Applied Mathematical Sciences, Vol. 6, 2012, no. 74, 3675 - 3686

An Application of Neural Networks Trained with Kalman Filter Variants (EKF and UKF) to Heteroscedastic Time Series Forecasting

Mauri Aparecido de Oliveira

Department of Quantitative Methods
Escola Paulista de Política, Economia e Negócios – EPPEN
Federal University of São Paulo – Brazil – UNIFESP
mauri.oliveira@unifesp.br

Abstract

In this work, two Kalman filters variants are applied to recurrent neural network training. The Unscented Kalman Filter (UKF) has been presented outperforming the Extended Kalman filter (EKF). Due to this a comparison between GARCH model and a neural network using EKF and UKF was implemented to heteroscedasticity time series prediction. Our experimental results and analysis confirm that a neural network using UKF perform better prediction than the other approach.

1. Introduction

The Extended Kalman Filter (EKF) was successfully applied to the estimation of parameters of neural networks [1] [2] [3]. It was shown that the statistics estimated by the EKF can be used to estimate sequentially the structure (number of hidden neurons and connections) and the parameters of feedforward networks [4], recurrent [5] and Radial Basis Function (RBF). The Unscented Kalman filter estimator has been presented [6] [7] with results that exceed the performance of the EKF state estimation in nonlinear. In the estimation of parameters of the feedforward neural networks UKF is comparable or slightly better than the EKF [8], with the significant advantage that it does not require the calculation of the Jacobian of the neural network.

Consider the dynamic system of nonlinear discrete time at which the signal (or series) unobserved $x(t)$ is modeled as a Markov process [9] with initial distribution $p(x(0))$ and transition equation:

$$x(t) = f\left(x(t-1), u(t-1)\right) + q(t), \tag{1a}$$

where $u(t)$ denotes the input that is exogenous known. The observations are assumed conditionally independent given the state $x(t)$:

$$y(t) = h(x(t)) + r(t). \tag{1b}$$

The noise process $q(t)$ guides the dynamic system, while the noise observation is given by $r(t)$. The Minimum Mean Squared Error (MMSE) of the state $x(t)$ of the system state discrete-time nonlinear (1) satisfies the conditions that the estimation error $\tilde{x}(t) = x(t) - \hat{x}(t)$ is not biased $E\left[\tilde{x}(t)\right] = 0$ and also orthogonal to the observation $y(t)$, ie. $E\left[\tilde{x}(t)\, y^T(t)\right] = 0$. The EKF and the UKF provide a MMSE of state $x(t)$ using the state predictor-corrector scheme as shown in Figure-1.

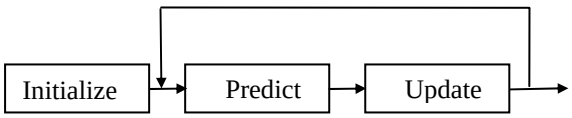

Figura-1 Structure recursive predictor-corrector of the Kalman filter.

Given the state estimator $\hat{x}(t-1)$ and its covariance $P_x(t-1)$, obtained from the information set until time step $(t-1)$: $Y_{t-1} = \{y(1), y(2), ..., y(t-1)\}$, the filter predicts the future state using the process model and knowledge about the distribution of the noise process. The predicted mean and covariance are ideally:

$$\hat{x}^-(t) = E\left[x(t) | Y_{t-1}\right], \tag{2a}$$

$$P_x^-(t) = E\left[\left(x(t) - \hat{x}^-(t)\right)\left(x(t) - \hat{x}^-(t)\right)^T | Y_{t-1}\right]. \tag{2b}$$

The estimator $\hat{x}(t)$ and its covariance $P_x(t)$ are obtained by the update (correction) of the state prediction $\left(\hat{x}^-(t), P_x^-(t)\right)$ with the current observation $y(t)$:

$$\hat{x}(t) = \hat{x}^-(t) + K(t)\left(y(t) - \hat{y}^-(t)\right), \tag{3a}$$

$$K(t) = P_{xy}(t) P_y^{-1}(t), \tag{3b}$$

$$P_x(t) = P_x^{-1}(t) - K(t)P_y^{-1}(t)K^T(t), \tag{3c}$$

where $\hat{y}^-(t) = E\left[y(t)|Y_{t-1}\right]$ and $P_y(t) = E\left[\left(y(t)-\hat{y}^-(t)\right)\left(y(t)-\hat{y}^-(t)\right)^T|Y_{t-1}\right]$ are the prediction of the observation $y(t)$ and its covariance. The conditioned correlation is given by $P_{xy}(t) = E\left[\left(x(t)-\hat{x}^-(t)\right)\left(y(t)-\hat{y}^-(t)\right)^T|Y_{t-1}\right]$. These equations depend on the predicted values of the first two moments $x(t)$ and $y(t)$, give the set of observations Y_{t-1}.

In this paper we will consider training a recurrent neural network autoregressive nonlinear endogenous inputs (Non-linear Autoregressive with exogenous inputs – NARX), using the EKF and UKF. The results of the predictions obtained from the NARX network training with the EKF and UKF are compared with the results of GARCH (Generalized Autoregressive Conditional Heteroscedasticity).

2. State space model of recurrent neural network NARX

NARX model of a dynamic system is given by:

$$y(t) = f\left(y(t-1),\ldots,y(t-\Delta_y),u(t-1),\ldots,u(t-\Delta_u)\right),$$

where $y(t)$ corresponds to the real output of the system (without noise), $u(t)$ the known entry at time *t*, Δ_u and Δ_y are the orders of inputs and outputs and $f(\cdot)$ is a nonlinear function. We will consider an autoregressive model with nonlinear exogenous inputs for which we give the name of NARX_RMPL, i.e. NARX Recurrent Multilayer Perceptron. Let us assume that the model has two layers of neurons (Figure-2), an output layer having a linear activation function. The output of the i-th neuron of the hidden NARX_RMPL network is given by (4).

$$\phi_i\left(\mathbf{y}_{t-1},\mathbf{u}_{t-1},\mathbf{b}_i\right) = \tanh\left\{b_{i0} + \sum_{l=1}^{\Delta_y} b_{il}y_{t-l} + \sum_{j=1}^{n_u}\sum_{\tau=1}^{\Delta_u} b_{ij\tau}u_{j,t-\tau}\right\}. \tag{4}$$

where $\mathbf{y}_{t-1} = \left[y(t-1)\cdots y(t-\Delta_y)\right]^T_{1\times\Delta_y}$ denotes the network vector of previous outputs, $\mathbf{u}_{t-1} = \left[u^T(t-1)\cdots u^T(t-\Delta_u)\right]^T_{1\times\Delta_y}$ is the vector of previous inputs and $\mathbf{b}_i = \left[b_{i0}b_{i1}\cdots b_{i\Delta_y}b_{i11}\cdots b_{in_u\Delta_u}\right]^T$ denotes the weight vector of hidden neurons.

The estimation of parameters of a recurrent neural network can be placed in a structural form of nonlinear state estimation, from the definition of state space model of dynamic networks. The state vector ***x*** is obtained by increasing the base state ***y*** [5], which in our case is defined as the outputs of previous recurrent

network, with the parameter vector ***w***. Since ***w*** denotes the vector of unknown size of the network weights.

$$x(t) = \Phi\left(x(t-1), u(t-1)\right) + q(t), \qquad q(t) \sim N\left(0, Q(t)\right), \tag{5a}$$

$$y(t) = H(t)\,x(t) + r(t), \qquad r(t) \sim N\left(0, R(t)\right). \tag{5b}$$

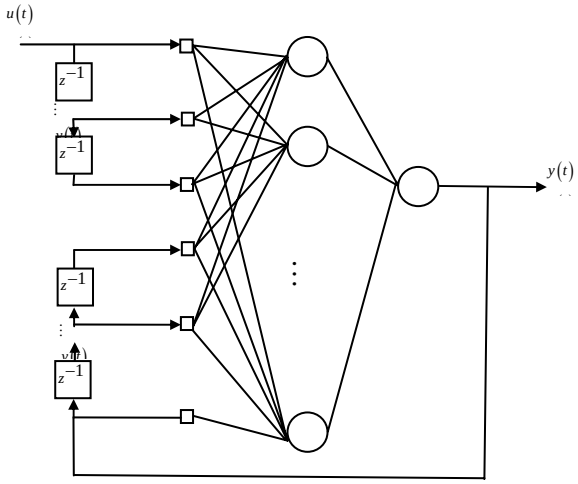

Figure-2 NARX recurrent neural network.

Equation (5a) describes the time evolution of the augmented state ***x***, while equation (5b) selects the current output of the network as the observation. The process noise $q(t)$ and observation noise $r(t)$ are assumed to be jointly independent, white, and Gaussian with known covariances $Q(t)$ and $R(t)$, respectively.

$$\hat{y}^{-}(t) = H(t)\,\hat{x}^{-}(t), \tag{6a}$$

$$P_y(t) = H(t)\,P_x^{-}(t)\,H^T(t) + R(t), \tag{6b}$$

$$P_{xy}(t) = P_x^{-}(t)\,H^t(t). \tag{6c}$$

Due to the linearity of the observation equation (5b), the prediction of the observation $\hat{y}^{-}(t)$, its covariance $P_y(t)$ and cross correlation $P_{xy}(t)$, necessary to update the filter in step (3) are given by:

$$x(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \vdots \\ y\left(t-\Delta_y+1\right) \\ \mathbf{w}(t) \end{bmatrix}, q_t = \begin{bmatrix} q_y(t) \\ 0 \\ \vdots \\ 0 \\ q_w(t) \end{bmatrix}, Q(t) = \begin{bmatrix} Q_y(t) & 0 \\ 0 & Q_w(t) \end{bmatrix},$$

$$\Phi\left(x(t),u(t)\right)=\begin{bmatrix} f\left(y(t),u(t),w(t)\right) \\ y(t-1) \\ \vdots \\ y\left(t-\Delta_y+1\right) \\ \mathbf{w}(t) \end{bmatrix} \text{ and } H=\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}^T_{1\times\left(\Delta_y+n_w\right)}$$

The problem of parameter estimation and state of the recurrent neural network NARX_RMLP is therefore reduced the spread of the state $x(t)$ through non-linear dynamic equation (5a) so as to obtain the prediction $\left(\hat{x}^-(t),\hat{P}^-(t)\right)$ [5].

3. Nonlinear Bayesian Filters

Let us consider two different approaches to the estimation of a non-linear and apply them to the estimation of NARX recurrent neural network. As we saw in the previous section, the remaining problem to be solved is the estimation of statistics of a random variable propagated through a nonlinear transformation. Let us define the problem more generally. Suppose *x* is a random variable with mean $\hat{x}$ and covariance P_x. A random variable *y* is related to *x* through the nonlinear function $y=f(x)$. We want to calculate the mean $\hat{y}$ and covariance P_y of *y*. Note that the derived solutions could be easily applied to the prediction of the state (2) introducing the substitutions $x\to x(t-1)$ and $y\to x(t)$.

3.1 Extended Kalman Filter

There are two basic assumptions in the derivation of the Kalman filter. The first is that the system is described by a model of linear state space and the second is that the noises are white and Gaussian with zero mean, and they are also assumed to be uncorrelated with each other and with the initial state. When these assumptions are satisfied, the Kalman filter is optimal in the sense of mean square error. When the system under consideration is nonlinear, the first condition is violated and the extended Kalman filter (EKF) is applied as a sub-optimal filter. In the EKF the nonlinear terms are approximated by linear terms of first order using Taylor expansion. To begin, consider a nonlinear system described as:

$$\mathbf{y}(t)=h_t(t)\mathbf{x}(t)+\mathbf{r}(t) \tag{7}$$

$$\mathbf{x}(t+1)=f_t(t+1,t)\mathbf{x}(t)+\mathbf{q}(t) \tag{8}$$

Equations (7) and (8) are the non-linear equivalents of the equations (1a) and (1b). They are the equations of measurement and process for the nonlinear case, where h_t and f_t are functions of nonlinear vectors of the state. Operating as stated

above, we have made two linear equations by using the Taylor series expansion of $\hat{\mathbf{x}}(t|t-1)$ and $\hat{\mathbf{x}}(t|t)$ as follows, respectively:

$$h_t(\mathbf{x}(t)) = h_t(\hat{\mathbf{x}}(t|t-1)) + \mathbf{H}_t(t)(\mathbf{x}(t) - \hat{\mathbf{x}}(t|t-1)) + \ldots \quad (9)$$

$$f_t(\mathbf{x}(t)) = f_t(\hat{\mathbf{x}}(t|t)) + \mathbf{F}_t(t+1,t)(\mathbf{x}(t) - \hat{\mathbf{x}}(t|t)) + \ldots \quad (10)$$

where the Jacobian matrices $\mathbf{F}_t(t+1,t)$ and $\mathbf{H}_t(t)$ are defined as:

$$\mathbf{F}_t(t+1,t) = \frac{\partial f_t(\hat{\mathbf{x}}(t|t))}{\partial \mathbf{x}} \quad \text{and} \quad \mathbf{H}_t(t) = \frac{\partial h_t(\hat{\mathbf{x}}(t|t-1))}{\partial \mathbf{x}}.$$

Ignoring higher order terms in Taylor expansions above, the measurement equations and nonlinear process can be approximated as follows:

$$\mathbf{y}(t) = \mathbf{H}_t(t)\mathbf{x}(t) + \mathbf{u}(t) + \mathbf{r}(t), \quad (11)$$

$$\mathbf{x}(t+1) = \mathbf{F}_n(t+1,t)\mathbf{x}(t) + \mathbf{v}(t) + \mathbf{q}(t), \quad (12)$$

where:

$$\mathbf{u}(t) = h_t(\hat{\mathbf{x}}(t|t-1)) - \mathbf{H}_n(t)\hat{\mathbf{x}}(t|t-1) \text{ and } \mathbf{v}(t) = f_t(\hat{\mathbf{x}}(t|t-1)) - \mathbf{F}_t(t+1,t)\hat{\mathbf{x}}(t|t).$$

The algorithm of the extended Kalman filter can be represented as:

Computing the Kalman gain

$$\mathbf{K}(t) = \frac{\mathbf{P}(t|t-1)\mathbf{H}_t^T(t)}{\mathbf{H}_t(t)\mathbf{P}(t|t-1)\mathbf{H}_t^T(t) + \mathbf{R}(t)} \quad (13)$$

Update (correct) measures

$$\hat{\mathbf{x}}(t|t) = \hat{\mathbf{x}}(t|t-1) + \mathbf{K}(t)\left[\mathbf{y}(t) - h_t\hat{\mathbf{x}}(t|t-1)\right] \quad (14)$$

$$\mathbf{P}(t|t) = \mathbf{P}(t|t-1) - \mathbf{K}(t)\mathbf{H}_t(t)\mathbf{P}(t|t-1) \quad (15)$$

Update time (prediction)

$$\hat{\mathbf{x}}(t+1|t) = f_t\hat{\mathbf{x}}(t|t) \quad (16)$$

$$\mathbf{P}(t+1|t) = \mathbf{F}_t(t+1,t)\mathbf{P}(t|t)\mathbf{F}_t^T(t+1,t) + \mathbf{Q}(t) \quad (17)$$

Tabela-1 Extended Kalman filter equations.

Comparing the equations of the Kalman filter with those of the EKF in Table-1, only a few differences are noted. First, the linear terms $\mathbf{H}(t)\hat{\mathbf{x}}(t|t-1)$ and $\mathbf{F}_t(t+1,t)\hat{\mathbf{x}}(t|t)$ presented in the Kalman filter are changed by $h_t(\hat{\mathbf{x}}(t|t-1))$ and $f_t(\hat{\mathbf{x}}(t|t-1))$ in the EKF, respectively. The state transition matrix $\mathbf{F}(t+1,t)$ and the measure matrix $\mathbf{H}(t)$ are also changed in the Kalman filter, respectively, by

the Jacobian matrices $\mathbf{F}_t(t+1,t)$ and $\mathbf{H}_t(t)$ in the EKF. The matrices of derivatives must be recalculated for each iteration of the Kalman filter.

3.2 Unscented Kalman Filter

Julier and Uhlmann [6][7] proposed the Unscented Transformation to calculate the statistics of a variable **x** propagated through a nonlinear function $\mathbf{y} = f(\mathbf{x})$. Consider the propagation of a random variable **x** (dimension L) through a non-linear function, $\mathbf{y} = f(\mathbf{x})$. Assuming that **x** has mean $\bar{\mathbf{x}}$ and covariance $\mathbf{P}_\mathbf{x}$. To calculate the statistics of **y**, we build the matrix $\aleph$ with 2L + 1 sigma vectors $\aleph_i$ according to the following equations:

$$\aleph_0 = \bar{\mathbf{x}}, \tag{18}$$

$$\aleph_i = \bar{\mathbf{x}} + \left(\sqrt{(L+\lambda)\mathbf{P}_\mathbf{x}}\right)_i, \quad i = 1,...,L, \tag{19}$$

$$\aleph_i = \bar{\mathbf{x}} - \left(\sqrt{(L+\lambda)\mathbf{P}_\mathbf{x}}\right)_{i-L}, \quad i = 1,...,2L, \tag{20}$$

where $\lambda = \alpha^2(L+\kappa) - L$ is the scaling parameter.

The constant α determines the dispersion of the sigma points around $\bar{\mathbf{x}}$, is usually set as a small positive value $\left(1 \le \alpha \le 10^{-4}\right)$. The constant κ is the second scaling parameter, which is usually set to be 3 – L [1], β is used to incorporate a priori knowledge of the distribution of **x** (to Gaussian distributions, $\beta = 2$ is optimum). $\left(\sqrt{(L+\lambda)\mathbf{P}_\mathbf{x}}\right)_i$ is the *i*-th column of the square root of the matrix(i.e., lower triangular Cholesky factorization).

Each sigma point is propagated through the function $f(\cdot)$ to produce the transformed set of sigma points $\Im_i = f(\aleph_0)$ and the mean $\bar{\mathrm{y}}$ of a transformed distribution is estimated by:

$$\bar{\mathbf{y}} = \sum_{i=0}^{2L} W_i \wp_i = \frac{\lambda}{L+\lambda} f(\bar{x}) + f\left(\bar{x} - \sqrt{L+\lambda}.s_{x,i}\right) + \frac{1}{2(L+\lambda)} \sum_{i=1}^{L} \left(f\left(\bar{x} + \sqrt{L+\lambda}.s_{x,i}\right)\right)$$

The covariance estimator obtained by the unscented transform is given by:

$$\mathbf{P}_\mathbf{y} = \sum_{i=0}^{2L} W_i \left(\wp_i - \bar{\mathbf{y}}\right)\left(\wp_i - \bar{\mathbf{y}}\right)^T = \frac{\lambda}{L+\lambda}\left(f(\hat{x}) - \bar{\mathbf{y}}\right)\left(f(\hat{x}) - \bar{\mathbf{y}}\right)^T +$$

$$+ \frac{1}{2(L+\lambda)} \sum_{i=1}^{L} \left(f\left(\bar{\mathbf{x}} + \sqrt{L+\lambda}.s_{x,i}\right) f\left(\bar{\mathbf{x}} + \sqrt{L+\lambda}.s_{x,i}\right)^T \right) +$$

$$+ \frac{1}{2(L+\lambda)} \sum_{i=1}^{L} \left(\left(f\left(\bar{\mathbf{x}} - \sqrt{L+\lambda}.s_{x,i}\right) - \bar{\mathbf{y}}\right)\left(f\left(\bar{\mathbf{x}} - \sqrt{L+\lambda}.s_{x,i}\right) - \bar{\mathbf{y}}\right)^T \right) \tag{21}$$

The estimation of states and parameters of NARX neural network (state space model given by (8)) using the unscented Kalman filter consists to apply the unscented transformation in the dynamic equation (5a) so that to obtain the prediction $\left(\hat{x}_k^-, \hat{P}_k^-\right)$. The predicted statistics are updated with the current observation $\mathbf{y}(t)$ substituting (6) in the data update steps equations (3).

4. GARCH Models

The generalized ARCH model, known as GARCH was first proposed by Bollerslev in 1986. This model is the most used model for the volatility and the GARCH (1,1) is the most common. The equation for the process GARCH (1,1) is given by:

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \tag{22}$$

The important point is that the conditional variance ε_t is given by $E_{t-1} = \left[\varepsilon_t^2\right] = h_t$.

4.1 Test for non-linearity

In the analysis of heteroscedastic series, before beginning a search for a general specification to find a model that fit your particular data set, it is important to test for nonlinearity [10] [11] [12] [13]. Tests to check non-linearity have been most used are the Brock, Dechert, and Scheinkman [1987], the test of McLeod and Li [1983], a test developed by Hsieh [1989] and a test suggested by Teräsvirta, Lin and Granger [1993]. In this paper we consider the test of McLeod and Li.

In the estimation of an ARMA model, the autocorrelation function (ACF) can help select the values of *p* and *q*, and the ACF of the residuals is an important diagnostic tool. Unfortunately, the ACF as used in linear models can lead to false conclusions in nonlinear models. The reason is that the autocorrelation coefficients measure the degree of linear association between y_t and y_{t-i}. Thus, the ACF may fail to detect important nonlinear relationships in the data. Having interest in nonlinear relationships of the data, a useful diagnostic tool is to examine the ACF of squares and cubes of a series of values.

The test McLeod-Li (1983) seeks to determine whether there are significant autocorrelations in the squared residuals of a linear equation. To perform the test is to estimate the series model using the best linear fit and identify the residuals $\hat{e}_t$.

As in a formal test for ARCH errors, we construct the autocorrelations of squared residuals. Making ρ_i denote the sample correlation coefficient between the residuals $\hat{e}_t^2$ and $\hat{e}_{t-i}^2$ we use the Ljung-Box statistic to determine whether the squared residuals exhibit serial correlation.

Consequently, we have:

$$Q = T(T+2)\sum_{i=1}^{n} \frac{\rho_i}{(T-i)}. \tag{23}$$

The value of Q has an asymptotic distribution χ^2 with *n* degrees of freedom if the sequence $\{\hat{e}_t^2\}$ is non-correlated. Reject the null hypothesis is equivalent to accepting that the model is nonlinear. Alternatively, one can estimate the regression: $\hat{e}_t^2 = \alpha_0 + \alpha_1 \hat{e}_{t-1}^2 + \ldots + \alpha_n \hat{e}_{t-n}^2 + \nu_t$. If there is no non-linearity, α_1 to α_n are statistically equal to zero. With a sample of *T* residues, if not non-linearities, the statistical test TR^2 converges to a distribution χ^2 with *n* degrees of freedom. This test has substantial power to detect various forms of non-linearity. However, the actual shape of the nonlinearity is not specified by the test. Reject the null hypothesis of linearity does not tell the nature of the nonlinearity present in the data.

5. Experiment

The data set is the set of values of the daily price of 60kg sack of soybeans in the period from 07/29/1997 to 11/28/2003, totaling 1575 values in Figure-3 is shown a representation of all data.
These data were separated into two parts: the training set and test set. The training set includes 1000 data from 07/29/1997 until 08/09/2001 and was used for the estimation of the GARCH model and the training of neural networks using the EKF and UKF. The test set includes data 08/10/2001 to 28/10/2003, totaling 575 values, and was used for comparison of different approaches.

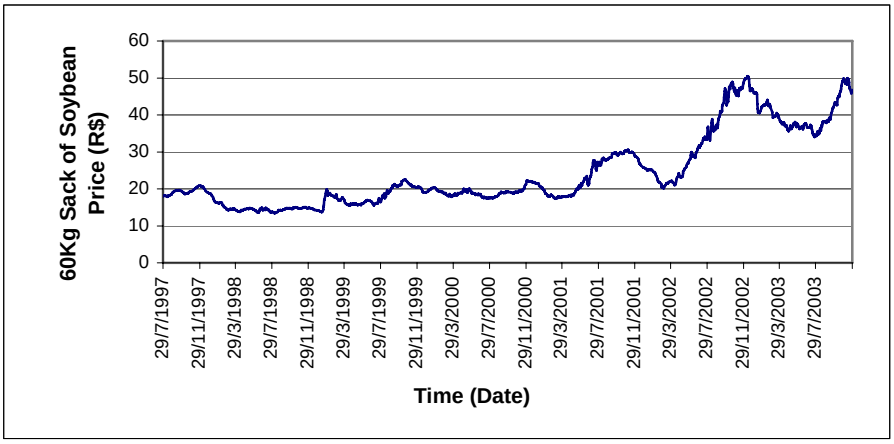

Figure-3 Series of values of the daily price of 60Kg sack of soybean. Source: CEPEA/ESALQ (R$ /sc 60 kg)

Figure 4 shows the logarithmic difference of the price of 60Kg soybean sack.

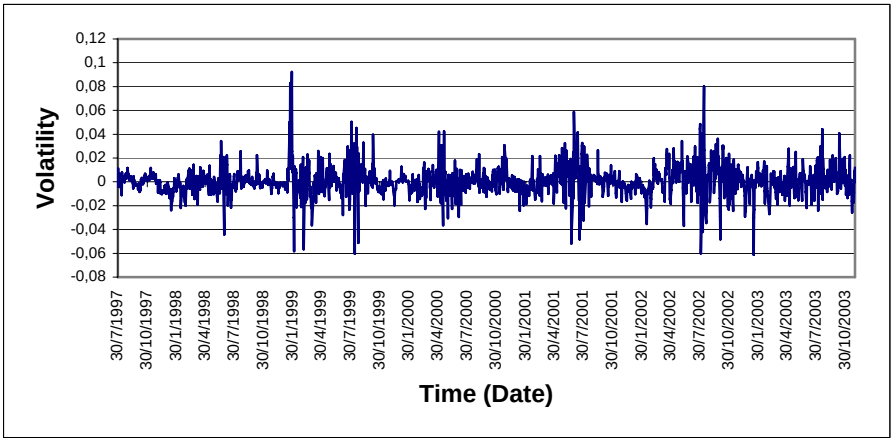

Figure-4 Series of daily returns of 60Kg sack of soybean.

Specifying an AR ([1]) model to represent the series of daily returns of the price of 60kg sack of soybeans, we found that this model is no correlation of the squared residuals.
The result of the McLeod-Li test for the series DLSOJA AR ([1]) for five lags:

n	Q	Sig.
1	25.15323	0.00000
2	155.0258	0.00000
3	161.9376	0.00000
4	173.8994	0.00000
5	205.2077	0.00000

Therefore, we reject the null hypothesis that is equivalent to accepting that the model is nonlinear. The estimation of parameters of an AR ([1])-GARCH (1,1) model for this series of daily returns of soybean RSOJA produces the following results:

$RSOJA_t = \phi_1 RSOJA_{t-1} + \varepsilon_t$, sendo $\varepsilon_t^2 \sim N(0; h_t)$,

$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1}$.

Parameters	Coefficient	Standard Error	Estat. z	Sig.
ϕ_1	0.309594	0.039205	7.897	0.0000
α_0	4.08E-06	1.56E-06	2.618	0.0088
α_1	0.261741	0.054845	4.772	0.0000
β	0.747039	0.041938	17.813	0.0000

Table-2 Parameters an model statistics AR([1])-GARCH(1,1).

A network NARX_RMPL (2-4-1) was sequentially trained as a predictor of one step ahead of the series h_t using 1500 samples. The training was stopped after 20,000 iterations. After training, the recurrent network was iterated and the outputs were compared using the test series the normalized root mean squared error (NRMSE):

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \tilde{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}},$$

where y_i is the current output, $\tilde{y}_i$ is the output of the model, $\bar{y}$ is the average of values y_i, (i = 1, 2, ..., N). The experiment was repeated 35 times with random re-initialization for each run. Table-3 shows the average of NRMSE of independents runs performed. The UKF parameters were chosen as $\alpha = 1$, $\beta = 0$ and $\kappa = 2$. These parameters are optimal for the scalar case [14].

Method	Mean (NRMSE)	Variance (NRMSE)
UKF	0.38765	0.0631
EKF	0.39658	0.0956
GARCH (1,1)	0.42127	0.0732

Tabela-3 Comparison of methods for predicting the heteroscedastic series (daily return of the sack of soybeans).

6. Conclusion

In this study, we found that a neural network trained with the unscented Kalman filter showed a better prediction than the RN trained with the EKF and the GARCH model. Although UKF does not require the calculation of the Jacobian, a limitation of the implementation of UKF is the need to choose the three unscented transformation parameters (α, β and κ). The optimal selection depends on the problem and still not fully understood.

References

[1] J. F. G. de Freitas, M. Niranjan, e A. H. Gee, “Hierarchical Bayesian-Kalman models for regularization and ARD in sequential learning”, Technical Report CUED/F-INFENG/TR 307, Cambridge University, 1997.

[2] S. Singhal, e L. Wu, “Training feedforward networks with the extended Kalman filter based pruning method for recurrent neural networks”, *Neural Computation* 10, 1481-1505.

[3] R. J. Williams, “Some observations on the use of the extended Kalman filter as recurrent network learning algorithm”, Technical Report NU_CCS_92-1. Boston: Northeastern University, College of Computer Science, 1992.

[4] B. Todorovic´, M. Stankovic´, S. Todorovic´-Zarkula, “Structurally adaptive RBF network in non-stationary time series prediction”, In *Proc. IEEE AS-SPCC*, Lake Luise, Alberta, Canada, Oct. 1-4 (2000) pp. 224-229.

[5] B. Todorovic´, M. Stankovic´, C. Moraga, “Extended Kalman Filter trained Recurrent Radial Basis Function Network in Nonlinear System Identification”, *Proc. Of ICANN 2002*, Spain, LNCS 2415, pp. 819-824, Springer, Agosto 2002.

[6] S. J. Julier, e J. K. Uhlmann, “A new extension of the Kalman filter to nonlinear systems. *Proceedings of AeroSense: The 11th international symposium on aerospace/defence sensing, simulation and controls*, Orlando, FL, 1997.

[7] S. J. Julier, e J. K. Uhlmann, “The Scaled Unscented Transformation”, *Proceedings of the IEEE American Control Conference*, 8-10 May, 2002.

[8] R. van der Merwe e E. A. Wan, “Efficient derivative-free Kalman filters for online learning”, *Proceedings of European Symposium on Artificial Neural Networks (ESSAN)*, Bruges, Belgium, Abril 2001.

[9] P. Ferrari e A. Galves,. Acoplamentos e processos estocásticos. IMPA, Rio de Janeiro, Brasil. http://www.ime.usp.br/ ~Pablo, 1997.

[10] A. McLeod, e W. Li, “Diagnostic checking of ARMA time series models using squared residuals autocorrelations”. *Journal of Time Series Analysis*, 4, 269–273, 1983.

[11] W. A. Brock, W. D. Dechert e J. A. Scheinkman, “A Test for Independence Based on the Correlation Dimension”. Work. pap., Department of Economics, University of Wisconsin at Madison, University of Houston, and University of Chicago, 1987.

[12] D. A. Hsieh, “Testing for nonlinear dependence in daily foreign exchange rates”, *Journal of Business*, 62(3), 339–368, 1989.

[13] T. Teräsvirta e C. Granger, “Modeling Nonlinear Economic Relationships”, Oxford University Press, 1993.

[14] S. Haykin, “Kalman Filtering and Neural Networks”, Wiley, 2001.

Received: April, 2012