Recursive Least Squares with ARCH Errors and Nonparametric Modelling of Environmental Time Series

Work supported by MURST'98 grant

Alessandro Fassò
University of Bergamo, Italy
fasso@unibg.it

Gruppo di Ricerca per le Applicazioni della Statistica ai Problemi Ambientali

Working paper n.6 - October, 2000

Abstract

After discussing some recent modeling approaches in environmental time series analysis, a new time-varying heteroscedastic model is introduced. It is based on exponentially weighted recursive least squares adjusted for estimated heteroscedasticity. Examples based on simulated data show some of its capabilities both for autoregressive conditional heteroscedasticity estimation and for robust recursive estimation in presence of smooth heteroscedasticity. An environmental application, related to ground ozone hourly data, shows that this method is capable of tracking heteroscedastic time varying environmental systems.

1 Introduction

In environmental time series analysis, a number of dynamical stochastic models have been used so far. In particular, for modelling high frequency air pollution and meteorological data, e.g. daily or hourly data, both parametric and nonparametric models of increasing complexity have been proposed.

For example, parametric conditionally autoregressive heteroscedastic models ($ARCH$) have been used by Tol (1996) for daily temperature modelling and by Graff-Jacottet & Jaunin (1998) for daily sulphur dioxide and ozone linear forecasting and by Fassò & Negri (2000a and b) for hourly ground ozone time series nonlinear modelling with long memory and heteroscedasticity.

Moreover, some nonparametric homoscedastic models have been considered in environmental literature. For example, Bordignon & Lisi (2000) used the nearest neighbor based smoothing for modelling hydrological time series. Cai & Tiwari (2000) considered BOD weekly data nonparametric modelling and, using local linear kernel smoothing over time, they found a time-varying homoscedastic autoregressive model description.

Nonparametric heteroscedastic models have been used in environmental regression analysis with independent data. For example in air pollution remote sensing, a semiparametric heteroscedastic kernel regression model has been proposed by Holst et al. (1996) and, recently, Lindström et al. (2000) used the approach of Ruppert et al. (1997) for local polynomial variance estimation of sulphur dioxide concentrations at volcano Etna, Sicily. In the Econometrics literature, local polynomial heteroscedastic time series models have been considered for example in Härdle et al. (1998) and Hafner (1998).

For framing the above models, it is important to distinguish between predictive and retrospective models. In this paper, we call predictive a time series model which uses only past values of the observed process y_t to compute both the forecasting function and the actual estimated forecast. This kind of model can be used in practice for online forecasting. Both in the parametric or nonparametric setup's, this usually leads to invariant causal models. In this case a model is fitted on an estimation data set and then used for forecasting as if it was the "true model" on new data.

Retrospective models use both past and future data to fit the forecasting function and/or the actual estimated forecasts. For example, Cai & Tiwari (2000) used smoothing over time models based on a symmetric kernel. Hence at each time t the process dynamics is explained using both future and past observations in a symmetric way.

As long as one is interested in time-varying or evolutionary predictive models, the parametric approach needs to be completed with some recursive estimation techniques. Recursive estimation and predictive modelling with homoscedastic innovations are inherently related to least squares (LS) and the Kalman filter. These methods have been recently discussed by Grillenzoni (1997a) and Xing-Qi Jiang (1999) with main emphasis on engineering applications. In environmental analysis, monitoring and forecasting, these methods may be useful for time-varying problems due to climatic and anthropic changes or for periodical moving of air pollution stations and the related re-start up.

In this paper, we consider recursive estimation of smoothly changing autoregressive models with heteroscedastic innovations. In particular in section 2, the recursive heteroscedastic algorithm is introduced and commented upon. In sections 3.1-3.3, some synthetic examples illustrate the techniques while, in section 3.4, an environmental case study based on tropospheric ozone data is presented. Finally, some conclusions and open problems are given in section 4.

2 RLS-ARCH Approach

In order to track and forecast a time-varying heteroscedastic model, we introduce the following $ARCH$ extension of the standard recursive least square approach (RLS).

Let the process of interest, y_t, be given by the following time-varying $ARX-ARCH$ equations

$$\begin{aligned} y_t &= \theta_t' \varphi_{t-1} + \varepsilon_t h_{t-1} \\ h_{t-1} &= \beta_{t-1}' \eta_{t-1}. \end{aligned} \tag{1}$$

In the first or mean equation, ε_t is and independent sequence with $E\varepsilon = 0$ and $E|\varepsilon| = 1$, φ_{t-1} is the regressor vector including both covariates and lagged process observations, namely $\varphi_t = (x_{t-1,1}, ..., x_{t-1,q}, y_{t-1}, ..., y_{t-p})'$. In the second or scedastic equation, we have $\eta_{t-1} = (1, |e_{t-1}|, \ldots, |e_{t-r}|)'$, and $e_t = y_t - \theta_t' \varphi_{t-1}$. The quantities θ_t and $\beta_t = (\beta_{0,t}, \beta_{1,t}, ..., \beta_{r,t})$ are unknown slowly time-varying coefficient vectors.

In this model, the time-varying conditional mean function $\hat{y}_t = \theta_t \varphi_{t-1}$ has time-varying precision given by its conditional mean absolute error h_t, which is a function of previous forecast errors.

When no heteroscedastic component is present, i.e. $h_t = h^*$ say, by assuming some stochastic linear dynamics for θ_t, e.g. θ_t is a random walk, it is possible to get the estimate $\hat{\theta}_t$ using the Kalman filter, see e.g. Xing-Qi Jiang (1999). This is optimal for the particular Gaussian model assumed. Alternatively, one can use exponentially weighted RLS tracking or robustified versions of it.

In our case, suppose for a moment that β_t is known. Hence the weighted RLS at time $t = 1, 2, ...$ is given by

$$\hat{\theta}_t | h_{s<t} = \arg\min \left(\sum_{s=1}^{t} (y_s - \theta' \varphi_{s-1})^2 \lambda_1^{t-s} \omega_{s-1} \right) \tag{2}$$

where

$$0 < \lambda_1 < 1, \quad \omega_{s-1} = \frac{1}{h_{s-1}^2}.$$

This is the solution of a weighted LS problem where the weights decrease both with time remoteness and size of previous forecast errors. The recursive solution of (2) can be handled by the Kalman filter algorithm as in section 6.3.1 of Mosca (1997). In particular, this is the state estimates of a state space representation where the state is $\theta_s = \theta$ and the observation equation is given by (1) with the scedastic equation replaced by $h_{s-1}^* = \sqrt{\lambda_1^{s-t} \omega_{s-1}^{-1}}$, say.

In practice, $\beta_{s<t}$ is unknown and we need some estimate for h_s. As in the empirical WLS for time-invariant $ARCH$ models, this can be done by using the following LS estimates

$$\hat{\beta}_t | \hat{\theta}_{s \leq t} = \arg\min \left(\sum_{s=1}^{t} (|e_s| - \beta' \eta_{s-1})^2 \lambda_2^{t-s} \right) \tag{3}$$

where

$$0 < \lambda_2 < 1.$$

Extending the standard Kalman filter techniques developed for homoscedastic RLS problems, estimates (2) and (3) have recursive formulations as follows.

2.1 $RLS - ARCH$ algorithm

Fix initial values for $R_0, S_0, h_0, \hat{y}_1, \varphi_0, \hat{\theta}_0$ and $\hat{\beta}_0$ then, for $t = 1, 2, ...,$ the recursive algorithm is as follows:

1. ARX computations

$$e_t = y_t - \hat{y}_t$$

$$\omega_{t-1} = \frac{1}{h_{t-1}^2}$$

$$R_t^{-1} = \frac{1}{\lambda_1}\left(R_{t-1}^{-1} - \omega_{t-1}\frac{R_{t-1}^{-1}\varphi_{t-1}\varphi_{t-1}'R_{t-1}^{-1}}{\lambda_1 + \omega_{t-1}\varphi_{t-1}'R_{t-1}^{-1}\varphi_{t-1}}\right)$$

$$\hat{\theta}_t = \hat{\theta}_{t-1} + R_t^{-1}\varphi_{t-1}e_t\omega_{t-1}$$

2. ARCH computations

$$\tilde{e_t} = y_t - \hat{\theta}_t'\varphi_{t-1}$$

$$\eta_{t-1} = (1, |\tilde{e}_{t-1}|, \ldots, |\tilde{e}_{t-r}|)'$$

$$S_t^{-1} = \frac{1}{\lambda_2}\left(S_{t-1}^{-1} - \frac{S_{t-1}^{-1}\eta_{t-1}\eta_{t-1}'S_{t-1}^{-1}}{\lambda_2 + \eta_{t-1}'S_{t-1}^{-1}\eta_{t-1}}\right)$$

$$\hat{\beta}_t = \hat{\beta}_{t-1} + S_t^{-1}\eta_{t-1}(|\tilde{e}_t| - \hat{\beta}_{t-1}'\eta_{t-1}).$$

Where the matrices R_t and S_t are estimates of the covariance matrices of $\hat{\theta}_t$ and $\hat{\beta}_t$ respectively. With these values the one step ahead forecast is simply given by

$$\hat{y}_{t+1} = \hat{\theta}_t'\varphi_t$$

and its precision can be evaluated by

$$\hat{h}_t = \hat{\beta}_t'\eta_t.$$

In particular, if ε_t in equation (1) is Gaussian distributed, then $E(\varepsilon_t^2) = \frac{\pi}{2}$ and the 95% approximated forecast interval is given by $\hat{y}_{t+1} \pm \sqrt{2\pi}\hat{h}_t$.

2.2 Robust Recursive Estimates

The $RLS-ARCH$ algorithm, has some connections to recursive M-estimates of Grillenzoni (1997) which are robust against innovation outliers. The innovation outlier scheme can be paraphrased by innovations with heavy tails or by some random mechanism which produces large innovations in an independent manner. For example we can use model (1) with the scedastic equation replaced by independent random variables as follows

$$h_t = \begin{cases} \sigma_0 & \text{with probability } 1-\varepsilon \\ k >> \sigma_0 & \text{with probability } \varepsilon \end{cases}$$

where σ_0 is the non outlier innovation standard deviation. In other situations, the scedastic function h_t varies smoothly in an unspecified way. In these cases nonparametric estimates of h_t may be obtained from the $RLS-ARCH$ algorithm with $r=0$. Hence we have $h_t=\beta_t$ and the $RLS-ARCH$ estimate of β_t is simply the exponentially weighted moving average $EWMA$ of $|e_t|$.

From this point of view, β_t may be considered both as a nuisance parameter for the (robust) estimation of θ_t or as an important parameter for predicting the precision of $\hat{y}_t$.

3 Case Studies

In order to illustrate the method proposed, we discuss both synthetic and real data cases. In particular in the first two synthetic examples the smoothing factors $\lambda_1 = \lambda_2 = \lambda = 0.99$ have been, the third one is an example of smooth heteroscedasticity with $\lambda_1 = 1 \neq \lambda_2 < 1$ and $\lambda = 0.995$ has been found appropriate in the last real data example. Although some estimation techniques have appeared in the literature (see e.g. Grillenzoni (1997*b*)), we have chosen these values by looking at the forecasting capability, given by the R^2 statistic, and at the periodic behavior of the data. A regular periodic path has been used for the first three simulation examples whilst the last real data example is characterized by a strong and irregular daily and seasonal pattern.

3.1 Simulation A

The first synthetic example is given by $n = 10,000$ observations from the slowly oscillating heteroscedastic $AR(1)$ system given by the following equations

$$\begin{aligned} y_t &= \theta_t y_{t-1} + e_t \\ \theta_t &= \cos\left(\frac{t}{n} 6\pi\right) \\ e_t &= \varepsilon_t h_{t-1} \\ h_{t-1} &= 1 + .2\cos\left(\frac{t}{n} 6\pi\right) \\ & \quad \varepsilon_t \; NID\,(0,1)\,. \end{aligned} \tag{4}$$

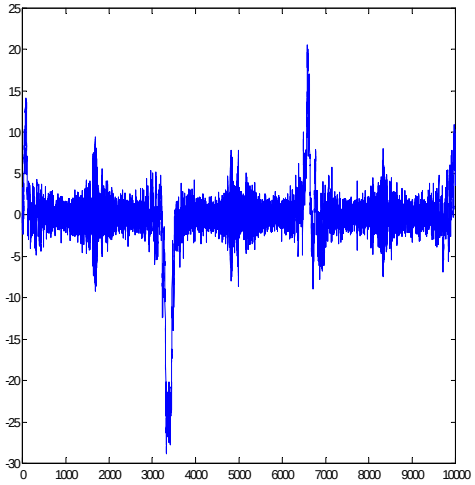

Figure 1: Simulation A. System observations

The corresponding highly nonlinear stochastic pattern is displayed in Figure 1. Moreover the resulting AR function θ_t and scedastic function h_t are given in Figures 2 and 4 respectively. Note that the scedastic function is deterministic and has been estimated by the $EWMA$ of $|e_t|$ reported in Figure 4 which results from the $RLS - ARCH$ algorithm of previous section with $p = 0$. Taking account of heteroscedasticity improves both parameter uncertainty and smoothness. This can be seen from Figure 3, where θ_t is estimated by the traditional homoscedastic RLS algorithm.

3.2 Simulation B

The second exercise is given by equations (4) with the scedastic equation replaced by the following time-varying $ARCH$ component:

$$h_{t-1} = 2 + \cos\left(\frac{t}{n}6\pi\right)^2 |\varepsilon_{t-1}h_{t-2}| .$$

The simulated data, residuals and studentized residuals, given by $\frac{e_t}{h_{t-1}}$, are reported in column 1 to 3 of Figure 5 respectively. In Figure 6, the β functions show that the estimation of the scedastic component is more erratic than $\hat{\theta}_t$ function and $\lambda_2 < 1$ prevents from consistent estimation of $\beta_{0,t} = 2$. As a matter of fact, the oscillations of the actual h_t are shared between $\hat{\beta}_{0,t}$ and $\hat{\beta}_{1,t}$ and the latter underscores the variations of the actual $\beta_{1,t}$. The estimate $\hat{\theta}_t$ is quite similar to Figure 2 of the previous case and has not been reported.

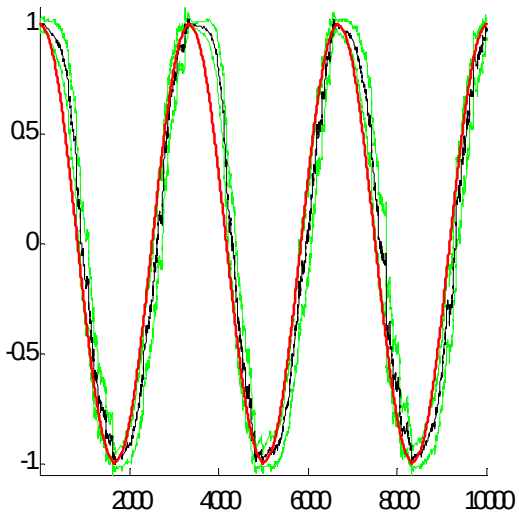

Figure 2: Simulation A. $\hat{\theta}_t$ with estimated 95% confidence intervals and actual θ_t.

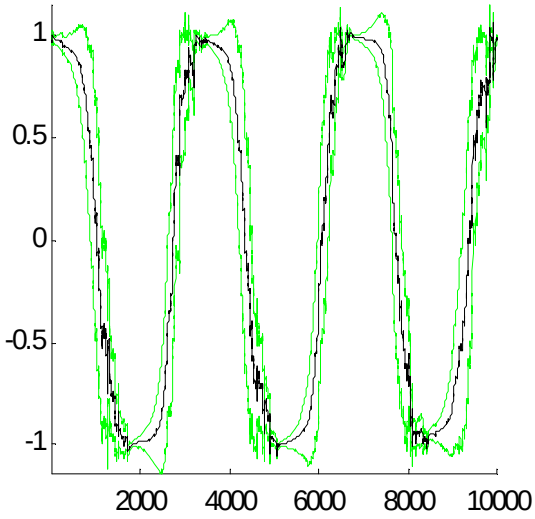

Figure 3: Simulation A. $\hat{\theta}_t$ and estimated 95% confidence intervals estimated without $ARCH$ component

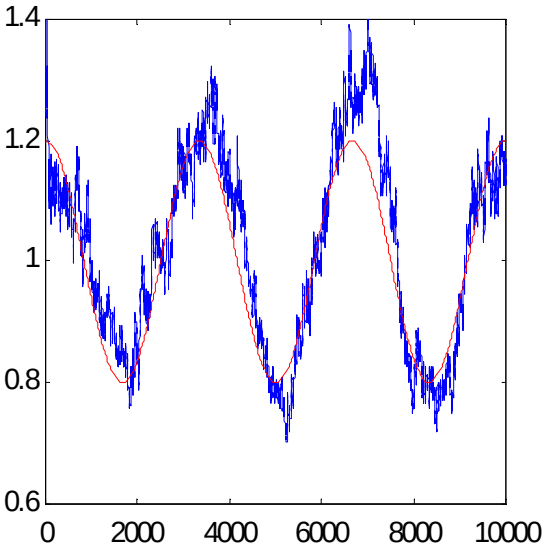

Figure 4: Simulation A. Observed volatility $\hat{h}_t\sqrt{\frac{\pi}{2}}$ and actual h_t.

3.3 Simulation C

The third synthetic example illustrates the use of $RLS-ARCH$ as a semi-parametric recursive estimator robust against smooth heteroscedasticity.

To see this, we consider the time-invariant predictor model given by equation (4) with $\theta_t = 0.7$ and scedastic function

$$h_{t-1} = 1 + 10\cos\left(\frac{t}{n}2\pi\right)^6.$$

We then use the $RLS-ARCH$ algorithm with $r=0$, $\lambda_1 = 1$ and $\lambda_2 = 0.95$. Figures 7 and 8 show the estimated path $\hat{\theta}_t$ for the usual homoscedastic RLS and the $RLS-ARCH$ respectively. It is clear that the high heteroscedasticity shown in Figure 9 has a reduced influence on the $RLS-ARCH$ estimator.

3.4 The Ozone Data

The third example is concerned with ground ozone hourly data collected in Bergamo, Italy, 1993-1997. This data set has already been discussed in details (see Fassò & Negri (2000a, b) and is now depicted in the first column of Figure 11 and summarized in Table 1.

	Mean	Variance	Skewness	Kurtosis
O_3	19,064	510,155	2,482	12,049
residuals	0,367	37,774	0,896	13,647
studentized residuals	0,037	1,287	0,344	4,339

Table 1. Ozone and residuals statistics.

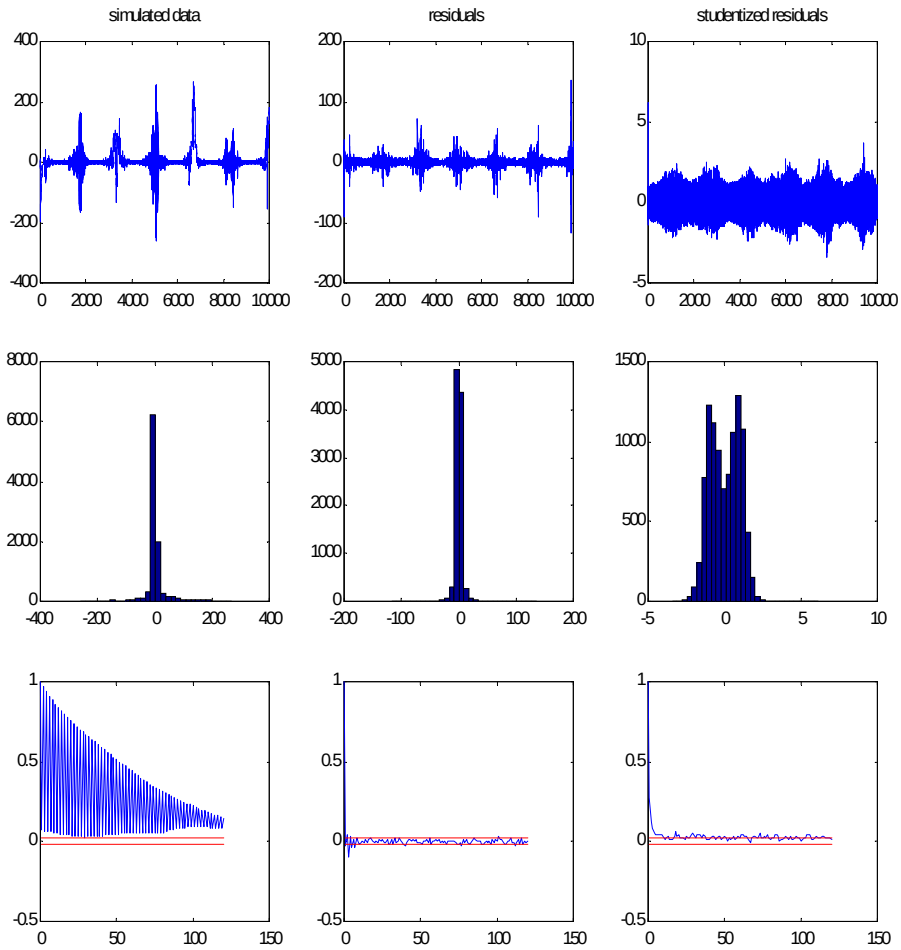

Figure 5: Simulation B. Time series, histogram and autocorrelation of simulated data, residuals and studentized residuals

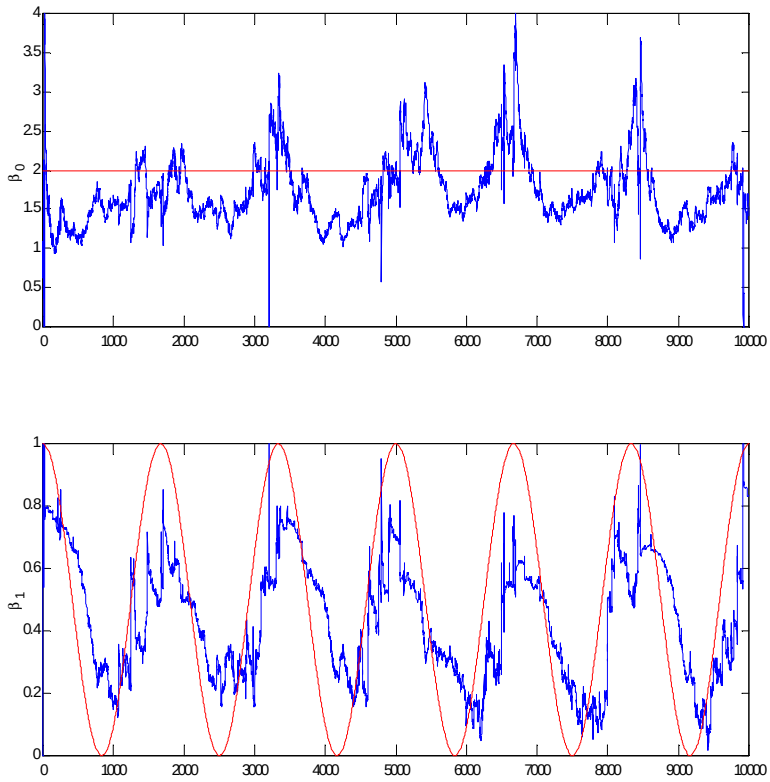

Figure 6: Simulation B. Estimated $ARCH$ functions with superimposed actual values

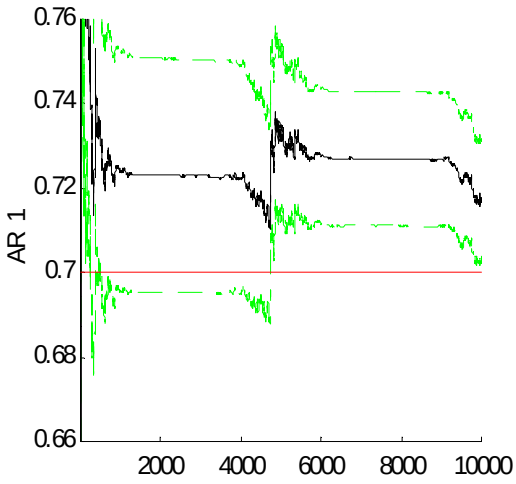

Figure 7: Simulation C. $\theta_t = 0.7$ estimated by homoscedastic RLS with $\lambda = 1$.

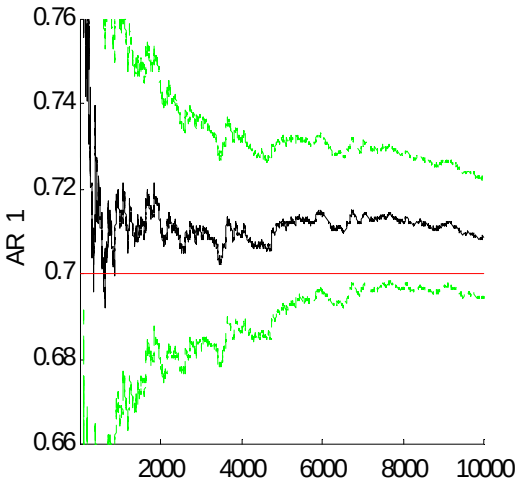

Figure 8: Simulation C. $\theta_t = 0.7$ estimated by robust $RLS - ARCH$ with $\lambda_1 = 1$ and $\lambda_2 = 0.95$.

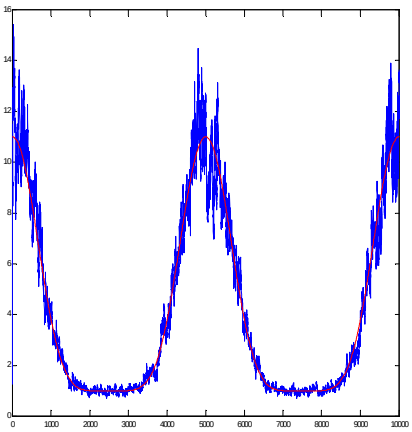

Figure 9: Simulation C. Estimated vs. actual scedastic function.

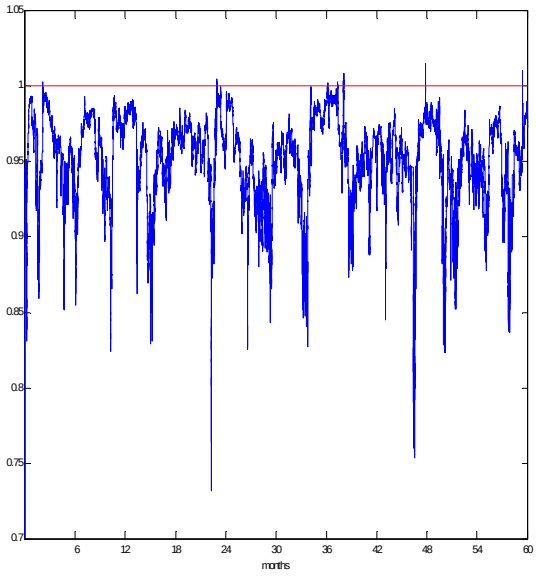

Figure 10: Ozone Data. Maximum absolute root function.

The model adopted here contains the four AR components of Figure 12 and the three covariates given by the lagged values of total solar radiation, relative humidity and nitrogen dioxide of Figure 13. These figures show a marked seasonal nonlinearity of the conditional mean. In particular, the seasonal variation of the θ coefficients is more apparent for the covariates then the AR components. Similarly the $ARCH$ components of Figure 14 indicate a strong seasonal nonlinear heteroscedasticity.

The ordinary and studentized residual graphical analysis is reported in column 2 and 3 of Figure 11. Although some daily periodicity is still present, the small residual autocorrelations and absolute studentized residual autocorrelations suggest that the model fitting is satisfactory. From top right of Figure 11, we see that this is generally true except for an outlier occurring in the studentized residuals during the highly polluted summer 1995. Moreover, in order to perform a stability analysis as discussed e.g. by Grillenzoni (1997a), we plotted the maximum of absolute roots of the AR polynomial at time t in Figure 10. This quantity is given by

$$\mu_t = \max |root(\alpha_t(z))|$$

where

$$\alpha_t(z) = 1 - z^{-1}\theta_{t,q+1} - \ldots - z^{-p}\theta_{t,q+p}$$

is the AR polynomial at time t. It follows that, being $\mu_t < 1$ except for a sparsely finite set, the system is stable. A similar diagnostic can be done on the $ARCH$ component by plotting $1 - \sum_{j=1}^{r} \beta_{j,t}$ as in the third line of Figure 14.

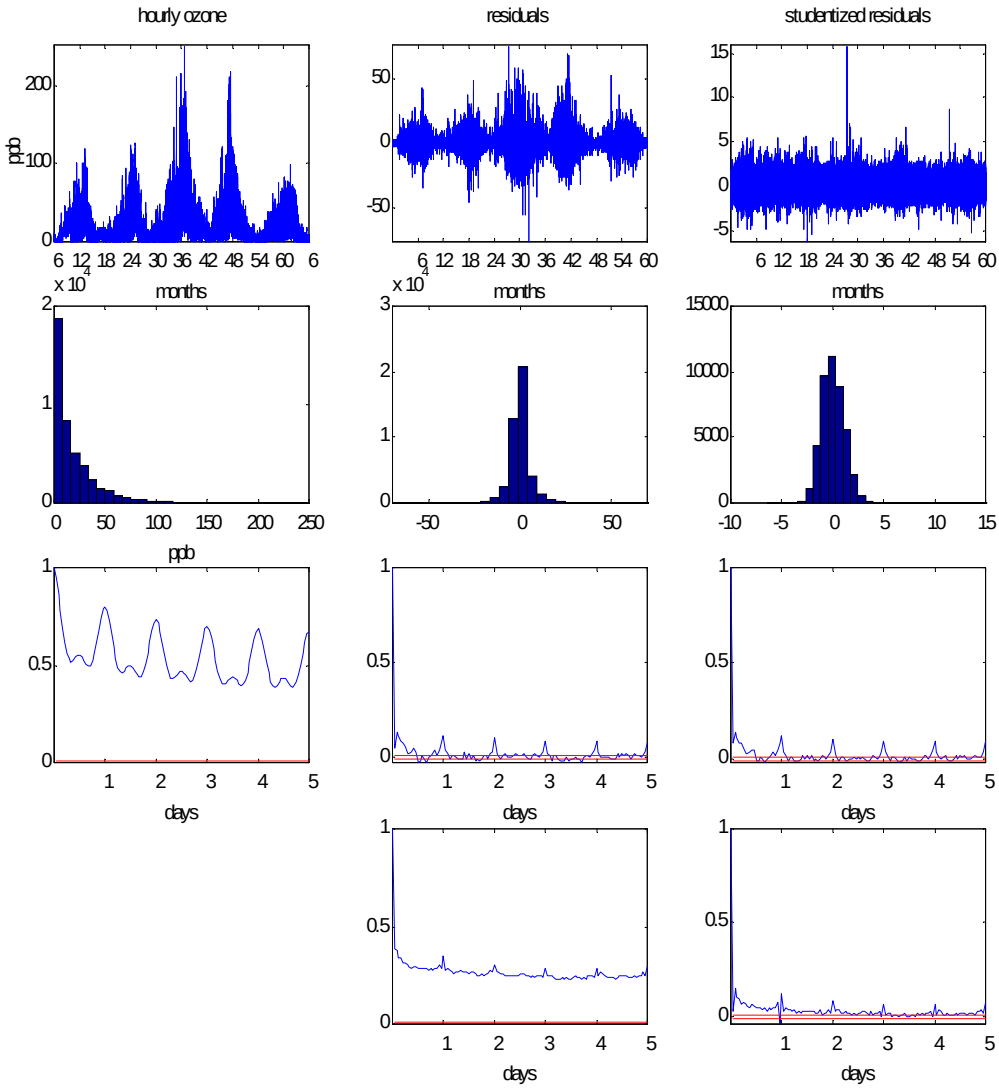

Figure 11: Ozone data. Time series, histograms, autocorrelations and autocorrelations of absolute values.

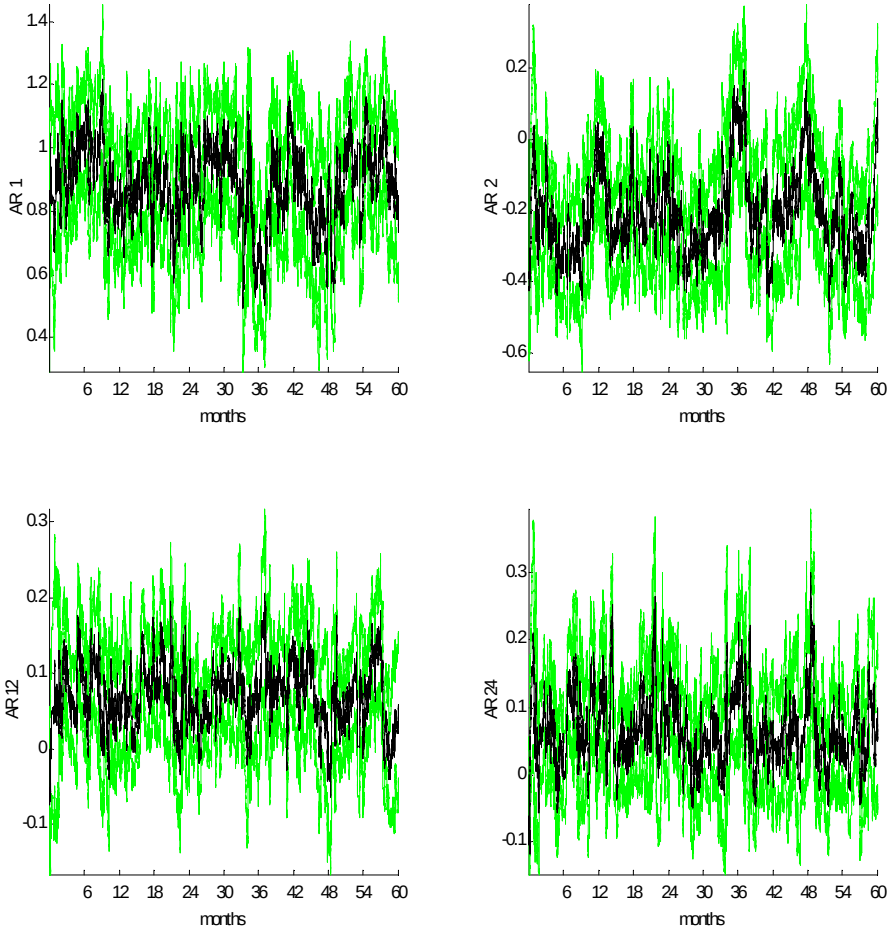

Figure 12: Ozone Data. θ_t function for AR component with approximated 95% intervals.

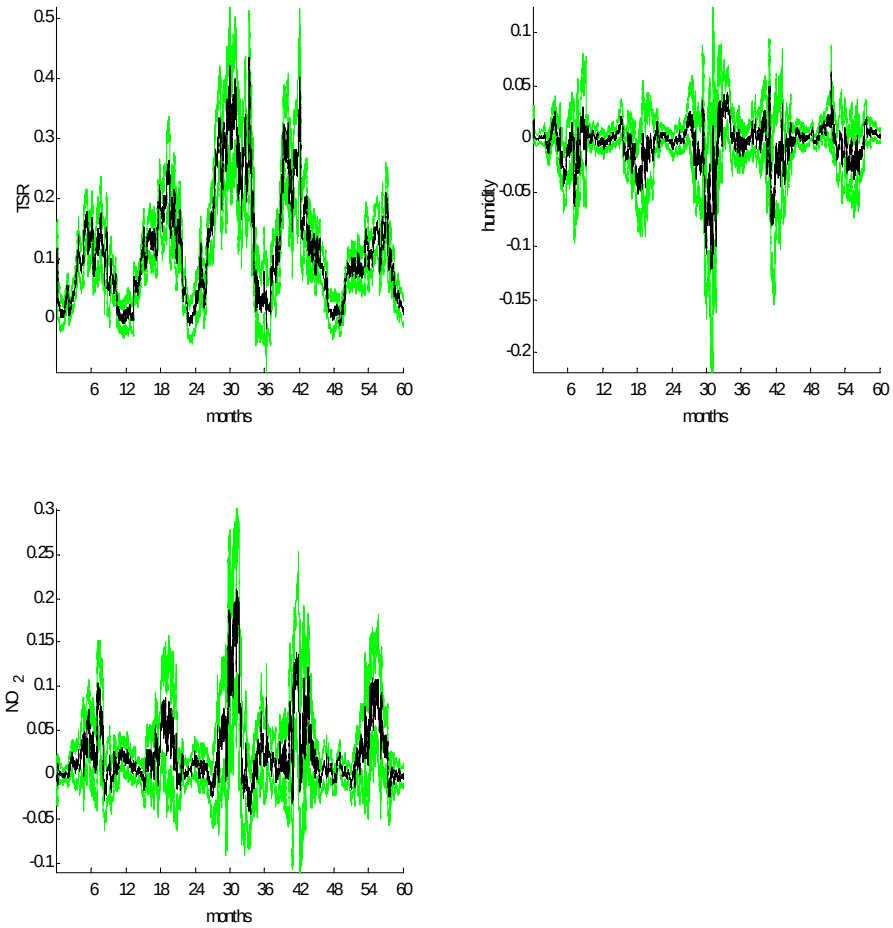

Figure 13: Ozone Data. θ_t functions for covariates with approximated 95% intervals

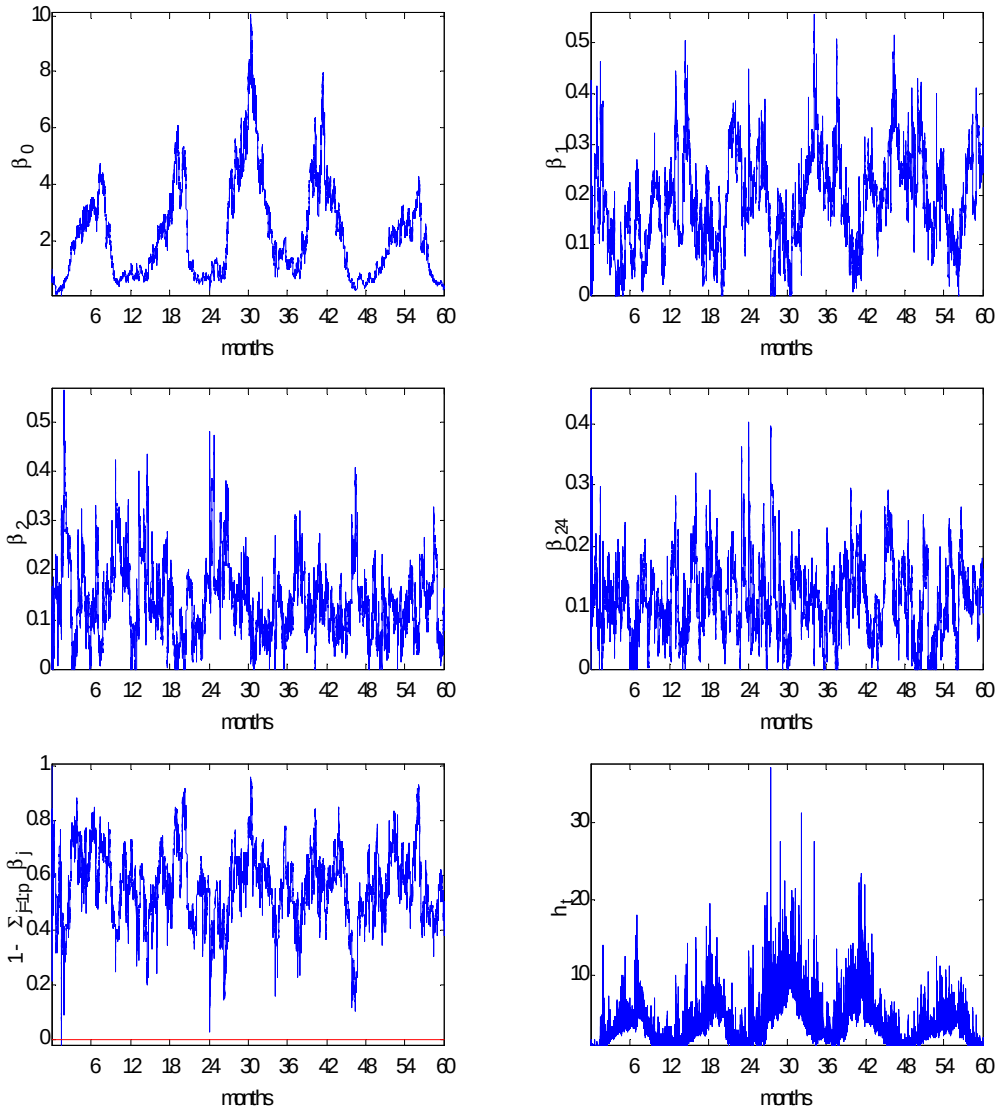

Figure 14: Ozone Data. β_t functions, Stability plot and estimated scedastic function.

4 Conclusions and Further Developments

In this paper, the problem of on-line prediction with time-varying heteroscedastic models has been considered using a time-varying scedastic function which may be of deterministic or $ARCH$ type. The model proposed seems adequate to track slowly time-varying heteroscedastic dynamical systems and has been applied to an air pollution high frequency monitoring problem.

The approach used here is essentially based on WLS and may be extended to fully cover model complexity estimation by means, for example, of time-varying AIC statistics.

Although our approach is related in some way to Kalman filtering, further extensions could lead to an appropriate heteroscedastic state space representation and connected nonlinear Kalman filter.

References

[1] Bordignon S., Lisi F. (2000) Nonlinear analysis and prediction of river flow time series, *Environmetrics*, **11**, 4, 463-477.

[2] Cai Z., Tiwari R. (2000) , Application of a local linear autoregressive model to BOD time series. *Environmetrics*, **11**, 3, 341-350.

[3] Fassò A., Negri I. (2000*a*) Nonlinear statistical modelling of high frequency ground ozone data. *Environmetrics* (to appear). GRASPA working paper (http://sirio.stat.unipd.it/graspa/working_papers.html).

[4] Fassò A., Negri I. (2000*b*) Multi step forecasting for nonlinear models of high frequency ground ozone data: a Monte Carlo approach. Submitted to *Quantitative Methods for Current Environmental Issues*, Anderson C., Barnett V., Chatwin, P., El Shaarawi A. ed.s., Springer.

[5] Graf-Jaccottet M., Jaunin M.H. (1998) Predictive models for ground ozone and nitrogen dioxide time series. *Environmetrics*, 9, 393-406.

[6] Grillenzoni (1997*a*) Optimized Adaptive Prediction. *J. Italian Statist. Soc.*, 6, 1, 37-58.

[7] Grillenzoni (1997*b*) Recursive Generalized M-Esitimators of Sustem Parameters. *Technometrics*. 39, 2, 211-224.

[8] Härdle W., Tsybakov A., Yang L. (1998) Nonparametric vector autoregression. *J. Statist. Planning and Inference*. 68, 221-245.

[9] Hafner C. (1998) Estimating high-frequency foreign exchange rate volatility with nonparametric $ARCH$ models. *J. Statist. Planning and Inference*. 68, 247-269.

[10] Holst U., Hössjer C., Björklund c., Ragnarson P., Edner H. (1996) Locally Weighted Least Squares Kernel Regression and Statistical Evaluation of LIDAR Measurements. *Environmetrics*, **7**, 4, 401-416.

[11] T. Lindstrom, U. Holst, P. Weibring and H. Edner (2000) Analysis of LIDAR measurements using nonparametric kernel regression methods , Draft 2000. (http://www.maths.lth.se/matstat/staff/torgny/).

[12] Ruppert D., Wand M.P., Holst U., Hössjer O. (1997) Local polynomial variance-function estimation. *Technometrics*, 39, 3, 262-273.

[13] Tol R.S.J.(1996) , Autoregressive conditional heteroscedasticity in daily temperature measurements, *Environmetrics*, 7, 67-76.

[14] Xing-Qi Jiang (1999) Time Varying AR and VAR Models. In "*The practice of time series*", Akaike & Kitagawa Ed.s. Springer-Verlag, New York.