Application of the Evidence Procedure to Linear Problems in Signal Processing

Dmitriy Shutin* and Heinz Köppl†

*Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c, A-8010, Graz, Austria
†Christian Doppler Laboratory for Nonlinear Signal Processing, Graz University of Technology, Inffeldgasse 16b, A-8010, Graz, Austria

Abstract. The presented work addresses application of the evidence procedure to the field of signal processing where ill-posed estimation problems are frequently encountered. We base our analysis on the Relevance Vector Machines (RVM) technique originally proposed by M. Tipping. It effectively locally maximizes the evidence integral for linear kernel-based models. We extend the RVM technique by considering correlated additive Gaussian observation noise and complex-valued signals. We also show that grouping model parameters $\vec{w}$, such that a single hyperparameter α_k controls the kth cluster can be very effective in practice. In particular, it allows to cluster parameters $\vec{w}$'s according to their potential relevance which in turns leads to highly improved generalization performance of the therewith parametrized models.

The developed scheme is then illustratively applied to the problem of nonlinear system identification based on a discrete-time Volterra model. Similar ideas are used to analyze wireless channels from the channel measurement data. Results for synthetic as well as real-world data are presented.

1. INTRODUCTION

Many problems in signal processing deal with linear signal expansions. Its simplicity and analytical tractability make them very attractive for modeling purposes. Generally, such an expansion can be represented in vector form as

$$\vec{y} = \vec{K}\vec{w} + \vec{\xi}, \tag{1}$$

where the vector $\vec{y}$ is obtained by stacking sampled values into a vector, $\vec{y} = [y[0], y[1], \ldots, y[N-1]]^T$. The matrix $\vec{K}$, also known as kernel matrix, contains L basis functions (kernels) $\vec{\kappa}_l = [\kappa_l[0], \ldots, \kappa_l[N-1]]^T$, such that $\vec{K} = [\vec{\kappa}_0, \vec{\kappa}_1, \ldots, \vec{\kappa}_L]$, where $\vec{\kappa}_0 \equiv 1$. The weight vector $\vec{w} = [w_0, \ldots, w_L]^T$ accumulates all individual kernel weights. The signal $\vec{y}$ is usually observed in the presence of some noise process $\vec{\xi}$. It is often convenient to assume it to be zero mean with a covariance matrix $\vec{\Xi}$. The form and properties of the set of basis functions determine the properties of the whole expansion. In particular, as a universal approximator, $\kappa_l[n]$ could be a radial basis function, or delayed samples of the impulse response, if (1) represents a linear convolution.

Inverse problem usually requires estimation of the parameter vector $\vec{w}$ given the observations $\vec{y}$, which is in many cases a nontrivial task especially if the number of parameters involved is very high or if it is known *a priori* that the solution is sparse.

CP735, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint

Examples are wireless multipath channels that can be represented by sparse models[1]. Measured channels, however, contain discrete multipath components that are always obscures by noise. Another example is identification of the Volterra kernels: due to the large number of parameters involved, direct solution using the pseudo-inverse is often impractical.

The evidence procedure[2] supplied with the automatic relevance determination (ARD) principle [3] is a Bayesian tool that offers a flexible way of solving inverse problems along with providing very compact representation of the data. In this paper we consider a special case of the evidence procedure, known as Relevance Vector Machines (RVM) and some extension thereof. This technique was first introduced by Tipping [4], and was originally proposed for regression and classification tasks.

However, straightforward application of the RVM to signal processing problems is not always possible since some of its basic assumptions are often violated. In particular, analysis of bandpass representations of wireless channels requires complex signals and the additive noise present in the measured data may also not be white. In addition, often several realizations of the measured signal are available, and it is desirable to somehow incorporate this knowledge in the RVM framework. In many applications the parameter vector $\vec{w}$ has some physical interpretation by which the parameters could be split into groups and controlled by a single hyperparameter. These requirements have motivated the extensions presented here.

For the ease of understanding the paper we have organized it as follows: Section 2 presents the modifications arising when the additive noise is no longer white and how to estimate the noise covariance matrix using the evidence procedure. Section 3 introduces model parameter clustering, and, finally, Section 4 shows some application results.

2. COMPLEX RVM WITH COLORED NOISE

In this section we will present modifications of the original RVM scheme arising when the data involved is complex and additive noise process is correlated. For the more detailed treatment of the standard RVM algorithm the interested reader is encouraged to read the original paper [4].

Estimating the parameters of interest consists in considering the likelihood function

$$p(\vec{y}|\vec{w},\vec{\Xi}) = \mathscr{N}(\vec{y}|\vec{K}\vec{\mu},\vec{\Xi}) = \frac{1}{(\pi|\vec{\Xi}|)^N}\exp\left\{-(\vec{y}-\vec{K}\vec{w})^H\vec{\Xi}^{-1}(\vec{y}-\vec{K}\vec{w})\right\}. \qquad (2)$$

Each model weight is controlled by means of a single evidence parameter α and it is described by the conditional normal density i.e., $p(\vec{w}|\vec{A}) \sim \mathscr{N}(\vec{w}|0,\vec{A}^{-1})$, where $\vec{A} = \mathrm{diag}\{\alpha_0,\ldots,\alpha_L\}$. Estimation of the α_l given the observations is the heart of the evidence procedure. The weight prior is in turn controlled by the corresponding hyperparameters. The hyperprior over the parameters $\vec{\alpha}$ is defined in the form of the Gamma distribution

$$p(\vec{\alpha}|a,b) = \prod_{l=0}^{L}\mathscr{G}(\alpha_l|a,b) = \prod_{l=0}^{L}\frac{b^a}{\Gamma(a)}\alpha_l^{a-1}\exp(-b\alpha_l). \qquad (3)$$

A similar hierarchy is build to describe the additive noise process $\xi[n]$. We will assume it to be stationary Gaussian process, with zero mean and covariance matrix $\vec{\Xi}$. From the conceptual standpoint this should not cause any difficulties, unless we want to estimate the matrix $\vec{\Xi}$ from the data. Optimizing with respect to the general Toeplitz matrix can be a difficult task. To ease the optimization, we restrict the search to circulant matrices $\vec{\Xi} \in \mathbb{C}^{N\times N}$ that can be diagonalized in the following form

$$\vec{\Xi} = \vec{U}\vec{\Lambda}\vec{U}^H, \vec{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\} \tag{4}$$

where $\vec{U}$ is the discrete Fourier transform (DFT) matrix [5]. Matrix $\vec{U}$ is scaled properly to make sure it is unitary. Just as in case of $\vec{\alpha}$ we specify the probabilistic structure for each eigenvalue λ_i. We will introduce $\vec{B} = \vec{\Xi}^{-1}$, such that $\beta_i = \lambda_i^{-1},\ i = 1, \ldots, N$, with each β_i being Gamma-distributed just as in (3), i.e., $p(\beta_i|c,d) \sim \mathscr{G}(\beta_i|c,d)$.

The hyperpriors over $\vec{\alpha}$ and β_i can be made non-informative, i.e., uniform by setting a,b, c, and d to very small values. Setting these values to zero will result in uniform hyperpriors (over the logarithmic scale), which makes them independent of scaling of variables involved in the processing.

RVM learning is the search for the hyperparameter posterior mode [4], i.e. maximization of $p(\vec{\alpha}, \vec{\Xi}|\vec{y}) \propto p(\vec{y}|\vec{\alpha}, \vec{\Xi})p(\vec{\alpha})p(\vec{\Xi})$. Unfortunately, the maximizing values can not be found in the closed form and iterative approaches are needed to solve the optimization task. If some good initial values of the hyperparameters $\vec{\alpha}^{[0]}$ and $\vec{\Xi}^{[0]}$ are known then the parameters of the posterior distribution over the weights $p(\vec{w}|\vec{y}, \vec{\alpha}, \vec{\Xi}) \sim \mathscr{N}(\vec{w}|\vec{\mu}, \vec{\Sigma})$ can be computed as follows:

$$\vec{\Sigma}^{[i]} = (\vec{A}^{[i]} + \vec{K}^H\vec{B}^{[i]}\vec{K})^{-1}, \quad \vec{B}^{[i]} = (\vec{\Xi}^{[i]})^{-1} \tag{5}$$

and

$$\vec{\mu}^{[i]} = \vec{\Sigma}^{[i]}\vec{K}^H\vec{B}^{[i]}\vec{y} \tag{6}$$

Having computed the former, the hyperparameters are updated as follows:

$$\alpha_l^{[i+1]} = \frac{1+a}{\Sigma_{ll}^{[i]} + |\mu_l^{[i]}|^2 + b} \tag{7}$$

$$(\lambda_n)^{[i+1]} = \frac{(\vec{y} - \vec{K}\vec{\mu}^{[i]})^H\vec{U}\vec{E}_{nn}\vec{U}^H(\vec{y} - \vec{K}\vec{\mu}^{[i]}) + \mathrm{Tr}[\vec{K}^H\vec{U}\vec{E}_{nn}\vec{U}^H\vec{K}\vec{\Sigma}^{[i]}] + d}{1+c}, \tag{8}$$

where the matrix $\vec{E}_{nn}$ is a matrix whose nnth element is 1 and all other elements are zero. Expression (5)-(8) are iterated until a certain convergence criterion is met.

3. PARAMETER CLUSTERING

In this section we present a promising extension of the RVM technique that involves parameter clustering or grouping. It turns out to be very effective if a group of parameters is controlled by a single hyperparameter α_l. A simple consequence of this approach is

the reduced number of computations since fewer hyperparameters are to be estimated. It should be noted that such clustering is highly problem-dependent and should be constructed with respect to the specific problem in mind.

We will assume that L model parameters w_l, $l = 1,\ldots,L$ are clustered according to a certain rule into R disjoint sets S_r, $r = 1,\ldots,R$, and $R < L$. As it has been already mentioned, learning RVM is equivalent to maximizing the evidence $p(\vec{\alpha},\vec{\vec{\Xi}}|\vec{y}) \propto p(\vec{y}|\vec{\alpha},\vec{\vec{\Xi}})p(\vec{\alpha})p(\vec{\vec{\Xi}})$. When maximizing $p(\vec{\alpha},\vec{\vec{\Xi}}|\vec{y})$ the derivative with respect to the parameters of interest is taken. Since the number of hyperparameters α_r is now less than the number of the data samples, some entries in the matrix $\vec{A} \in \mathbb{R}^{N\times N}$ will contain the repeating hyperparameters, which will result in the corresponding modification of expression (7):

$$\alpha_r^{[i+1]} = \frac{|S_r| + a}{\sum_{j\in S_r}\left(|\mu_j^{[i]}|^2 + \Sigma_{jj}^{[i]}\right) + b} \tag{9}$$

3.1. Parameter clustering in case of multiple observations.

Weight grouping could also be used to accommodate several signal observations. Consider a situation when the signal from, let us say, a sensor has been observed M times. We will assume the noise for different realizations to have the same mean and covariance matrix, and to be statistically independent. To accommodate this case in the RVM framework we redefine the vector form of the problem as follows:

$$\vec{\vec{\Xi}} = \underbrace{\begin{bmatrix} \vec{\vec{\Xi}} & & \\ & \ddots & \\ & & \vec{\vec{\Xi}} \end{bmatrix}}_{\text{M times}}, \quad \vec{\vec{A}} = \underbrace{\begin{bmatrix} \vec{A} & & \\ & \ddots & \\ & & \vec{A} \end{bmatrix}}_{\text{M times}}, \quad \vec{\vec{K}} = \underbrace{\begin{bmatrix} \vec{K} & & \\ & \ddots & \\ & & \vec{K} \end{bmatrix}}_{\text{M times}}$$
$$\vec{\vec{y}} = \begin{bmatrix} \vec{y}_1 \\ \vdots \\ \vec{y}_M \end{bmatrix}, \quad \vec{\vec{w}} = \begin{bmatrix} \vec{w}_1 \\ \vdots \\ \vec{w}_M \end{bmatrix} \tag{10}$$

Although, at first glance it seems that we simply split the problem into M independent RVMs this not so. The number of hyperparameters α_l remains the same and independent of the M. Moreover such a setup is equivalent to grouping the weights $\vec{w}$ along the snapshots. Also the covariance matrix $\vec{\vec{\Xi}}$ is shared by all the snapshots, which is a sort of 'noise clustering'. This simple representation results in basically the same re-estimation equation of the evidence parameters α_l:

$$\alpha_l^{[i+1]} = \frac{M + a}{\sum_{m=0}^{M-1}\left(\Sigma_{ll}^{[i]} + |\mu_{m,l}^{[i]}|^2\right) + b}, \tag{11}$$

which is a special form of (9) with parameter grouping along snapshots, and

$$(\lambda_n)^{[i+1]} = \frac{1}{M+c}\left(\sum_{m=0}^{M-1}(\vec{y}_m - \vec{K}\vec{\mu}_m^{[i]})^H\vec{U}\vec{E}_{nn}\vec{U}^H(\vec{y}_m - \vec{K}\vec{\mu}_m^{[i]}) + \right. \\ \left. + \sum_{m=0}^{M-1}\mathrm{Tr}[\vec{K}^H\vec{U}\vec{E}_{nn}\vec{U}^H\vec{K}\vec{\Sigma}^{[i]}] + d\right). \tag{12}$$

The latter is a rewritten form of (8) resulting from the special block-diagonal structure of the matrices (10) involved. Equations (5) and (6) are modified accordingly:

$$\vec{\Sigma}^{[i+1]} = (\vec{A}^{[i+1]} + \vec{K}^H\vec{B}^{[i+1]}\vec{K})^{-1}, \tag{13}$$

$$\vec{\mu}_m^{[i+1]} = \vec{\Sigma}^{[i+1]}\vec{K}^H\vec{B}^{[i+1]}\vec{y}_m \tag{14}$$

It can be seen, that the estimation of the noise covariance in case of multiple observations is equivalent to power spectral density estimation, when the matrix $\vec{U}$ is chosen to be the DFT matrix. Terms $(\vec{y}_m - \vec{K}\vec{\mu}_m^{[i]})$ are noise estimates for each signal snapshot transformed in the Fourier domain by the projection onto the column-space of the $\vec{U}$. In fact, the estimated eigenvalues in (12) are the resulting powers of the corresponding eigenvectors after the projection. Matrix $\vec{E}_{nn}$ simply selects the proper frequency bin (or the corresponding complex exponential) that results in the estimation of the nth eigenvalue.

4. APPLICATIONS

4.1. Nonlinear System Identification

This section considers the problem of finding a discrete-time model V for the sampled input-output (i/o) characteristic of a forced ordinary differential equation. A general system identification setup is depicted in Fig. 1. The applied model structure is the doubly truncated discrete-time Volterra series [6]

$$z[n] = (\mathsf{V}u)[n] = h_0 + \sum_{p=1}^{P}\sum_{q_1=0}^{Q_p}\cdots\sum_{q_p=0}^{Q_p} h_p[q_1,\ldots,q_p]u[n-q_1]\cdots u[n-q_p], \tag{15}$$

denoted as Volterra model V subsequently. In this setup system identification is the estimation of the multivariate Volterra kernels $h_p[q_1,\ldots,q_p]$ from the observation of i/o data samples $\{u[n],y[n]\}$. The problem belongs to the class of inverse problems, which leads to ill-posed estimation. Thus, to obtain robust estimates $\vec{\mu}$ of the model parameters

$$\vec{w} = [\,h_0, h_1[0],\ldots,h_1[Q_1], h_2[0,0],\ldots,h_2[Q_2,Q_2],\ldots,h_p[Q_p,\ldots,Q_p]\,]^T$$

the obviously linear estimation problem needs to be regularized. Furthermore, (15) indicates that the computational complexity of the Volterra model increases dramatically

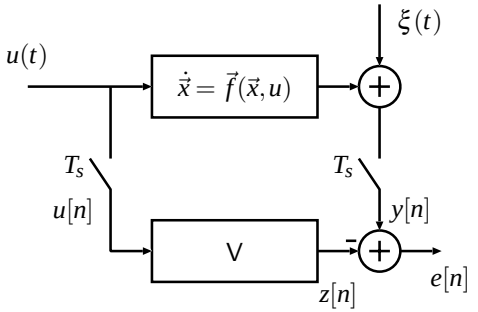

FIGURE 1. Basic system identification setup applied in this work.

with the memory length Q_p and the order of nonlinearity P. Subset selection algorithms could be applied to regulate the model complexity. Fortunately, ARD performs regularization as well the subset selection and it is applied with the above introduced extensions of parameter clustering to this identification problem. In contrast to standard regression the model parameter w_k allow some physical interpretation. Roughly speaking, if the continuous-time nonlinear system is global asymptotically stable the kernels $h_p[q_1,\ldots,q_p]$ decay to zero for increasing index q_k with $k = 1,\ldots,p$. Thus, it is reasonable to assume that kernels that are close in the index space $\vec{q} \in \mathbb{Z}^p$ with $\vec{q} = [q_1,\ldots,q_p]^T$, do have similar relevance for the model performance. The idea of clustering is that such a group of parameters is now controlled by a single relevance parameter α_k.

In the following, the parameter clustering is applied to the identification of a simple nonlinear system. The system is governed by

$$\begin{aligned} \dot{x}_1 &= a_1 x_1 + bu \\ \dot{x}_2 &= a_2 x_2 + d x_1 u \\ r &= c x_2 \quad \text{with} \quad \vec{x}(0) = \vec{0}, \end{aligned}$$

which corresponds to a second order homogeneous bilinear system[7]. Thus, its discrete-time model consists of a second order homogeneous Volterra model characterized by the kernel $h_2[q_1,q_2]$. As excitation signal a superposition of 512 sinusoids at different frequencies are applied. The identification of the system is performed in the presence of an additive white Gaussian perturbation with a signal-to-noise-ratio of 40dB. Results of the identification in terms of mean-square generalization error and model complexity can be found in Figure 2. The results indicate that the method with clustering shows a more robust generalization performance than the method with no clustering or the application of the pseudo-inverse.

4.2. Detecting multipath components in measured wireless channel

Estimation of the wireless channel is in some sense similar to the problem of system identification describe above with the distinction that the channel itself is linear. The

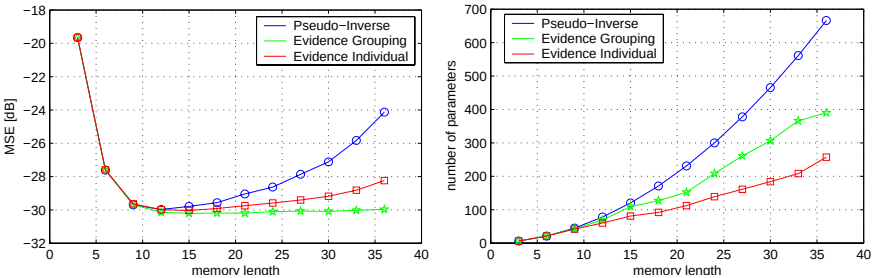

FIGURE 2. Mean-square generalization error (left) and model complexity (right) versus different memory lengths of the discrete-time Volterra system; pseudo-inverse (circle), automatic relevance determination with no clustering (square), with clustering (star).

sparse nature of the channel makes RVM a perfect candidate to be used for channel identification and multipath detection. Consider a channel sounding scheme shown in Fig.3 It can be easily shown that the received signal $z[n]$ is a linear combination of

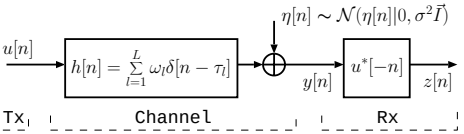

FIGURE 3. A communication system with matched receiver.

L autocorrelation sequence of the sounding sequence $u[n]$, embedded in the colored noise $\xi[n] = \eta[n] \otimes u^*[-n]$ which is the convolution of $\eta[n]$ and matched filter impulse response $u*[-n]$.

$$z[k] = \sum_{l=1}^{L} w_l R_{uu}[k-\tau_l] + \xi[k] \tag{16}$$

Assuming that there are N samples of $z[k]$ available, (16) can be easily rewritten in the form (1). From this point, the application of the RVM technique is straightforward, with kernels being the sampled autocorrelation function $R_{uu}[k]$. Figure 4 shows the result for detecting multipath components within the RVM framework for the measured multipath channel. There are in total five channel realizations. By visual inspection of the Fig. 4 it can be seen that the algorithm successfully detects the position of the multipath components.

5. CONCLUSION

In this contribution we have considered applications of the evidence procedure to selected problems in Signal processing. Motivated by the practical applications we have considered the extension of the RVM technique to the cases when the additive noise is no longer white and thus described by the full covariance matrix. In order to simplify

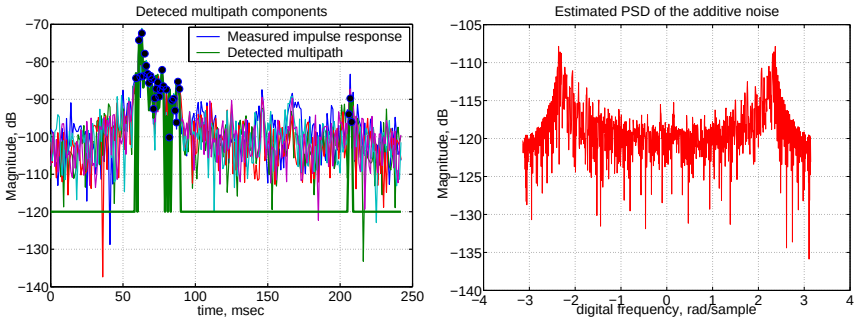

FIGURE 4. Detection of the multipath components with RVM technique.

the optimization involved only the eigenvalues of the covariance matrix were updated, with eigenvectors chosen to be complex exponential. This renders the whole scheme to be equivalent to the Fourier analysis of the additive noise within the RVM framework.

We also have shown that in many situations it turns to be quite effective to employ a certain parameter grouping. In particular it can be used to accommodate multiple observations of the signal as well as to parametrize Volterra kernels. We have shown applicability of the idea for two practical applications: nonlinear system identification and multipath detection from wireless channel measurement data. In both cases results look quite promising.

REFERENCES

1. Rappaport, T. S., *Wireless communications. Principles and practice.*, Prentice Hall PTR, 2002.
2. MacKay, D. J. C., *Neural Computation*, **4**, 415–447 (1992).
3. Neal, R., *Bayesian Learning for Neural Networks*, vol. 118 of *Lecture Notes in Statistics*, New York: Springer-Verlag, 1996.
4. Tipping, M., *Journal on Machine Learning Research*, pp. 211–244 (2001).
5. Strang, G., *Linear Algebra and its Applications.*, Brooks/Cole Publishing Company, 1988.
6. Rugh, W. J., *Nonlinear System Theory*, Johns Hopkins University Press, London, 1981.
7. V. J. Mathews, G. L. S., *Polynomial Signal Processing*, Wiley-Interscience, 2000.