ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Gaussian process regression for the estimation of stable univariate time-series processes

Georgios Birpoutsoukis[1] and Julien M. Hendrickx[1]

1- Université Catholique de Louvain - ICTEAM
Avenue Georges Lemaître 4-6/L4.05.01 1348 Louvain-la-Neuve - Belgium

Abstract. In this paper, estimation of AutoRegressive (AR) and AutoRegressive Moving Average (ARMA) models is proposed in a Bayesian framework using a Gaussian Process Regression (GPR) approach. Impulse response properties of the underlying process to be modeled are exploited during the parameter estimation. As such, models of enhanced predictability can be consistently obtained, even in the case of large model orders. It is also proved that the proposed approach is strongly linked with the Prediction Error (PE) model estimation approaches, if the estimated parameters are regularized. Simulations are provided to illustrate the efficiency of the proposed approach.

1 Introduction

In the field of time-series analysis, AR and ARMA models have been widely studied for the description of dynamic stochastic processes, with the majority of the applications to be related to prediction and forecasting [3]. Estimation of these models has been proposed based on different approaches (Maximum Likelihood estimation [7], PE approaches [5], Bayesian methods [8]).

In this work, we show that in a Bayesian framework, it is possible to estimate the model parameters of AR and ARMA processes when taking into account prior information related to the behavior of the impulse response of the inverse model. In this way, unlike in the typical PE method, models of enhanced predictability can be consistently obtained, even in the case of large model orders. Moreover, unlike other methods, we use prior information about the estimated model which originates from system theory. The method proposed is quite general for many processes under estimation and has been inspired by the impulse response estimation method for dynamic systems proposed in [9].

2 Problem statement

Based on data measured from a stochastic process, the objective is to describe its dynamics and predict the process evolution. It is assumed that the true process is exactly described by the discrete-time ARMA model structure:

$$\phi(q)\ y_t = \theta(q)\ e_t \Rightarrow (1 + \sum_{j=1}^{n_{\mathrm{AR}}} \phi_j q^{-j}) y_t = (1 + \sum_{i=1}^{n_{\mathrm{MA}}} \theta_i q^{-i}) e_t \ \Rightarrow\ y_t = H_{\mathrm{ARMA}}\ e_t \tag{1}$$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

where $y_t \in \mathbb{R}$ is the measured output at t, $e_t \in \mathbb{R}$ is the innovation at t, $n_{\text{AR}}, n_{\text{MA}} \in \mathbb{N}_0$ are the AR and MA orders, respectively, $\phi_j \in \mathbb{R}, j = 1, \ldots, n_{\text{AR}}$ and $\theta_i \in \mathbb{R}, i = 1, \ldots, n_{\text{MA}}$ are the coefficients of the AR ($\phi(q)$) and MA ($\theta(q)$) part, respectively, q is an operator such that $q^{-j} y_t = y_{t-j}$, and $H_{\text{ARMA}} = \frac{1+\sum_{i=1}^{n_{\text{MA}}} \theta_i q^{-i}}{1+\sum_{j=1}^{n_{\text{AR}}} \phi_j q^{-j}}$. In case $n_{\text{MA}} = 0$, the AR process is recovered. The innovation process e_t is assumed to be a realization of an identically and independently distributed (i.i.d.) Gaussian white noise process with variance σ_e^2. Moreover, both polynomials $\phi(q)$ and $\theta(q)$ have roots inside or on the unit disc, i.e. both linear filters H and H^{-1} are stable. The filters H and H^{-1} can be written as infinite series:

$$H = 1 + \sum_{t=1}^{\infty} h_t \; q^{-t}, H^{-1} = 1 + \sum_{t=1}^{\infty} h_t^{\text{inv}} \; q^{-t}$$

where h_t and $h_t^{\text{inv}}, t = 1, \ldots, \infty$ correspond to the impulse responses of H and H^{-1}, respectively. For stable linear filters, such as H and H^{-1}, the impulse response is decaying exponentially to zero. Since estimation of a finite order AR process is equivalent to the estimation of a finite impulse response, certain properties are a priori known for the AR coefficients and will be used during estimation. In the next section, AR process estimation is formulated as a GPR problem, and this framework is further extended to the ARMA process estimation case.

3 GPR for AR process estimation

Consider (1) with $n_{\text{MA}} = 0$, which corresponds to a univariate AR process. If N points of the process output $y_t, t = 1, \ldots, N$ are collected, then the AR process can be written as follows:

$$\begin{aligned} Y &= K\beta + E, \; Y = [y_{n_{\text{AR}}+1} \; \ldots \; y_N]^T, \; K = [\kappa(0) \; \kappa(1) \; \ldots \; \kappa(n_{\text{AR}} - 1)] \\ \kappa(r) &= [-y_{n_{\text{AR}}-r} \; - y_{n_{\text{AR}}+1-r} \; \cdots \; - y_{N-1-r}]^T, \; \beta = [\phi_1 \; \ldots \; \phi_{n_{\text{AR}}}]^T \\ E &= [e_{n_{\text{AR}}+1} \; \ldots \; e_N]^T \end{aligned}$$

where $Y \in \mathbb{R}^{N-n_{\text{AR}}}$ is the output vector, $K \in \mathbb{R}^{(N-n_{\text{AR}}) \times n_\beta}$ is the regressor, $E \in \mathbb{R}^{N-n_{\text{AR}}}$ is the innovation vector and $\beta \in \mathbb{R}^{n_\beta}$ contains the parameters of the model (in this case $n_\beta = n_{\text{AR}}$). The objective is to estimate β using the measured output data. We will adopt a Bayesian framework in order to include prior knowledge in the estimation step. Consider the AR process:

$$y_t = H_{\text{AR}} \; e_t, \quad H_{\text{AR}} = \frac{1}{1 + \sum_{j=1}^{n_{\text{AR}}} \phi_j q^{-j}}, \quad H_{\text{AR}}^{-1} = 1 + \sum_{j=1}^{n_{\text{AR}}} \phi_j q^{-j}$$

The first prior used during estimation is related to the fact that, since H_{AR} and H_{AR}^{-1} are stable, the impulse responses of both filters will be decaying exponentially to zero. In case of AR process, the parameters to estimate $\phi_j \in \mathbb{R}, j =$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

$1, \ldots, n_{\mathrm{AR}}$ correspond in fact to the impulse response coefficients of the filter H_{AR}^{-1}. The second prior that will used is related to the fact that the impulse response coefficients of linear systems exhibit often a certain level of correlation between them. Practically speaking, the coefficients that are closer to each other in time are more correlated than the ones further away. These two properties of stability and correlation can be incorporated into the modeling procedure in order to obtain parameter estimates of increased accuracy, and further models of enhanced predictability.

3.1 Parameter estimation and prior covariance

Towards this direction, we assume that the parameter vector β in (3) is a zero mean Gaussian random variable with covariance $\Sigma_\beta \in \mathbb{R}^{n_\beta \times n_\beta}$, i.e. $\beta \sim \mathcal{N}(0, \Sigma_\beta)$. Assuming that the elements in K are known and given that $Y = K\beta + E$, $E \sim \mathcal{N}(0, \sigma_e^2 I_N)$, I_x denoting the identity matrix of size x, then β and Y are jointly Gaussian variables [6]:

$$\begin{bmatrix} \beta \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_\beta & \Sigma_\beta K^T \\ K\Sigma_\beta & K\Sigma_\beta K^T + \sigma_e^2 I_{(N-n_{\mathrm{AR}})} \end{bmatrix} \right)$$

The posterior distribution of β conditioned on the process data Y is given by $\beta | Y \sim \mathcal{N}(\hat{\beta}^{apost}, \Sigma_\beta^{apost})$ where:

$$\hat{\beta}^{apost} = (K^T K + \sigma_e^2 \Sigma_\beta^{-1})^{-1} K^T Y, \quad \Sigma_\beta^{apost} = ((\sigma_e^2 \ (K^T K)^{-1})^{-1} + \Sigma_\beta^{-1})^{-1} \quad (2)$$

Different types of structures can be used for the prior covariance matrix Σ_β of the parameter vector β, in order to impose exponential decay (stability) and correlation. Among other choices proposed in the literature [9], a typical structure towards this direction is the Diagonal Correlated structure (DC) [2] given by:

$$\Sigma_\beta(i,j) = e^{-\alpha_\rho(|i-j|)} e^{-\alpha_\lambda(i+j)}$$

The hyperparameters $\alpha_\rho > 0$, $\alpha_\lambda > 0$ will be learned from data and will define the prior knowledge on the process parameters. The value of α_ρ determines the level of correlation between the impulse coefficients and the value of α_λ determines the rate of decay of the impulse response to zero. These values, together with the innovation variance σ_e^2, can be learned by means of the following method, also known as Empirical Bayes: Given the prior covariance Σ_β, the pdf f_Y of the observed output vector is given by

$$f_Y(\Sigma_\beta) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma_Y)}} e^{-\frac{1}{2} Y_N^T \Sigma_Y^{-1} Y_N}$$

with $\Sigma_y = K\Sigma_\beta K^T + \sigma_e^2 I_{(N-n_{\mathrm{AR}})}$. The optimal values of $\alpha_\rho, \alpha_\lambda, \sigma_e$ are the ones that maximize the marginal likelihood of the observed output, i.e.:

$$\hat{\alpha}_\rho, \hat{\alpha}_\lambda, \hat{\sigma}_e = \arg \min_{\alpha_\rho, \alpha_\lambda, \sigma_e} -2 \log(f_Y) = \arg \min_{\alpha_\rho, \alpha_\lambda, \sigma_e} Y_N^T \Sigma_Y^{-1} Y_N + \mathrm{logdet}(\Sigma_Y)$$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Optimizing the marginal likelihood of the observed output is a non-convex optimization problem, however it can be efficiently solved as shown already in [1]. It should be noted that other hyper-parameter tuning methods are also available in the literature, such as the residual analysis approach (optimal hyper-parameters render the estimated residuals white and independent [5], [4]) and the cross-validation technique (splitting the data set into three parts to use for parameter estimation, hyper-parameter estimation and model validation, respectively [10]).

3.2 The link to the prediction error approach

There exists a link between the GPR method and the classical prediction error (PE) framework used for time-series estimation. For an AR process described by $y_t = H_{\text{AR}}\, e_t$, it can be shown that the best one-step ahead predictor is given by $\hat{y}_{t/t-1} = (1 - H_{\text{AR}}^{-1})\; y_t = (\sum_{j=1}^{n_{\text{AR}}} \phi_j q^{-j})\; y_t$ [5]. Consider the following PE criterion for parameter estimation:

$$\hat{\beta}_{\text{PE}} = \arg\min_{\beta} \sum_{t=1}^{N} (y_t - \hat{y}_{t/t-1})^2 = \arg\min_{\beta} \|Y - K\beta\|^2 = (K^T K)^{-1} K^T Y \quad (3)$$

The least squares (LS) problem in (3) can often be ill-conditioned when the ratio N/n_β is not sufficiently large and it does not also provide a unique optimal solution when $N/n_\beta < 1$. However it is possible to overcome these issues by considering the regularized LS criterion:

$$\hat{\beta}_{\text{reg}} = \arg\min_{\beta} \|Y - K\beta\|^2 + \beta^T D\beta = (K^T K + D)^{-1} K^T Y \quad (4)$$

It is easy to see that $\hat{\beta}^{\text{reg}} = \hat{\beta}^{\text{apost}}$ in (2), if $D = \sigma_e^2\; \Sigma_\beta^{-1}$. As such, the GPR method can be seen as minimizing a regularized PE criterion in order not only to get a lower prediction error, but also to constrain the parameter vector β according to the quadratic penalty $\beta^T D\beta$, with the prior knowledge about β contained in $D = \sigma_e^2\; \Sigma_\beta^{-1}$.

4 Extension to the ARMA process estimation

First, recall the well-known approach where, based on N measured data y_t, a sufficiently long AR model of the ARMA process (1) is initially estimated. This model is used to estimate the unknown innovation values and obtain $\hat{e}_1, \ldots, \hat{e}_N$. Using these estimates, the ARMA process can be written as:

$$Y = [K_{\text{AR}} \;\; K_{\text{MA}}]\; \beta + E,\; \beta = \begin{bmatrix} \beta_{\text{AR}} \\ \beta_{\text{MA}} \end{bmatrix}$$

where $Y, E, K_{\text{AR}}, \beta_{\text{AR}}$ are constructed as in the AR case while K_{MA} is an appropriately constructed matrix, similar to K_{AR}, but containing values of $\hat{e}_t$. In that case the GPR approach can be applied by considering $\beta \sim \mathcal{N}(0, \Sigma_\beta)$ with:

$$\Sigma_\beta = \begin{bmatrix} \Sigma_{\beta_{\text{AR}}} & 0 \\ 0 & \Sigma_{\beta_{\text{MA}}} \end{bmatrix}$$

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

As such it is proposed for the ARMA case to use first the GPR method described in Section 3 for the estimation of a long AR process, obtain the estimates the unknown innovations as $\hat{e}_t = \hat{H}_{\text{AR}}^{-1}\, y_t$, and then use them to estimate the ARMA parameters with the GPR method described in the current section. Again, as in the AR estimation case, it is easy to show that the solution of the GPR method will lead numerically to the solution of the regularized PE problem.

5 Numerical illustration

The GPR method is first illustrated on the estimation of the 7^{th} order ARMA process (1) with $\theta(q) = [0.8374\ 0.5237\ 0.2765\ 0.1312\ 0.0572\ 0.0230\ 0.0085]^T$ and $\phi(q) = [0.2355\ -0.2291\ -0.1199\ 0.0434\ 0.0454\ -0.0034\ -0.0144]^T$. A sufficiently long AR model (15 lags) is first estimated using the GPR method described in Section 3 and the unknown innovation vector $e_1, e_2, \ldots, e_N$ is estimated. Then, the ARMA process is estimated using the GPR method described in Section 4. This method is compared to the PE method (3), both for the long AR model step and for the ARMA process. Comparison of the two methods is done through 50 Monte Carlo (MC) simulations where, at each iteration a new realization of a Gaussian process with variance $\sigma_e^2 = 1$ and length 1000 samples is used to generate the ARMA output data, and further estimate the model parameters using these data. Different values of the data length N are considered in order to observe the evolution of the performance for both methods. To evaluate

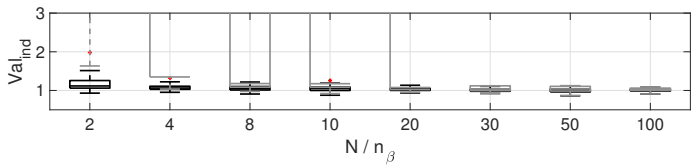

Fig. 1: Monte Carlo simulation results for the comparison of the GPR with the PE method. Black: Performance of models using GPR. Grey: Performance of models using PE.

the performance of the models, a new validation dataset is considered of length $N_{\text{val}} = 1000$ samples, of data not used during model estimation. The model performance is evaluated with the index $\text{Val}_{\text{ind}} = \frac{1}{N_{\text{val}}} \sum_1^{N_{\text{val}}} (y_t - \hat{y}_{t/t-1})^2$. The results for the 7^{th} order ARMA process are depicted in Fig.1. It is clear that the GPR method outperforms the PE method in terms of predictability, measured using Val_{ind}. It should be noted that a number of models obtained with the PE method are not in the figure due to their very bad performance and the very high value of Val_{ind}.

The GPR method is expected to perform even better than the PE method the higher the ARMA process order. To illustrate this, we perform another simulation with 100 random ARMA processes of orders $6, 7$. For each order and random process, 50 MC simulations are executed in the same way as for the first simulation example above. Each time, the ARMA parameters are estimated once

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Table 1: Global performance measured in terms of median values of $\mathrm{Val}_{\mathrm{ind}}$. The first value corresponds to the PE and the second to the GPR method. The percentage of error decrease is reported in the parenthesis.

N/n_β	4	5	10
Order 6	1.45 / 1.13 (-22%)	1.28 / 1.08 (-15.6%)	1.1 / 1.06 (-3.6%)
Order 7	2.28 / 1.13 (-50.4%)	1.47 / 1.08 (-26.5%)	1.13 / 1.05 (-7%)

using the GPR method and once with the PE method (3). To evaluate the model performance, a validation dataset is again considered of length 10000 samples and the index $\mathrm{Val}_{\mathrm{ind}}$ is calculated. The global behavior of the two methods is compared through the median value of $\mathrm{Val}_{\mathrm{ind}}$ among the 50 MC simulations and in turn the median value among the 100 random processes of same order and N/n_β. It should be noted that the innovation variance has also been set here to 1. The results are shown in Table 1 where it is clear that the benefit of the GPR method of the PE classical approach increases with the order of the ARMA process considered (percentage value in the table).

6 Conclusions

In this paper, a GPR method is used for the estimation of AR and ARMA processes. It is proposed to use prior knowledge about the impulse response of the inverse model in order to increase the efficiency of the PE methods. Numerical simulations are provided to illustrate this fact. Currently the work is getting extended to the estimation of multivariate AR and ARMA processes.

References

[1] Chen, T., & Ljung, L. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49(7):2213–2220, 2013.

[2] Chen, T., Ohlsson, H., & Ljung, L. On the estimation of transfer functions, regularizations and Gaussian processes - revisited. *Automatica*, 48(8):1525–1535, 2012.

[3] De Gooijer, J. G., & Hyndman, Rob J. 25 years of time series forecasting. *International journal of forecasting*, 22(3):443–473, 2006.

[4] Ljung, L. *Model validation and model error modeling.* Linköping University Electronic Press, 1999.

[5] Ljung, L. System Identification: Theory for the User, PTR Prentice Hall Information and System Sciences Series, 1999.

[6] Ljung, L., & Chen, T. What can regularization offer for estimation of dynamical systems? In *16th IFAC International Workshop on Adaptation and Learning in Control and Signal Processing*, Caen, France, 2013.

[7] Mauricio, J. A. Exact maximum likelihood estimation of stationary vector ARMA models. *Journal of the American Statistical Association*, 90(429):282–291, 1995.

[8] Monahan, J. F. Fully Bayesian analysis of ARMA time series models. *Journal of Econometrics*, 21(3):307–331, 1983.

[9] Pillonetto, G., & De Nicolao, G. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

[10] Pintelon, R., & Schoukens, J. *System identification: a frequency domain approach.* John Wiley & Sons, 2012.