Mixture-Based Extension of the AR Model and its Recursive Bayesian Identification

Václav Šmídl*, Anthony Quinn,
Department of Electronic & Electrical Engineering,
University of Dublin, Trinity College,
Dublin 2. IRELAND.
Tel: +353–1–6081863; Fax: +353–1–6772442;
E-mail: smidl@mee.tcd.ie,aquinn@tcd.ie

EDICS: 2–SREP, 2–ESTM

***Abstract*— An extension of the AutoRegressive (AR) model is studied, which allows transformations and distortions on the regressor to be handled. Many important signal processing problems are amenable to this Extended AR (i.e. EAR) model. It is shown that Bayesian identification and prediction of the EAR model can be performed recursively, in common with the AR model itself. The EAR model does, however, require that the transformation be known. When it is unknown, the associated transformation space is represented by a finite set of candidates. What follows is a Mixture-based EAR model, i.e. the MEAR model. An approximate identification algorithm for MEAR is developed, using a restricted Variational Bayes (VB) procedure. It preserves the efficient recursive update of sufficient statistics. The MEAR model is applied to the robust identification of AR processes corrupted by outliers and burst noise respectively, and to click removal for speech.**

***Index Terms*— Bayesian identification, probabilistic mixtures, sufficient statistics, recursive identification, Variational Bayes, adaptive systems, filter-bank, burst noise.**

I. Introduction

Bayesian identification of a model is defined as evaluation of the posterior distribution of the model parameters [1], [2]. Recursive Bayesian identification is concerned with updating the parameter distribution as new data become available. A numerically efficient solution is possible for the class of models with conjugate parameter priors. The linear AutoRegressive (AR) model belongs to this class. Linear AR processes are widely applied in filtering [3], speech analysis [4], spectrum analysis [5], control [6], etc. However, the underlying assumptions (i.e. linear combination of measured values, and Gaussian distribution for the residue) are rarely met in practice. Physical models, typically requiring complex non-linear modelling, may then be used to fit the observed data. Attempts to extend the AR model itself have also been made [7], [8]. However, these solutions are computationally expensive and thus unsuitable for processing of large amounts of data or for on-line (real-time) identification. Typically, therefore, AR models continue to be used even in these cases.

In this text, we propose an extension to the AR model that preserves analytical tractability, allowing fast, recursive, on-line identification of the model. Recursive algorithms are important in on-line control applications [9], and for adaptive filtering [10]. In off-line cases, the emphasis on computational issues and recursive methods can also pay off, for example in the off-line processing of massive datasets [11].

The link between conjugacy and recursive Bayesian identification is reviewed in Section II and applied to the AR model in Section III. In Section IV, we extend to all possible models whose posterior distribution on parameters is as given in Section III. In Section V, we further extend the model to allow for *unknown* transformations of data. The price paid is loss of recursive conjugacy in the resulting probabilistic mixture. Conjugacy is restored via an approximation, optimal in the sense of Kullback-Leibler distance. It yields a numerically efficient identification procedure for this set of models. This is used in Section VI to de-noise AR processes corrupted by isolated outliers, and by burst noise, respectively. An application in speech analysis is also given. Discussion and Conclusions follow in Sections VII and VIII respectively.

II. Bayesian Recursive Identification

Our concern is with the inference of unknown model parameters at all observation times, $n = 1, 2, 3, \ldots$. The Bayesian perspective requires evaluation of a probability distribution on these unknowns at all n. This contrasts with the point estimation task, where unknowns are represented by a decision-theoretic certainty equivalent. Tractability of the identification task is assured when the parameter distribution is confined to the family of distributions that is *conjugate* to the observation model. A full review of the concept of conjugacy is available in [12] and briefly summarized next.

The data measured at time n are denoted by x_n, and the *history* of the system is defined as $\boldsymbol{X}_n = [x_1, x_2, \ldots, x_n]$. Let the data be generated by an *observation model* formalized as a probability density function (distribution), $f\left(x_n|\theta, \boldsymbol{X}_{n-1}\right)$, with $\boldsymbol{X}_0 = \{\}$ by assignment. This model is parameterized by unknown θ. Identification of the model is equivalent to evaluation of the posterior distribution of θ, $\forall n$. From Bayes' rule:

$$f\left(\theta|\boldsymbol{X}_n\right) \propto f\left(x_n|\theta, \boldsymbol{X}_{n-1}\right) f\left(\theta|\boldsymbol{X}_{n-1}\right). \qquad (1)$$

Since (1) is recursive, analytical tractability of the update is assured if distributions $f(\theta|\boldsymbol{X}_n)$ and $f(\theta|\boldsymbol{X}_{n-1})$ are of the same form. This is achieved if there exists a mapping, $s_n = s(\boldsymbol{X}_n), s_n \in \Re^q$, satisfying the condition

$$f(\theta|\boldsymbol{X}_n) = f(\theta|s_n). \tag{2}$$

$s(\cdot)$ is time-invariant and finite-dimensional ($q < \infty$), and s_n are known as the *sufficient statistics* [12]. Then, $f(\theta|\cdot)$ is said to be *conjugate* to the observation model, $f(x_n|\theta, \boldsymbol{X}_{n-1})$. Since (2) must be valid for $n = 0$, the prior, $f(\theta) = f(\theta|s_0)$, must also be conjugate. A conjugate distribution exists for every observation model in the exponential family [13]. Under (2), functional recursion (1) can be replaced by an algebraic recursion on s_n, achieving Bayesian identification of θ, $\forall n$, and guaranteeing a numerically stable procedure.

III. Review of Bayesian Identification for the Autoregressive (AR) Model

A univariate time-invariant AR model is of the form

$$x_n = -\sum_{k=1}^{p} a_k x_{n-k} + \sigma e_n, \tag{3}$$

where $p \geq 1$, e_n denotes the input and x_n the output (observation) of the system, as illustrated in Fig. 1 (left). The problem is to estimate fixed, unknown, real parameters, σ and $\mathbf{a} = [a_1, \ldots, a_p]'$, of this model. Here, $'$ denotes transposition. The classical solution to this problem is based on the Wiener criterion. Point estimates are obtained by solution of the *normal equations*. Two principal approaches to its solution are the covariance and correlation methods respectively [14]. Recursive solutions exist, such as the Recursive Least Squares (RLS) algorithm [9].

The Bayesian approach assumes that (3) is driven by white noise of Gaussian distribution, i.e. $f(e_n) = \mathcal{N}(0, 1)$. Then,

$$f(x_n|\mathbf{a}, \sigma, \mathbf{x}_n) = \mathcal{N}\left(-\mathbf{a}'\mathbf{x}_n, \sigma^2\right), \tag{4}$$

where $n > p$, and $\mathbf{x}_n = [x_{n-1} \ldots x_{n-p}]'$ is the regression vector at time n.

(4) belongs to the exponential family, and so both a conjugate prior and sufficient statistics are available. The parameter distribution which is conjugate to (4) is of the *Normal-inverse-Gamma* ($\mathcal{N}i\mathcal{G}$) type [12]:

$$\mathcal{N}i\mathcal{G}_{\mathbf{a},\sigma}(V, \nu) \equiv \frac{\sigma^{-\nu}}{\mathcal{I}_{\mathcal{N}i\mathcal{G}}(V, \nu)} \times \exp\left\{-\frac{1}{2}\sigma^{-2}\left[-1, \mathbf{a}'\right] V \left[-1, \mathbf{a}'\right]'\right\}, \tag{5}$$

$$\mathcal{I}_{\mathcal{N}i\mathcal{G}}(V, \nu) = \Gamma(0.5\nu)\, \lambda^{-0.5\nu} \left|V_{\mathbf{aa}}\right|^{-0.5} 2^{0.5p}, \tag{6}$$

$$V = \begin{bmatrix} V_{11} & V_{\mathbf{a}1}' \\ V_{\mathbf{a}1} & V_{\mathbf{aa}} \end{bmatrix}, \quad \lambda = V_{11} - V_{\mathbf{a}1}' V_{\mathbf{aa}}^{-1} V_{\mathbf{a}1}. \tag{7}$$

$\Gamma(\cdot)$ denotes the Gamma function [15], and (7) denotes a partitioning of $V \in \Re^{(p+1)\times(p+1)}$ into blocks, isolating V_{11}, the $(1,1)$ element. V, ν are the sufficient statistics of $\mathcal{N}i\mathcal{G}_{\mathbf{a},\sigma}(\cdot)$.

The statistics of the conjugate prior distribution, V_0, ν_0, are chosen to reflect our initial parameter knowledge. If we do not have any preferences, we use a diffuse (non-committal) distribution. Typically, $V_0 = \rho I_{p+1}$, $\nu_0 = \rho$, where I_{p+1} is the $(p+1)\times(p+1)$ identity matrix, and ρ is a small positive scalar. Substituting (4) into (1), and using (5) at time $n-1$, the posterior distribution at time $n > p$ is

$$f(\mathbf{a}, \sigma|\boldsymbol{X}_n) = \mathcal{N}i\mathcal{G}_{\mathbf{a},\sigma}(V_n, \nu_n), \tag{8}$$

$$V_n = V_{n-1} + \bar{\mathbf{x}}_n \bar{\mathbf{x}}_n' = V_0 + \sum_{i=p+1}^{n} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i', \tag{9}$$

$$\nu_n = \nu_{n-1} + 1 = \nu_0 + (n - p). \tag{10}$$

Here, $\bar{\mathbf{x}}_n = [x_n, \mathbf{x}_n']'$ is the extended regression vector. The outer product, $\bar{\mathbf{x}}_n \bar{\mathbf{x}}_n'$, will be called a *dyad* in this paper. Since the recursion begins at $n = p+1$, (9) and (10) are initialized with $V_p = V_0$ and $\nu_p = \nu_0$ respectively. This is equivalent to choosing a stationary distribution for $n \leq p$.

For many practical tasks, we need to derive moments of these distributions. They are:

$$\mathcal{E}(\mathbf{a}|\boldsymbol{X}_n) = V_{\mathbf{aa},n}^{-1} V_{\mathbf{a}1,n} = \hat{\mathbf{a}}_n, \tag{11}$$

$$\mathcal{E}\left(\sigma^2|\boldsymbol{X}_n\right) = \frac{\lambda_n}{\nu_n - p + 2} = \hat{\sigma^2}_n, \tag{12}$$

$$\mathcal{E}\left((\mathbf{a} - \hat{\mathbf{a}}_n)(\mathbf{a} - \hat{\mathbf{a}}_n)'|\boldsymbol{X}_n\right) = \frac{\lambda_n}{\nu_n - p} V_{\mathbf{aa},n}, \tag{13}$$

where the quantities in (7) have been evaluated at time n. These (and subsequent) formulae are also valid when x_n is a vector of measured data. The distributions of unknown variables are then the corresponding multivariate forms.

The Bayesian posterior moments (11)–(13) correspond to point estimates employed in the signal processing literature. (11), (12) are algorithmically identical to the results of the *covariance method* [14], and are valid $\forall n > p$, as derived. The Bayesian identification framework above yields the following extensions.

Computational Issues: a numerically efficient solution to (9), (11) is based on the LD decomposition [16]; i.e. $V_n = L_n D_n L_n'$, where L_n is lower triangular and D_n is diagonal. The update of the sufficient statistics (9) is replaced by recursions on L_n and D_n [17]. The resulting identification algorithm is then compact, efficient, and numerically stable.

Prediction: the one-step-ahead predictive distribution is given by the ratio of normalizing coefficients (6), a result established in general for the exponential family in [12]. For the AR model,

$$f(x_{n+1}|\boldsymbol{X}_n) = \frac{\mathcal{I}_{\mathcal{N}i\mathcal{G}}\left(V_n + \bar{\mathbf{x}}_{n+1}\bar{\mathbf{x}}_{n+1}', \nu_n + 1\right)}{\sqrt{2\pi}\mathcal{I}_{\mathcal{N}i\mathcal{G}}(V_n, \nu_n)}, \tag{14}$$

using (6). This is the Student t-distribution with $\nu_n - p + 2$ degrees of freedom. The mean value of this distribution is readily found to be

$$\mathcal{E}(x_{n+1}|\boldsymbol{X}_n) = \hat{\mathbf{a}}_n \mathbf{x}_{n+1} = \hat{x}_{n+1}, \tag{15}$$

using (11), and is therefore equal to the intuitively appealing result from classical theory [14].

Fig. 1. Block diagrams of the AutoRegressive (AR) (left) and Extended AR (EAR) (right) models.

Model Order Determination: when the model order, p, in (3) is unknown, it must be handled as a discrete random variable in the Bayesian identification framework, and represented as such in the notation. Application of the chain rule $n-p$ times, using (14), and Bayes' rule yield the following posterior distribution of p:

$$f\left(p|\boldsymbol{X}_n\right) \propto \mathcal{I}_{\mathcal{N}i\mathcal{G}}\left(V_n, \nu_n\right) f\left(p\right). \tag{16}$$

$f\left(p\right)$ denotes the prior distribution of model order, typically chosen as uniform on integer support $p \in \{1, \ldots, p_{\max}\}$, where $p_{\max} \geq 1$ is a hyperparameter.

Exponential forgetting: the assumption of constant parameter values is rarely met in practice. In many applications, however, a complete model of parameter variations—such as that required in [1], [18]—is not known. The full Bayesian parametric solution must then be replaced by heuristic techniques. The standard batch (off-line) algorithm uses windowing [19]. Alternatively, the concept of forgetting [20] is used in adaptive signal processing [21] and recursive estimation [9].

A Bayesian treatment of forgetting was developed in [22]. There, the missing model of parameter evolution is handled via a probabilistic operator:

$$f\left(\theta_n|\boldsymbol{X}_{n-1}\right) \propto \left[f\left(\theta_{n-1}|\boldsymbol{X}_{n-1}\right)_{\theta_n}\right]^{\phi_n} \tilde{f}\left(\theta_n\right)^{1-\phi_n}. \tag{17}$$

The notation $f\left(\cdot\right)_{\theta_n}$ indicates the replacement of the argument of $f\left(\cdot\right)$ by θ_n, where θ_n is the time-variant unknown parameter set at time n. $\tilde{f}\left(\cdot\right)$ is a chosen (known) alternative distribution, expressing auxiliary knowledge about θ_n at time n. Coefficient ϕ_n, $0 \leq \phi_n \leq 1$ is known as the forgetting factor. The $\mathcal{N}i\mathcal{G}$ conjugate family (5) is closed under the convex combination (i.e. geometric mean) in (17), yielding another member of the $\mathcal{N}i\mathcal{G}$ family. Hence, we choose $\tilde{f}$, to be $\mathcal{N}i\mathcal{G}\left(\tilde{V}, \tilde{\nu}\right)$, with time-invariant parameters, $\tilde{V}$ and $\tilde{\nu}$. The posterior distribution is then given by (8), with statistics

$$V_n = \phi_n V_{n-1} + \bar{\mathbf{x}}_n \bar{\mathbf{x}}_n' + \left(1-\phi_n\right)\tilde{V}, \tag{18}$$

$$\nu_n = \phi_n \nu_{n-1} + 1 + \left(1-\phi_n\right)\tilde{\nu}, \tag{19}$$

with $n > p$. Note that $\tilde{V}$, $\tilde{\nu}$ play an important role in these identification recursions, as they are injected at each step. In order to minimize their influence, we can choose $\tilde{f}$ diffuse, having, for example, the same parameter values as the prior: $\tilde{V} = V_0, \tilde{\nu} = \nu_0$.

IV. The Extended AutoRegressive (EAR) Model

In this section, the largest set of models is proposed for which the algorithms in Section III remain valid. The favourable algorithmic properties for the AR model are based on the elegant recursive form (9), (10) of the $\mathcal{N}i\mathcal{G}$ sufficient statistics (5). This distribution remains $\mathcal{N}i\mathcal{G}$ if there is a change in the conditioning variables of (8), or even an increase in the number of variables in the conditioning set by an auxiliary vector of *known* exogenous terms, W_n:

$$f\left(\mathbf{a}, \sigma|G(\boldsymbol{X}_n, W_n)\right) = f\left(\mathbf{a}, \sigma|\bar{\mathbf{y}}_n\right) = \mathcal{N}i\mathcal{G}_{\mathbf{a},\sigma}\left(V_n, \nu_n\right). \tag{20}$$

Here, the *known* transformation, G, is in general a set of $p+1$ nonlinear functions:

$$y_n = g_0\left(\boldsymbol{X}_n, W_n\right), \tag{21}$$

$$y_{i;n} = g_i\left(\boldsymbol{X}_{n-1}, W_n\right),\ i = 1, 2, \ldots, p, \tag{22}$$

where the latter are the p transformed regressors at time n. The regression vector is therefore $\mathbf{y}_n = [y_{1;n}, y_{2;n}, \ldots, y_{p;n}]'$, and the extended regression vector at time n is

$$\bar{\mathbf{y}}_n = \left[y_n, \mathbf{y}_n'\right]' = G\left(\boldsymbol{X}_n, W_n\right). \tag{23}$$

This is illustrated in Fig. 1 (right). Auxiliary variable, W_n, may contain any known variables, such as the time index, n, for time-variant systems, or a measured external (exogenous) signal, etc. (20) implies an AR structure (3) defined with respect to an internal variable, y_n. Hence, the distribution of observations is obtained by transformation of (4):

$$f\left(x_n|\mathbf{a}, \sigma, \boldsymbol{X}_{n-1}, G\right) = \left|J_n\left(x_n\right)\right| \mathcal{N}\left(-\mathbf{a}'\mathbf{y}_n, \sigma^2\right), \tag{24}$$

where $J_n\left(\cdot\right)$ is the Jacobian of transformation g_0 (21); i.e. $J_n\left(x_n\right) = \frac{\partial g_0}{\partial x_n}$. This creates an additional restriction that g_0 be a differentiable, one-to-one mapping for each setting of W_n. Moreover, g_0 must explicitly be a function of x_n in order that $J_n \neq 0$. This ensures uncertainty propagation from e_n to x_n (Fig. 1).

Bayesian identification with this model is, by design, of the same form as for the AR model (4)–(13), replacing the dyadic update of V_n in (9) with one in terms of $\bar{\mathbf{y}}_n$:

$$V_n = V_{n-1} + \bar{\mathbf{y}}_n \bar{\mathbf{y}}_n',\ n > p. \tag{25}$$

The update for ν_n (10) is unchanged.

The EAR model class includes the following important cases [2]: (i) the ARMA model with a *known* MA part; (ii) the ARX model, i.e. AR with exogenous observed input, w_n; (iii) an AR process, y_n, observed via a *known* bijective non-linear transformation, $x_n = g_0^{-1}(y_n)$; (iv) the incremental AR process with the regression defined on increments of the measurement process.

Prediction: this is given by (14) with $\bar{\mathbf{x}}_{n+1}$ replaced by $\bar{\mathbf{y}}_{n+1}$ and V_n given by (25). Using (21):

$$f\left(x_{n+1}|\boldsymbol{X}_n, G\right) = \left|J_{n+1}\left(x_{n+1}\right)\right| f\left(g_0\left(x_{n+1}, \boldsymbol{X}_n\right)|\boldsymbol{X}_n, G\right). \tag{26}$$

Note that (26) remains a Student t-distribution (14) iff Jacobian J_{n+1} is independent of x_{n+1}, i.e. for linear transformations, g_0 (21).

Model Structure Determination: the structure of the EAR model is no longer dependent solely on p (Fig. 1) but on the whole transformation, G. The structure determination problem is then one of calculating the *a posteriori* probabilities of choices, $\{G_1, G_2, \ldots, G_c\}$, from a finite set. From Bayes' rule:

$$f\left(G_i|\boldsymbol{X}_n\right) \propto f\left(\boldsymbol{X}_{n\backslash p(i)}|\boldsymbol{X}_{p(i)}, G_i\right) f\left(G_i\right), \; i = 1, \ldots, c, \tag{27}$$

where the first term on the right-hand-side is formed from $n - p(i)$ terms of the kind in (26), and $\boldsymbol{X}_{n\backslash p(i)} = \left[x_{p(i)+1}, \ldots, x_n\right]$, where $p(i)$ is the regression length (i.e. order (3)) of the ith EAR model.

V. The Mixture-Based EAR (MEAR) Model

We now relax the EAR assumption (20) which requires G to be known. Instead, we consider a finite set, $\mathbf{G}$, of possible transformations, called the *filter-bank*:

$$\mathbf{G} = \left\{G_i, i = 1, \ldots, c\right\}. \tag{28}$$

We assume that the observation, x_n, at each time n was generated by one element of $\mathbf{G}$. (24) can be rewritten as

$$f\left(x_n|\mathbf{a}, \sigma, \boldsymbol{X}_{n-1}, \mathbf{G}, \boldsymbol{l}_n\right) = \prod_{i=1}^{c} f\left(x_n|\mathbf{a}, \sigma, \boldsymbol{X}_{n-1}, G_i\right)^{l_{i;n}}. \tag{29}$$

Here, the active transformation is labelled by a new discrete auxiliary variable, $\boldsymbol{l}_n = \left[l_{1;n}, \ldots, l_{c;n}\right]'$, with possible states $\boldsymbol{l}_n \in \{\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_c\}$. Here, $\boldsymbol{\varepsilon}_i$ is the ith elementary basis vector:

$$\boldsymbol{\varepsilon}_i = \boldsymbol{\delta}_c\left(i\right) = \left[\delta\left(i-1\right), \ldots, \delta\left(i-c\right)\right]', \; i = 1, \ldots, c, \tag{30}$$

$$\delta\left(\rho\right) = \begin{cases} 1, & \text{if } \rho = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

$\boldsymbol{l}_n$ constitutes a hidden field which we model via a first-order homogeneous Markov chain, with transition matrix $T \in [0,1]^{c\times c}$:

$$f\left(\boldsymbol{l}_n|T, \boldsymbol{l}_{n-1}\right) = Mu_{\boldsymbol{l}_n}\left(T\boldsymbol{l}_{n-1}\right) = \prod_{i=1}^{c}\prod_{j=1}^{c} t_{i,j}^{l_{i;n}l_{j;n-1}}; \tag{32}$$

i.e. $\Pr\left(\boldsymbol{l}_n = \boldsymbol{\varepsilon}_i|T, \boldsymbol{l}_{n-1} = \boldsymbol{\varepsilon}_j\right) = t_{i,j}$, the ijth element of T. $Mu_{\boldsymbol{l}_n}(\cdot)$ denotes the multinomial distribution, whose conjugate distribution is Dirichlet [23] with matrix parameter $\Psi \in \left(\Re^+\right)^{c\times c}$:

$$f\left(T|\Psi\right) = \mathcal{D}i_T\left(\Psi\right) = \frac{1}{\mathcal{I}_{\mathcal{D}i}\left(\Psi\right)}\prod_{i=1}^{c}\prod_{j=1}^{c} t_{i,j}^{\psi_{i,j}-1}, \tag{33}$$

$$\mathcal{I}_{\mathcal{D}i}\left(\Psi\right) = \frac{\prod_{i=1}^{c}\prod_{j=1}^{c}\Gamma\left(\psi_{i,j}\right)}{\Gamma\left(\sum_{i=1}^{c}\sum_{j=1}^{c}\psi_{i,j}\right)}, \tag{34}$$

$$\hat{T} = \mathcal{E}\left(T|\Psi\right) = \frac{1}{\sum_{i=1}^{c}\sum_{j=1}^{c}\psi_{i,j}}\Psi. \tag{35}$$

The extended observation model is

$$f\left(x_n, \boldsymbol{l}_n|\mathbf{a}, \sigma, T, \boldsymbol{X}_{n-1}, \mathbf{G}, \boldsymbol{l}_{n-1}\right) = f\left(x_n|\mathbf{a}, \sigma, \boldsymbol{X}_{n-1}, \mathbf{G}, \boldsymbol{l}_n\right) f\left(\boldsymbol{l}_n|T, \boldsymbol{l}_{n-1}\right), \tag{36}$$

which may be evaluated via (29), (32). Hence, the conditioning model parameter set is $\mathbf{a}, \sigma$ augmented by $T, \boldsymbol{l}_{n-1}$. Marginalization over $\boldsymbol{l}_n$ yields an observation model in the form of a probabilistic Mixture of EAR components with common AR parameterization, $\mathbf{a}, \sigma$:

$$f\left(x_n|\mathbf{a}, \sigma, T, \boldsymbol{X}_{n-1}, \mathbf{G}, \boldsymbol{l}_{n-1}\right) = \sum_{i=1}^{c} f\left(x_n, \boldsymbol{l}_n = \boldsymbol{\varepsilon}_i|\mathbf{a}, \sigma, T, \boldsymbol{X}_{n-1}, \mathbf{G}, \boldsymbol{l}_{n-1}\right). \tag{37}$$

This defines the MEAR model. Next, consider the posterior distribution of model parameters at time $n-1$, i.e. $f\left(\mathbf{a}, \sigma, T, \boldsymbol{l}_{n-1}|\boldsymbol{X}_{n-1}, \mathbf{G}\right)$. This is updated by (36) according to Bayes' rule:

$$f\left(\mathbf{a}, \sigma, T, \boldsymbol{l}_n, \boldsymbol{l}_{n-1}|\boldsymbol{X}_n, \mathbf{G}\right) \propto f\left(\mathbf{a}, \sigma, T, \boldsymbol{l}_{n-1}|\boldsymbol{X}_{n-1}, \mathbf{G}\right) \times f\left(x_n, \boldsymbol{l}_n|\mathbf{a}, \sigma, T, \boldsymbol{X}_{n-1}, \mathbf{G}, \boldsymbol{l}_{n-1}\right). \tag{38}$$

The update introduces the extra random variable, $\boldsymbol{l}_n$. Hence, the parameter distributions at times n and $n-1$ have different functional forms, violating conjugacy. After m updates, m random variables will have been generated, with c^m possible states. This renders the update (38) unsuitable for on-line identification. We overcome this problem via the following conditional independence approximation of the posterior distribution at time n (38):

$$\overline{f}\left(\mathbf{a}, \sigma, T, \boldsymbol{l}_n, \boldsymbol{l}_{n-1}|\boldsymbol{X}_n, \mathbf{G}\right) = \overline{f}\left(\mathbf{a}, \sigma, T|\boldsymbol{X}_n, \mathbf{G}\right) \times \overline{f}\left(\boldsymbol{l}_n|\boldsymbol{X}_n, \mathbf{G}\right)\overline{f}\left(\boldsymbol{l}_{n-1}|\boldsymbol{X}_n, \mathbf{G}\right), \tag{39}$$

where the $\overline{f}(\cdot)$ denote approximating distributions. Using (39) at both n and $n-1$ (i.e. for the first two terms in (38) respectively), we see that $\overline{f}\left(\mathbf{a}, \sigma, T|\boldsymbol{X}_{n-1}, \mathbf{G}\right)$ is updated in the step from $n-1$ to n independently of the label sequence $\boldsymbol{l}_n$, avoiding the exponential explosion referred to above. An optimal approximation within the class (39) may be found via minimization of the Kullback-Leibler (KL) distance [24]:

$$KL\left(\overline{f}\left(\theta\right)||f\left(\theta\right)\right) = \int_{\theta}\overline{f}\left(\theta\right)\ln\frac{\overline{f}\left(\theta\right)}{f\left(\theta\right)}d\theta. \tag{40}$$

Here, for convenience, θ denotes $\mathbf{a}, \sigma, T, \boldsymbol{l}_n, \boldsymbol{l}_{n-1}$. KL optimization of a conditional independence model (39) defines

the *Variational Bayes* (VB) procedure [7]. The VB procedure requires iterations between the optimizing functions (39) at each time n, rendering it unsuitable for on-line identification. To avoid this, we assign the third and second terms on the right-hand-side of (39) respectively, as follows:

$$\overline{f}\left(\boldsymbol{l}_{n-1}|\boldsymbol{X}_n,\mathbf{G}\right)=f\left(\boldsymbol{l}_{n-1}|\boldsymbol{X}_{n-1},\mathbf{G}\right), \tag{41}$$

$$\overline{f}\left(\boldsymbol{l}_{n}|\boldsymbol{X}_n,\mathbf{G}\right)=f\left(\boldsymbol{l}_{n}|\boldsymbol{X}_n,\mathbf{G}\right). \tag{42}$$

These assignments are obtained from marginalization of the *exact* model (38). Then, the only functional variant in (39) is the first term on the right-hand-side. It may be optimized by minimization of the KL distance (40) from (39) to (38), subject to constraints (41), (42):

$$\overline{f}\left(\mathbf{a},\sigma,T|\cdot\right)=\arg\min_{\overline{f}(\mathbf{a},\sigma,T|\cdot)} KL\left[\overline{f}\left(\mathbf{a},\sigma,T,\boldsymbol{l}_n,\boldsymbol{l}_{n-1}|\cdot\right)||f\left(\mathbf{a},\sigma,T,\boldsymbol{l}_n,\boldsymbol{l}_{n-1}|\cdot\right)\right]. \tag{43}$$

It is easy to show that the unique solution of (43) is

$$\overline{f}\left(\mathbf{a},\sigma,T|\boldsymbol{X}_n,\mathbf{G}\right)\propto \exp\mathcal{E}_{f(\boldsymbol{l}_n|\cdot)f(\boldsymbol{l}_{n-1}|\cdot)}\left[\ln f\left(\mathbf{a},\sigma,T,\boldsymbol{l}_n,\boldsymbol{l}_{n-1}|\boldsymbol{X}_n,\mathbf{G}\right)\right]. \tag{44}$$

Here, $\mathcal{E}_{..}\left[\cdot\right]$ denotes expectation with respect to (41) and (42). Substituting (29) and (32), via (36), into (38), and the result into (44), we obtain

$$\overline{f}\left(\mathbf{a},\sigma,T|\boldsymbol{X}_n,\mathbf{G}\right)\propto f\left(\mathbf{a},\sigma,T|\boldsymbol{X}_{n-1},\mathbf{G}\right)\times \prod_{i=1}^{c} f\left(x_n|\mathbf{a},\sigma,\boldsymbol{X}_{n-1},G_i\right)^{w_{i;n}}\prod_{i=1}^{c}\prod_{j=1}^{c}t_{i,j}^{w_{i;n}w_{j;n-1}}, \tag{45}$$

where

$$\boldsymbol{w}_n=\mathcal{E}\left(\boldsymbol{l}_n|\boldsymbol{X}_n,\mathbf{G}\right). \tag{46}$$

(45) is a KL-optimized approximate update of the MEAR model parameter inference from time $n-1$ to n. We now design the parameter distribution to be self-replicating under this update. The second term on the right-hand-side of (45) is a geometric mean of Gaussian distributions (24), being therefore Gaussian with $\mathcal{N}i\mathcal{G}$ conjugate distribution (5). The third term is multinomial (32), with Dirichlet conjugate distribution (33). The required conjugate distribution at time $n-1$, subject to approximation (39), is therefore

$$f\left(\mathbf{a},\sigma,T|\boldsymbol{X}_{n-1},\mathbf{G}\right)=\overline{f}\left(\mathbf{a},\sigma,T|\boldsymbol{X}_{n-1},\mathbf{G}\right)= \mathcal{N}i\mathcal{G}_{\mathbf{a},\sigma}\left(V_{n-1},\nu_{n-1}\right)Di_T\left(\Psi_{n-1}\right). \tag{47}$$

Substituting (47) into (45) yields the update equations:

$$\overline{f}\left(\mathbf{a},\sigma,T|\boldsymbol{X}_n,\mathbf{G}\right)=\mathcal{N}i\mathcal{G}_{\mathbf{a},\sigma}\left(V_n,\nu_n\right)Di_T\left(\Psi_n\right), \tag{48}$$

$$V_n=V_{n-1}+\sum_{i=1}^{c}w_{i;n}\bar{\mathbf{y}}_{i,n}\bar{\mathbf{y}}_{i,n}', \tag{49}$$

$$\nu_n=\nu_{n-1}+1, \tag{50}$$

$$\Psi_n=\Psi_{n-1}+\boldsymbol{w}_n\boldsymbol{w}_{n-1}', \tag{51}$$

with $n>p$. Update (49) is a weighted sum of dyads, each generated respectively from the extended regressor, $\bar{\mathbf{y}}_{i,n}$, of the ith filter in (28). Note that (49) is similar to the result derived in [25], using the *Quasi-Bayes* (QB) principle [26].

It remains to evaluate $\boldsymbol{w}_n$ (46). The first term on the right-hand-side of (38) may be expanded using the chain rule, and (41), (47) then substituted. Furthermore, (29) and (32) may be substituted into the second term on the right-hand-side of (38). Then, integrating over $\mathbf{a},\sigma,T,\boldsymbol{l}_{n-1}$, it follows that

$$\overline{f}\left(\boldsymbol{l}_n|\boldsymbol{X}_n,\mathbf{G}\right)=Mu_{\boldsymbol{l}_n}\left(\boldsymbol{w}_n\right), \tag{52}$$

$$w_{i;n}\propto\mathcal{I}_{\mathcal{N}i\mathcal{G}}\left(V_{n-1}+\bar{\mathbf{y}}_{i,n}\bar{\mathbf{y}}_{i,n}',\nu_{n-1}+1\right)\times \sum_{j=1}^{c}w_{j;n-1}\mathcal{I}_{\mathcal{D}i}\left(\Psi_{n-1}+\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'\right),\ \ i=1\dots c. \tag{53}$$

Finally, (41) is found by replacing n by $n-1$ in (52).

Computational Issues: the implied MEAR identification algorithm (49)–(51) requires about $2c$ times more operations—c for (53) and c for (49)—than the standard AR procedure (9), (10). Once again, the efficient LD decomposition of V_n may be exploited (Section III).

MEAR-based Prediction: the MEAR predictor can be found by marginalization, using (36) (replacing n by $n+1$), (48), and the chain rule:

$$f\left(x_{n+1}|\boldsymbol{X}_n,\mathbf{G}\right)=\sum_{i=1}^{c}\sum_{j=1}^{c}\hat{t}_{i,j}w_{j;n-1}f\left(x_{n+1}|\boldsymbol{X}_n,G_i\right), \tag{54}$$

where $\hat{t}_{i,j}$ is the ijth element of (35). (54) is therefore a mixture of EAR predictors (26). All non-central moments—e.g. the expected value, $\hat{x}_{n+1}$—of (54) can be obtained as the weighted algebraic mean of non-central moments of the candidates. This does not hold for the central moments [27].

MEAR Model Structure Determination: the key restriction of the MEAR model—namely, common AR parameters $\mathbf{a},\sigma$ (37)—implies that all filter candidates, $G_i\in\mathbf{G}$ (28), must have the same dimension, p (Fig. 1). The identification of the MEAR model does not provide inference of p, and additional treatment is required. However, various choices of filter-bank, $\mathbf{G}$, can be tested via (54), in the same way as for the EAR model (27).

Exponential forgetting: the assumption of a stationary parameter set, $\mathbf{a},\sigma,T$, can be relaxed by means of the probabilistic operator in (17). The prior for the update (47) is then replaced by

$$\overline{f}\left(\mathbf{a}_n,\sigma_n,T_n|\boldsymbol{X}_{n-1},\mathbf{G}\right)\propto \left[\overline{f}\left(\mathbf{a}_{n-1},\sigma_{n-1}|\boldsymbol{X}_{n-1},\mathbf{G}\right)_{\substack{\mathbf{a}_n\\ \sigma_n}}\right]^{\phi_{\mathcal{N}i\mathcal{G}}}\tilde{f}\left(\mathbf{a}_n,\sigma_n\right)^{1-\phi_{\mathcal{N}i\mathcal{G}}}\times \left[\overline{f}\left(T_{n-1}|\boldsymbol{X}_{n-1},\mathbf{G}\right)_{T_n}\right]^{\phi_{\mathcal{D}i}}\tilde{f}\left(T_n\right)^{1-\phi_{\mathcal{D}i}}, \tag{55}$$

where we have used the notation in (17). Two time-invariant forgetting factors, $\phi_{\mathcal{N}i\mathcal{G}}$ and ϕ_{Di}, are chosen to reflect the conditional independence in (47). The alternative distributions, $\tilde{f}$, are assumed to be time-invariant, and are chosen as $\mathcal{N}i\mathcal{G}$ (5) and $\mathcal{D}i$ (33), respectively, to ensure that (55) is self-replicating (conjugate) under the KL-optimized update (45). The recursions on sufficient statistics (49)–(51) are then reformulated

as ($n > p$):

$$V_n = \phi_{\mathcal{N}i\mathcal{G}} V_{n-1} + \sum_{i=1}^{c} w_{i;n} \bar{\mathbf{y}}_{i,n} \bar{\mathbf{y}}'_{i,n} + (1 - \phi_{\mathcal{N}i\mathcal{G}}) \tilde{V}, \quad (56)$$

$$\nu_n = \phi_{\mathcal{N}i\mathcal{G}} \nu_{n-1} + 1 + (1 - \phi_{\mathcal{N}i\mathcal{G}}) \tilde{\nu}, \quad (57)$$

$$\Psi_n = \phi_{\mathcal{D}i} \Psi_{n-1} + \boldsymbol{w}_n \boldsymbol{w}'_{n-1} + (1 - \phi_{\mathcal{D}i}) \tilde{\Psi}, \quad (58)$$

where $\tilde{V}, \tilde{\nu}$ and $\tilde{\Psi}$ are the parameters of the alternative distributions, and $\boldsymbol{w}_n$ is given by (53) adapted appropriately via (56)–(58).

VI. Applications in the Robust Identification of Corrupted AR Processes

A. Identification of an Outlier-corrupted AR Process

We consider the problem of isolated outliers. These are not modelled by (3), since the outlier-affected observed value does not become an element of any future regression. Hence, the autoregressive variable, z_n, is *unobserved*, and the observation process is

$$x_n = z_n + \omega_n, \quad (59)$$

where ω_n denotes a possible outlier at time n. For an *isolated* outlier, it holds that

$$\Pr[\omega_{n\pm i} = 0 | \omega_n \neq 0] = 1, \ i = 1, \ldots, p. \quad (60)$$

The AR model is identified via $f(\mathbf{a}, \sigma | \boldsymbol{X}_n)$ (8) (i.e. *not* via $f(\mathbf{a}, \sigma | \mathbf{Z}_n)$) and so the outlier has influence if and only if it enters the extended regressor $\bar{\mathbf{x}}_n$ (9). Since $\bar{\mathbf{x}}_n$ is of finite length, $p+1$, and since the outliers are isolated, a finite number of mutually exclusive scenarios can be defined. Each of these scenarios can be expressed via an EAR model and combined together using the MEAR approach, as follows.

a) None of the values in $\bar{\mathbf{x}}_n$ is affected by an outlier: i.e. $x_{n-i} = z_{n-i}, i = 0, \ldots p$. G_1 is then the unity transformation: $\bar{\mathbf{y}}_n = \bar{\mathbf{x}}_n$. For ease of notation, we have dropped the component index, i, in $\bar{\mathbf{y}}_{i;n}$, here and in what follows.

b) The observed value, x_n, is affected by an outlier: from (60), all delayed values are unaffected; i.e. $x_{n-i} = z_{n-i}, i = 1, \ldots p$. For convenience, ω_n can be expressed as $\omega_n = h_n \sigma e_n$, where h_n is an unknown multiplier of the realized AR residual (3). From (3), (59):

$$x_n = -\mathbf{a}' \mathbf{x}_n + (1 + h_n) \sigma e_n.$$

Dividing across by $(1 + h_n)$ reveals the appropriate EAR transformation (23):

$$G_2: \ \bar{\mathbf{y}}_n = \frac{1}{1 + h_n} \bar{\mathbf{x}}_n. \quad (61)$$

G_2 is parameterized by h_n, with constant Jacobian, $J_2 = \frac{1}{1+h_n}$ (24).

c) The k-steps-delayed observation, x_{n-k}, is affected by an outlier, $k \in \{1, \ldots, p\}$: in this case, the known transformation should replace this value by an interpolant, $\hat{z}_{n-k}$, which is known at time n. The set of transformations for each k is then

$$G_{2+k}: \ \bar{\mathbf{y}}_n = \bar{\mathbf{x}}_n + \boldsymbol{\delta}_{p+1}(k+1)(\hat{z}_{n-k} - x_{n-k}),$$

where vector $\boldsymbol{\delta}_{\cdot}(\cdot)$ is defined in (30). G_{2+k} is parameterized by $\hat{z}_{n-k}$, with Jacobian $J_{2+k} = 1$.

We have described an exhaustive set of $c = p + 2$ filters, G_i, transforming the observation regressors, $\bar{\mathbf{x}}_n$, to EAR regressors, $\bar{\mathbf{y}}_n$, for which the AR model (3) is valid, via (24). Parameters h_n and $\hat{z}_{n-k}$ must be defined. We choose the parameter of G_2 to be a known fixed $h_n = h$. Alternatively, if the variance of outliers is known to vary significantly, we can split G_2 into $u > 1$ candidates with respective fixed values $h_{(1)} < h_{(2)} < \ldots < h_{(u)}$. Next, $\hat{z}_{n-k}$ is chosen as the k-steps-delayed value of the following causal reconstruction:

$$\hat{z}_n = \sum_{j=1}^{c} \mathcal{E}\{z_n | \boldsymbol{l}_n = \boldsymbol{\varepsilon}_j\} f(\boldsymbol{l}_n = \boldsymbol{\varepsilon}_j | \boldsymbol{X}_n, \mathbf{G})$$

$$= x_n \left(\sum_{j=1, j\neq 2}^{c} w_{j;n} \right) - w_{2;n} \hat{\mathbf{a}}'_{n-1} \mathbf{x}_n. \quad (62)$$

Here, we are using (15), (52), and the fact that $z_n = x_n$ for all transformations except G_2.

A second-order, stable, stationary AR process with parameters $\mathbf{a} = [-1.8, 0.98]'$, $\sigma = 0.01$, was simulated with a random outlier at every 30th sample. The total number of samples was $N = 200$, and $u = 1$ (i.e. $c = 4$), with $h = 10$. Identification results (using stationary identification (49)–(51) with non-informative priors) are illustrated in Fig. 2 along with the reconstruction (62). When an outlier occurs, all candidate filters are sequentially used (middle diagram). Thus, the regressors, $\bar{\mathbf{y}}_{i,n}$, containing the outlier are sequentially removed from (49) very effectively. The marginal distribution of $\mathbf{a}$ is Student-t, $\forall n > p$, with moments given by (11) and (13). The terminal moments are illustrated in Fig 2 (right), via the 95% Highest Posterior Density (HPD) ellipses [12]. The MEAR inference of $\mathbf{a}$ is close that using (8) with *uncorrupted* data. Robust identification has therefore been achieved. In effect, the procedure has unified the pre-processing and identification tasks for the AR model.

B. Identification of a Burst-noise-corrupted AR Process

A burst noise scenario requires more than one outlier to be considered in the regressor. We transform the underlying AR model (3) into state-space form [9]:

$$\mathbf{z}_{n+1} = A\mathbf{z}_n + \mathbf{r}\sigma e_n, \quad (63)$$

$$A = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_p \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}, \ \mathbf{r} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (64)$$

such that $A \in \Re^{p \times p}$ and $\mathbf{r} \in \Re^{p \times 1}$. The observation process with burst noise is modelled as

$$x_n = \boldsymbol{c}' \mathbf{z}_n + h_n \sigma \xi_n, \quad (65)$$

where $\boldsymbol{c} = [1, 0, \ldots, 0]' \in \Re^{p \times 1}$, and ξ_n is distributed as $\mathcal{N}(0, 1)$, independent of e_n. $h_n\sigma$ denotes the time-dependent standard deviation of the noise which is assumed strictly positive during any burst, and is zero otherwise. Note that (63),

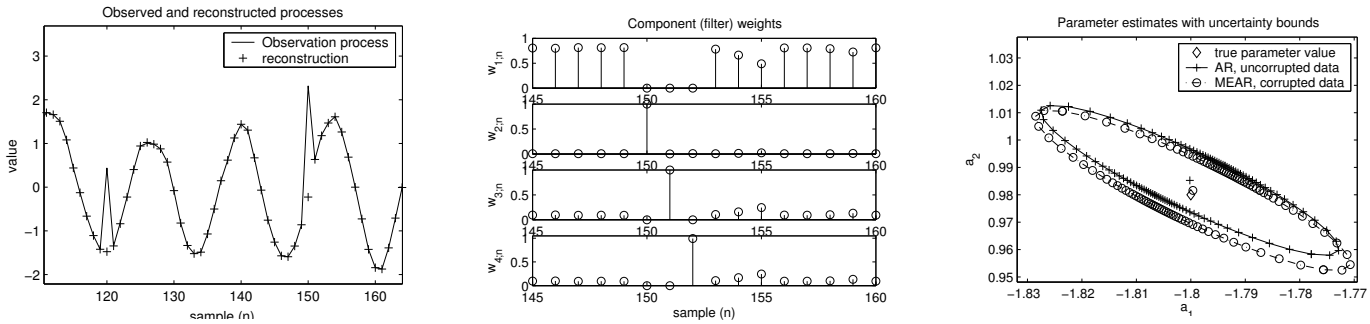

Fig. 2. Reconstruction and identification of an AR(2) process corrupted by isolated outliers. **Left**: comparison of the measured and reconstructed signal. **Middle**: weights (53) of the MEAR components (outlier occurred at $n = 150$). **Right**: comparison of the terminal ($n = N = 200$) moments of the marginal posterior distributions of $\mathbf{a}$.

(64) is identical to the AR model in the previous example (3). The only modelling difference is in the observation process (65) compared to (59), (60). We identify a finite number of mutually exclusive scenarios, each of which can be expressed using an EAR model:

a) The AR process is observed without distortion: i.e. $h_n = h_{n-1} = \ldots = h_{n-p} = 0$. Formally, $G_1 : \ \bar{\mathbf{y}}_n = \bar{\mathbf{x}}_n$.

b) The measurements are all affected by constant-deviation burst noise: i.e. $h_n = h_{n-1} = \ldots = h_{n-p} = h$. The state-space model (63), (65) is now defined by the joint distribution

$$f\left(\mathbf{z}_n, x_n | \mathbf{a}, \sigma, \mathbf{z}_{n-1}, h\right) = \mathcal{N}\left(\begin{bmatrix} A\mathbf{z}_{n-1} \\ \boldsymbol{c}'\mathbf{z}_n \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{r}\mathbf{r}' & 0 \\ 0 & h^2 \end{bmatrix}\right). \quad (66)$$

(66) cannot be directly modelled as an EAR process because it contains unobserved state vector $\mathbf{z}_n$. Using standard Kalman Filter (KF) theory [2], [9], we can multiply terms together of the kind in (66), and then integrate over the unobserved trajectory—i.e. over $\{\mathbf{z}_{p+1}, \ldots, \mathbf{z}_n\}$— to obtain the direct observation model:

$$f\left(x_n | \mathbf{a}, \sigma, \boldsymbol{X}_{n-1}, h\right) = \mathcal{N}\left(\mathbf{a}\hat{\mathbf{z}}_n, \sigma^2 q_n\right). \quad (67)$$

The moments in (67) are defined recursively as follows:

$$q_n = h^2 + \boldsymbol{c}' S_{n-1} \boldsymbol{c}, \quad (68)$$

$$Z_n = S_{n-1} - q_n^{-1} \left(S_{n-1}\boldsymbol{c}\right)\left(S_{n-1}\boldsymbol{c}\right)', \quad (69)$$

$$\hat{\mathbf{z}}_n = A\hat{\mathbf{z}}_{n-1} + h^{-2} Z_n \boldsymbol{c}\left(x_n - \boldsymbol{c}' A \hat{\mathbf{z}}_{n-1}\right), \quad (70)$$

$$S_n = \mathbf{r}\mathbf{r}' + A Z_n A'. \quad (71)$$

(67) can be expressed as a valid EAR model (24), if $\hat{\mathbf{z}}_n$ and q_n are independent of the unknown AR parameters, $\mathbf{a}, \sigma$. From (68) and (70) however, both q_n and $\hat{\mathbf{z}}_n$ are functions of $A\left(\mathbf{a}\right)$ (64). In order to obtain a valid EAR model, we replace $A\left(\mathbf{a}\right)$ in (70), (71) by its expected value, $\hat{A}_{n-1} = A\left(\hat{\mathbf{a}}_{n-1}\right)$, using (11). Then, (67) is a valid EAR model defined by the set of transformations

$$G_2 : \ \bar{\mathbf{y}}_n = \frac{1}{\sqrt{q_n}} \left[x_n, \hat{\mathbf{z}}_n'\right]', \quad (72)$$

with time-variant Jacobian, $J_n = q_n^{-\frac{1}{2}}\left(\boldsymbol{X}_{n-1}\right)$, evaluated recursively using (68). G_2 is parameterized by unknown h, each setting of which defines a distinct candidate transformation. Note that $\bar{\mathbf{y}}_n$ in (72) depends on $\hat{\mathbf{a}}_{n-1}$ (11). Parameter updates are therefore correlated with previous estimates, $\hat{\mathbf{a}}_{n-1}$.

c) Remaining cases: cases a) and b) do not address the case where h_k is not constant on a regression interval $k \in \{n-p, \ldots, n\}$. Complete modelling for such cases is prohibitive, since $[h_{n-p}, \ldots, h_n]$ exists in a continuous space. Nevertheless, it is anticipated that such cases might be accommodated via a weighted combination of the two cases above.

A *non-stationary* AR(2) process was studied, with $a_{1;n}$ in the interval $[-0.98, -1.8]$ (as displayed in Fig. 3 (right)), $a_{2;n} = a_2 = 0.98$, $\sigma_n = \sigma = 0.01$, and $N = 200$. Realizations are displayed in Fig 3 (left). For $n < 95$, $a_{1;n}$ is increasing, corresponding to faster signal variations. Thereafter, $a_{1;n}$ decreases, yielding slower variations. The process was corrupted by two noise bursts (samples 50–80 and 130–180), with parameters $h = 8$ and $h = 6$ respectively (65). The process was estimated using $c = 3$ filter candidates: namely the unity transformation, G_1, along with G_2 $(h = 5)$ and G_2 $(h = 10)$. Identification results, using (48), (56)–(58), are displayed in Fig. 3 (middle). Specifically, the 95% HPD interval, via (11) and (13), of the marginal Student t-distribution of $a_{1;n}$ and $a_{2;n}$ respectively, is displayed. The process was identified using forgetting factors (55) $\phi_{\mathcal{N}i\mathcal{G}} = 0.92$, $\phi_{\mathcal{D}i} = 0.9$, and non-informative, stationary, alternative $\mathcal{N}i\mathcal{G}$ distribution, $\tilde{f}\left(\mathbf{a}, \sigma\right)$. Furthermore, the matrix parameter, $\tilde{\Psi}$, of the stationary, alternative Di distribution, $\tilde{f}\left(T\right)$ (55), was chosen to be diagonally dominant with ones on the diagonal. This discourages frequent transitions between filters.

The identification results (Fig. 3 (middle)) indicate better detection of the first burst than the second. As already noted, $\hat{\mathbf{z}}_{i,n}, i = 2, \ldots, c$, (which denotes the reconstructed state vector (70) with respect to the ith filter), is correlated with $\hat{\mathbf{a}}_{n-1}$, which may undermine the tracking of time-varying AR parameters, $\mathbf{a}_n$. In this case, each Kalman component predicts observations poorly, and receives low weights, $w_{2;n}$ and $w_{3;n}$ (53), in (56). This means that the first component—which does not pre-process the data—has a significant weight, $w_{1;n}$. Clearly then, the Kalman components have not spanned the space of necessary pre-processing transformations well, and need to be supplemented. Extra filters can be 'plugged in' in a naïve manner (in the sense that they *may* improve the spanning

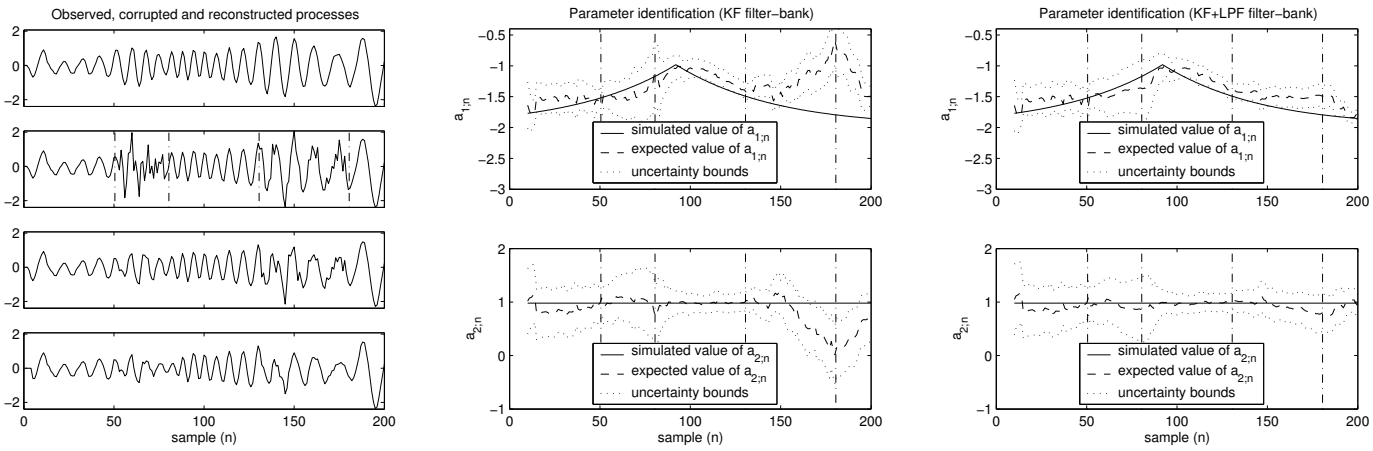

Fig. 3. Identification of a non-stationary AR(2) process corrupted by burst noise. **Left**: comparison of simulated, corrupted (dash-dotted vertical lines delimit beginning and end of each burst), reconstructed values (KF filter-bank), and reconstructed values (KF +LPF filter-bank). **Middle**: recursive identification of parameters $\mathbf{a}_{1;n}$ and $\mathbf{a}_{2;n}$ (KF filter-bank). **Right**: recursive identification of parameters $\mathbf{a}_{1;n}$ and $\mathbf{a}_{2;n}$ (KF+LPF filter-bank).

of the pre-processing space, but should simply be rejected, via (53) if poorly designed). During the second burst (Fig. 3), the process is slowing down. Therefore, we have extended the bank of KF filters by a simple arithmetic mean Low-Pass Filter (LPF) on the observed regressors:

$$G_3: \quad \bar{\mathbf{y}}_n = \frac{1}{3}\left(\bar{\mathbf{x}}_n + \bar{\mathbf{x}}_{n-1} + \bar{\mathbf{x}}_{n-2}\right). \tag{73}$$

(72) and (73) yield EAR models with the same AR parameterization, and so they can be combined within the MEAR framework.

Identification of the process using the KF+LPF filter-bank is displayed in Fig. 3 (right). Identification is improved during the second burst. The observed signal is compared with the reconstruction obtained using both variants (i.e. the KF and KF+LPF filter-banks) in Fig. 3 (left). Reconstructed values for the KF variant are derived from (62):

$$\hat{z}_n = w_{1;n} x_n - \sum_{i=2}^{3} w_{i;n} \hat{\mathbf{a}}'_n \hat{\mathbf{z}}_{i,n}, \tag{74}$$

using (11), (56). For the KF+LPF variant, the term $\frac{w_{4;n}}{3}\left(x_n + x_{n-1} + x_{n-2}\right)$ is added to (74), where $w_{4;n}$ is the estimated weight of the LPF component (53), (56)–(58).

Speech reconstruction: the MEAR filter-bank for the burst noise case (KF variant) was applied in the reconstruction of speech. A $c = 4$ MEAR model was used, involving G_1 ($\bar{\mathbf{y}}_n = \bar{\mathbf{x}}_n$), G_2 ($h = 3$), G_2 ($h = 6$) , G_2 ($h = 10$). The speech was modelled as AR with $p = 8$ (3). The forgetting factors (55) were $\phi_{\mathcal{N}i\mathcal{G}} = \phi_{Di} = 0.95$. Once again, a diagonally-dominant $\tilde{\Psi}$ was chosen for $\tilde{f}(T)$.

During periods of silence in speech, statistics (56) are effectively not updated, creating difficulties for adaptive identification. Therefore, we use an *informative* stationary alternative distribution, $\tilde{f}(\mathbf{a}, \sigma)$, of the $\mathcal{N}i\mathcal{G}$ type (5) for the AR parameters in (55). We identify the time-invariant alternative statistics, $\tilde{V}$, $\tilde{\nu}$, using 1800 samples of unvoiced speech. $\tilde{f}(\mathbf{a}, \sigma)$ was then flattened to reduce $\tilde{\nu}$ from 1800 to 2. This choice moderately influences the accumulating statistics at each step, via (56). Specifically, after a long period of silence, the influence of data in (56) becomes negligible, and V_n is reduced to $\tilde{V}$.

Three sections of the `bbcnews.wav` speech file, sampled at 11kHz, were corrupted by additive noise. Since we are particularly interested in performance in non-stationary epochs, we have considered three transitional cases: (i) voiced-to-unvoiced transition corrupted by zero-mean, white, Gaussian noise, with a realized Signal-to-Noise Ratio (SNR) of -1 dB during the burst; (ii) an unvoiced-to-voiced transition corrupted by zero-mean white uniform noise at -2 dB; and (iii) a silence-to-unvoiced transition corrupted by a click of type $0.25\cos(3n)\exp(-0.3n)$, superimposed on the silence period. In the first two cases, the noise burst was successfully suppressed. In the third case, the click was suppressed, but with some suppression also of the unvoiced speech.

VII. Discussion

The MEAR model (37) proposes a relatively rich extension of the classical AR model. It allows transformations on regressors, which relates it to semi-physical modelling [28]. Being a mixture-based extension, it is also related to the multiple model approach [29], to mixtures of AR processes [30], and to the Generalized AR (i.e. GAR) approach [7]. It must be remembered, though, that the MEAR model is a *single* AR model subject to an unknown transformation of observations. This is formalized as a mixture with common AR parameters (37). There are two main consequences. Firstly, the MEAR model is appropriate in cases where the transformation/distortion process is independent of the underlying AR process. Secondly, the AR parameter inference (48) requires a *single* sufficient statistic matrix, V_n (49), updated via a linear combination of c dyads, each calculated from one component in turn. This is expressed in the associated computational structure (Fig. 5). V_n is therefore updated by a structure of rank c.

The restriction to common AR parameterization across all components can easily be relaxed via obvious changes to the recursive algorithm (49)–(51). Each AR component would

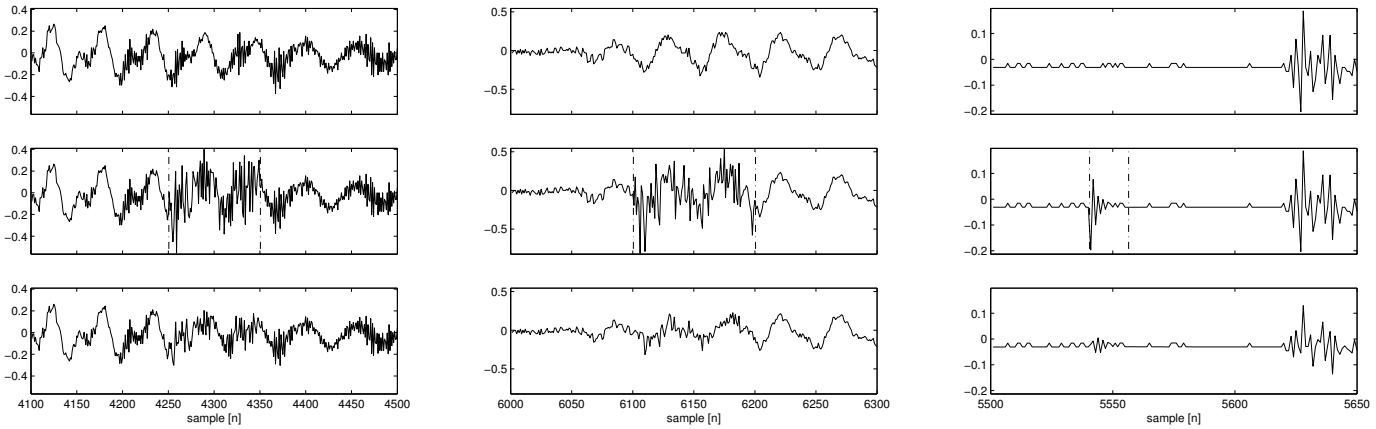

Fig. 4. Reconstruction of three sections of the `bbcnews.wav` speech file. In each column, the top figure is the original speech, the middle figure is the corrupted speech, and the reconstruction is in the bottom figure. Dash-dotted vertical lines delimit beginning and end of each burst.

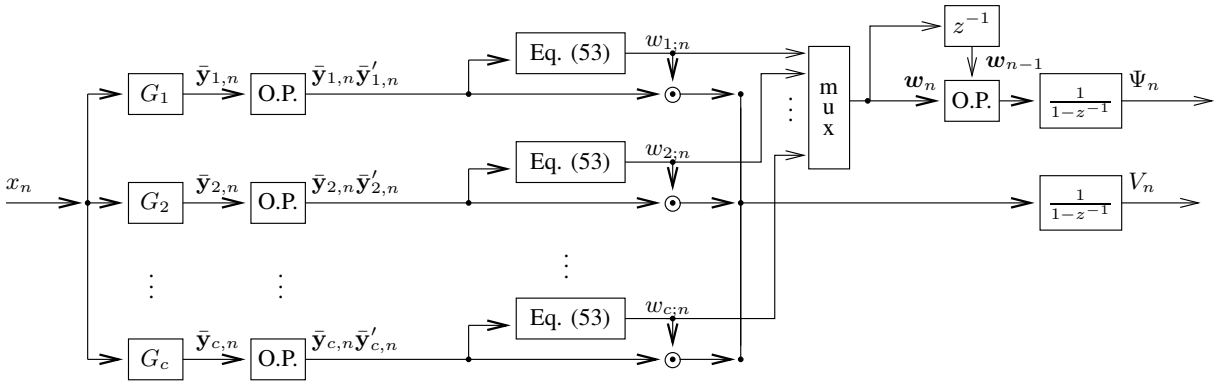

Fig. 5. The recursive summed-dyad computational scheme for identification of the MEAR model. 'O.P.' denotes outer product (dyad), 'mux' denotes the organization of the component weights, $w_{i;n}$ into vector form, $\boldsymbol{w}_n$. Dependence of Eq. (53) on V_{n-1}, $\boldsymbol{w}_{n-1}$ and Ψ_{n-1} is not shown, for clarity.

then experience a local rank-1 update, and there would be no inter-component interaction. Such a model would be over-parameterized, as each component would then have unknown AR parameters *and* an unknown transformation G_i, causing identification problems. The common AR parameterization in the MEAR model overcomes this problem. Moreover, the MEAR rank-c update implies an interaction of regressors from each component. This appears to be a key benefit of the MEAR model, as it allows a small number of candidate models to span a larger transformation space. In Section VI, for example, a small number of discrete values, h, drawn from a potentially large continuous range, could handle bursts generated by a model not explicitly represented by any one component.

Interaction between a finite set of components has been implemented by other techniques. The Kalman-based Interacting Multiple Models (IMMs) [29] linearly combine state vectors (i.e. certainty equivalents) evaluated using each filter, before using it in the Kalman updates. Again, however, this corresponds to a rank-1 update in our framework. The MEAR algorithm (Fig. 5) only propagates sufficient statistics, V_n, and not certainty equivalents.

The Bayesian identification method presented in this paper unifies all tasks of inference into a single, model-consistent framework. In the burst noise example of Section VI-B, the MEAR algorithm combines the pre-processing tasks (of burst detection and signal reconstruction) with on-line identification. It is the dynamic weights (53) which balance the dyadic update contributed by each component at every step (49). This contrasts with the previously reported methods. For example, in [31], a Boolean detection decision is made concerning presence of outliers. During a detected burst, a Kalman filter is used for reconstruction, and updating of statistics is interrupted. In our work, the updating of statistics is never interrupted. Components which, in effect, pre-process noisy data, contribute dyads constructed from *filtered* data. Furthermore, exponential forgetting is used to handle time-varying AR parameters, in place of the extended Kalman filter in [31]. In difficult cases, such as silence regions of speech (Section VI-B), forgetting with informative alternative distributions (55) can be used.

A Quasi-Bayes (QB)-based approximate update of sufficient statistics was employed in [25], for estimating an ARMA model using a mixture-based extension (ARMMAX). The ARMMAX model is a special case of the MEAR model, but with time-invariant component weights, instead of (32), and with moving-average whitening filters as candidate transformations (23). The candidates, $\mathbf{G}$ (28), used to represent the continuous multidimensional transformation space, were designed using a simplex method. This is an example of a technique for filter-bank design, achieved at the price of loss of recursivity in the

identification method. Further work on filter-bank design is required.

In our work, we model the possible degradations of the AR process, and design the filter-bank, $\mathbf{G}$ (28), in an attempt to span these possibilities. The task is facilitated by interaction between the filters, via the rank-c update (49). The parallel architecture of the summed-dyad algorithm (Fig. 5) permits extra candidates to be 'plugged in' with ease, in order to supplement the set. We saw in Section VI-B, for instance, how this can improve identification. When the extra candidate is not relevant, its contributing dyads are weighted by low component weights in (49), and become negligible.

VIII. Conclusion

We have introduced a mixture-based extension of the AR model, and derived an associated recursive Bayesian identification scheme. The resulting MEAR model is a mixture of AR components with common AR parameterization, each component modelling the AR process defined with respect to one possible data transformation. These transformations can be interpreted as a bank of filters, used to pre-process a single AR process.

The principal design aim of the MEAR model was to extend the modelling abilities of the classical AR model without losing the recursive computational properties of its identification. The recursive update was optimized at each time step in the sense of Kullback-Leibler (KL) distance. Conjugacy and sufficient statistics were preserved using a conditional independence assumption. This resulted in an on-line Variational Bayes (VB) approximation, which was further restricted in order to yield a non-iterative solution, known as the Quasi-Bayes (QB) formulation. This step-wise optimization of the parameter inference is important in non-stationary processing, yielding, for example, optimized point estimates along with measures of their uncertainty. The computational load of the MEAR identification procedure is light, increasing only linearly with the number of components (i.e. the number of filters in the filter-bank), and so real-time implementation is feasible.

The MEAR model is expected to be useful in situations where AR models are already used, but where there are now various distortions present. A correctly designed filter-bank for the MEAR model permits on-line recursive identification of the AR process, robust to these distortions. The MEAR model does not impose any specific form of filter on the filter-bank. Thus, it can be seen as a flexible framework for on-line comparison and cooperation between various *ad hoc* candidate pre-processing filters. Key to the computational flow of the proposed algorithm (Fig. 5) is the rank-c updating of parameter statistics via a weighted sum of dyads formed from the regressors of each transformation. The model can therefore perform well even in situations where the filter-bank does not include the true underlying data transformation.

Acknowledgements: This work was supported by grants: AVCR S1075102, GACR 102/03/0049.

References

[1] M. West, P. J. Harrison, and H. S. Migon, "Dynamic generalized linear models and Bayesian forecasting," *Journal of the American Statistical Association*, vol. 80, no. 389, 1985.

[2] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System identification* (P. Eykhoff, ed.), pp. 239–304, Oxford: Pergamon Press, 1981.

[3] B. Porat, *Digital processing of random signals: theory and methods*. Englewood Cliffs, N.J.: Prentice-Hall, 1994.

[4] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[5] S. Kay, *Modern Spectral Estimation*. New Jersey: Prentice-Hall, 1988.

[6] P. Wellstead and M. Zarrop, *Self-tuning Systems*. Chichester: John Wiley & Sons, 1991.

[7] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002.

[8] J. Rajan, P. Rayner, and S. Godsill, "Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler," *Vision, Image and Signal Processing, IEE Proceedings*, vol. 144, no. 4, pp. 249–256, 1997.

[9] L. Ljung and T. Söderström, *Theory and practice of recursive identification*. Cambridge; London: MIT Press, 1983.

[10] B. Widrow and S. Stearns, *Adaptive Signal Processing*. Prentice-Hall, 1985.

[11] A. Quinn, P. Ettler, L. Jirsa, I. Nagy, and P. Nedoma, "Probabilistic advisory systems for data-intensive applications," *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 2, pp. 133–148, 2003.

[12] J. Bernardo and A. Smith, *Bayesian Theory*. Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons, 1997. 2nd edition.

[13] B. Koopman, "On distributions admitting a sufficient statistic," *Transactions of American Mathematical Society*, vol. 39, p. 399, 1936.

[14] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[15] M. Abramowitz and I. Stegun, *Handbook of mathematical functions*. New York: Dover Publications, Inc., 1972.

[16] G. Golub and C. VanLoan, *Matrix Computations*. Baltimore – London: The John Hopkins University Press, 1989.

[17] G. Bierman, *Factorization Methods for Discrete Sequential Estimation*. New York: Academic Press, 1977.

[18] G. Kitagawa and W. Gersch, "A smoothness priors time-varying AR coefficient modelling on nonstationary covariance time series," *IEEE Transactions on Automatic Control*, vol. 30, no. 1, 1985.

[19] R. H. Middleton, G. C. Goodwin, D. J. Hill, and D. Q. Mayne, "Design issues in adaptive control," *IEEE Transactions on Automatic Control*, vol. 33, no. 1, pp. 50–58, 1988.

[20] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1979.

[21] G. V. Moustakides, "Locally optimum adaptive signal processing algorithms," *IEEE Transactions on Signal Processing*, vol. 46, no. 12, pp. 3315–3325, 1998.

[22] R. Kulhavý and M.B.Zarrop, "On general concept of forgetting," *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.

[23] S. Kotz and N. Johnson, *Encyclopedia of statistical sciences*. New York: John Wiley, 1985.

[24] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–87, 1951.

[25] L. He and M. Kárný, "Estimation and prediction with ARMMAX model: a mixture of ARMAX models with common ARX part," *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 4, pp. 265–283, 2003.

[26] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixtures*. New York: John Wiley & Sons, 1985.

[27] M. G. Kendall, A. Stuart, and K. Ord, *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Edward Arnold, 6th ed., 1998.

[28] T. Söderström and R. Stoica, *System Identification*. Prentice-Hall, 1989.

[29] X. R. Li and Y. Bar-Shalom, "Multiple-model estimation with variable structure," *IEEE Transactions on Automatic Control*, vol. 41, no. 4, pp. 478–493, 1996.

[30] M. Kárný, J. Böhm, T. V. Guy, and P. Nedoma, "Mixture-based adaptive probabilistic control," *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 2, pp. 119–132, 2003.

[31] M. Niedźwiecki and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals," *IEEE Transactions on Signal Processing*, vol. 44, no. 3, 1996.