi}_t$ up to a horizon $t+h$, with the information contained in $\mathcal{F}_t$. We therefore need to estimate probability density function $f(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t)$, characterizing the distribution. Since we know the stochastic process evolves as in (6), if the parameters are known we obtain a predicted random variable

$$\boldsymbol{\Phi}_{t+h} = \mathbf{A}^h \boldsymbol{\Phi}_t + \sum_{i=0}^{h-1} \mathbf{A}^i \boldsymbol{\Sigma} \mathbf{e}_{t+h-i} \tag{15}$$

and we can also compute the mean value (point prediction) and covariance of this random variable

$$\begin{aligned} &\boldsymbol{\mu}_h = \mathbb{E}\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t, \boldsymbol{\theta}\right] = \\ &\mathbb{E}\left[\mathbf{A}^h \boldsymbol{\Phi}_t|\mathcal{F}_t, \boldsymbol{\theta}\right] + \mathbb{E}\left[\sum_{i=0}^{h-1} \mathbf{A}^i \boldsymbol{\Sigma} \mathbf{e}_{t+h-i}|\mathcal{F}_t, \boldsymbol{\theta}\right] = \\ &\mathbf{A}^h \boldsymbol{\Phi}_t + \sum_{i=0}^{h-1} \mathbf{A}^i \boldsymbol{\Sigma} \mathbb{E}\left[\mathbf{e}_{t+h-i}|\mathcal{F}_t, \boldsymbol{\theta}\right] = \mathbf{A}^h \boldsymbol{\Phi}_t \end{aligned} \tag{16}$$

$$\begin{aligned} &\mathbf{R}_h = \operatorname{cov}\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t, \boldsymbol{\theta}\right] = \\ &\operatorname{cov}\left[\mathbf{A}^h \boldsymbol{\Phi}_t|\mathcal{F}_t, \boldsymbol{\theta}\right] + \operatorname{cov}\left[\sum_{i=0}^{h-1} \mathbf{A}^i \boldsymbol{\Sigma} \mathbf{e}_{t+h-i}|\mathcal{F}_t, \boldsymbol{\theta}\right] = \\ &\sum_{i=0}^{h-1} \left(\mathbf{A}^i \boldsymbol{\Sigma}\right) \underbrace{\operatorname{cov}\left[\mathbf{e}_{t+h-i}|\mathcal{F}_t, \boldsymbol{\theta}\right]}_{\mathbf{I}} \left(\mathbf{A}^i \boldsymbol{\Sigma}\right)' = \\ &\sum_{i=0}^{h-1} \left[\mathbf{A}^i \boldsymbol{\Sigma} \left(\mathbf{A}^i \boldsymbol{\Sigma}\right)'\right] \end{aligned} \tag{17}$$

where $\mathbf{A}^0 \equiv \mathbf{I}$, since we consider a model with $\operatorname{cov}\left[\mathbf{e}_t, \mathbf{e}_s|\, \mathcal{F}_t\right] = \mathbf{0}$ for $s > t$. These first two moments of distribution of $\boldsymbol{\Phi}_{t+h}$ will be important in the next paragraph. Since $\mathbf{e}_t$ are normally distributed $\forall t \in \mathcal{T}$, we obtain the pdf of random variable on the left-hand side in (15) as

$$f(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t) = \boldsymbol{\varphi}_{\boldsymbol{\mu}_h, \mathbf{R}_h}(\boldsymbol{\phi}_{t+h}) \tag{18}$$

where $\varphi_{\boldsymbol{\mu}, \mathbf{R}}$ is the normal multivariate probability density fuction of a random variable with mean value $\boldsymbol{\mu}$ and covariance $\mathbf{R}$.

If instead of the delta function, the parameter values are uncertain, we can proceed with the multi-step ahead prediction by drawing N *Monte Carlo* samples $\boldsymbol{\theta}_i$, where $i \in \{1, \ldots, N\}$ from their probability distribution characterized by $f(\boldsymbol{\theta}|\mathcal{F}_t)$. After having estimated the predictions, we reconstruct the predictive probability density by simple numerical integration.

In an honest computation, we should consider the predicted value for $t+1$ to estimate new probability density $f(\boldsymbol{\theta}|\mathcal{F}_{t+1})$ and we should perform the forgetting (13). Instead, for computational feasibility, we can use a so called *receding horizon* or *moving window* approximation. In this approximation

$$f(\boldsymbol{\theta}|\mathcal{F}_s) = f(\boldsymbol{\theta}|\mathcal{F}_t) \qquad t < s \leq t+h \tag{19}$$

for purposes of prediction up to time $t+h$ – we do not update the parameter distribution with the use of *estimated* data. Once we obtain new real data at time $t+1$, we proceed with parameter estimation (7), forget (13), again fix the distribution, draw new N samples and predict up to horizon $t+1+h$ and so on. This approximation allows us to use the previously obtained result (18) for multistep-ahead predictive probability density function, except now we obtain N such result, each conditioned on the drawn parameter value $\boldsymbol{\theta}_i$.

We now reconstruct the final predictive probability density function $f(\phi_{t+h}|\mathcal{F}_t)$ by integrating out the parameters, which in the Monte Carlo approximation results in an averaging

$$f_N(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t) = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\phi}_{t+h}|\boldsymbol{\theta}_i, \mathcal{F}_t) \tag{20}$$

In the moving window approximation such a probability density function should converge pointwise to the distribution obtained by a general integration for $N \to \infty$.

From the distribution obtained we can generally compute its central moments, which characterize the distribution of the predicted values of the stochastic process $\boldsymbol{\Phi}_t$. For illustration we now compute the first two central moments of the distribution. With the first two moments known, we could fit a normal distribution to the prediction, although it is certain, that the uncertainty in parameters $\boldsymbol{\theta}$ causes the real predictive distribution to be heavy–tailed. For the mean value we get

$$\begin{aligned} &\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t\right] = \int_{\mathbb{R}^{kp+1}} \boldsymbol{\phi}_{t+h} f_N(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t)\, d\boldsymbol{\phi}_{t+h} = \\ &\int_{\mathbb{R}^{kp+1}} \boldsymbol{\phi}_{t+h} \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t, \boldsymbol{\theta}_i)\, d\boldsymbol{\phi}_{t+h} = \\ &\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t, \boldsymbol{\theta}_i\right] \end{aligned} \tag{21}$$

if the sum and integral can be transposed, which we assume. From knowledge of the covariance matrices $\mathbf{R}_{h;i} = \operatorname{cov}(\boldsymbol{\phi}_{t+h}|\boldsymbol{\theta}_i, \mathcal{F}_t)$ we can compute

$$\begin{aligned} &\operatorname{cov}\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t\right] = \\ &\mathbb{E}_N\Big[\left(\boldsymbol{\phi}_{t+h} - \mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t\right]\right)\left(\boldsymbol{\phi}_{t+h} - \mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t\right]\right)'|\mathcal{F}_t\Big] \\ &= \int_\Omega \left(\boldsymbol{\phi}_{t+h} - \frac{1}{N} \sum_j \mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t, \boldsymbol{\theta}_j\right]\right) \cdot \\ &\left(\boldsymbol{\phi}_{t+h} - \frac{1}{N} \sum_k \mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t, \boldsymbol{\theta}_k\right]\right)' \cdot \\ &\frac{1}{N} \sum_i f(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t, \boldsymbol{\theta}_i)\, d\boldsymbol{\phi}_{t+h} = \end{aligned} \tag{22}$$

$$\frac{1}{N^3}\sum_i\int_\Omega\sum_j\left(\boldsymbol{\phi}_{t+h}-\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_j\right]\right)\cdot$$
$$\sum_k\left(\boldsymbol{\phi}_{t+h}-\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_k\right]\right)'f(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i)\,d\boldsymbol{\phi}_{t+h}=$$
$$\frac{1}{N^3}\sum_{i,j,k}\int_\Omega(\boldsymbol{\phi}_{t+h}\boldsymbol{\phi}'_{t+h}-\boldsymbol{\phi}_{t+h}\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_k\right]'-$$
$$\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_j\right]\boldsymbol{\phi}'_{t+h}+$$
$$\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_j\right]\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_k\right]')f(\boldsymbol{\phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i)\,d\boldsymbol{\phi}_{t+h}=$$
$$\frac{1}{N^3}\sum_{i,j,k}(\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}\boldsymbol{\phi}'_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i\right]-$$
$$\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i\right]\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_k\right]'-$$
$$\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_j\right]\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i\right]'+$$
$$\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_j\right]\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_k\right]')=$$
$$\frac{1}{N}\sum_i\mathbf{R}_{h;i}+\frac{1}{N}\sum_i\left(\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i\right]\right)\left(\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i\right]\right)'-$$
$$\frac{1}{N^2}\left(\sum_i\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_i\right]\right)\left(\sum_j\mathbb{E}_N\left[\boldsymbol{\Phi}_{t+h}|\mathcal{F}_t,\boldsymbol{\theta}_j\right]\right)'$$

where $\Omega=\mathbb{R}^{kp+1}$ and we again suppose, we can transpose sums and integrals. Higher moments could be computed to obtain a more accurate approximation of the final distribution.

III. CONCLUSIONS AND FUTURE WORKS

In the article a prediction procedure has been proposed, that can be applied to wide variety of problems, where the use of AR models is appropriate. In case of financial time series, the model has to be extensively tested. The usual practice in Financial Mathematics [15], [9], [16] is to construct a model in agreement with so called Efficient Market Hypothesis (EMH), proposed by E.Fama in his PhD. thesis in 1960 later supported in [20], [21]. This hypothesis, especially in its stronger forms is in contradiction with presented model. Tests proposal should play a considerable part of future work on the project.

Although the Monte Carlo approximation should converge to the distribution obtained by original integration, it can be quite computationally demanding. Possibly the moments of the predictive distribution (21) and (22) and even further moments could be computed analytically under the moving window approximation using moment generating function of the GiW distribution.

Except for normally distributed data channels, the expert can assume, some of the data should be rather log-normally distributed, when the mean value is subtracted. Such channels can be incorporated into the model, but bring a few computational difficulties. Computations have been already made, leading to incorporation of such channels into the model and should be presented soon.

REFERENCES

[1] G. Box and G. Jenkins, *Time Series Analysis*. San Francisco, U.S.A.: Holden-Day, 1970.

[2] T. Cipra, *Finanční Ekonometrie*. EKOPRESS, 2008, in Czech.

[3] J. Ghosh, M. Delampady, and T. Samanta, *An Introduction to Bayesian Analysis, Theory and Methods*. Springer, 2006.

[4] C. Robert, *The Bayesian Choice, From Decision Theoretic Foundations to Computational Implementation*. Springer, 2007.

[5] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Athena Scientific, 2000, vol. 1.

[6] T. Yamamoto, "Asymptotic mean square prediction error for an autoregressive model with estimated coefficients," *Applied Statistics*, vol. 25, pp. 123–127, 1976.

[7] C. Ing, "Multistep prediction in autoregressive processes," *Econometric theory*, vol. 19, pp. 254–279, 2003.

[8] H. Fölmer and A. Schied, *Stochastic Finance, An Introduction in Discrete Time*. Walter de Gruyter, 2002.

[9] S. Shreve, *Stochastic Calculus for Finance II., Continuous-Time Models*. Springer, 2004.

[10] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1986.

[11] M. Kárný, J. Böhm, T. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař, *Optimized Bayesian Dynamic Advising, Theory and Algorithms*. Springer, 2005.

[12] P. Brockwell and R. Davis, *Introduction to Time Series and Forecasting*. Springer, 1996.

[13] J. Anděl, *Základy matematické statistiky*. Matfyz-Press, 2005, in Czech.

[14] I. Nagy, L. Pavelková, E. Suzdaleva, J. Homolová, and M. Kárný, *Bayesian Decision Making*. Czech Academy of Sciences, 2005.

[15] S. Shreve, *Stochastic Calculus for Finance I., The Binomial Asset Pricing Model*. Springer, 2004.

[16] I. Karatzas and S. Shreve, *Methods of Mathematical Finance*. Springer, 1998.

[17] M. Kárný and R. Kulhavý, "Structure determination of regression-type models for adaptive prediction and control," in *Bayesian Analysis of Time Series and Dynamic Models*, J. C. Spall, Ed. Marcel Dekker, 1988, pp. 313–345.

[18] R. Horn and J. C.R., *Matrix Analysis*. Cambridge University Press, 1985.

[19] A. Votava, "*Estimation of forgetting factor in the frame of dynamic decision making* Bachelor's thesis, Faculty of Mathematics and Physics, Charles University," Prague, 2009.

[20] E. Fama, "The behavior of stock market prices," *Journal of Business*, vol. 38, p. 34105, 1965.

[21] P. Samuelson, "Proof that properly anticipated prices fluctuate randomly," *Industrial Management Review*, vol. 6, pp. 44–49, 1965.