EQUIVALENCE-MOTIVATED NON-LINEAR RECURSIVE ESTIMATION

Miroslav Kárný, Josef Andrýsek *

* *Adaptive Systems Department*
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
P. O. Box 18, 182 08 Prague, Czech Republic

Abstract: Recursive non-linear Bayesian estimation is addressed using equivalence approach as motivating framework. Its specific form – tailored to a model class covering non-normal ARX (auto-regression with exogenous variables) models, models with discrete outputs and continuous-valued regression vectors and their dynamic mixtures – is presented. The resulting algorithms provide efficient solutions of difficult and practically important estimation problems.

Keywords: non-linear estimation, recursive estimation, non-gaussian autoregressive models, complex models, probabilistic models.

1. INTRODUCTION

Data processing has many aims ranging from noise suppression (Anderson and Moore 1979) up to design of models (Kashyap and Rao 1976), serving to a subsequent decision making. Predominantly, their efficient reaching depends on estimation (Ljung 1987) of explicitly or implicitly specified models parameterized by an unknown finite-dimensional parameter $\Theta \in \Theta^*$. Adopted Bayesian paradigm (Bernardo and Smith 1997) treats unknown parameters as random. It modifies the prior probability density function (pdf) $f(\Theta)$, expressing prior knowledge about Θ, to the posterior pdf $f(\Theta|D)$ by the observed data D. The modification is determined by the Bayes rule $f(\Theta|D) \propto f(D|\Theta) f(\Theta)$. The proportionality symbol $\propto$ means that right-hand side of this expression has to be normalized to unit integral in order to get equality. The pdf $f(D|\Theta)$ of data D, viewed as a function of Θ in its condition, is known as likelihood function if the measured data D are inserted in it. The symbol f refers to pdfs distinguished by arguments' identifiers.

As a rule, complexity of the likelihood function increases with the extent of data D. Consequently, the pdf $f(\Theta|D)$ can be treated neither analytically nor numerically. Restriction to parametric models admitting the finite-dimensional sufficient statistic $\mathcal{V} = \mathcal{V}(D)$ (Koopman 1936) is the common remedy used. For such models, $f(\Theta|D) = f(\Theta|\mathcal{V}(D))\,,\ \forall\Theta \in \Theta^*$. The statistic $\mathcal{V}(D)$ belongs to a set $\mathcal{V}^*$ whose dimension is constant and finite even when the extent of data grows without limits. Under rather general conditions, the class of models with finite-dimensional sufficient statistic coincides with the *exponential family* (EF), (Barndorff-Nielsen 1978),

$$f(D|\Theta) = \exp\langle \mathcal{V}(D), C(\Theta)\rangle\,. \qquad (1)$$

The sufficient statistic $\mathcal{V}(D)$ is related to the unknown quantity Θ through the scalar product $\langle\cdot,\cdot\rangle$ with a fixed function $C(\Theta)$ of a compatible dimension. In the addressed *recursive* processing, the data D consist of a sequence $d^t \equiv (d_1, \ldots, d_t)$ of data records d_t obtained at discrete time instances labelled by $t \in t^* \equiv \{1, 2, \ldots\}$. In recursive setting, evaluation of the statistic $\mathcal{V}$ must not require re-processing of whole data sequence. This

restricts the EF further on. With $\mathcal{V}_t = \mathcal{V}(d^t)$, the chain rule for pdfs (Peterka 1981) gives

$$\exp\langle \mathcal{V}_t, C(\Theta)\rangle = f\left(d_t\middle| d^{t-1}, \Theta\right) f\left(d^{t-1}\middle|\Theta\right) \quad (2)$$
$$= f\left(d_t\middle| d^{t-1}, \Theta\right) \exp\langle \mathcal{V}_{t-1}, C(\Theta)\rangle .$$

Identity (2) implies that recursive evaluation of the likelihood is only possible if $\mathcal{V}_t - \mathcal{V}_{t-1} = B(\Psi_t)$, where $B(\cdot)$ is a function of a fixed dimensional *data vector* Ψ_t that can be updated recursively using a known function $\tilde{\Psi}(\cdot)$

$$\Psi_t = \tilde{\Psi}(\Psi_{t-1}, d_t). \quad (3)$$

The parametric models

$$f\left(d_t\middle| d^{t-1}, \Theta\right) = \exp\langle B(\Psi_t), C(\Theta)\rangle \quad (4)$$

together with the function $\tilde{\Psi}(\cdot)$ define the *dynamic exponential family* whose only non-trivial members are generalized (Peterka 1981) normal autoregressive models with exogenous variables (ARX) and controlled Markov chains. Narrowness of this class calls for an *approximate recursive estimation* applicable to more general parametric models $f\left(d_t\middle| d^{t-1}, \Theta\right) \equiv M(\Psi_t, \Theta)$ with a known updating (3) of Ψ_t but with the *model* $M(\Psi_t, \Theta)$ *out of the EF*. Naturally, a range of attempts has been made in this respect. In our opinion, the equivalence approach (Kulhavý 1996) is still the most advanced one and motivated this paper.

Here, an approximate posterior pdf in the class of pdfs conjugated to the EF is constructed. The approximate pdf is designed so that it is asymptotically close to an equivalence class containing the exact posterior pdf, see Section 2. The algorithm is applicable to non-normal ARX models, models relating discrete values to continuous ones and mixtures with components whose factors and *dynamic* weights belong to the EF, see Section 3.

2. RECURSIVELY FEASIBLE REPRESENTATIONS AND EF

Always limited computer resources call for a reduced representation of the evolving posterior pdfs. It is a peculiar task as the posterior pdfs concentrate quickly on a small support at a priori unknown position in Θ^*. Thus, for instance, interpolation on a fine grid that does not miss the final position becomes soon computationally prohibitive. Hence, more sophisticated approximations are needed. The approach proposed here is motivated by Proposition 2.1 presented below. It characterizes the recursively updated non-sufficient statistics that are compatible with the recursive evaluation of the posterior pdf.

Proposition 2.1. (Equivalence-preserving $\mathcal{V}$). Let the posterior pdf $f\left(\Theta\middle| d^{t-1}\right) \in f^*\left(\Theta\middle|d^{t-1}\right) \equiv$ a set of pdfs with a common, time, data and parameter invariant support – the set of arguments on which $f\left(\Theta\middle| d^{t-1}\right) > 0$. Let the mapping

$$\mathsf{V}_{t-1} : f^*(\Theta|d^{t-1}) \to \mathcal{V}^*_{t-1} \quad (5)$$

assign to $f(\Theta|d^{t-1})$ a finite-dimensional statistic $\mathcal{V}_{t-1} \equiv \mathsf{V}(d^{t-1})$, not necessarily sufficient. Then, $\mathcal{V}_{t-1}$ can be *exactly recursively updated* using the value $\mathcal{V}_{t-1}$ and model $f\left(d_t\middle| d^{t-1}, \Theta\right) = M(\Psi_t, \Theta)$, with Ψ_t (3), iff V_t is *time-invariant linear mapping* $\mathsf{V} \equiv \mathsf{V}_t$ of logarithms of the posterior pdfs. V has to map Θ-independent functions to zero.

Proof: Proof of necessity is in (Kulhavý 1990*a*, Kulhavý 1990*b*). To prove sufficiency, it suffices to apply V to the logarithmic version of the Bayes rule and use both time-invariance and linearity of V. The Θ-independent normalizing term $\ln\left(f\left(d_t\middle| d^{t-1}\right)\right)$ is mapped to zero and $\mathcal{V}_t = \mathsf{V}\left[\ln\left(M(\Psi_t, \Theta)\right)\right] + \mathcal{V}_{t-1}$, $\mathcal{V}_0 = \mathsf{V}(\ln(f(\Theta))) \equiv \mathsf{V}(\ln(\text{prior pdf}))$. □

Choice of the mappings V_t *(5) that hopefully converge to a mapping* V *from Proposition 2.1 is addressed for a class of models enriching the EF*

$$f\left(d_t\middle| d^{t-1},\Theta\right) \equiv M(\Psi_t, \Theta) \equiv \mathcal{M}\left(B(\Psi_t), C(\Theta)\right) \quad (6)$$

with $\mathcal{M}(B(\Psi_t), C(\Theta))$ being smooth in $C(\Theta)$.

Example 2.1. (Examples of models in (6)). In all cases, except the last one, one-dimensional d_t is considered. It does not restrict generality of the problem formulation as in the multivariate case the chain rule for pdfs implies that $f\left(d_t\middle|\bullet\right) = \prod_{i=1}^{d^\ell} f\left(d_{t;i}\middle| d_{t;i+1}, \ldots, d_{t;d^\ell}, \bullet\right)$. Hereafter, x^ℓ denotes below length of the vector x.

Normal regression model: It models

$$d_t = \underbrace{\theta'}_{\substack{\text{regression}\\ \text{coefficients}}} \times \underbrace{\psi_t}_{\substack{\text{regression}\\ \text{vector}}} + \underbrace{\text{zero-mean noise}}_{\substack{\text{white normal}\\ \text{with variance } r}}$$

with $'$ being transposition. $\ln\left(\mathcal{M}(B(\Psi_t), C(\Theta))\right) = -0.5\left[\ln(2\pi r) + \text{tr}\left(\Psi_t\Psi_t'\frac{[-1,\theta'][-1,\theta']'}{r}\right)\right]$. It belongs to the dynamic EF (4) and thus to the class (6). It is given by $\Theta = (\theta, r)$, $\Psi_t' = [d_t, \psi_t']$, $C(\Theta) = -0.5\left[\ln(2\pi r), \text{vec}\left(\frac{[-1,\theta'][-1,\theta']'}{r}\right)\right]$ and

$$B(\Psi_t) = [1, \text{vec}[\Psi_t\Psi_t']]\,. \quad (7)$$

The operation vec maps the symmetric matrix on a vector so that $\langle a, b\rangle \equiv \text{vec}(a)'\text{vec}(b)$.

Cauchy regression model: It is given by

$$\mathcal{M}(B(\Psi_t), C(\Theta)) \propto \frac{1}{r + \frac{(d_t - \theta'\psi_t)^2}{r}} = \frac{1}{B'(\Psi_t)C(\Theta)}$$

$\Theta = (\theta, r)$, scaling factor $r > 0$, for $B(\Psi_t)$ see (7),

$$C(\Theta) = \left[r, \text{vec}\left(\frac{[-1,\theta'][-1,\theta']'}{r}\right)\right]. \quad (8)$$

Regression with discrete outputs: It predicts discrete data $d_t \in d^* \equiv \left\{1, \ldots, d^\ell\right\}$. Their proba-

bilities are high if Gaussian-like regression, given by the regression vector ψ_t, predicts well the value $d_t \in d^*$, i.e., $\mathcal{M}(B(\Psi_t), C(\Theta))$

$$= \frac{r_{d_t}^{-0.5} \exp\left\{-0.5([-1, \theta'_{d_t}]\Psi_{t;d_t})^2 / r_{d_t}\right\}}{\sum_{d \in d^*} r_d^{-0.5} \exp\left\{-0.5([-1, \theta'_d]\Psi_{t;d})^2 / r_d\right\}}$$

$$\Theta \equiv [\Theta_1, \ldots, \Theta_{d^\ell}],\ \Theta'_d \equiv [r_d, \theta'_d],\ \Psi'_{t;d} \equiv [d, \psi'_t],$$

$$B(\Psi_{t;d}) \equiv [B_1(\Psi_{t;d}), \ldots, B_{d^\ell}(\Psi_{t;d})]$$

$$B_k(\Psi_{t;d}) \equiv \delta_{kd}\left[1, \mathrm{vec}[\Psi_{t;d}\Psi'_{t;d}]\right] \quad (9)$$

$$\delta_{kd} = \begin{cases} 1 \text{ if } k = d \\ 0 \text{ if } k \neq d \end{cases},\ C(\Theta) \text{ has entries}$$

$$C_d(\Theta_d) \equiv -0.5\left[\ln(r_d), \mathrm{vec}\left([-1, \theta'_d]'[-1, \theta'_d]/r_d\right)\right].$$

Fully dynamic mixture with factorized components in the EF: It is parametric model $M(\Psi_t, \Theta)$

$$= \frac{\sum_{c \in c^*} \prod_{i=1}^{\Psi^\ell} \exp\left\langle B_{ic}(\Psi_{t;i}), C_{ic}(\Theta_{ic})\right\rangle}{\sum_{c \in c^*} \prod_{i=d^\ell+1}^{\Psi^\ell} \exp\left\langle B_{ic}(\Psi_{t;i}), C_{ic}(\Theta_{ic})\right\rangle}$$

$$\Psi_{t;i} = [\text{scalar},\ \Psi_{t;i+1}],\ \Psi_{t;0} \equiv \Psi_t \equiv \text{data vector}$$

$$(\Theta, B(\Psi_t), C(\Theta)) \equiv \{\Theta_{ic},\ B_{ic}(\Psi_{t;i}),\ C_{ic}(\Theta_{ic})\}$$

$$i \in i^* \equiv \{1, \ldots, \Psi^\ell\},\ c \in c^* \equiv \text{a finite set.} \quad (10)$$

For the class (6) and the prior pdfs conjugated to the EF, the posterior pdfs have the form $f(\Theta|d^t) = g(d^t, C(\Theta))$. This leads to the choice of approximate posterior pdfs conjugated to (1).

Let us consider models (6) and, for a reference point ${}^0\Theta$ from Θ^*, define the mapping (5)

$${}^{0\Theta}\mathsf{V}[g(\cdot,\cdot)] \equiv \frac{\partial \ln\left(g\left(d^t, C\left({}^0\Theta\right)\right)\right)}{\partial C({}^0\Theta)} \equiv {}^0\mathcal{V}_t. \quad (11)$$

Proposition 2.2. (Mapping ${}^{0\Theta}\mathsf{V}$). Let us consider the class (6) of models and conjugated pdfs $f(\Theta|\mathcal{V}) \propto \exp\langle \mathcal{V}, C(\Theta)\rangle$ with almost surely linearly independent entries of $C(\Theta)$. The mapping (11) acting on the posterior pdfs meets conditions of Proposition 2.1 and defines equivalence classes $g^*_{{}^0\mathcal{V}_t} \equiv \left\{g(d^t, C(\Theta)) :\ {}^{0\Theta}\mathsf{V}[g] = {}^0\mathcal{V}_t\right\}$. Any class contains at most one member conjugated to (1).

Proof: Meeting of conditions of Proposition 2.1 is obvious. An application of ${}^{0\Theta}\mathsf{V}$ to $\ln(\exp\langle \mathcal{V}_t, C(\Theta)\rangle)$ shows that only this conjugated pdf belongs to the equivalence class $g^*_{{}^0\mathcal{V}_t}$. It is well defined pdf iff $\int \exp\langle \mathcal{V}, C(\Theta)\rangle\, d\Theta < \infty$. □

Proposition 2.2 implies that ${}^{0\Theta}\mathsf{V}$ determines at most a single member in the EF whose statistics ${}^0\mathcal{V}_t$ can be recursively updated while preserving equivalence ${}^{0\Theta}\mathsf{V}(\exp\langle \mathcal{V}_t, C(\Theta)\rangle) = {}^{0\Theta}\mathsf{V}[f(\Theta|d^t)]$ on the class (6). The choice of the reference point ${}^0\Theta$ determining ${}^{0\Theta}\mathsf{V}$ (11) decides whether the pdf $\exp\left\langle {}^0\mathcal{V}_t, C(\Theta)\right\rangle$ in the EF and equivalent to the correct posterior pdf $f(\Theta|d^t)$ is close to $f(\Theta|d^t)$ or not. The following proposition, proved in (Kárný *et al.* 2005), guides its choice. Within it, the support Θ^* of the prior pdf $f(\Theta)$ is assumed closed set.

Proposition 2.3. (Asymptotic of posterior pdfs).
(1) The correct posterior pdfs $f(\Theta|d^t)$ converges almost surely for any point $\Theta \in \Theta^*$ for which $f(\Theta|d^t)$ stay bounded while $t \to \infty$.
(2) Let pdfs $f\left(d_\tau \middle| d^{\tau-1}\right)$ be derived from the joint pdf describing correctly the data generator. For any fixed $\Theta \in \Theta^*$, let exist the finite conditional expectations $\mathsf{E}\left[\ln\left(\frac{f\left(d_\tau|d^{\tau-1}\right)}{M(\Psi_\tau,\Theta)}\right)\middle| d^{\tau-1}\right]$ defined by $f\left(d_\tau\middle|d^{\tau-1}\right)$. Then, the zero mean, mutually uncorrelated, innovations $\varepsilon(d^{\tau-1}, \Theta) \equiv \ln\left(\frac{f\left(d_\tau|d^{\tau-1}\right)}{M(\Psi_\tau,\Theta)}\right) - \mathsf{E}\left[\ln\left(\frac{f\left(d_\tau|d^{\tau-1}\right)}{M(\Psi_\tau,\Theta)}\right)\middle| d^{\tau-1}\right]$ are well defined. If, for all $\Theta \in \Theta^*$ and all $\tau \in t^*$, these innovations have finite variances then, the support of the posterior pdf $f(\Theta|d^t)$ concentrates (almost surely) on minimizers ${}^\infty\Theta \in {}^\infty\Theta^*$ of the entropy rate $\mathcal{H}(d^\infty, \Theta)$

$${}^\infty\Theta^* \equiv \lim_{t\to\infty} \mathrm{Arg}\inf_{\Theta\in\Theta^*} \mathcal{H}(d^{t-1}, \Theta) \equiv \lim_{t\to\infty} \mathrm{Arg}\inf_{\Theta\in\Theta^*} \frac{1}{t}\sum_{\tau=1}^{t} \mathsf{E}\left[\ln\left(\frac{f\left(d_\tau\middle|d^{\tau-1}\right)}{M(\Psi_\tau,\Theta)}\right)\middle| d^{\tau-1}\right].$$

Proposition 2.3 also implies that the posterior pdf concentrates on a "small" set ${}^\infty\Theta^*$ of extreme points ${}^\infty\Theta \in {}^\infty\Theta^*$. The points Θ for which $f(\Theta|d^t)$ are unbounded as $t \to \infty$ belong to this set. Thus, $\frac{\partial \ln\left(f\left({}^\infty\Theta|d^t\right)\right)}{\partial\Theta}$ is expected to converge to zero. This indicates the expected behavior of the mapping ${}^{\infty\Theta}\mathsf{V}$ (11)

$$\underbrace{\frac{\partial \ln\left(g\left(d^t, C\left({}^\infty\Theta\right)\right)\right)}{\partial C\left({}^\infty\Theta\right)}}_{\mathcal{V}_t} \frac{\partial C\left({}^\infty\Theta\right)}{\partial\Theta} \to 0. \quad (12)$$

The property (12) applies only to the narrow set ${}^\infty\Theta^*$ enriched by (often empty) set of asymptotic local extremes. Knowing the set ${}^\infty\Theta^*$ the pdf in the EF equivalent to the correct posterior pdf $f(\Theta|d^t)$ could be constructed, whose support contains the same asymptotic stationary points as $f(\Theta|d^t)$. This is a strong property, as – for non-over-parametric models and sufficiently rich σ algebra generated by data d^∞ the set ${}^\infty\Theta^*$ is singleton and no other stationary point exists.

Practically, reference points $\underline{\Theta}_t \in \Theta^*$ approaching ${}^\infty\Theta^*$ have to be constructed. Let us denote $\overline{\Theta}_t$ maximizer of the approximate pdf $\exp\langle \underline{\mathcal{V}}_t, C(\Theta)\rangle$, $\underline{\mathcal{V}}_t \approx {}^{\underline{\Theta}_t}\mathsf{V}(\ln(f(\Theta|d^t)))$. Ideally, $\overline{\Theta}_t$ and $\underline{\Theta}_t$ should (almost) coincide. During recursive estimation, they differ and $\overline{\Theta}_t$ is a candidate for the definition of the considered equivalence class. It

should be better than $\underline{\Theta}_t$, i.e., closer to ${}^{\infty}\Theta^*$. At the same time, swapping from $\underline{\Theta}_t$ to $\overline{\Theta}_t$ gives up the accumulated information and introduces an error into learning. The error gradually diminishes if the points $\overline{\Theta}_t$, $\underline{\Theta}_t \equiv \overline{\Theta}_{t-1}$ at which derivatives $\frac{\partial}{\partial C(\Theta)}$ are taken converge to a constant parameter. At present, this property has to be checked in each particular case. The above exposition motivated:

Algorithm 1. (Nonlinear recursive estimation).
Initial phase

- Select the learnt model $\mathcal{M}(B(\Psi_t), C(\Theta))$.
- Set the learning time $t = 0$.
- Select the statistics $\mathcal{V}_t$ determining the prior pdf $\exp\langle \mathcal{V}_t, C(\Theta)\rangle$ in the EF for $t = 0$.
- Select $\Theta_t \in \mathrm{Arg}\max_{\Theta\in\Theta^*} \exp\langle \mathcal{V}_t, C(\Theta)\rangle$, determining ${}^{\Theta_t}\mathsf{V}$ (signs $\_^{-}$ are dropped).

Recursive phase

- Increase time $t = t + 1$ and collect Ψ_t.
- Evaluate the trial update of the statistics

$$\mathcal{V}_t = \mathcal{V}_{t-1} + \frac{\partial \mathcal{M}(B(\Psi_t), C(\Theta_{t-1}))}{\partial C(\Theta_{t-1})}.$$

- Find the new $\Theta_t \in \mathrm{Arg}\max_{\Theta\in\Theta^*} \langle \mathcal{V}_t, C(\Theta)\rangle$ and go to the beginning of Recursive phase if the statistics $\mathcal{V}_t$ is well defined, i.e., $\int \exp\langle \mathcal{V}_t, C(\Theta)\rangle \, d\Theta < \infty$. Otherwise, correct the trial statistics to get the finite normalizing integral and repeat this item.

Algorithm is close to a recursive version of expectation-maximization algorithm but it does not try to find a point estimate of Θ only. The corrective actions, needed when $\mathcal{V}_t$ is not well defined, is its critical step. Design of these corrective actions is specific for each type of the model.

3. APPLICATIONS

Cauchy regression model has the form (8). The function $C(\Theta)$ has the entries r, $\frac{[-1,\theta'][-1,\theta']'}{r}$. It is useful to decompose the stored statistics $\mathcal{V}_t$ into the scalar ν_t and the symmetric positive semi-definite matrix V_t, so that $\langle \mathcal{V}_t, C(\Theta)\rangle = \nu_t r + \left\langle V_t, \frac{[-1,\theta'][-1,\theta']'}{r}\right\rangle$. For a fixed $\Theta_{t-1} \equiv \overline{\Theta}_{t-1} \in \Theta^*$, the statistics, determining approximately the equivalence class, evolve

$$\nu_t = \nu_{t-1} + w_t, \quad V_t = V_{t-1} + w_t \Psi_t \Psi_t' \quad (13)$$
$$w_t \equiv \left(r_{t-1} + \frac{([-1, \theta_{t-1}'] \Psi_t)^2}{r_{t-1}} \right)^{-1}.$$

The posterior pdf in the EF, approximately equivalent to the correct posterior pdf, is

$$f(\Theta|V_t, \nu_t) \propto \exp\left\{ -\frac{1}{2} \left[\nu_t r + \frac{[-1,\theta'] V_t [-1,\theta']'}{r} \right] \right\}. \quad (14)$$

Due to the non-negativity of the weights w_t, the statistics $\nu_t > 0, V_t > 0$ (positive definite) for all $t \in t^*$ whenever these inequalities hold for the prior statistics. Thus, no corrective actions are needed. Maximizer of the pdf (14) on Θ^* can be found explicitly. It has the form well-known from least squares (LS) (Kárný *et al.* 2005). The split factorization $V = L'DL$, with unitary lower triangular L and positive diagonal D,

$$L = \begin{bmatrix} 1 & 0 \\ {}^{d\psi}L & {}^{\psi}L \end{bmatrix}, \; D = \begin{bmatrix} {}^{d}D & 0 \\ 0 & {}^{\psi}D \end{bmatrix}, \; {}^{d}D \text{ is scalar},$$
$$\text{gives } \; \hat{\Theta} \equiv \left(\hat{\theta}, \hat{r}\right) \equiv \left({}^{\psi}L^{-1\,d\psi}L, \sqrt{\frac{{}^{d}D}{\nu}} \right). \quad (15)$$

Algorithm ldupdt (Nedoma *et al.* 2005) updates efficiently factors L, D by the dyad $w_t \Psi_t \Psi_t'$ (13).

Example 3.1. (Cauchy regression model). Data $d_t = (y_t, u_t)$ ≡(scalar system output,scalar system input) were simulated and recorded. The output y_t dependent on the input u_t and past data was generated by the Cauchy system (8). The scaling factor $r = 0.27$ and the regression vector $\psi_t = [u_t, y_{t-1}, u_{t-1}, y_{t-2}]'$ were used. The input u_t was white normal noise with zero mean and unit-variance. The chosen $\theta = [0.2, 0.8, 0.07, -0.07]'$ describes the second-order auto-regression with poles 0.7, 0.1. The coefficients at inputs and scaling factor are chosen so that their respective static gain equals to 1. The output and weigh realizations (13) are shown in Fig. 1.

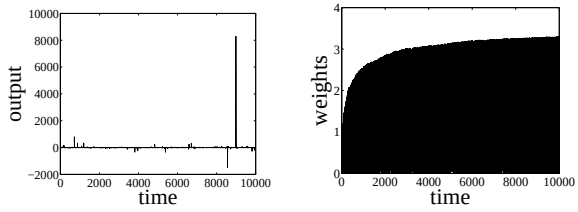

Fig. 1. Simulated outputs (left) and weights (13).

Point estimates of the regression coefficients θ are in Fig. 2 together with their estimates obtained by recursive least squares (LS). Similar comparison of the scaling- factor estimates is in Fig. 3.

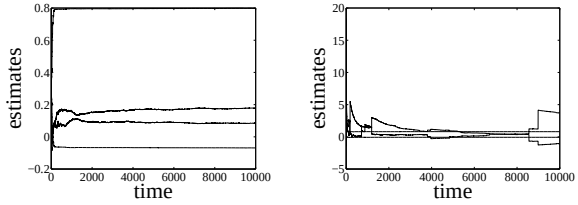

Fig. 2. Estimates of $\theta = [0.2, 0.8, 0.07, -0.07]$ (left) and their LS counterpart.

The results confirm positive effects of the weighted recursive LS updating (13) with the weights tailored to the assumed distribution of innovations.

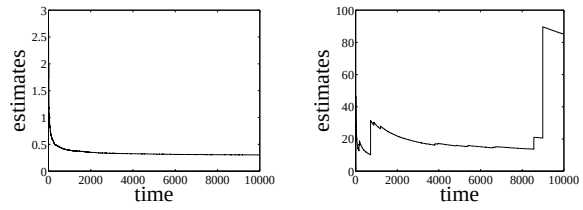

Fig. 3. Estimates of $r = 0.27$ (left) and its LS counterpart.

The estimation is more robust to outlying innovations and enables good estimation of all parameters, including the scaling one. It is in a strong contrast with non-weighted LS.

Regression with discrete outputs is described by (9). The derivatives needed in Algorithm 1, are

$$\frac{\partial \ln\left(\mathcal{M}(B(\Psi_t), C(\Theta_{t-1}))\right)}{\partial[-0.5\ln(r_{t-1;d})]} \equiv w_{t;d} = \delta_{dd_t}$$

$$-\frac{r_{t-1;d}^{-0.5}\exp\left\{-\frac{\left([-1,\theta'_{t-1;d}]\Psi_{t;d}\right)^2}{2r_{t-1;d}}\right\}}{\sum_{d=1}^{d^\ell} r_{t-1;d}^{-0.5}\exp\left\{-\frac{\left([-1,\theta'_{t-1;d}]\Psi_{t;d}\right)^2}{2r_{t-1;d}}\right\}}$$

$$\frac{\partial \ln\left(\mathcal{M}(B(\Psi_t), C(\Theta_{t-1}))\right)}{\partial\left[-\frac{[-1,\theta'_{t-1;d}]'[-1,\theta'_{t-1;d}]}{2r_{t-1;d}}\right]} = w_{t;d}\Psi_{t;d}\Psi'_{t;d}$$

$f(\Theta|\mathcal{V}_t) \equiv \prod_{d=1}^{d^\ell} GiW_{\theta_d,r_d}(V_{t;d},\nu_{t;d})$ with Gauss-inverse-Wishart factors (Kárný *et al.* 2005)

$$GiW_{\theta,r}(V,\nu) \propto \frac{\exp\left\{-\frac{[-1,\theta']V[-1,\theta']'}{2r}\right\}}{r^{\frac{\nu+\theta^\ell+2}{2}}} \quad \text{with}$$

$$\nu_{t;d} = \nu_{t-1;d} + w_{t;d},\ V_{t;d} = V_{t-1;d} + w_{t;d}\Psi_{t;d}\Psi'_{t;d}.$$

The normalization is finite iff $\nu_{t;d} > 0$, $V_{t;d} > 0$ for $d \in d^*$, $t \in t^*$. The found weights do not guarantee this. For dth, for which the normalization is finite (surely, for $d = d_t$), the updating needs no corrections. Maximizing θ is computed according to formula (15), using the same $L'DL$ decomposition of V_d statistics. The estimate of r is here square of the estimate in (15). For some d's corrective actions may be needed. Even if they correspond to a change of the reference point, it is sufficient to change the weight $w_{t;d}$ and compute directly a new reference point. It is reasonable to require $\nu_{t;d} \geq \nu_{0;d}$ and $V_{t;d} > 0$. Thus, the corrected $w_{t;d}$ should $\nu_{t-1;d} + w_{t;d} \geq \nu_{0;d} > 0$, $V_{t-1;d} + w_{t;d}\Psi_{t;d}\Psi'_{t;d} > 0$. A sufficient condition for this reads $V_{t-1;d} + w_{t;d}\Psi_{t;d}\Psi'_{t;d} \geq \omega V_{t-1;d}$ with optional $\omega \in (0,1)$, i.e., $-\frac{1-\omega}{\zeta_{t;d}} \equiv -\frac{1-\omega}{\Psi'_{t;d}V_{t-1;d}^{-1}\Psi_{t;d}} \leq w_{t;d}$. The scalar $\zeta_{t;d}$ is by-product of updating the factorized $V_{t-1;d}$, which motivated this condition.

Example 3.2. (Regression with discrete outputs). The regression model with the output $y_t \in y^* \equiv \{1,2,3\}$ and white normal input u_t with mean 0 and variance 1 was simulated with the options:
2nd order ARX for $y = 1$
$[r_1^{0.5}; \theta'_1] = [0.03; 1.3, 0.15, -0.36, -0.05]$
$\Psi_{t;1} = [1, u_t, y_{t-1}, u_{t-1}, y_{t-2}, u_{t-2}]'$
2nd order auto-regression for $y = 2$
$[r_2^{0.5}; \theta'_3] = [0.08; 1, 0.15]$, $\Psi_{t;2} = [2, y_{t-1}, y_{t-2}]'$
Regression with memory 1 for $y = 3$
$[r_3^{0.5}; \theta'_3] = [0.1; 0.3, 2]$, $\Psi_{t;3} = [3, u_t, u_{t-1}]'$.

Estimation results concern simulation run with 2000 samples characterized by the histogram of outputs, Fig. 4. The estimates of respective scaling factors $r_d^{0.5}$ in Fig. 4 give feeling about their behavior. Estimates of respective parameters in Fig. 5, Fig. 6 complement the picture. The most interesting and informative is right-hand side of Fig. 6 and Fig. 7 that show histograms of relative errors $\kappa_{t;y}$, $y = 1,2,3$, $\kappa_{t;y} \equiv$

$$\frac{f\left(y_t = y\middle|\, u_t, d^{t-1}\right) - f\left(y_t = y\middle|\, u_t, d^{t-1}, \text{true}\,\Theta\right)}{f\left(y_t = y\middle|\, u_t, d^{t-1}, \text{true}\,\Theta\right)}. \tag{16}$$

Fig. 8 shows relative errors on realized outputs only. Values out of the range $[-3,3]$ are aggregated. Quality of these estimates seems to be good, especially, taking into account that $y_t = 3$ occurred 34 times only. The additional information about the experiment is:

- The system and model structures coincided.
- Prior $V_{0;d} = 1e-3\times$ unit matrix were used.
- Guess 0.1 of $r_d^{0.5}$ and $\nu_{0;d} = 1$ were chosen.
- Prior guesses of regression coefficients were $\hat{\theta}_{0;1} = [0.5, 0, 0, 0]$, $\hat{\theta}_{0;2} = [0.5, 0]$, $\hat{\theta}_{0;3} = [0, 0]$;
- $40, 28, 114$ weight corrections were needed for $y = 1, 2, 3$.

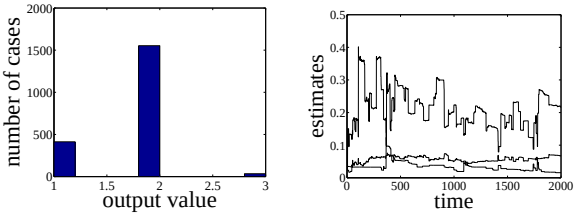

Fig. 4. Histogram of simulated outputs (left) and estimates of scaling factors $[r_1^{0.5}, r_2^{0.5}, r_3^{0.5}] = [0.03, 0.08, 0.1]$.

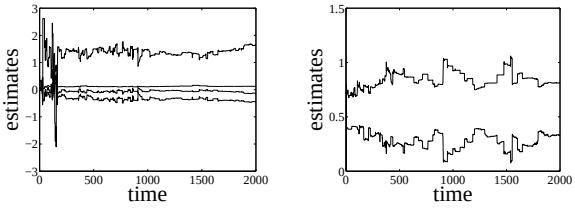

Fig. 5. Estimates of $\theta'_1 = [1.3, 0.15, -0.36, -0.05]$ for $y = 1$ (left) and estimates of $\theta'_2 = [1, 0.15]$ for $y = 2$.

Fully dynamic mixture (10) is estimated via Algorithm 1 by updating of statistics

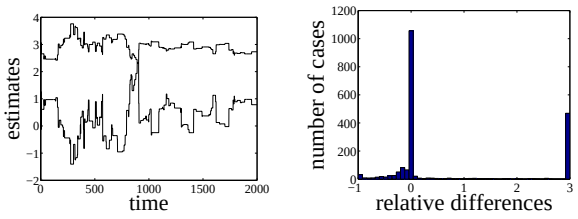

Fig. 6. Estimates of $\theta'_3 = [-0.09, -1]$ for $y = 3$ (left) and relative errors $\kappa_{t;y}$ (16) for $y = 1$.

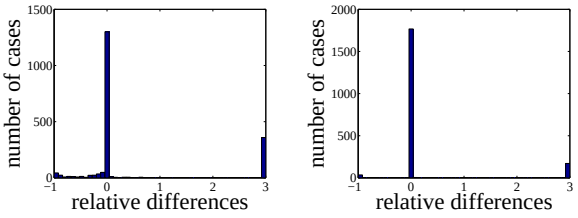

Fig. 7. Relative errors $\kappa_{t;y}$ (16) for $y = 2, 3$.

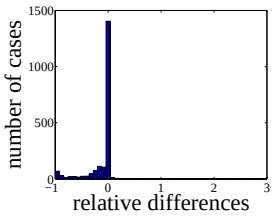

Fig. 8. Relative errors $\kappa_{t;y_t}$ (16) for realized y_t.

$$V_{t;ic} = V_{t-1;ic} + w_{t;ic} B_{ic}(\Psi_{t;i})$$

$$w_{t;ic} = \frac{\prod_{i=1}^{\Psi^\ell} \exp\left\langle B_{ic}(\Psi_{t;i}), C_{ic}\left(\Theta_{t-1;ic}\right)\right\rangle}{\sum_{c\in c^*}\prod_{i=1}^{\Psi^\ell} \exp\left\langle B_{ic}(\Psi_{t;i}), C_{ic}\left(\Theta_{t-1;ic}\right)\right\rangle} - \frac{\chi(i > d^\ell)\prod_{i=d^\ell+1}^{\Psi^\ell} \exp\left\langle B_{ic}(\Psi_{t;i}), C_{ic}\left(\Theta_{t-1;ic}\right)\right\rangle}{\sum_{c\in c^*}\prod_{i=d^\ell+1}^{\Psi^\ell} \exp\left\langle B_{ic}(\Psi_{t;i}), C_{ic}\left(\Theta_{t-1;ic}\right)\right\rangle}$$

where $\chi(\cdot)$ denotes indicator of the set in argument. The approximate posterior pdf has the form of product of pdfs conjugated to respective factors $\prod_{i=1}^{\Psi^\ell}\prod_{c\in c^*} \exp\left\langle V_{t;ic}, C_{ic}(\Theta_{ic})\right\rangle$. The weights for $i > d^\ell$ can be negative. Thus, the corrective actions, changing the weighting, have to be applied. The way proposed in connection with regression modelling discrete outputs is applicable. Its detailed development and corresponding experiments exceed the scope of this paper and will be published independently.

4. CONCLUDING REMARK

The paper presents an open-ended attempt to design good recursive estimators based on a common methodology. In this respect, it relies on a) motivating equivalence approach; b) asymptotic properties of the posterior pdfs and c) focus on a rich model class extending the useful but narrow dynamic exponential family. Our limited experience indicates that the attempt is promising but a lot of work remains to be done. For instance, i) relationships to expectation-maximization algorithm and stochastic approximations should be established; ii) rich experience of statistical research with batch and asymptotic versions of the problem exploited; and iii) numerically safe algorithms developed ...

5. ACKNOWLEDGEMENT

This research has been partially supported by AV ČR 1ET 100 750 401, MŠMT ČR 2C06001.

REFERENCES

Anderson, B.D.O. and J.B. Moore (1979). *Optimal Filtering*. Prentice Hall.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. Wiley. New York.

Bernardo, J.M. and A.F.M. Smith (1997). *Bayesian Theory*. 2 ed.. John Wiley & Sons. Chichester, New York, Brisbane, Toronto, Singapore.

Kárný, M., J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma and L. Tesař (2005). *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer. London.

Kashyap, R.L. and A.R. Rao (1976). *Dynamic Stochastic Models from Empirical Data*. Academic Press. New York.

Koopman, R. (1936). On distributions admitting a sufficient statistic. *Transactions of American Mathematical Society* **39**, 399.

Kulhavý, R. (1990*a*). A Bayes-closed approximation of recursive nonlinear estimation. *International Journal Adaptive Control and Signal Processing* **4**, 271–285.

Kulhavý, R. (1990*b*). Recursive Bayesian estimation under memory limitations. *Kybernetika* **26**, 1–20.

Kulhavý, R. (1996). *Recursive Nonlinear Estimation: A Geometric Approach*. Vol. 216 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag. London.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall. London.

Nedoma, P., M. Kárný, J. Böhm and T. V. Guy (2005). Mixtools Interactive User's Guide. Technical Report 2143. ÚTIA AV ČR. Praha.

Peterka, V. (1981). Bayesian system identification. In: *Trends and Progress in System Identification* (P. Eykhoff, Ed.). pp. 239–304. Pergamon Press. Oxford.