Improved Bayesian Modelling and Estimation with Mixtures

Li He, Miroslav Kárný

Adaptive System Department
Institute Information Theory and Automation
Academy of Sciences of the Czech Republic
P. O. Box 18, 182 08 Prague, Czech Republic
Tel: (4202) 6605 2274, Fax: (4202) 6605 2268,
E-mail: heli@utia.cas.cz, school@utia.cas.cz

Abstract: In this paper, improved Bayesian Modelling and estimation of dynamic systems is concerns by means of Bayesian mixture methodology: First, Bayesian Modelling has been reexamined from the view of probabilistic mixture; Next, under the mixture modelling framework, a novel type of mixture model, called ARMMAX, has been introduced and studied. ARMMAX defines as a finite mixture of ARMAX having common ARX part. Efficient estimation of ARMMAX model with fixed MA parts is provided as well by a specific version of recursive quasi-Bayes estimation algorithm; Then, one possible idea to Bayesian-related estimate single ARMAX under the general assumption of unknown stochastic MA term is discussed.

Keywords: Bayesian Modelling and estimation, ARMAX, finite mixtures.

1 Introduction

ARMAX model – auto-regression (AR) with moving average noise (MA) and external input (X) – is commonly used for description of linear stochastic systems. It is equivalent to the linear state-space model that forms the corner stone of so-called modern estimation and control theory. Therefore the problem of estimation of ARMAX has been an active research area and remains as an issue which is not solved completely yet. Among the multitude of estimation variants of ARMAX, it seems that an approximate minimization of prediction error (PE) method (Ljung, 1987), has become a golden standard. It is, however, oriented only to point estimation. Consequently, information on precision of estimates and other tasks like structure estimation are weakly supported. Moreover, PE is restricted on ARMAX models with strictly stable condition of its MA part.

From Bayesian viewpoint, the difficulty to estimate ARMAX stems from the lack of sufficient statistic with dimension smaller than the number of data. Peterka (1981a) relaxed the stability restriction on MA part and provided real-time Bayesian estimation of ARX part when MA is fixed. Essentially, he shown that LD factorization of the known correlation matrix acts as a time-varying filter on the observed data. The filtered data are then used in standard Bayesian estimation of ARX part. Consequently, its uncertainties are under the control, Bayesian structure estimation can be used (Kárný and Kulhavý, 1988) , etc.

Unfortunately, the mentioned properties are reached under the impractical assumption that MA part is known. A Bayesian comparison of hypothesis was ever proposed for relaxing this assumption (Peterka 1989) to evaluates posterior probabilities on hypotheses that a specific MA part is the best one in a finite set of candidates. This improvement gives, however, no guide how to selects the candidates. Moreover, the posterior probabilities converge to a zero-one vector in generic case, the quality of the original choice of candidates therefore is out of objective control. The extensively Bayesian use of ARMAX thus is hindered because of the difficult MA term. One

of aim of the paper is to discuss the possible estimation of MA part so that make improvement on Bayesian solution of ARMAX.

Meanwhile, the literature review and practical applications study have indicated that a single model cannot deal with some complex problems, such as the issues of heterogeneity. As one of options to be able to combine the features of different models, probabilistic finite mixtures have become popular modelling tools in a widespread practical use (Titterington et al, 1985). This has motivated us to open a wider overview of Bayesian modelling through probabilistic finite mixture.

Under the mixture Bayesian modelling, we then have introduced and investigated a novel mixture model, a finite mixture of ARMAX models with a common ARX part but different fixed MA parts. We call it as ARMMAX model. ARMMAX model is richer than ARMAX model since it assumes invariant deterministic ARX part but allows variations of the stochastic MA part. One efficient estimation of ARMMAX with given MA parts is provided also by the proposed numerical ARMMAX-QB algorithm with computational burden well comparable to that of single ARMAX. The estimation provides a quantitative measure of descriptive quality of the ARMAX models forming mixture components. Thus, several ARMAX models are estimated in parallel.

At the end, the possibility to estimate single ARMAX model with unknown MA term is then presented. What is interested is that it shows that estimation of ARMAX is possible to be made by mean of estimation of more complex ARMMAX. The basic idea is that difficult point estimation of unknown MA part could be searched by *multi-directional search* (MDS) method (Torczon, 1989) in ARMMAX "algorithmic" parallel environment to produce a implementable algorithm so that it generates a sequence of points, denoted as the best vertices of the simplex or defined by one of the components in ARMMAX, that convergence to a critical point–ideally a "true" MA part. With the searched MA C-parameters, the rest Bayesian estimation of ARX part of ARMAX model is relied on running several filters of Peterka (1981a) in parallel.

2 Bayesian Modelling

To interact with the system, a description of the system properties, a *model*, is needed. Modelling of relationships among a sequence of observations of systems is to provide a means to help us to get better understanding about the interested complex systems. Here we shall first briefly review the Bayesian modelling (Peterka, 1981b) and then discuss the mixture view of Bayesian modelling in the next section.

Consider a stochastic system on which a time-oriented discrete data sequence $d_1, d_2, \cdots, d_t, \cdots$ are observed at discrete time instant $t = 1, 2, \cdots$. The sequence of data observed on the system up to time t is denoted by

$$d(t) = (d_1, \cdots, d_t)$$

at each time instant t, data d_t is composed of a pair of random variables: $d_t = (u_t, y_t)$, u_t is defined as a direct manipulated input to the system and y_t is the output, i.e., response of the system at time t to the past history of data $d(t-1)$ and current input u_t.

From Bayesian viewpoint, all forms of uncertainty can be inherently quantified by means of *probability*. Instead of being interpreted in terms of limits of relative frequencies or other *objective* ways, in Bayesian inference the concept of probability is used to describe uncertainty about the unknown random quantities, like input-output data, model parameters, hypotheses, etc. Here by *random*, it means *uncertainty* rather than description of a process of observing a repeatable event. With assuming the existence of the unknown parameter Θ, *system model* parameterized by an finite set of parameter Θ is given by a set of conditional probability distributions

$$f(y_t|d(t-1), u_t, \Theta) \tag{1}$$

to describe the dependence of the output y_t on the known past history of input-output data including the last input and unknown parameter set.

Usually if the output y_t is a random continuous variable, it may be useful to introduce a related random variable e_t as a difference between the output y_t and its mean value conditioned on the past history of the input-output process

$$e_t = y_t - \hat{y}_t(d(t-1), u_t)$$

where $\hat{y}_t(d(t-1), u_t) = \mathcal{E}[y_t|d(t-1), u_t] = \int y_t f(y_t|d(t-1), u_t)dy_t$. One of important properties of the sequence $e_t, t = 1, 2, \cdots$ is its *whiteness*.

By means of the above relation, the system model then can be given in the form of a stochastic equation

$$y_t = \hat{y}_t(d(t-1), u_t) + e_t \tag{2}$$

Further, if the normality (Gaussian type) of e_t is assumed and the conditional mean value $\hat{y}_t$ is expressed as a function of the past history of the input-output data through a finite set of parameter, (1) can be specified as

$$f(y_t|d(t-1), u_t, \Theta) \sim \mathcal{N}(\hat{y}_t, r) \tag{3}$$

where if the quantities of the stochastic noise, like standard derivation, are unknown, they should be included into the parameter collection Θ of the studied system and estimated as well.

2.1 ARMAX models

Here we give the description of ARMAX model, since it is the basic model used in the paper.

If assume the mean value of output $\hat{y}_t(d(t-1), u_t)$ is a function of the entire past history of input-output data, then to express the conditional mean $\hat{y}_t$ through a finite number of parameters, it has to be assumed that $\hat{y}_t$ is defined recursively as following

$$1 + \sum_{i=1}^{n} c_i \hat{y}_t = \sum_{i=1}^{m} g_i y_{t-i} + \sum_{i=0}^{m} b_i y_{t-i}$$

where $\hat{y} = \hat{y}_t(d(t-1), u_t)$. For simplicity, a constant term is intentionally ignored, although such a constant term usually can not be eliminated if the parameters are unknown.

Applying it into the stochastic equation (2), gives ARMAX model

$$y_t = \theta' \psi_t + v_t \tag{4}$$

with $a_i = c_i - g_i$, $\psi_t' = [u_t, y_{t-1}, u_{t-1}, \cdots, y_{t-n}, u_{t-n}]$ is regression vector. $\theta = [b_0, a_1, b_1, \cdots, a_n, b_n]$ is the unknown parameter vectors. v_t is a colored noise with zero mean and the finite correlation span

$$\begin{aligned} \mathcal{E}[v_t v_{t-i}'] &= r_i \text{ for } i = 0, 1, \cdots, n \\ &= 0 \text{ for } |i| > n \end{aligned}$$

It can be considered as a moving average defined on the sequence of mutually uncorrelated white noise $\{e_t\}$

$$v_t = \sum_{i=0}^{n} c_i e_{t-i}$$

Clearly, there is the relation $r_i = r_e s_i, i = 1, ..., n$ with $c_0 = 1, s_i = \sum_{k=i}^{n} c_i c_{k-i}$.

If the known past history of input-output is long enough so that the influence of initial conditions $\hat{y}_i (i = 1, \cdots, n)$ may be negligible. With the normality of e_t, an ARMAX is fully parameterized through the following parameter set $\Theta = \{\theta, r_e, c = [c_1, \cdots, c_n]\}$.

As one of generally stable requirement of ARMAX representation, the zeros (roots) of C-polynomial of MA term in ARMAX is needed to lie outside the unit circle, otherwise, as showed by the most filters, it cannot produce optimal predictions. However, the condition can be relaxed as long as the filter is constructed based on LD factorization of the covariance matrix by the filter of Peterka (see Apendix).

3 Mixture view of Bayesian modelling

As a flexible probabilistic model, finite mixture has become a fruitful branch of Bayesian approach. Markov chain Monte Carlo methods make it easy to fit mixtures to Bayesian approach. However a consistently overview of Bayesian modelling so far has not been well built yet. Several different types of views are possible to be adopted to introduce mixture into Bayesian modelling. Here we shall build such a wider mixture overview of Bayesian modelling along the traditional way.

By general Bayesian modelling, to provide a *system model* is to specify a conditional probability density function $f(y_t|d(t-1), u_t, \Theta)$. We then face a distribution approximation problem. Since mixtures can be interpreted as universal approximators of probability density functions (Titterington et al, 1985):

> *Provided the number of component densities is not bounded, certain forms of mixture can be used to provide arbitrarily close approximations to a given probability distribution*

Thus it is conceptually straightforward to extend mixture to the above Bayesian modelling

$$f(y_t|d(t-1), u_t, \Theta) \quad = \quad \sum_{i=1}^{k} \alpha_i f(y_{i,t}|d(t-1), u_t, \Theta_i) \tag{5}$$

with mixture weights $\alpha = [\alpha_1, \cdots, \alpha_k]$ satisfying $\alpha_p \geq 0,\ \ p = 1, \cdots, k,\ \ \sum_{p=1}^{k} \alpha_p = 1$. Θ_i is the parameter set associated with the i-th component of the mixture while $f(y_{i,t}|d(t-1), u_t, \Theta_i)$ describes its the density. And the parameters Θ of the model is now the collection of all components‘ parameters and mixing weights, $\Theta \equiv \left\{\alpha = [\alpha_1, \cdots, \alpha_k], \{\Theta_p\}_{p=1}^{k}\right\}$.

In practical, most often the condition $k < \infty$ is assumed, i.e. a finite mixture is in use. With limitation on the number of components, it is not always able to arbitrarily closely approximate a probability density functions by means of finite mixtures. In this sense, the mixture view of Bayesian modelling can not substitute the general Bayesian modelling as long as finite mixture is used.

To allow single component in mixture, i.e. $k = 1$, some general single models shall be derived in similarly way as the traditional way. For example, with the assumptions i) There is only one component $k = 1$; ii) The mean value of output $\hat{y}_t(d(t-1), u_t)$ is a function of the entire past history of input-output data, the mixture (5) is specified as a single ARMAX model.

With the number of components of mixture, $k > 1, k < \infty$, some more complex model structure can be introduced, for example non-linearity. At present, the most research of mixture field is mainly focused on extending the basic linear models classes to the corresponding mixtures cases. In next section, we have introduced a new type of finite ARMAX mixture with common ARX part in all ARMAX components.

4 ARMMAX model

Consider the mixture description(5), and assume:

i) There is more than one but finite components $k > 1, k < \infty$; ii) Each component of mixture is described by one ARMAX; iii) There is a common deterministic ARX part in all ARMAX components while the characteristics of the stochastic noise parts vary.

Then, it gives a novel type model class, we call it ARMMAX model. For a given d_t at each time instance t, the *probability density function* (pdf) of ARMMAX model is given as

$$f\left(y_t|u_t, d(t-1), \Theta_{ARX}, \Theta_c, \alpha\right) \;=\; \sum_{p=1}^{k} \alpha_p f(y_t|u_t, d(t-1), \Theta_{ARX}, C_p) \tag{6}$$

with mixture weights $\alpha_p \geq 0,\;\; p = 1, \cdots, k,\;\; \sum_{p=1}^{k} \alpha_p = 1$.The individual components $f(y_t|u_t, d(t-1), \Theta_{ARX}, C_p)$ are ARMAX models. If the normal ARMAX type mixture is used, each component is described by the following filter regression (see Appendix)

$$f(y_{p,t}|d(t-1), \Theta_{ARX}, C_p) \;=\; \mathcal{N}_{\tilde{y}_t}(\theta'\tilde{\psi}_{p;t}, rD_t) \tag{7}$$

Notice that one of important features of the model is that all components have the same parameters related to AR part Θ_{ARX} in common and different C-parameter for MA parts $\{C_p\}_{p=1}^{k}$. Thus ARMMAX model is parameterized by parameter set

$$\{\Theta_{ARX} \equiv (\theta, r), \alpha \equiv [\alpha_1, \cdots, \alpha_k], \Theta_c \equiv \{C_p\}_{p=1}^{k}\}$$

Basically, there are the following advantages to investigate ARMMAX models

- Flexibility in system description of a single deterministic dynamic driven by noise with characteristics varying. ARMMAX describes well the cases when the common ARX part has a physical meaning of interest. It meanwhile provides more freedom in describing stochastic part of the input-output relationship.
- Natural parallelism to check several ARMAX in parallel. ARMMAX can be interpreted as a parallel realization of several ARMAX. It is the key property of ARMMAX we are going to exploit when investigating the possibility of the estimation of ARMAX by means of ARMMAX at the end of paper.
- Implementable algorithm (see next section) can be provided when MA parts are known with Computational demand comparable to that of single ARMAX estimation.

ARMMAX is more flexible and richer for modelling unmeasured disturbances compared to a single ARMAX model. It is intuitively obvious but it can be shown also formally as below.

Proposition 4.1 (Moments of ARMMAX model) *For an ARMMAX model with a given selection of C-parameters $\Theta_c = \{C_p\}_{p\in p^*}$, an equivalent single ARMAX model exists in terms of the first moment. The equivalence does not hold with respect to variance.*

Proof: Its is straightforward implied by the mixture definition, linearity of the expectation and identity $E[y^2] = \mathrm{var}[y] + E^2[y]$.

For simplicity, the proof is only presented for the simple case $k = 2$. With a pair $\Theta_c = (C_1, C_2)$, the filters generate the filtered data $\tilde{\psi}_{p;t},\ \Delta_{p;t},\ p \in \{1, 2\}$, c.f. Appendix, Proposition 6.1. The mixing weights are $\alpha = [\alpha, 1-\alpha]$, then the corresponding conditional expectation $E[\cdot|\cdot]$ and variance $\mathrm{var}[\cdot|\cdot]$ of the output y_t are

$$\begin{aligned} E\left[y_t|u_t, d(t-1), \Theta_{ARX}, \alpha, \Theta_c\right] &= \alpha[\theta'\tilde{\psi}_{1;t} + \Delta_{1;t}] + (1-\alpha)[\theta'\tilde{\psi}_{2;t} + \Delta_{2;t}] \\ &= \theta'\left(\alpha\tilde{\psi}_{1;t} + (1-\alpha)\tilde{\psi}_{2;t}\right) + \alpha\Delta_{1;t} + (1-\alpha)\Delta_{2;t} \\ \mathrm{var}\left[y_t|u_t, d(t-1), \Theta_{ARX}, \alpha, \Theta_c\right] &= \alpha(1-\alpha)\left(\theta'(\tilde{\psi}_{1;t} - \tilde{\psi}_{2;t}) + \Delta_{1;t} - \Delta_{2;t}\right)^2 + \\ &+ \; r\left(\alpha D_{1;t} + (1-\alpha)D_{2;t}\right). \end{aligned} \tag{8}$$

Smooth dependence of the filtered data vectors on Θ_c implies that a single equivalent C can be found generating a single filtered data vector equivalent to the convex combination describing the first moment. However, the dependence of the conditional variance on data obviously cannot be neglected. □

5 ARMMAX-QB Estimation

The possibility to use ARMMAX model is supported by a recent progress in estimation of finite mixtures. The *Quasi-Bayes* (QB) estimation of mixtures with ARX components (Kárný et al 1998), throw light on how to estimate ARMMAX model based on the filter of Perterka. QB algorithm is a slight extension of classical mixture-estimation algorithms (Titterington et al, 1985). It has good properties, Bayesian motivation as well as predictable and feasible computational complexity. Estimation of Gaussian ARMMAX model with fixed MA parts is solved by a specific version of recursive quasi-Bayes (ARMMAX-QB) estimation algorithm.

The same as standard QB, we assume the existence of some discrete random pointers $\{p_t\}, p_t \in p^* = \{1, 2, \ldots, k\}$ to indicate the active component that takes the value $p \in p^*$ with the probability α_p. Consider $\{p_t\}$ as the unmeasured data together with measured data $\{d_t\}$, we have

Proposition 5.1 (Joint Description) *Let the modelled system be described by ARMMAX model (6) with individual normal ARMAX components having known MA parts $\Theta_c = \{C_p\}_{p\in p^*}$. Considering the pointers$\{p_t\}$, the joint pdf of data then can be described as following*

$$\begin{aligned} & f(y_t, p_t|u_t, d(t-1), \Theta_{ARX}, \alpha, \Theta_c) \\ = \; & f(y_t|p_t, u_t, d(t-1), \Theta_{ARX}, \alpha, \Theta_c) f(p_t|u_t, d(t-1), \Theta_{ARX}, \alpha, \Theta_c) = \mathcal{N}_{\tilde{y}_{p;t}}(\theta' \tilde{\psi}_{p;t}, rD_{p;t})\alpha_{p_t}. \end{aligned}$$

by marginalization over p_t, its marginal pdf does $f(y_t|u_t, d(t-1), \Theta_{ARX}, \alpha, \Theta_c)$ give ARMMAX model (6).

Since there is common ARX part in all components, the probability of current system represented by p-th component is now depended on its MA parameter $\{C_p\}$

$$\alpha_p(d(t-1), \{C_p\}_1^k) = f(p_t = p|d(t-1), \{C_p\}_1^k) \tag{9}$$

The distribution on probabilistic vector $\alpha = [\alpha_1, \cdots, \alpha_k]$ is usually specified as *Dirichlet* pdf $f(\alpha|d(t-1)) = Di_\alpha(\kappa)$,

$$Di_\alpha(\kappa_{t-1}) \equiv \prod_{p=1}^{k} \frac{\alpha_p^{\kappa_{p;t-1}-1}}{\Gamma(\kappa_{p;t-1})} \Gamma\left(\sum_{p=1}^{k} \kappa_{p;t-1}\right), \; \kappa_{p;t-1} > 0,$$

with $\Gamma(x) = \int_0^\infty z^{x-1} \exp(-z)\, dz, \; x > 0$.

If we assume the pdf on unknown parameters Θ_{ARX} of the ARX part of the components in the conjugate GiW (*Gauss-inverse-Wishart*) form,

$$GiW_\Theta(V, \nu) \equiv \frac{r^{0.5(-\nu+n_a+2)}}{\mathcal{I}(V, \nu)} \exp\left\{-\frac{1}{2r} \mathrm{tr}\left(V[-1, \theta']'[-1, \theta']\right)\right\}$$

where n_a is the order of AR part and $\mathcal{I}(V, \nu)$ is the normalising integral.

Thus we have

$$f\left(\Theta_{ARX}, \alpha|d(t-1), \Theta_c\right) = GiW_{\Theta_{ARX}}(V_{t-1}, \nu_{t-1}) Di_\alpha(\kappa_{t-1}) \tag{10}$$

where individual ARMA normal component can be described through the filter (15) (see Appendix) and efficiently evaluated.

Then according to joint description (9), the assumption (10) and Bayes rule, it gives

$$
\begin{aligned}
& f(\Theta_{ARX}, \alpha, p_t | d(t), \Theta_c) \qquad (11)\\
= \; & GiW_{\Theta_{ARX}}\left(V_{t-1} + \sum_{p\in p^*} \delta_{p,p_t} \tilde{\Psi}_{p;t}\tilde{\Psi}'_{p;t}, \nu_{t-1}+1\right) Di_\alpha\left(\kappa_{t-1} + \sum_{p\in p^*} \delta_{p,p_t}[\underbrace{0,\ldots,0}_{p-1},1,0,\ldots]'\right).
\end{aligned}
$$

where the Kronecker symbol $\delta_{p,p_t} = \begin{cases} 1 & \text{if } p = p_t \\ 0 & \text{otherwise} \end{cases}$

To preserves the above assumed form (10), let us approximate δ_{p,p_t} by its expectation $w_{p;t}$

$$
\delta_{p,p_t} \approx w_{p;t} \equiv E\left[\delta_{p,p_t} | d(t), \Theta_c\right] \propto \mathcal{I}\left(V_{t-1} + \tilde{\Psi}_{p;t}\tilde{\Psi}'_{p;t}, \nu_{t-1}+1\right)(\kappa_{p;t-1}+1). \qquad (12)
$$

Then the followings proposition is given

Proposition 5.2 (Quasi-Bayes estimation of ARMMAX model) *Under natural condition of control (Peterka, 1981b), consider ARMMAX model (6) with individual components having known MA parts $\Theta_c = \{C_p\}_{p\in p^*}$, and assume at the time $t-1$ prior pdf on unknown parameters $\Theta_{ARX} = [\theta, r]$ of the common ARX part of all components (15) in the conjugate GiW form and that of α in* Dirichlet *form,*

$$
f\left(\Theta_{ARX}, \alpha | u_t, d(t-1), \Theta_c\right) = GiW_{\Theta_{ARX}}(V_{t-1}, \nu_{t-1}) Di_\alpha(\kappa_{t-1})
$$

Then, the posterior at time t is approximately preserved in the same form

$$
f(\Theta_{ARX}, \alpha | u_{t+1}, d(t), \Theta_c) = GiW(V_t, \nu_t) Di(\kappa_t)
$$

with the associated statistics updated according to the recursions

$$
V_t \;=\; V_{t-1} + \sum_{p\in p^*} w_{p;t}\tilde{\Psi}_{p;t}\tilde{\Psi}'_{p;t}, \quad \nu_t = \nu_{t-1}+1, \quad \kappa_{p;t} = \kappa_{p;t-1} + w_{p;t}.
$$

where $w_{p,t}$ are approximates (12). $\tilde{\Psi}_{p;t}$ is the filtered data vectors generated by the filters (16) determined by $C = C_p$, $p \in p^*$. k filters run in parallel to generate them. After new data item d_t is observed, their predictive ability is checked and expressed by approximated weights $w_{p,t}$.

Proof: The exact updating with unknown value δ_{p,p_t} uses the normal filtered ARX version of the ARMAX model (15), the form of the GiW pdf and the Bayes rule applied under the natural conditions of control.

Marginalization over Θ_{ARX} and α giving the weight $w_{p;t}$ exploits the predictive Student pdf and the elementary property of Dirichlet pdf stating that $Di_\alpha(\kappa)\; E[\alpha_p|\kappa] \propto \kappa_p$.

The final updating is implied directly by the adopted approximation of δ_{p,p_t}. □

As showed above, the assumption that all components have the the common AR part parameters implies that we shall update a single *extended information matrix* V for all components through all filtered data vectors weighted by $w_{p,t}$ instead of updating one extended information matrix for each component individually.
The estimation requires about $2n^2_{ARX} + kn(2n + n_{ARX})$ flops so that the computational burden increases linearly with the number of components k. The dimension n_{ARX} of the ARX part is often much larger than the order n of MA. Since a single common ARX part is estimated, the computational complexity connected with estimation of ARMMAX model is slightly larger than that needed for estimation of a single ARMAX model.

6 Discussions and Conclusions

Now let us discuss the possibility of Bayesian-related estimation of ARMAX by means of ARMMAX.

Consider an ARMAX and assign a prior pdf $f(C)$ to the possible candidates C's of the "true" MA part. Observations $d(t)$ then correct it to the posterior pdf $f(C|d(t))$ through the Bayes rule. Under the natural conditions of control, it reads

$$f(C|d(t)) \propto \prod_{\tau=1}^{t} f(y_t|u_t, d(t-1), C) f(C) \equiv \mathcal{L}(d(t), C) f(C). \tag{13}$$

For any fixed C, the value of the introduced *likelihood function* $\mathcal{L}(d(t), C)$ is simply product of values of the predictive pdf. Formally, the posterior pdf can be used for estimating the unknown C. Problem is that $\mathcal{L}(d(t), C)$ can be evaluated value-wise only. Peterka (Peterka, 1989) made the formula (13) usable by restricting the competing C's to a finite set. The posterior pdf then serves for selecting the most promising ones among them. No rule was, however, given how to select the finite set of suitable competitors. Numerical maximization procedures offer themselves for generating the most interesting competitors around the maximum of the posterior pdf.

For simplicity, we restrict ourselves to uniform prior pdf on C and search for the candidates in a neighborhood of

$$\operatorname{Arg}\max_{\{C\}} f(C|d(t)) = \operatorname{Arg}\max_{\{C\}} \mathcal{L}(d(t), C). \tag{14}$$

it amounts to nonlinear programming and numerical analysis. Since efficient evaluation of the gradient is inhibited by the complex nature of the likelihood $\mathcal{L}(d(t), C)$, we have to restrict ourselves to derivative-free search methods. Meanwhile, we have to respect the fact that each evaluation of the objective function $\mathcal{L}(d(t), C)$ coincides with a run of least squares estimation over the whole data set so that it is expensive in a way to evaluate the objectives. This seems to exclude Monte-Carlo-based maximization. Moreover, the optimized likelihood function may be multi-modal with quite sharp but smooth modes. Thus, we can rely at most on continuous differentiability of $\mathcal{L}(d(t), C)$ with respect to C.

The mentioned facts reduce the set of options more or less to simplex type methods. They evaluate values of the optimized function at vertices of a simplex that is iteratively modified. The Nelder-Mead simplex is their most popular example. It lacks, however, convergence proof. *Multidirectional search* (MDS) (Torczon, 1989) is a simplex-based search method for parallel machines with convergence analysis basis and robustness with respect to errors in function evaluation (Torczon, 1991). These favorite properties are achieved by keeping a fixed shape of simplex that is moved through the search space. Low demands of on prior knowledge of the maximized function and guaranteed convergence are paid by a rather slow convergence, especially, in the terminal phase of the optimization when MDS behaves like a gradient method. We resort to MDS procedure for generating suitable candidates of MA C-parameters around maximizing arguments of the likelihood $\mathcal{L}(d(t), C)$ in a way that is implementable even for a higher dimensional cases.

Meanwhile, possibility to check k ARMAX in parallel in one ARMMAX gives us a chance combine it with the MDS to produce the parallel search procedure implementable on a single-processor machine. The main idea is that unknown MA C-parameters could be searched by MDS in "algorithmic" ARMMAX parallel environment to generate a sequence of points that convergence to a critical point–ideally a "true" MA part. More exactly, by estimating ARMAX whose C-parameters form simplex recommended by the MDS method, we evaluate quality of several competitive ARMAX models in *parallel*. Moreover, likelihood values assigned to individual component can be interpreted as approximations of the values the likelihood $\mathcal{L}(d(t), C_p)$ corresponding to single ARMAX given by C_p.

Determining the number of components of mixture, one of important issues of in most mixture analysis, is out of question for our ARMMAX in the context of (MDS), since it is the same as the

number of vertices in the used simplex of MDS search and defined by the order of MA part of ARMAX. Although a general mixture of ARMAX without the assumption of common ARX part can provides such kind of parallelism as well, the associated computational burden keep us from it. With estimating a single common ARX part, ARMMAX can be estimated with computational demands close to those needed for estimation of a single ARMAX model.

The above discussion has open a way to deal with the improved Bayesian estimation and prediction of standard single ARMAX system under the general assumption of unknown stochastic MA term. However to develop the practical efficient implementable solution of ARMAX following the proposed idea, it needs both more elaborated work on the implementation of optimization and further exploring the power of mixture. They will be the subject of future work. In particular, the stronger theoretical support on parallelism of ARMMAX is preferably provided, since the parallelism to check several ARMAX in parallel is the key property we are going to exploit when use ARMMAX in the estimation of ARMMAX. If the proposed idea shall be proven work well on ARMAX, then it is possible be further applied to the case of output error models.

Acknowledgment

This work has been supported by CZ-SLO grant KONTAKT 2001/020, AV CR S1075102 and Project MIAPS 6458-3 from Internal Grant Agency of Ministry of Health of the Czech Republic.

Appendix

Estimation and prediction of ARMAX model with known MA part

Consider ARMAX model, if its MA part C-parameters are known, then the covariance matrix G of the noise w_t is Toeplitz matrix with $rs_i,\ s_i = \sum_{k=i}^{n} c_i c_{k-i},\ c_0 = 1$ on i-th sub- and super-diagonals, $i \leq n$.

Consider LD decomposition of the covariance $G = rLDL'$, where L is a lower triangular matrix with unit diagonal and $D = \mathrm{diag}[D_1, D_2, \ldots, D_t]$ is a diagonal matrix with positive diagonal entries D_τ, $\tau = 1, \ldots, t$. The factors L, D are functions of MA parameters. Their entries can be evaluated recursively as follows

$$D_1 = s_0, \quad L_{1,2} = s_1 D_1^{-1}, \quad D_2 = s_0 - L_{1,2}^2 D_1$$

and for $\tau = 3, 4, ..,$ and $i = n, n-1, ..., 1$ with $n_\tau = \min(n, \tau)$

$$L_{i,\tau} = \left(s_i - \sum_{k=i+1}^{n_\tau} L_{k,\tau} D_{\tau-k} L_{k-i,\tau-i} \right) D_{\tau-i}^{-1}, \quad D_\tau = s_0 - \sum_{k=i+1}^{n_\tau} L_{k,\tau} D_{\tau-k} L_{k,\tau}. \tag{15}$$

The recursive evaluation requires to store $n+1$ numbers.

Using this decomposition, Peterka (1981a) proved the following Proposition.

Proposition 6.1 (Relationship of ARMAX with known MA to ARX model) *Using the filter (15), the* probability density function *(pdf) of the normal ARMAX model (4) equals to the pdf describing ARX model defined on filtered data (marked by ˜)*

$$f(y_t|u_t, d(t-1), \Theta_{ARX}, C) = \mathcal{N}_{y_t}(\mu_t, rD_t) = (2\pi r D_t)^{-0.5} \exp\left[-\frac{(y_t - \mu_t)^2}{2rD_t} \right] \tag{16}$$

$$\mu_t = \theta' \tilde{\psi}_t + \underbrace{\sum_{i=1}^{n_t} L_{t,i} \tilde{y}_{t-i}}_{\Delta_t}, \quad n_t = \min(n, t)$$

$$\tilde{y}_1 = y_1, \quad \tilde{y}_t + \sum_{i=1}^{n_t} L_{t,i} \tilde{y}_{t-i} = y_t, \quad \tilde{\psi}_t + \sum_{i=1}^{n_t} L_{t,i} \tilde{\psi}_{t-i} = \psi_t, \quad \tilde{\psi}_1 = \psi_1.$$

ARX model is parameterized by the unknown parameters $\Theta_{ARX} = (\theta, r)$ and acts on the filtered data vector $\tilde{\Psi} = [\tilde{y}, \tilde{\psi}']'$ obtained by passing the observed data vector $\Psi = [y, \psi']'$ through the filter determined by L, D entries.

The number of flops needed for filtering per single data sample is about $n(2n + n_a)$, where n_a is dimension of regression vector and n is the order of MA part.

The time-evolution of the filter, specified by $\{L_{t,i}\}$, and $\{D_t\}$, coincides with a spectral factorization of the MA-part. Consequently, no restriction on stability of the spectral factor determined by C is imposed. The time variations take place in spite of the fact that the noise covariance G is time invariant. The variations are data independent and are driven only by the time-invariant MA parameters. The evaluation of the filter is computationally cheap but the variations hinder the attempts to estimate the unknown C-parameters recursively. These simple properties are vital when we address the case of unknown C.

The transformation of the ARMAX model to ARX model allows us to use effectively Bayesian estimation that can be performed in real-time as the general functional recursion reduces to updating of fixed dimensional sufficient statistics. Note that the updating of the extended information matrix V_t, together with the degrees of freedom ν_t forming sufficient statistics for estimation of Θ_{ARX}, is equivalent to well-known recursive least squares. The updating is often poorly conditioned and its $L'DL$ *decomposition* has to be updated in order to counteract the induced numerical troubles (Peterka, 1981b).

References

[1] M. Kárný, J. Kadlec, and E.L. Sutanto. Quasi-bayes estimation applied to normal mixture. In *Preprints of the 3rd European IEEE workshop on Computer-Intensive Methods in Control and Data Processing, J. Rojìcek, M. Valeckovà, M. Kárný and K. Warwick,Eds.*, pages 77–82. ÚTIA AV ČR, Prague, September 1998.

[2] M. Kárný and R. Kulhavý. Structure determination of regression-type models for adaptive prediction and control. In J.C. Spall, editor, *Bayesian Analysis of Time Series and Dynamic Models*. Marcel Dekker, New York, 1988. chapter 12.

[3] L. Ljung. *System Identification-Theory for the User*. D. van Nostrand Company Inc., Prentice-hall. Englewood Cliffs, N.J, 1987.

[4] V. Peterka. Bayesian approach to system identification. In P. Eykhoff, editor, *Trends and Progress in System identification*, pages 239–304. Pergamon Press, Oxford, 1981.

[5] V. Peterka. Real-time parameter estimation and output prediction for ARMA type system models. *Kybernetika*, 17(6):526–533, 1981.

[6] V. Peterka. Self–tuning control with alternative sets of uncertain process models. In *Proceedings of IFAC Symposium on Adaptive Systems in Control and Signal Processing*, pages 409–414. Glasgow, UK, 1989.

[7] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixtures*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1985. ISBN 0 471 90763 4.

[8] V. Torczon. *Multi-directional Search: A Direct Search Algorithm for Parallel Machines*. Ph.D thesis, Rice University, Houston, Texas, USA, 1989.

[9] V. Torczon. On the convergence of the multidirectional search algorithm. *SIAM journal on optimization*, 1:123–145, 1991.