30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

Bayesian model order selection for nonlinear system function expansions

Georgios D. Mitsis and Saad Jbabdi

Abstract— **Orthonormal function expansions have been used extensively in the context of linear and nonlinear systems identification, since they result in a significant reduction in the number of required free parameters. In particular, Laguerre basis expansions of Volterra kernels have been used successfully for physiological systems identification, due to the exponential decaying characteristics of the Laguerre orthonormal basis and the inherent nonlinearities that characterize such systems. A critical aspect of the Laguerre expansion technique is the selection of the model structural parameters, i.e., polynomial model order, number of Laguerre functions in the expansion and value of the Laguerre parameter α, which determines the rate of exponential decay. This selection is typically made by trial-and-error procedures on the basis of the model prediction error. In the present paper, we formulate the Laguerre expansion technique in a Bayesian framework and derive analytically the posterior distribution of the α parameter, as well as the model evidence, in order to infer on the expansion structural parameters. We also demonstrate the performance of the proposed method by simulated examples and compare it to alternative statistical criteria for model order selection.**

I. INTRODUCTION

The study of many nonlinear physiological systems has been pursued in the context of Volterra-Wiener models, which apply to a very broad class of systems and are well-suited to the inherent nonlinearities and the complexity of physiological systems [1], [2]. The Volterra-Wiener framework yields a rigorous description of the system nonlinearities in the form of a hierarchy of Volterra or Wiener kernel functions, which are valid over the entire bandwidth and dynamic range of system operation. Among several approaches suggested for the estimation of these kernel functions from input-output data [1], [2], [3], [4], an approach that results in a significant reduction in the number of free parameters is to use function expansions in terms of an orthonormal basis, first suggested by Wiener [5]. For example, the Laguerre [6] and Kautz [7] bases have been employed for linear systems identification. In the case of nonlinear systems, expansion of the Volterra kernels in terms of the Laguerre basis has also proven successful [8], [9], [10]. An efficient way to estimate the Laguerre expansion coefficients utilizing least-squares estimation in connection with discrete-time Laguerre expansions [9] has been shown to be efficient, e.g., in several applications to physiological systems [2]. The combination of Laguerre expansions with feedforward networks with polynomial activation functions was proposed in [10].

The performance of the Laguerre expansion technique has been shown to be excellent, as long as the Laguerre parameter α and the number of basis functions are selected properly. The rate of convergence for truncated Laguerre series depends on the selection of α - it was suggested that the latter should be selected according to the dominant time constant of the system [6]. In many applications to date of Laguerre expansions, the selection of the model structural parameters, i.e., the value of α, the number of Laguerre functions L and model order Q in the case of nonlinear systems, has been empirical (based on trial-and-error procedures). These procedures are typically based on least-squares based cost functions and may also require prior assumptions about some of the system characteristics (e.g., system memory) or the use of out-of-sample data, which poses limitations for short input/output data sets. However, some theoretical work has been published on the determination of the optimal pole position of Laguerre filters for linear systems [11], [12]. In the case of nonlinear systems, a closed-form optimal value for the Laguerre poles of the expansion of a given set of Volterra kernels was derived in [13]. Finally, an efficient computational approach was introduced in order to train the Laguerre parameter α by gradient descent on the basis of the input-output data [10].

In the present paper, we formulate the Laguerre expansion technique for nonlinear systems in the Bayesian framework and we analytically calculate the posterior distribution of α, as well as the Bayesian model evidence, allowing us to infer on the values of α, L and Q on the basis of the input-output observations. We present the performance of this approach in various simulation scenarios.

II. METHODS

A. Laguerre expansion of Volterra kernels

The general Volterra model for a Q-th order nonlinear system is given below in discrete time [2]:

$$\begin{aligned} y(n) &= \sum_{q=0}^{Q} \sum_{m_1,\ldots,m_q} k_q(m_1,\ldots,m_q)x(n-m_1)\ldots x(n-m_q) \\ &= k_0 + \sum k_1(m)x(n-m) \\ &\quad + \sum_{m_1}\sum_{m_2}^{m} k_2(m_1,m_2)x(n-m_1)x(n-m_2) + \ldots \end{aligned} \quad , \tag{1}$$

Manuscript received April 16, 2008. This work was supported in part by the European Social Fund (75%) and National Resources (25%) - Operational Program Competitiveness - General Secretariat for Research and Development (Program ENTER 04 - GDM) and by the Dr Hadwen Trust for Humane Research (SJ).

G.D. Mitsis is with the Institute of Communication and Computer Systems, Dept. of Electrical and Computer Engineering, National Technical University of Athens, Zografou 15780, Greece (e-mail: gmitsis@esd.ntua.gr)

S. Jbabdi is with the Centre for Functional Magnetic Resonance Imaging of the Brain, University of Oxford, Oxford OX3 9DU, United Kingdom (e-mail: saad@fmrib.ox.ac.uk)

where $x(n)$ and $y(n)$ are the system input and output respectively and k_q denotes the q-th order Volterra kernel of the system. The Volterra kernels describe the linear ($q = 1$) and nonlinear ($q > 1$) dynamic effects of the input on the output. The sum of eq. (1) can be viewed as a generalization of the convolution sum, with the Volterra kernels quantifying the effect of past input values (linear kernel), as well as their q-th order products (nonlinear kernels) on the output at present time n. For causal and finite memory systems the sums in (1) are defined from $m_i = 0$ to M, where M is the system memory.

An efficient method to obtain estimates of the Volterra kernels k_q using input-output observations is to utilize function expansions of the kernels in terms of the discrete-time Laguerre orthonormal basis [9]:

$$k_q(m_1, \ldots, m_q) = \sum_{j_1=0}^{L} \cdots \sum_{j_q=j_{q-1}}^{L} c_{j_1,\ldots,j_q} b_{j_1}(m_1) \ldots b_{j_q}(m_q). \tag{2}$$

where $b_j(m)$ is the j-th order discrete-time Laguerre function given by:

$$b_j(m) = \alpha^{(m-j)/2}(1-\alpha)^{1/2} \sum_{k=0}^{j} (-1)^k \binom{m}{k} \binom{j}{k} \alpha^{j-k}(1-\alpha)^k \tag{3}$$

The Laguerre parameter α ($0 < \alpha < 1$) determines the exponential decay of the Laguerre functions and is critical for the efficiency and parsimony of the expansion. For example, using large α values leads to more efficient representations for systems with slow dynamics and/or large memory M.

By combining equations (1) and (2) and using matrix notation, the output can be expressed in terms of the expansion coefficients as:

$$\mathbf{y} = \mathbf{V}\mathbf{c} + \epsilon, \tag{4}$$

where $\mathbf{y}$ is the $(N \times 1)$ vector of output observations, $\mathbf{V}$ is a $(N \times d)$ matrix containing the convolution of the input with the Laguerre functions $v_j = x * b_j$, as well as all possible higher-order products $v_{j_1} v_{j_2} \ldots v_{j_Q}$ for $j_1, j_2, \ldots, j_Q = 0, ..., L$, and $\mathbf{c}$ is the $(d \times 1)$ vector of the unknown expansion coefficients. Note that the number of free parameters d is equal to $L + 1$ for $Q = 1$ and $L(L + 1)/2$ for $Q = 2$, due to the symmetry with respect to $j_1, j_2, \ldots, j_Q$.

The least-squares estimate of $\mathbf{c}$ is given by:

$$\hat{\mathbf{c}} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{y}. \tag{5}$$

B. Bayesian inference for Laguerre function expansions

Bayesian inference is based on fitting a probabilistic model to a set of data, and summarising the results in terms of probabilities [14]. The main steps of the Bayesian approach are: (i) setting up a full probabilistic data generative model, (ii) conditioning on observed data, and (iii) evaluating the performance of the model. The fundamental Bayes equation relates the full posterior distribution of the model parameters $p(\Omega|\mathbf{y})$ given the observed data to the likelihood function $p(\mathbf{y}|\Omega)$ and the prior parameter PDF $p(\Omega)$ as:

$$p(\Omega|\mathbf{y}) = \frac{p(\mathbf{y}|\Omega)p(\Omega)}{p(\mathbf{y})}. \tag{6}$$

Often, we are interested in considering only one or some of the model parameters in isolation. For this, the corresponding marginal posterior distribution, which accounts for the uncertainty in the remaining parameters, needs to be calculated. The marginal posterior for a set of parameters ω, given the data $\mathbf{y}$ and the (complementary) set of the remaining parameters Ω^- is given by:

$$\begin{aligned} p(\omega|\mathbf{y}) &= \int p(\Omega|\mathbf{y}) d\Omega^- \\ &\propto \int p(\mathbf{y}|\Omega)p(\Omega) d\Omega^-. \end{aligned} \tag{7}$$

The normalising factor $p(\mathbf{y})$ in equation (6) is of particular importance. This quantity is termed model evidence and is equal to:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\Omega)p(\Omega) d\Omega. \tag{8}$$

Model evidence takes into account both model accuracy and complexity, serving as a basis for Bayesian model comparison. We often write $p(\mathbf{y}|\mathcal{M})$, as the evidence of the model $\mathcal{M}$, given the observed data $\mathbf{y}$, and use this quantity to select model structure.

In the present case, we will infer on the expansion coefficients $\mathbf{c}$, the Laguerre parameter α and the noise variance σ^2, therefore our concatenated parameter vector is $\Omega = (\mathbf{c}, \alpha, \sigma^2)$. In the case of white Gaussian noise (i.e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$), we can write the likelihood function as:

$$p(\mathbf{y}|\mathbf{c}, \sigma^2, \alpha) \sim \mathcal{N}(\mathbf{y}|\mathbf{V}(\alpha)\mathbf{c}, \sigma^2\mathbf{I}). \tag{9}$$

Note that the dependence of the likelihood on α is incorporated in the expression for $\mathbf{V}$. We considered non-informative priors for the expansion coefficients $\mathbf{c}$ and the Laguerre parameter α, and an inverse gamma prior with shape a and scale $b(a, b > 0)$ for the noise precision:

$$\begin{aligned} p(\mathbf{c}, \sigma^2) &\sim \text{inv}{-}\Gamma(\sigma^2; a, b) \\ p(\alpha) &\sim \mathcal{U}(0, 1). \end{aligned} \tag{10}$$

The posterior distributions for $\mathbf{c}$ and σ^2, conditioned on the value for α, are:

$$\begin{aligned} p(\mathbf{c}|\mathbf{y}, \sigma^2, \alpha) &\sim \mathcal{N}(\mathbf{c}|\hat{\mathbf{c}}, \sigma^2\hat{\Sigma}) \\ p(\sigma^2|\mathbf{y}, \alpha) &\sim \text{inv}{-}\Gamma(\sigma^2; \tfrac{N-d}{2} + a, A(\alpha)/2 + b), \end{aligned} \tag{11}$$

where $\hat{\mathbf{c}}$ and $\hat{\Sigma}$ denote the least squares estimates for the expansion coefficients and noise covariance matrix respectively:

$$\begin{aligned} \hat{\mathbf{c}} &= (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{y} \\ \hat{\Sigma} &= (\mathbf{V}^T\mathbf{V})^{-1}, \end{aligned} \tag{12}$$

and the additional parameter appearing in the noise variance posterior is given by:

$$A(\alpha) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{V}\hat{\Sigma}\mathbf{V}^T\mathbf{y}. \tag{13}$$

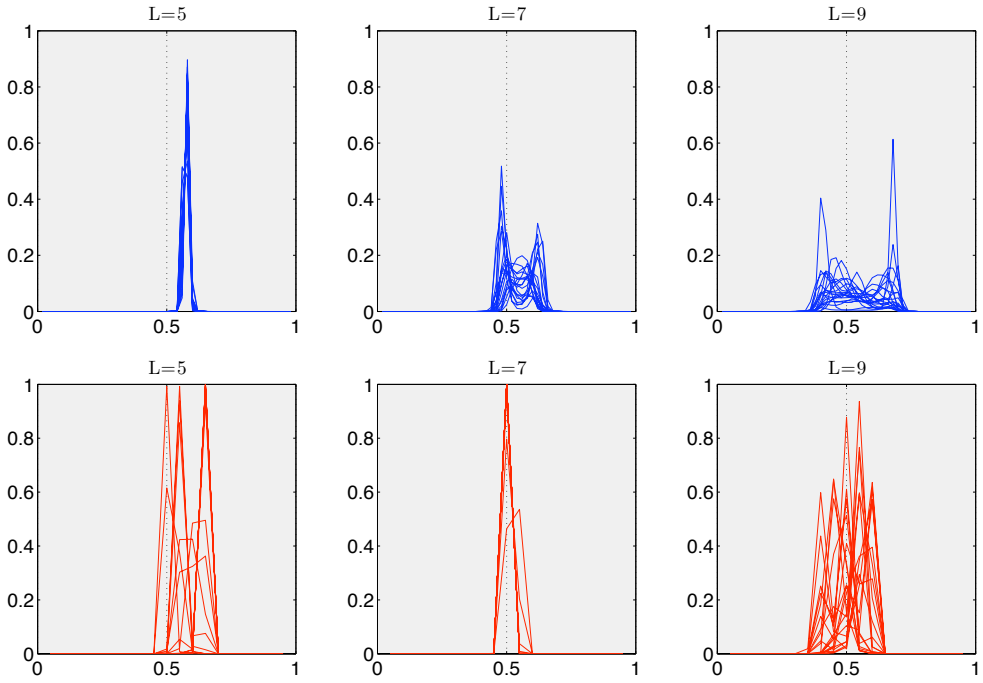

Fig. 1. Posterior distribution on α for twenty input-output realizations for SNR=10 dB in the case of linear (top) and quadratic (bottom) models.

The model complexity is specified by the number of Laguerre functions L and order Q. Let us for now consider α fixed, and denote the model by $\mathcal{M}_{LQ}(\alpha)$. The model evidence in such case writes:

$$p(\mathbf{y}|\mathcal{M}_{LQ}(\alpha)) = \iint p(\mathbf{y}|\mathbf{c},\sigma^2)p(\mathbf{c},\sigma^2)d\mathbf{c}d\sigma^2. \quad (14)$$

This integral can be calculated analytically, leading to the following formula:

$$p(\mathbf{y}|\mathcal{M}_{LQ}(\alpha)) = (2\pi)^{(d-N)/2}|\hat{\Sigma}|^{1/2}\frac{b^a}{\Gamma(a)}\frac{\Gamma(a+\frac{N-d}{2})}{(A(\alpha)+b)^{a+\frac{N-d}{2}}}. \quad (15)$$

One can see from this formula that the term $A(\alpha)$ accounts for model *accuracy*, while *complexity* is accounted for in the remaining factors. This quantity depends on the value of α, due to the dependence of the above equation on $\mathbf{V}$; therefore, let us denote it with $\mathcal{E}(\alpha)$. Considering the uniform prior of eq. (10) for α, we can write the marginal posterior distribution of α, which is in this case proportional to $\mathcal{E}(\alpha)$, as:

$$\begin{aligned} p(\alpha|\mathbf{y},\mathcal{M}_{LQ}) &= \frac{p(\mathbf{y}|\alpha,\mathcal{M}_{LQ})p(\alpha|\mathcal{M}_{LQ})}{p(\mathbf{y},\mathcal{M}_{LQ})} \\ &= \frac{\iint p(\mathbf{y}|\mathbf{c},\sigma^2,\alpha)p(\mathbf{c},\sigma^2)d\mathbf{c}d\sigma^2}{p(\mathbf{y}|\mathcal{M}_{LQ})} \\ &\propto \mathcal{E}(\alpha). \end{aligned} \quad (16)$$

where we have kept the conditioning on a model $\mathcal{M}_{LQ}$ for clarity. In other words, the conditional posterior distribution on α is the previously defined model evidence in the case where α is assumed constant. Hence, given a model $\mathcal{M}_{LQ}$, one can estimate the maximum a posteriori α, i.e., as the value that maximises $\mathcal{E}(\alpha)$.

Now, having in mind that the evidence of a model $\mathcal{M}_{LQ}$ is $p(\mathbf{y}|\mathcal{M}_{LQ})$, and as the integral $\int p(\alpha|\mathbf{y},\mathcal{M}_{LQ})d\alpha$ must sum to one, this evidence is simply:

$$p(\mathbf{y}|\mathcal{M}_{LQ}) = \int \mathcal{E}(\alpha)d\alpha. \quad (17)$$

This integral has no simple analytical form. Therefore, we use numerical integration to integrate over α and compare models with different values for (L,Q).

III. RESULTS AND DISCUSSION

We consider a data-generating model of the same class as the models under test, i.e., a Volterra system where the true system kernels are linear combinations of discrete-time Laguerre functions. Thus the true model order is known. Initially, we simulated a linear system using eq. (4) with $\alpha = 0.5$, $L = 7$, $Q = 1$ and $\mathbf{c} = \begin{pmatrix} 0 & 1 & 0.5 & -1 & 1.5 & -0.5 & 0.25 & -0.1 \end{pmatrix}^T$. The "true" linear system kernel k_1 is given by eq. (2) for these parameter values. We used a gaussian white noise input of length $N = 1024$ and added independent white noise to the output for SNR values of 10 and 1000 dB. The posterior distributions of α, obtained from twenty noise realisations, are shown in the top panel of fig. 1 for SNR=10 dB and $3 \leq L \leq 9$. Note that for less complex models than the true system ($L < 7$), we get sharp peaks in the posterior α distribution at values larger than 0.5. This reflects the fact that larger α values are favored in order to account for the slower dynamics of the higher-order basis functions that are not included in the model. On the other hand, we get a sharp

peak at or around 0.5 for $L = 7$, as well as peaks both below and above 0.5 for $L > 7$.

The optimal model $\mathcal{M}_{LQ}$ is selected by integrating $\mathcal{E}(\alpha)$ over α and comparing its values between different models. $p(\mathbf{y}|\mathcal{M}_{LQ})$ is shown in fig. 2 for 20 different input-additive noise realizations (SNR=10 dB) where it can be seen that linear models are favored vs. nonlinear models in all cases. While the true value of $L = 7$ was selected in all 20 cases for SNR=1000 dB by the model evidence criterion (the respective values for AIC and BIC were 13/20 and 19/20, with the remaining cases being biased towards more complex models), $L = 6, 7$ was selected in 14/20 and 6/20 cases for SNR=10 dB (the AIC yielded L values of 6,7,8,9 in 8,4,4,4 cases and the BIC yielded $L = 6, 7$ in 19 and 1 cases respectively). This is partly due to that the coefficient of the 7-th Laguerre function had a smaller magnitude (0.1) compared to the lower order coefficients. The model evidence criterion was found to be more robust overall, while the AIC and BIC tended to overestimate and underestimate the value of L respectively.

Finally, we simulated this system for $Q = 2$ for the same values of α and L, whereby the second-order expansion coefficients were drawn from a uniform distribution in $[-0.5, 0.5]$. The resulting posterior α distributions and model evidence values are shown in the bottom panels of fig. 1 and 2 respectively, for SNR=10 dB. Model evidence favors nonlinear vs. linear models and the true value of $L = 7$ was selected in all cases both for SNR=10 and 1000 dB (the AIC tended to overestimate L as above). In terms of computation speed, the proposed approach requires 0.64 s for $Q = 1$ and 9.8 s for $Q = 2$ for the above system on a Pentium 4 computer with 1 GB of RAM. An issue that deserves further attention is ill-conditioning of the Gram matrix $\mathbf{V}^T\mathbf{V}$, which occurs for large α values, short input/output records and/or low SNRs, which may bias the α posterior distribution.

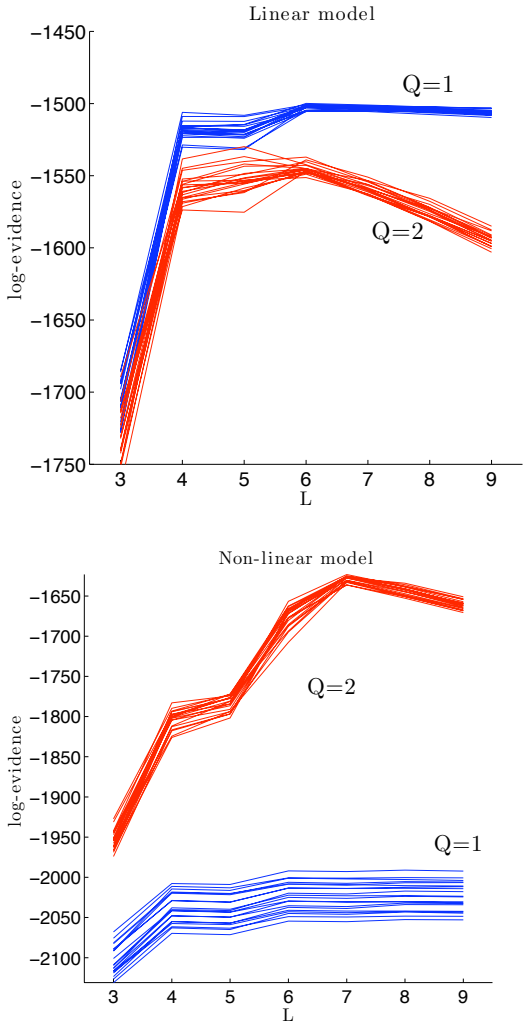

Fig. 2. Model evidence values for $Q = 1$ (blue) and $Q = 2$ (red) for 20 different realizations.

IV. CONCLUSION

We formulated Laguerre expansions of Volterra kernels in a Bayesian framework and used the results to infer on the expansion structural parameters, with the results suggesting better overall performance than the AIC and BIC. The calculation of uncertainty bounds for the expansion coefficient and Volterra kernel estimates, as well as the evaluation of the proposed approach using physiological data are currently underway.

REFERENCES

[1] M. Korenberg and I. Hunter, "The identification of nonlinear biological systems: Volterra kernel approaches," Annals of Biomedical Engineering, vol. 24, pp. 250–258, 1996.

[2] V. Marmarelis, Nonlinear Dynamic Modeling of Physiological Systems. Piscataway, NJ: Wiley-Interscience & IEEE Press, 2004.

[3] W. Rugh, Nonlinear System Theory: the Volterra/Wiener Approach. Baltimore, MD: Johns Hopkins University Press, 1981.

[4] S. Chen and S. Billings, "Neural networks for nonlinear dynamic system modelling and identification," International Journal of Control, vol. 51, no. 2, pp. 1191–1214, 1992.

[5] N. Wiener, Nonlinear Problems in Random Theory. New York, NY: Wiley, 1958.

[6] B. Wahlberg, "System identification using laguerre models," IEEE Transactions on Automatic Control, vol. 36, pp. 551–562, 1991.

[7] B. Wahlberg and P. Makila, "On approximation of stable linear dynamical systems using laguerre and kautz functions," Automatica, vol. 32, pp. 693–708, 1996.

[8] H. Ogura, "Estimation of wiener kernels of a nonlinear system and a fast algorithm using digital laguerre filters," in 15th NIIB Conference, Okazaki, Japan, 1985, pp. 14–62.

[9] V. Marmarelis, "Identification of nonlinear biological systems using laguerre expansions of kernels," Annals of Biomedical Engineering, vol. 21, pp. 573–589, 1993.

[10] G. Mitsis and V. Marmarelis, "Modeling of nonlinear physiological systems with fast and slow dynamics. i. methodology," Annals of Biomedical Engineering, vol. 30, pp. 272–281, 2002.

[11] Y. Fu and G. Dumont, "An optimum time scale for discrete laguerre network," IEEE Transactions on Automatic Control, vol. 38, pp. 934–938, 1993.

[12] T. O. e Silva, "On the determination of the optimal pole position of laguerre filters," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 43, pp. 2079–2087, 1995.

[13] R. Campello, G. Favier, and W. do Amaral, "Optimal expansions of discrete-time volterra models using laguerre functions," Automatica, vol. 40, pp. 815–822, 2004.

[14] A. Gelman, J. Carlin, H. Stern, and D. Rubin, Bayesian Data Analysis. London: Chapman et Hall/CRC, 2004.