Computational Statistics & Data Analysis 17 (1994) 509–536
North-Holland

Methods for recursive robust estimation of AR parameters

Ken Sejling and Henrik Madsen
The Technical University of Denmark, Lyngby, Denmark

Jan Holst, Ulla Holst and Jan-Eric Englund
University of Lund, Lund, Sweden

Received October 1992
Revised February 1993

Abstract: In many technical applications, such as automatic control or supervision of systems, on-line predictions are required. Since the system of interest might change as time is passing, the model used for predictions must follow these changes. The estimation method therefore has to be adaptive implying the use of a recursive estimation algorithm. Furthermore, because of the possibility of outliers among the observations, it must be required that the applied estimation algorithm minimizes the influence of any sort of outliers.

In this paper two recursive robust estimation algorithms for estimation of AR models are derived. One of them implements a recursive minimization of a criterion function, in which prediction errors enter through the weight function proposed by Huber (1964). The other algorithm is a recursive version of the bounded-influence estimator proposed by Krasker and Welsch (1982). This estimator is an extension of the Huber estimator where a measure of the amount of aberrant information in each observation is used to down-weight the influence of observations that stand out among the rest. By these derivations a general procedure for obtaining recursive algorithms is demonstrated. In a simulation study the proposed methods are compared with ordinary recursive least squares, as well as a modification of this, in which classified outliers imply the corresponding observations to be left out of the estimation.

Keywords: Recursive estimation; Robust estimation; Bounded-influence estimation; Time series analysis; Additive and innovation outliers

1. Introduction

In many different engineering situations, like adaptive control of an industrial process, on-line prediction of power demand or predictive coding in adaptive

Correspondence to: K. Sejling, The Institute of Mathematical Statistics and Operations Research, The Technical University of Denmark, DK-2800 Lyngby, Denmark.

SSDI 0167-9473(93)E0012-S

speech transmission, it is necessary to carry out on-line estimation of parametric models in order to have the parameters in the model follow the changes in the system. However, gross errors in the observations can occur. A desirable quality of the algorithm used to estimate the parameters is that it is able to protect the estimates from the destructive influence from such gross errors, while retaining the ability to track changes in the system. Indeed, this is a demanding quality, since it can obviously be difficult to separate abnormal prediction errors caused by gross errors in the data from prediction errors arising as a consequence of abrupt changes in the system.

Solutions to the robustness issue have mainly appeared as off-line estimation methods. Due to the nonlinear nature of the estimation problem, these robust algorithms turn out to be iterative. Such off-line techniques are discussed in, e.g., Martin and Yohai (1985), Huber (1981) and Hampel et al. (1986). The fact that these methods utilize the whole batch of data implies that all past data, possibly numerous, must be stored in order to obtain new robust estimates when also the effect of added measurements is to be included.

The time-variation of the system and/or the abundance of data motivate a study of recursive algorithms. Such algorithms are adaptable to time-varying systems, and they also have considerable computational advantages compared to off-line techniques. Recursive algorithms are thoroughly investigated in Ljung and Söderström (1983). They are also discussed in Söderström and Stoica (1989) and Ljung (1987). However, in these references the robustness issue is only briefly discussed in connection with choice of criterion function.

Robust estimation based on stochastic approximation is studied in Martin and Masreliez (1975), and used in Campbell (1982) for M-estimation of the parameters in an AR process. Recursive robust estimation of static systems is discussed by Poljak and Tsypkin (1980). In Kuh and Samarov (1986) recursive robust estimation is used in connection with detection of shifts in a regression. Masreliez (1975) introduces a certain robustification of the Kalman filter, and his results are used by West (1981) in a study of sequential estimation of a location vector in linear regression.

Recently Allende and Heiler (1992) have proposed a multi-stage procedure for robust estimation of ARMA processes in the presence of additive outliers. In the first stage innovation estimates are found as the residuals of a robust autoregressive fit of high order. These residuals are then cleaned through a weight function, and subsequently a generalized M-estimation is carried out repeatedly, inserting the residuals found in the preceding step in the MA part of the regressor vector, until the parameter estimates have stabilized.

Recursive M-estimators of location and scale are treated in the papers by Englund et al. (1988, 1989). The results are mostly theoretical concerning the asymptotic behaviour of the estimators. In Englund (1991) linear regression models are studied, and, by using stochastic approximation, recursive algorithms, based on the bounded-influence estimator in Krasker and Welsch (1982), are proposed. Poulsen and J. Holst (1982) study recursive robust estimation of parameters and scale in adaptive control systems. They use a recursive version

of the Huber Criterion for the parameter estimation, and show that it is possible to separate the effects of outliers from systematic changes in the system description.

Cipra (1992) considers robust modifications of simple and double exponential smoothing by use of the L_1 norm, as well as recursive M-estimation with exponential forgetting of AR processes. The recursive M-estimation is obtained by introducing approximations in the off-line estimator, which allows for the formulation of the algorithm as the recursive solution to a set of normal equations.

In this paper a general method for obtaining recursive robust parameter estimation algorithms is used to derive two different recursive robust estimation algorithms. The method, which is a generalization of a method proposed in Sejling (1987), originates from a formulation of the estimation problem as a recursive minimization of a criterion with respect to the parameters. The two algorithms are based on the M-estimator proposed by Huber (1973) and the bounded-influence estimator by Krasker and Welsch (1982), respectively. These algorithms are compared both to the recursive least squares (RLS) algorithm and a modified RLS algorithm, in which observations are treated as missing if the corresponding prediction errors exceed a specified bound.

2. Outlier models and robust estimation

Traditionally two distinct kinds of outlier models are considered, viz. the innovation outlier model (IO) and the additive outlier model (AO), see e.g., Denby and Martin (1979). The two kinds of outliers can both be described by the following model

$$y(t) = z(t) + w(t), \tag{1}$$

where $\{y(t)\}$ is the observations and $\{z(t)\}$ is the process, which is restricted to the class of pure autoregressive processes. That is, $\{z(t)\}$ is given by

$$z(t) - \theta_1 z(t-1) - \cdots - \theta_p z(t-p) = e(t), \tag{2}$$

where p is the order of the AR process. $\{w(t)\}$ and $\{e(t)\}$ are mutually independent sequences of independent random variables. Innovation outliers are present when the distribution of the innovations is different from the assumed distribution, and the process is observed perfectly ($w(t) \equiv 0$). For instance, when $e(t)$ has a heavy-tailed non-Gaussian density, such as a mixture of two Gaussian densities, innovation outliers are traditionally assumed to be present due to the inconsistence with the assumption of Gaussian innovations. Additive outliers are present when the observations can differ from the process due to additive effects. A model describing the situation where additive outliers are present can be that $e(t)$ is Gaussian, and $w(t)$ is zero most of the time and when different from zero given by a suitable density function, for instance, a zero mean Gaussian density. Of course, both kinds of outliers can be present.

Indeed, this is an idealized description of outliers occurring in real systems. For instance, single outliers cannot always be classified to be either IO or AO, and often outliers occur in bursts or are correlated in some way. In model (1)–(2) this can be described by allowing the outliers to follow some dependence structure.

Additive outliers in data cause severe problems for the estimation of model parameters. For traditional estimation techniques, e.g., least squares (LS), the result will be estimates with bias, and, as demonstrated in Allende and Heiler (1992), iterative methods involving detection and filtering of the additive effects are required.

Several results exist in the area of off-line robust estimation in linear regression models and ARMA models. In Martin and Yohai (1985) the M-estimate of the parameters in an ARMA(p, q) model is discussed, where the estimate, $\hat{\theta}$, minimizes the criterion

$$V(n, \theta) = \sum_{i=1}^{n} \rho_c\left(\frac{\epsilon(i, \theta)}{\sigma}\right). \tag{3}$$

$\epsilon(i, \theta)$ is the prediction error, $\epsilon(i, \theta) = y(i) - x^{\mathrm{T}}(i)\theta$, and σ is the scale parameter. $\rho_c(u)$ is a weight function, which, for instance, could be chosen as proposed by Huber (1964)

$$\rho_c(u) = \begin{cases} \frac{1}{2}u^2, & |u| \leqslant c, \\ c|u| - \frac{1}{2}c^2, & |u| > c. \end{cases} \tag{4}$$

c is the parameter which determines the level of influence of large prediction errors. Different alternative weight functions are suggested in the literature (Martin and Yohai, 1985). For a simultaneous estimation of the scale parameter, σ, Huber (1964) proposes to use the equation (Proposal 2)

$$0 = \sum_{i=1}^{n} \chi_c\left(\frac{\epsilon(i, \hat{\theta})}{\hat{\sigma}}\right), \tag{5}$$

where

$$\chi_c(u) = \psi_c^2(u) - b \tag{6}$$

with

$$\psi_c(u) = \frac{\mathrm{d}}{\mathrm{d}u}\rho_c(u) = \max(\min(c, u), -c);$$

$$b = E\{\psi_c^2(z)\}, \quad z \sim \mathrm{N}(0, 1). \tag{7}$$

This choice of correction term, b, ensures that, if u is N(0, 1) the scale estimate is consistent.

In Krasker and Welsch (1982) the more elaborate bounded-influence estimator (θ_n, σ_n, A_n) is considered, where A_n is the dispersion of the regressors. This estimator applies to the linear regression model

$$y(t) = x^{\mathrm{T}}(t)\theta + e(t), \tag{8}$$

where $(y(i), x(i))$ are independent samples and the innovations are zero mean Gaussian variables. In this investigation we apply the following form of the off-line estimator

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi_c\left(\sqrt{x^{\mathrm{T}}(i)\hat{A}_n^{-1}x(i)}\,\frac{\epsilon(i,\hat{\theta})}{\hat{\sigma}}\right)\frac{x(i)}{\sqrt{x^{\mathrm{T}}(i)\hat{A}_n^{-1}x(i)}}\hat{\sigma}, \tag{9}$$

$$0 = \frac{1}{n}\sum_{i=1}^{n} \chi_c\left(\frac{\epsilon(i,\hat{\theta})}{\hat{\sigma}}\right), \tag{10}$$

$$\hat{A}_n = \frac{1}{n}\sum_{i=1}^{n} g_1\left(\frac{a}{\sqrt{x^{\mathrm{T}}(i)\hat{A}_n^{-1}x(i)}}\right)x(i)x^{\mathrm{T}}(i), \tag{11}$$

with

$$g_1(u) = E\{\min(z^2, u^2)\}, \quad z \sim \mathrm{N}(0, 1). \tag{12}$$

Instead of having only one tuning parameter for all three parts of the estimator (Krasker and Welsch, 1982) we use one parameter, c, in the part of the estimator which measures the prediction error (9)–(10) and another tuning parameter, a, in the covariance estimation. Furthermore, in (10) we have decided upon Huber's Proposal 2 for estimation of the scale parameter.

(9)–(10) can be considered as the estimator of Huber (1973), in which a measure of information in the regressor vector, $\sqrt{x^{\mathrm{T}}(i)\hat{A}_n^{-1}x(i)}$, has been introduced to impose an upper limit on the influence of each observation. (11) defines a robust estimate of the covariance of the regressors, $\hat{A}_n$, by means of a down-weighting of aberrant regressors through $g_1(u)$.

Krasker and Welsch (1982) analyze the appropriate choice of the boundary parameter a (when used in all three estimator equations) with respect to asymptotic efficiency compared to the LS estimator for Gaussian error structure (relative efficiency). To obtain a relative efficiency of, e.g., 0.95 or 0.99, they propose to choose a as 1.596 or 2.093 times $\sqrt{\dim x(i)}$, respectively. However, this cannot be carried over directly to the situation of dependent sequences and recursive estimation.

3. Recursive and adaptive estimation

The recursive estimation algorithms, considered in this paper, demand that the model, in which we wish to estimate the parameters, is linear in these parameters. Hence the model can be written as in (8). In addition, it is assumed that $x(t)$ and $e(t)$ are mutually independent.

Four algorithms for recursive estimation of an AR model are considered:

RLS, the traditional Recursive Least Squares algorithm.

RMO, a modification of the RLS algorithm in which large prediction errors, classified as outliers, imply the corresponding observations to be treated as missing.

RHU, an algorithm based on recursive minimization of Huber's criterion (3) combined with a recursive solution to Huber's Proposal 2 (5) for scale estimation.

RKW, an algorithm based on a recursive solution to the estimator by Krasker and Welsch (9)–(12).

For all of them a forgetting factor has been introduced to obtain exponential discounting of remote information.

RLS. In the off-line case the LS estimates are found as the minimizing argument of the LS criterion function, i.e.,

$$\hat{\theta}(t) = \operatorname{argmin}_{\theta} \tfrac{1}{2} \sum_{i=1}^{t} \lambda^{t-i} \epsilon^2(i, \theta). \tag{13}$$

The parameter estimates can be found analytically as the explicit solution to the derivative of the criterion (the Normal Equations), see Ljung and Söderström (1983). This solution can easily be transformed into a recursive formulation, thus appearing as the algorithm

$$P_{\lambda}(t) = \frac{P_{\lambda}(t-1)}{\lambda} - \frac{1}{\lambda} \frac{P_{\lambda}(t-1)x(t)x^{\mathrm{T}}(t)P_{\lambda}(t-1)}{\lambda + x^{\mathrm{T}}(t)P_{\lambda}(t-1)x(t)}, \tag{14}$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P_{\lambda}(t)x(t)\epsilon\big(t, \hat{\theta}(t-1)\big). \tag{15}$$

$P_{\lambda}(t)$ is a matrix containing information about the covariance of the regressors observed until time t. The inverse of $P_{\lambda}(t)$ is given by

$$P_{\lambda}^{-1}(t) = \lambda^t P_{\lambda}^{-1}(0) + \sum_{i=1}^{t} \lambda^{t-i} x(i) x^{\mathrm{T}}(i) \tag{16}$$

with $P_{\lambda}(0)$ being the initial value.

RMO. This algorithm is an extension of RLS. Each residual is compared to a recursive estimate of the residual variance. If the numerical value of the residual exceeds c times the variance estimate, neither the estimate of the models parameters, the estimate of the P-matrix nor the residual variance estimate are updated. Hence, the algorithm is

$$k_{\sigma}(t) = \max\left(\frac{1}{t}, 1 - \lambda\right), \tag{17}$$

$$\hat{\sigma}^2(t) = \hat{\sigma}^2(t-1) + k_{\sigma}(t)\big(d_c \epsilon^2\big(t, \hat{\theta}(t-1)\big) - \hat{\sigma}^2(t-1)\big) \times I_{\{|\epsilon(t, \hat{\theta}(t-1))| < c\hat{\sigma}(t-1)\}}, \tag{18}$$

$$P_\lambda(t) = \frac{P_\lambda(t-1)}{\lambda} - \frac{1}{\lambda}\frac{P_\lambda(t-1)x(t)x^{\mathrm{T}}(t)P_\lambda(t-1)}{\lambda + x^{\mathrm{T}}(t)P_\lambda(t-1)x(t)} I_{\{|\epsilon(t,\hat{\theta}(t-1))| < c\hat{\sigma}(t-1)\}}, \tag{19}$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P_\lambda(t)x(t)\epsilon\big(t, \hat{\theta}(t-1)\big) I_{\{|\epsilon(t,\hat{\theta}(t-1))| < c\hat{\sigma}(t-1)\}}. \tag{20}$$

The gain $k_\sigma(t)$ is introduced to enable tracking of a slowly varying scale. Likewise, the scale factor d_c is used to obtain a consistent scale estimate, when the residuals are zero mean Gaussian random variables. This is obtained with

$$d_c^{-1} = E\{\phi_c^2(z)\}, \quad z \sim \mathrm{N}(0, 1),$$

$$\phi_c(u) = \begin{cases} u, & |u| \leqslant c, \\ 0, & |u| > c. \end{cases} \tag{21}$$

This method has the obvious disadvantage that if the estimates are too far from the true parameter values, then the prediction errors will always be too big, and neither the parameters nor the P-matrix will be updated.

RHU. The RLS algorithm can be regarded either as a recursive formulation of the off-line estimator (the Normal Equations) or as the solution to the recursive minimization of the LS criterion. The two approaches are equivalent since they both give exact solutions for the LS criterion and the model assumptions given above. However, when considering non-quadratic criterion functions the two approaches can lead to different algorithms depending on the nature of the approximations made. In the recursive minimization of Huber's criterion

$$V_{\mathrm{Hu}}^\lambda(t, \theta) = \sum_{i=1}^{t} \lambda^{t-i} \rho_c\left(\frac{\epsilon(i, \theta)}{\sigma}\right)\sigma^2, \tag{22}$$

with $\rho_c(u)$ defined in (4), our approach is to minimize (22) using approximations, which are justifiable, when the parameter estimates are close to their optimum. These approximations and the change of the criterion between consecutive steps are then used to obtain the connection between the parameter estimates at consecutive steps. The derivation of the RHU algorithm is found in Section 3.1 below.

RKW. The Recursive Krasker and Welsch algorithm is obtained using the same approach as for RHU by recursive minimization of the criterion

$$V_{\mathrm{KW}}^\lambda(t, \theta) = \sum_{i=1}^{t} \lambda^{t-i} \frac{\sigma^2}{x^{\mathrm{T}}(i)\hat{A}_t^{-1}x(i)} \rho_c\left(\sqrt{x^{\mathrm{T}}(i)\hat{A}_t^{-1}x(i)}\,\frac{\epsilon(i, \theta)}{\sigma}\right). \tag{23}$$

It is necessary to supplement this by a recursive estimation of the covariance matrix. The RKW algorithm is derived in Section 3.2.

3.1. Recursive minimization of the Huber Criterion

In this section the algorithm, implementing a recursive approximative minimization of the Huber Criterion (22), is derived. It is assumed that the scale

parameter, σ, is known. In Section 3.3 it is shown how a recursive estimation of the scale parameter is obtained from the off-line estimator (5)–(7).

Let $\hat{\theta}(t-1)$ be the estimate at time $t-1$. Then the result of applying the algorithm on an observation at time t should be the estimate $\hat{\theta}(t)$ making an approximative minimization of the criterion $V_{\mathrm{Hu}}^{\lambda}(t, \theta)$ (22). A Taylor expansion of the criterion around $\hat{\theta}(t-1)$ gives

$$\begin{aligned} V_{\mathrm{Hu}}^{\lambda}(t, \theta) = {} & V_{\mathrm{Hu}}^{\lambda}\big(t, \hat{\theta}(t-1)\big) + \big[\theta - \hat{\theta}(t-1)\big]^{\mathrm{T}} \nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t, \theta)_{|\hat{\theta}(t-1)} \\ & + \tfrac{1}{2}\big[\theta - \hat{\theta}(t-1)\big]^{\mathrm{T}} \nabla_{\theta}\nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t, \theta)_{|\hat{\theta}(t-1)} \big[\theta - \hat{\theta}(t-1)\big] \\ & + \mathrm{o}\Big(\big|\theta - \hat{\theta}(t-1)\big|^{2}\Big), \end{aligned} \tag{24}$$

where ∇_{θ} denotes differentiation with respect to θ, and $\mathrm{o}(x)$ is a function for which $\mathrm{o}(x)/|x| \to 0$ as $|x| \to 0$. Taking the derivative of this expression with respect to θ, and applying the assumption of the gradient of the criterion being zero in every time step when evaluated at the recursively obtained parameter estimates, i.e.,

$$\forall t : \nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t, \theta)_{|\hat{\theta}(t)} = 0, \tag{25}$$

lead to the following parameter update

$$\begin{aligned} \hat{\theta}(t) = {} & \hat{\theta}(t-1) - \big[\nabla_{\theta}\nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t, \theta)_{|\hat{\theta}(t-1)}\big]^{-1} \nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t, \theta)_{|\hat{\theta}(t-1)} \\ & + \mathrm{o}\big(\big|\hat{\theta}(t) - \hat{\theta}(t-1)\big|\big). \end{aligned} \tag{26}$$

Taking the derivative of the criterion (22) written as a recursion,

$$V_{\mathrm{Hu}}^{\lambda}(t, \theta) = \lambda V_{\mathrm{Hu}}^{\lambda}(t-1, \theta) + \rho_{c}\left(\frac{\epsilon(t, \theta)}{\sigma}\right)\sigma^{2}, \tag{27}$$

and using the assumption (25) at $t-1$ gives

$$\nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t, \theta)_{|\hat{\theta}(t-1)} = -x(t)\psi_{c}\left(\frac{\epsilon\big(t, \hat{\theta}(t-1)\big)}{\sigma}\right)\sigma. \tag{28}$$

Evaluation of the second derivative of (27) in $\hat{\theta}(t-1)$ gives

$$\begin{aligned} \nabla_{\theta}\nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t, \theta)_{|\hat{\theta}(t-1)} = {} & \lambda \nabla_{\theta}\nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t-1, \theta)_{|\hat{\theta}(t-1)} \\ & + x(t)x^{\mathrm{T}}(t)\psi_{c}'\left(\frac{\epsilon\big(t, \hat{\theta}(t-1)\big)}{\sigma}\right) \end{aligned} \tag{29}$$

with $\psi_{c}'(u) = (\mathrm{d}/\mathrm{d}u)\psi_{c}(u)$.

If it is assumed that $\hat{\theta}(t)$ is close to $\hat{\theta}(t-1)$ the following two approximations can be justified:

- The term $\mathrm{o}(|\hat{\theta}(t) - \hat{\theta}(t-1)|)$ in (26) can be neglected.
- $\nabla_{\theta}\nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t-1, \theta)_{|\hat{\theta}(t-1)}$ can be replaced by $\nabla_{\theta}\nabla_{\theta} V_{\mathrm{Hu}}^{\lambda}(t-1, \theta)_{|\hat{\theta}(t-2)}$ in (29).

Finally, by denoting the inverse of the second-order derivative of the criterion as

$$P_\lambda(t) \equiv \left[\nabla_\theta \nabla_\theta V_{\mathrm{Hu}}^\lambda(t, \theta)_{|\hat\theta(t-1)}\right]^{-1}, \tag{30}$$

and using the Matrix Inversion Lemma on (29), the algorithm is

$$P_\lambda(t) = \frac{P_\lambda(t-1)}{\lambda} - \frac{1}{\lambda}\psi_c'\left(\frac{\epsilon\big(t, \hat\theta(t-1)\big)}{\sigma}\right)\frac{P_\lambda(t-1)x(t)x^{\mathrm{T}}(t)P_\lambda(t-1)}{\lambda + x^{\mathrm{T}}(t)P_\lambda(t-1)x(t)}, \tag{31}$$

$$\hat\theta(t) = \hat\theta(t-1) + P_\lambda(t)x(t)\psi_c\left(\frac{\epsilon\big(t, \hat\theta(t-1)\big)}{\sigma}\right)\sigma. \tag{32}$$

Note that the expression in (31) has been simplified by use of $\psi_c'(u)$ being either zero or one.

3.2. Recursive minimization of the Krasker and Welsch Criterion

In this section a recursive algorithm based on the off-line estimator by Krasker and Welsch (1982) is derived. The algorithm is obtained as the recursive and approximative minimization of the criterion (23) accompanied by a recursive and approximative expression for update of the robust covariance estimator (11). As for the RHU algorithm, the scale estimation is separated from the model parameter estimation.

The recursive minimization of (23) is derived using the same procedure as in the derivation of RHU. The first step is to write down the second order Taylor expansion of the criterion and use the assumption that the criterion derivative at time t is zero when evaluated at $\hat\theta(t)$. As outlined below, however, it will be necessary to introduce a modified criterion, $V_{\mathrm{KW}}^{\lambda,r}(t, \theta)$, to be able to obtain a recursive expression for the criterion. For this reason the second order expansion is applied on the modified criterion leading to the following expression for the parameter update

$$\begin{aligned}\hat\theta(t) = \hat\theta(t-1) &- \left[\nabla_\theta \nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t, \theta)_{|\hat\theta(t-1)}\right]^{-1}\nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t, \theta)_{|\hat\theta(t-1)} \\ &+ o\big(\big|\hat\theta(t) - \hat\theta(t-1)\big|\big).\end{aligned} \tag{33}$$

To obtain expressions for the first- and second-order derivative of the criterion in (33), the criterion (23) must be written as a recursive expression. This requires that the batch covariance estimate $\hat A^{-1}(t)$ in the criterion (23) is replaced by the, at time i, available covariance estimate, i.e.,

$$V_{\mathrm{KW}}^{\lambda,r}(t, \theta) = \sum_{i=1}^{t} \lambda^{t-i}\frac{\sigma^2}{x^{\mathrm{T}}(i)\hat A^{-1}(i)x(i)}\rho_c\left(\sqrt{x^{\mathrm{T}}(i)\hat A^{-1}(i)x(i)}\,\frac{\epsilon(i, \theta)}{\sigma}\right). \tag{34}$$

This criterion can be written recursively as

$$V_{\mathrm{KW}}^{\lambda,r}(t,\theta)=\lambda V_{\mathrm{KW}}^{\lambda,r}(t-1,\theta)+\frac{\sigma^2}{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}\rho_c\left(\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}\,\frac{\epsilon(t,\theta)}{\sigma}\right). \tag{35}$$

Computing the gradients in (35), and using the assumption of the criterion gradient being zero in the parameter estimate gives

$$\nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t,\theta)_{|\hat{\theta}(t-1)}=-x(t)\frac{\sigma}{\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}}\times\psi_c\left(\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}\,\frac{\epsilon\big(t,\hat{\theta}(t-1)\big)}{\sigma}\right). \tag{36}$$

The second-order derivative of the criterion expression, evaluated at $\hat{\theta}(t-1)$, becomes

$$\nabla_\theta\nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t,\theta)_{|\hat{\theta}(t-1)}=\lambda\nabla_\theta\nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t-1,\theta)_{|\hat{\theta}(t-1)}+x(t)x^{\mathrm{T}}(t)\psi_c'\left(\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}\,\frac{\epsilon\big(t,\hat{\theta}(t-1)\big)}{\sigma}\right). \tag{37}$$

As in the derivation of RHU the following approximations are made

- The term $o(|\hat{\theta}(t)-\hat{\theta}(t-1)|)$ in (33) can be neglected.
- $\nabla_\theta\nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t-1,\theta)_{|\hat{\theta}(t-1)}$ can be replaced by $\nabla_\theta\nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t-1,\theta)_{|\hat{\theta}(t-2)}$ (37).

Applying the Matrix Inversion Formula on (37) and introducing the notation

$$P_\lambda(t)\equiv\left[\nabla_\theta\nabla_\theta V_{\mathrm{KW}}^{\lambda,r}(t,\theta)_{|\hat{\theta}(t-1)}\right]^{-1} \tag{38}$$

yields (42). Inserting the expression (36) in (33) gives the final expression of the parameter update (43).

To obtain the recursive formula for the covariance estimate, the batch estimate, $\hat{A}_n$, entering g_1 in summand i in (11) is replaced by the estimate which is available at time $i-1$. This gives

$$\hat{A}(t)=\frac{1}{n}\sum_{i=1}^{n}g_1\left(\frac{a}{\sqrt{x^{\mathrm{T}}(i)\hat{A}^{-1}(i-1)x(i)}}\right)x(i)x^{\mathrm{T}}(i), \tag{39}$$

which can be written recursively as

$$\hat{A}(t)=\hat{A}(t-1)+\frac{1}{t}\left(g_1\left(\frac{a}{\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t-1)x(t)}}\right)x(t)x^{\mathrm{T}}(t)-\hat{A}(t-1)\right). \tag{40}$$

By use of the Matrix Inversion Formula (40 is transformed into (41). This completes the derivation of the RKW algorithm.

$$\hat{A}^{-1}(t) = \frac{t}{t-1}\left(\hat{A}^{-1}(t-1) - g_1\left(\frac{a}{\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t-1)x(t)}}\right) \times \frac{\hat{A}^{-1}(t-1)x(t)x^{\mathrm{T}}(t)\hat{A}^{-1}(t-1)}{t + x^{\mathrm{T}}(t)\hat{A}^{-1}(t-1)x(t)g_1\left(a/x^{\mathrm{T}}(t)\hat{A}^{-1}(t-1)x(t)\right)}\right), \tag{41}$$

$$P_\lambda(t) = \frac{P_\lambda(t-1)}{\lambda} - \frac{1}{\lambda}\psi_c'\left(\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}\,\frac{\epsilon\left(t, \hat{\theta}(t-1)\right)}{\sigma}\right) \times \frac{P_\lambda(t-1)x(t)x^{\mathrm{T}}(t)P_\lambda(t-1)}{\lambda + x^{\mathrm{T}}(t)P_\lambda(t-1)x(t)}, \tag{42}$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P_\lambda(t)x(t)\frac{\sigma}{\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}} \times \psi_c\left(\sqrt{x^{\mathrm{T}}(t)\hat{A}^{-1}(t)x(t)}\,\frac{\epsilon\left(t, \hat{\theta}(t-1)\right)}{\sigma}\right). \tag{43}$$

Note that (42) has been simplified by using that $\psi_c'(u)$ is either zero or one.

3.3. Recursive estimation of scale

The algorithm for recursive estimation of scale is based on Huber's Proposal 2 given in (5). To develop the recursive scale estimation the function

$$X_\lambda(t, \sigma) = \sum_{i=1}^{t} \lambda^{t-i}\chi_c\left(\frac{\epsilon(i, \theta)}{\sigma}\right) \tag{44}$$

is introduced, in which the set of model parameters, θ, is assumed to be known.

The function X_λ can be written recursively as

$$X_\lambda(t, \sigma) = \lambda X_\lambda(t-1, \sigma) + \chi_c\left(\frac{\epsilon(t, \theta)}{\sigma}\right). \tag{45}$$

By applying a first-order Taylor expansion of the X_λ-function around $\hat{\sigma}(t-1)$, and assuming that the scale estimate at time t satisfies the estimator definition, i.e., $X_\lambda(t, \hat{\sigma}(t)) = 0$, give

$$0 = X_\lambda(t, \hat{\sigma}(t-1)) + [\hat{\sigma}(t) - \hat{\sigma}(t-1)]\nabla_\sigma X_\lambda(t, \sigma)_{|\hat{\sigma}(t-1)} + o(|\hat{\sigma}(t) - \sigma(t-1)|). \tag{46}$$

The first term on the right-hand side of (46) is found by evaluating (45) in $\hat{\sigma}(t-1)$ and using the assumption $X_\lambda(t-1, \hat{\sigma}(t-1)) = 0$. This gives

$$X_\lambda(t, \hat{\sigma}(t-1)) = \chi_c\left(\frac{\epsilon(t, \theta)}{\hat{\sigma}(t-1)}\right). \tag{47}$$

Evaluating the derivative of (45) in $\hat{\sigma}(t-1)$ gives

$$\nabla_\sigma X_\lambda(t, \sigma)_{|\hat{\sigma}(t-1)} = \lambda \nabla_\sigma X_\lambda(t-1, \sigma)_{|\hat{\sigma}(t-1)} - 2\frac{\epsilon^2(t, \theta)}{\hat{\sigma}^3(t-1)} I_{\{|\epsilon(t,\theta)| \leq c\sigma(t-1)\}}. \tag{48}$$

Now the following approximations are made

- The term $o(|\hat{\sigma}(t) - \hat{\sigma}(t-1)|)$ in (46) can be neglected.
- $\nabla_\sigma X_\lambda(t-1, \sigma)_{|\hat{\sigma}(t-1)}$ can be replaced by $\nabla_\sigma X_\lambda(t-1, \sigma)_{|\hat{\sigma}(t-2)}$ in (48).

By introducing $h_\lambda(t) \equiv -\nabla_\sigma X_\lambda(t, \sigma)_{|\hat{\sigma}(t-1)}$ this completes the derivation of the recursive scale estimator

$$h_\lambda(t) = \lambda h_\lambda(t-1) + 2\frac{\epsilon^2(t, \theta)}{\hat{\sigma}^3(t-1)} I_{\{|\epsilon(t,\theta)| \leq c\hat{\sigma}(t-1)\}} \tag{49}$$

$$\hat{\sigma}(t) = \hat{\sigma}(t-1) + \frac{\chi_c(\epsilon(t, \theta)/\hat{\sigma}(t-1))}{h_\lambda(t)}. \tag{50}$$

Due to the symmetry of χ_c the scale estimate can converge to both a positive value and its negative counterpart. To avoid a negative scale estimate the set of permissible values is restricted to the positive real line. In practice this is only of significance in the initial phase.

4. Simulation and estimation results

All programming in the simulation and estimation study was performed in Pascal on a HP-9000/835 using double precision arithmetic. For the generation of Gaussian and uniform random variables the pseudo random number generators of the IMSL Stat/Library (1987) DRNNOF and DRNUNF were used. For calculation of the solution to the off-line estimation using (9)–(12) the nonlinear-equation solver DNEQNF of the IMSL Math/Library (1987) was used.

4.1. Simulation of data

The simulation results in the present paper are all based on data simulated from the following two pure autoregressive models

$$(1 - 0.8q^{-1})z(t) = e(t) \tag{51}$$

and

$$(1 - 1.20q^{-1} + 0.52q^{-2})z(t) = e(t) \tag{52}$$

with the poles

$$z = 0.6 \pm 0.4i \qquad |z| = 0.72. \tag{53}$$

Each data sequence has a length of 3005 observations and is simulated either without outliers, with innovation outliers or with additive outliers. If the sequence is simulated with outliers, these will not occur in the first 5 observations. Due to this and the way the estimation algorithms are applied, the observations are numbered from -4 to 3000. The three types of simulated noise acting on the models are described in the following three paragraphs.

No Outliers. When the observations are computed without outliers, the innovations $\{e(t)\}$ are independent and identically distributed Gaussian random variables with zero mean and variance $\sigma^2 = 1$.

Innovation Outliers. When the observations $\{y(t)\}$ are simulated with innovation outliers they are still identical to the process values $\{z(t)\}$, but the innovations $\{e(t)\}$ are created from a contaminated Gaussian distribution. The innovations are independent, but with some minor probability p they are sampled from a distribution with a larger variance. In this simulation study, the sequences of innovations with outliers are computed from the following density

$$e(t) \sim 0.95 \times \text{“NID(0, 1)”} + 0.05 \times \text{“NID(0, 6.25)”}, \tag{54}$$

where NID(μ, σ^2) denotes independent Gaussian random variables with mean μ and variance σ^2.

Additive Outliers. When the observations are contaminated with additive outliers, they can differ from the process due to additive noise $\{w(t)\}$,

$$y(t) = z(t) + w(t). \tag{55}$$

$\{z(t)\}$ is simulated as in the situation without outliers, and the additive noise $\{w(t)\}$ is zero with probability 0.95 and NID(0, 6.25) with probability 0.05,

$$w(t) \sim 0.95 \times \text{“0”} + 0.05 \times \text{“NID(0, 6.25)”}. \tag{56}$$

Figure 1 shows an example of observations created with the first-order model (51) and with the three schemes outlined above. With both kinds of outliers (IO or AO), it is possible to observe the existence of some outlying observations. Also the different characteristics of the influence on the observations resulting from the different types of outliers can be observed. The innovation outliers affect the process itself and hereby imply a lasting excitation, whereas the additive outliers are additive spikes in the very moment they occur.

4.2. *Application of algorithms*

For each simulated data sequence the four estimation algorithms RLS, RHU(c), RKW(c, a) and RMO(c), with the parameters c and a as specified in earlier

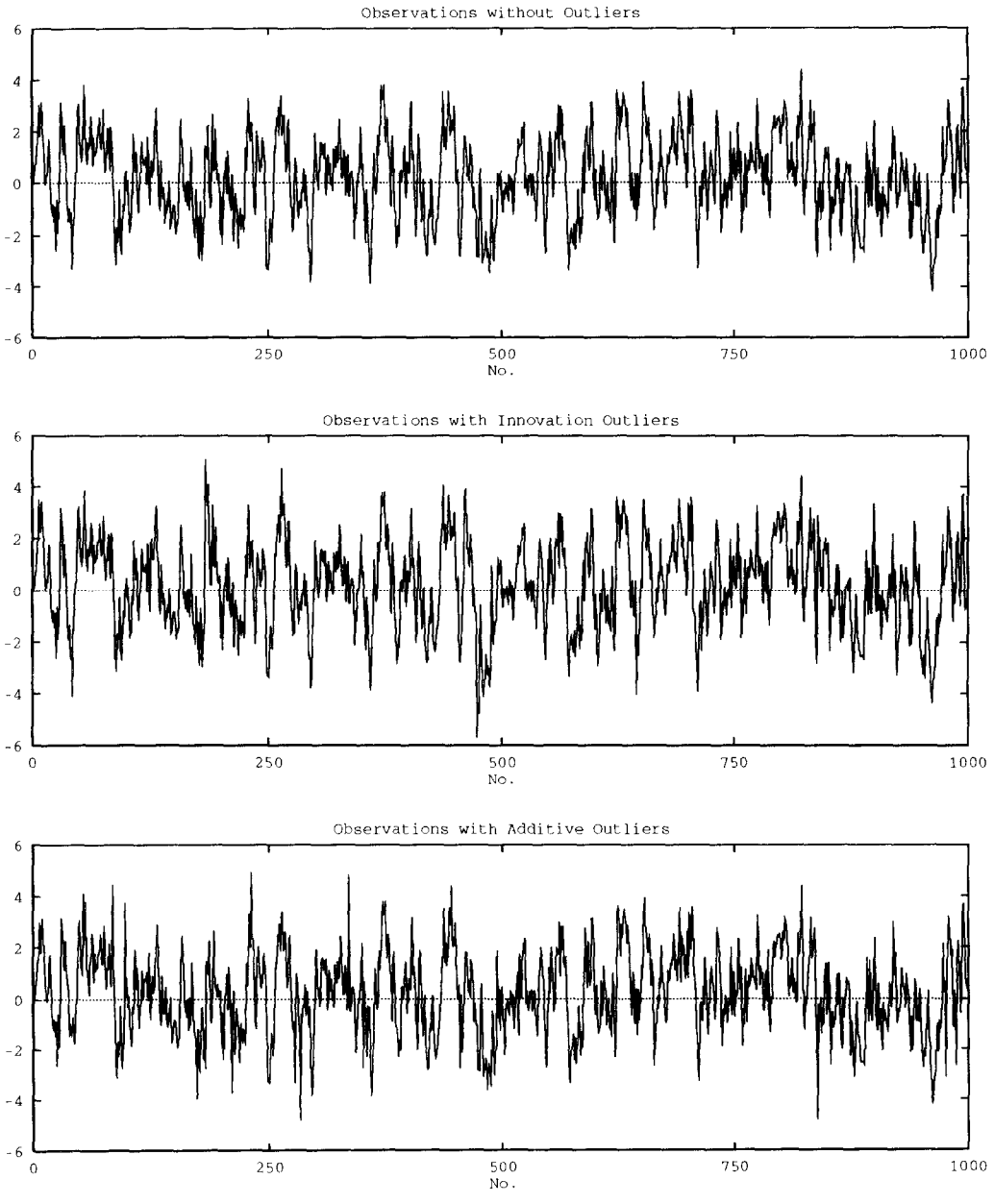

Fig. 1. Observations simulated with the first order model and using the three different outlier types with the same basic noise realization.

sections, are applied. The recursive estimation in all of the algorithms is started by using RLS on the first 5 observations. At $t = 1$ the algorithm to be studied is initialized with the parameters, $\hat{\theta}(0)_{\mathrm{RLS}}$, and covariance matrix, $P(0)_{\mathrm{RLS}}$, obtained from RLS. RLS is initialized with the covariance matrix equal to 100 times the unit matrix, and so is the robust covariance estimate of RKW, i.e., $\hat{A}^{-1}(0) = 100I$.

The starting value of $\hat{\sigma}^2(0)$ is the known simulation variance of the outlier free innovations, and the starting value of $h(0)$ in the recursive version of Huber's Proposal 2 is set to 1. For the first order model (51), RHU(c), RKW(c, a) and RMO(c) is applied with $c = 2$, 3 and for each value of c, RKW(c, a) is applied with $a = 2$, 3, 4. For comparison the off-line Krasker and Welsch estimator (9)–(12), denoted KW(c, a), is applied with the combinations $(c, a) = (2, 3)$, (3, 3) and (3, 4). The off-line estimation is based on observations for $t \in (1, 3000)$, and is applied to all three types of data sequences.

Estimation of the second-order model (52) is carried out with the recursive algorithms with $c = 2$ and $a = 5$ on all three types of data sequences.

Since the data are simulated with constant parameters, all algorithms are applied with the forgetting factor $\lambda = 1$.

The performance of the algorithms is illustrated in Figure 2, which shows the trajectories of the estimated parameter in the first-order model (51) for different kinds of observations and different estimation algorithms. The basic sequence of innovations was the same in all of the estimations shown. The upper part of the figure shows RLS and RKW(2, 3) applied to a data sequence without outliers. There is almost no difference to be observed between RLS and RKW(2, 3), and the estimates converge nicely to the simulation parameter. The same pattern applies to RHU(2) and RMO(2) (not shown).

The following two plots show the same algorithms used on an observation sequence simulated with innovation outliers as given in Section 4.1. Again there is almost no difference between RLS and RKW(2, 3). There is a slight tendency that the estimates from both algorithms lie even closer to the simulation parameters, than when data were simulated without outliers. This harmonize with the fact that the higher variability allows for an increased efficiency. Again the same pattern applies for RHU(2) and RMO(2) (not shown).

The bottom half of the figure shows the application of all four algorithms on a sequence simulated with additive outliers. Clearly the additive outliers have a rather drastic effect on the estimates for all algorithms. The most clear effect is observed on the RLS estimates, but also the RHU(2) algorithm seems to allow rather considerable steps in the estimates. On the contrary the RKW(2, 3) algorithm shows a smooth course of the estimates, and it gives a considerable improvement compared to RHU(2). The RMO(2) algorithm is in this case doing even better than RKW(2, 3) although more clear steps are seen. RMO(2) is obviously doing an easy job here classifying some of the additive outliers, but this simple algorithm still suffers from the possibility of the estimates becoming that bad that the prediction residuals are too big compared to the estimated variance of the prediction errors, implying that the estimates will not be updated.

The trajectories in Figure 2 were obtained using a basic innovation sequence, which was chosen at random among the simulation sequences. Although the specific trajectories have a clear dependence on the set of innovations and outliers, Figure 2 gives a representative illustration of the performance of the algorithms. The general experience from the simulation study is that the

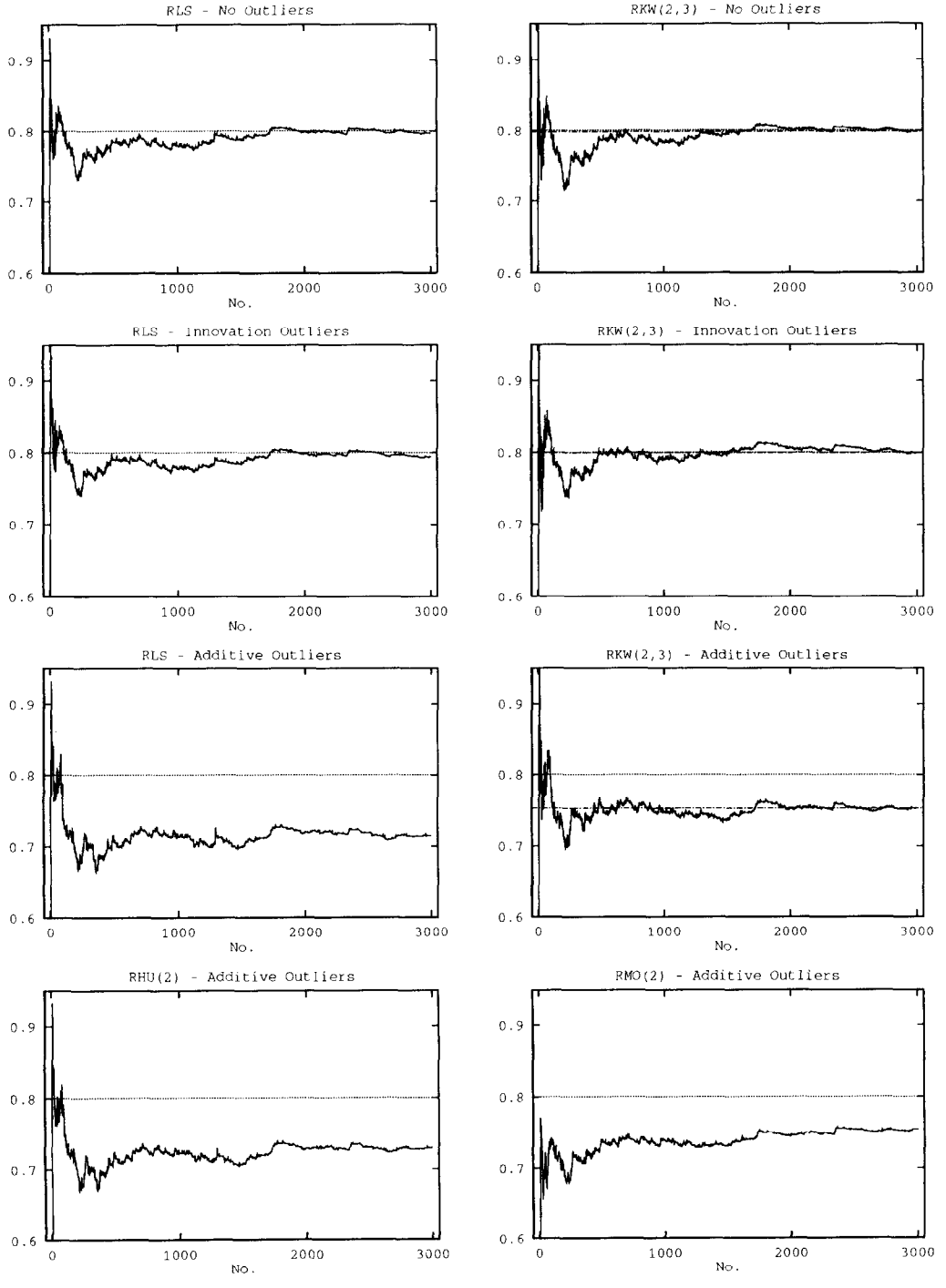

Fig. 2. Trajectories of the parameter estimates in the first order system for different outlier types and algorithms. The basic innovation sequence was the same in all simulations and the same as in Figure 1. The off-line Krasker and Welsch result is also marked on the figures.

Table 1
Results with no outliers and first-order model

No Outliers $\theta = 0.8$		$t = 2000$					$t = 3000$				
Algorithm	$\overline{SSD}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.95}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.95}$
RLS	0.143	0.799	0.013	0.775	0.800	0.819	0.799	0.011	0.780	0.800	0.817
RHU(2)	0.145	0.799	0.013	0.775	0.800	0.819	0.799	0.011	0.780	0.800	0.816
RHU(3)	0.143	0.799	0.013	0.775	0.800	0.819	0.799	0.011	0.780	0.800	0.817
RKW(2, 2)	0.180	0.799	0.015	0.771	0.800	0.822	0.799	0.012	0.778	0.800	0.819
RKW(2, 3)	0.166	0.799	0.014	0.772	0.800	0.820	0.799	0.012	0.779	0.799	0.818
RKW(2, 4)	0.163	0.799	0.014	0.773	0.800	0.820	0.799	0.012	0.780	0.799	0.818
RKW(3, 2)	0.164	0.799	0.014	0.772	0.800	0.820	0.799	0.012	0.779	0.799	0.818
RKW(3, 3)	0.154	0.799	0.014	0.773	0.800	0.820	0.799	0.011	0.780	0.800	0.817
RKW(3, 4)	0.152	0.799	0.014	0.774	0.800	0.820	0.799	0.011	0.780	0.799	0.817
RMO(2)	0.215	0.798	0.016	0.770	0.798	0.823	0.799	0.013	0.776	0.799	0.820
RMO(3)	0.147	0.799	0.013	0.775	0.800	0.820	0.799	0.011	0.780	0.800	0.817

Table 2
Results with innovation outliers and first-order model

Innovation $\theta = 0.8$		$t = 2000$					$t = 3000$				
Algorithm	$\overline{SSD}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.95}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.95}$
RLS	0.144	0.799	0.013	0.774	0.799	0.820	0.799	0.011	0.780	0.800	0.817
RHU(2)	0.130	0.799	0.013	0.776	0.799	0.818	0.799	0.010	0.781	0.800	0.816
RHU(3)	0.136	0.799	0.013	0.775	0.799	0.819	0.799	0.011	0.781	0.800	0.816
RKW(2, 2)	0.156	0.799	0.014	0.774	0.799	0.820	0.799	0.011	0.781	0.800	0.818
RKW(2, 3)	0.147	0.799	0.014	0.774	0.800	0.819	0.799	0.011	0.781	0.800	0.817
RKW(2, 4)	0.145	0.799	0.013	0.775	0.799	0.819	0.799	0.011	0.781	0.800	0.817
RKW(3, 2)	0.145	0.799	0.013	0.774	0.800	0.819	0.799	0.011	0.781	0.800	0.817
RKW(3, 3)	0.140	0.799	0.013	0.774	0.800	0.819	0.799	0.011	0.781	0.799	0.816
RKW(3, 4)	0.139	0.799	0.013	0.775	0.799	0.819	0.799	0.011	0.781	0.799	0.816
RMO(2)	0.187	0.798	0.015	0.772	0.798	0.823	0.799	0.012	0.779	0.798	0.819
RMO(3)	0.132	0.799	0.013	0.777	0.799	0.819	0.799	0.010	0.782	0.799	0.816

algorithms, for the chosen initializations, in all cases performed nicely showing convergence to limit values.

4.3. Performance evaluation

To evaluate the performance of the algorithms six experiments were carried out. One experiment consists of 1000 simulations with one of the models and one of the three noise simulation algorithms. All of the estimation algorithms are applied to each data sequence. For each sequence the sum of squared deviations *SSD* from the true parameter from observation no. 2001 till no. 3000 is computed

$$SSD = \sum_{i=2001}^{3000} \left(\hat{\theta}(i) - \theta\right)^2. \tag{57}$$

as is the average $\overline{SSD}$ of the 1000 values of *SSD*. Likewise, the average $\overline{\hat{\theta}(t)}$, the standard deviation $s_{\hat{\theta}(t)}$ and the 5, 50 and 95% fractiles $\hat{\theta}(t)_{0.05}$, $\hat{\theta}(t)_{0.50}$ and $\hat{\theta}(t)_{0.95}$ of the recursively obtained estimates at $t = 2000$ and $t = 3000$ are computed. The average, the standard deviation and the same fractiles are also computed for the estimates obtained from the off-line Krasker and Welsch algorithm.

To give a visual interpretation of the distribution of parameter estimates a non-parametric kernel density estimator (Silverman, 1986)

$$\hat{f}(u) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{u - U_i}{h}\right) \tag{58}$$

has been applied. h is the bandwidth (smoothing parameter), n is the number of estimates used in calculation and K is the kernel (weight function). In this study the Epanechnikov Kernel is used

$$K(u) = \begin{cases} \frac{3}{4}\left(1 - \frac{1}{5}u^2\right)/\sqrt{5}, & |u| < 5, \\ 0, & |u| \geqslant 5. \end{cases} \tag{59}$$

4.4. Discussion of the results

Table 1 shows the results of estimations carried out on data simulated with the first-order model without outliers. There is almost no difference in the results for the different algorithms, but it can be seen that there is a slight deterioration for lower values of c and a. This is reasonable since no erroneous information is present. Figure 3 shows the kernel estimate of the density of the estimates at $t = 3000$ obtained with $c = 2$ and $a = 3$. RLS obviously has the most narrow density, but RHU(2) is only slightly inferior. RKW(2, 3), however, shows a density with a little wider tails and a lower top level, and this is even more pronounced for RMO(2). This can be ascribed to the down-weighting or exclusion of information that actually is in accordance with the model.

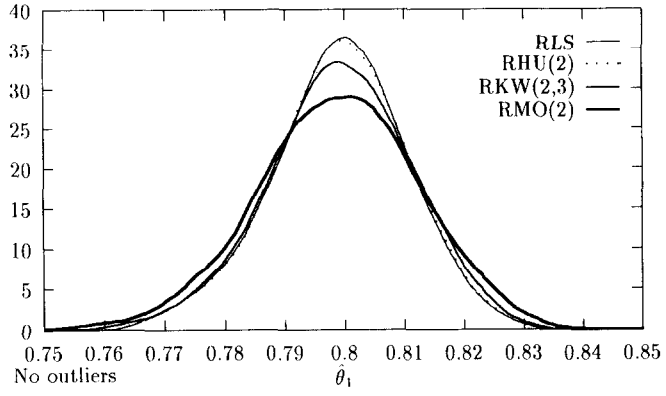

Fig. 3. Kernel estimate of the density of the parameter estimate at $t = 3000$ for data without outliers. The kernel is Epanechnikov with bandwidth $h = 0.004$.

In Table 2 is seen that with innovation outliers in data RHU, RKW and RMO improve compared to the situation with no outliers, whereas for RLS the result is very much the same without outliers and with innovation outliers. This is in accordance with Martin and Yohai (1985), who show that the asymptotic covariance of the LS estimates of AR and MA parameters is independent of the innovation distribution. However, for a heavy-tailed innovation distribution LS estimates can be inefficient compared to maximum likelihood (ML) estimates for the reason that LS does not make the most out of the increased precision attainable with heavy-tailed innovation distributions. When the distribution is not known, the M-estimator (3) can be applied for better utilization of the increased excitation for generally not known heavy-tailed innovation distributions (Martin and Yohai, 1985). Certainly, this is the reason that RHU, RKW and RMO, for some values of c and a, do better than RLS.

Kernel estimates of the recursive parameter estimates at $t = 3000$ in the first-order system with innovation outliers present is given in Figure 4. The

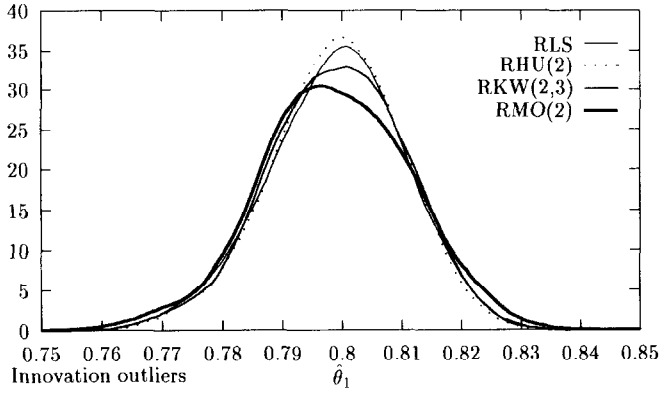

Fig. 4. Kernel estimate of the density of the parameter estimate at $t = 3000$ for data with innovation outliers. The kernel is Epanechnikov with bandwidth $h = 0.004$.

Table 3
Results with additive outliers and first-order model

Additive $\theta = 0.8$		$t = 2000$					$t = 3000$				
Algorithm	$\overline{SSD}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.95}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\theta(t)_{0.95}$
RLS	7.132	0.718	0.023	0.678	0.719	0.752	0.718	0.018	0.687	0.719	0.747
RHU(2)	4.364	0.736	0.021	0.699	0.737	0.767	0.737	0.017	0.708	0.738	0.763
RHU(3)	6.138	0.724	0.022	0.685	0.724	0.758	0.724	0.018	0.693	0.725	0.753
RKW(2, 2)	1.416	0.766	0.018	0.735	0.767	0.794	0.766	0.014	0.741	0.767	0.788
RKW(2, 3)	1.666	0.762	0.018	0.731	0.764	0.789	0.762	0.014	0.738	0.763	0.785
RKW(2, 4)	1.747	0.761	0.018	0.730	0.762	0.788	0.761	0.014	0.737	0.762	0.784
RKW(3, 2)	1.797	0.761	0.018	0.729	0.762	0.788	0.761	0.014	0.736	0.761	0.783
RKW(3, 3)	2.285	0.755	0.018	0.721	0.756	0.783	0.755	0.015	0.729	0.756	0.779
RKW(3, 4)	2.445	0.753	0.018	0.720	0.754	0.782	0.753	0.015	0.726	0.754	0.777
RMO(2)	0.968	0.775	0.021	0.738	0.777	0.809	0.776	0.017	0.747	0.777	0.803
RMO(3)	2.377	0.755	0.021	0.718	0.756	0.787	0.755	0.017	0.724	0.757	0.782

picture is almost the same as without outliers, except that RHU(2) is slightly superior to RLS.

Turning to the results for the estimation on data simulated with additive outliers in Table 3, a large bias is seen with all algorithms. The smallest bias is obtained with RMO(2). A plausible reason for this is that it is easy for the algorithm to classify the outliers as such. RKW(2, 3) also shows a good performance evidently because it takes care of the outliers both by comparing the residuals to estimated variance and by calculating the length of the regressor, $x(t)$, measured in the Euclidean norm in the space expanded by the estimated dispersion, $\hat{A}^{-1}(t)$. RHU(2) also reduces the effect of additive outliers but not quite enough, and as expected the bias is largest for RLS. It is evident that smaller values of c and a now give the best results. Comparing the results for the different types of outliers a reasonable choice seems to be $(c, a) = (2, 3)$. Note that a large amount of $\overline{SSD}$ stems from the bias in the estimates.

Obviously, it seems to be impossible to completely remove the bias of the parameter estimates with any of these recursive methods when additive outliers are present. In the off-line setting multi-stage methods, involving outlier detection and filtering, can be applied. However, it is problematic to carry this over to a recursive estimation algorithm.

It is seen that the estimated parameter corresponds to a faster system than the simulation system. This seems reasonable, because the additive outliers occur without being affected by the dynamics of the system, and therefore the contaminated observations seems to belong to a system with a smaller time constant.

The non-parametric density estimates of the parameter estimates given in Figure 5 for the case with additive outliers clearly show the improvement obtained by using RKW and RMO.

In Figure 6 the density of the RKW(2, 3) estimates is compared to the density of its off-line counterpart. The similarity between the densities is evident indicating that the approximations in the recursive version of the estimator by

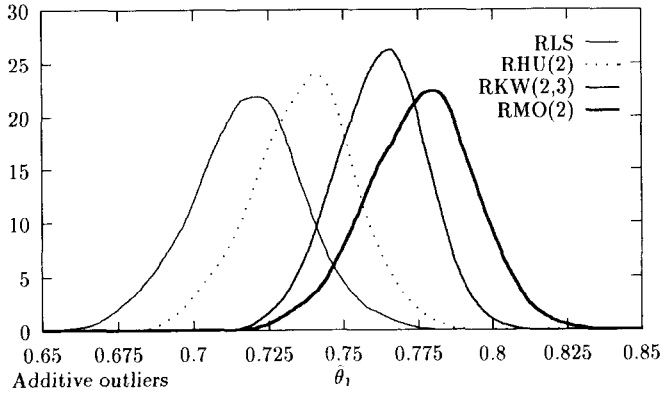

Fig. 5. Kernel estimate of the density of the parameter estimate at $t = 3000$ for data with additive outliers. The kernel is Epanechnikov with bandwidth $h = 0.005$.

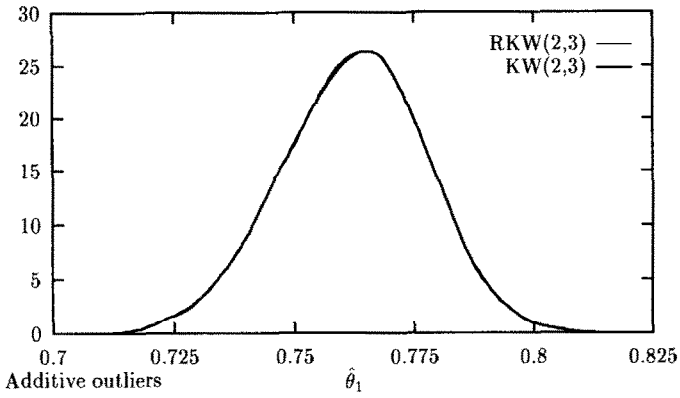

Fig. 6. Kernel estimate of the density of the recursive and off-line Krasker and Welsch parameter estimate for data with additive outliers. The kernel is Epanechnikov with bandwidth $h = 0.005$.

Krasker and Welsch is vanishing for an increasing number of observations. The same agreement is found between the densities of the estimates obtained with RKW(2, 3) and KW(2, 3), respectively, for both cases without outliers and with innovation outliers (not shown). The results in Table 4 confirm that the estimates for all three outlier settings are very much the same whether the off-line estimator proposed by Krasker and Welsch (1982) or the recursive algorithm is used.

To illustrate the performance of the recursive scale estimator proposed in Section 3.3, kernel estimates of the density of both the recursive ($t = 3000$) and the off-line scale estimates obtained with the Krasker and Welsch estimator for $(c, a) = (2, 3)$ are shown in Figures 7–9. It is seen that the recursive estimates have densities being similar to the densities of the off-line estimates. For data without outliers, the method gives a reasonable estimate of the scale. However, with innovation outliers in data the estimated dispersion is larger than for the outlier free observations, and this is even more distinct with additive outliers.

By determining σ in $E[\chi_2(e/\sigma)] = 0$ for e given by the three noise models, respectively (assuming that the model parameter is known), the expected scale parameter estimates, $\tilde{\sigma}_{NO} = 1.0$, $\tilde{\sigma}_{IO} = 1.054$ and $\tilde{\sigma}_{AO} = 1.088$, are obtained. For no outliers and for innovation outliers the center of the densities of the scale parameter estimates are close to the corresponding expected values. However, for additive outliers the center lies above the expected value. The explanation is most likely that for additive outliers the model parameter estimate is biased for which reason the prediction errors differ from the simulated noise components and therefore give rise an increased scale parameter estimate.

Since our interest is to use the standard deviation of the outlier free innovations in measuring the prediction errors in the RHU and RKW algorithms it is not desirable that the scale parameter estimate is different from this standard deviation. Neither is it desirable that the scale parameter estimate increases with increasing outlier contamination since the optimal value of c thus depends on the contamination degree. However, for little contamination the

Table 4
Results of Krasker–Welsch estimation of first order AR model

$\theta = 0.8$			Off-line					Recursive at $t = 3000$				
Outliers	c	a	$\bar{\hat{\theta}}$	$s_{\hat{\theta}}$	$\hat{\theta}_{0.05}$	$\hat{\theta}_{0.50}$	$\hat{\theta}_{0.95}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.95}$
None	2	3	0.799	0.012	0.780	0.799	0.818	0.799	0.012	0.779	0.799	0.818
None	3	3	0.799	0.011	0.780	0.800	0.817	0.799	0.011	0.780	0.800	0.817
None	3	4	0.799	0.011	0.780	0.800	0.817	0.799	0.011	0.780	0.799	0.817
Innovation	2	3	0.799	0.011	0.781	0.800	0.817	0.799	0.011	0.781	0.800	0.817
Innovation	3	3	0.799	0.011	0.780	0.800	0.816	0.799	0.011	0.781	0.799	0.816
Innovation	3	4	0.799	0.011	0.780	0.800	0.816	0.799	0.011	0.781	0.799	0.816
Additive	2	3	0.762	0.014	0.739	0.763	0.786	0.762	0.014	0.738	0.763	0.785
Additive	3	3	0.755	0.015	0.729	0.756	0.778	0.755	0.015	0.729	0.756	0.779
Additive	3	4	0.753	0.015	0.727	0.754	0.777	0.753	0.015	0.726	0.754	0.777

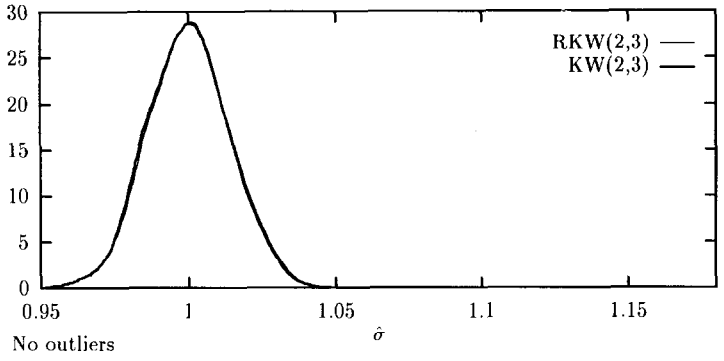

Fig. 7. Kernel estimate of the density of the recursive and off-line Krasker and Welsch scale estimate for data without outliers. The kernel is Epanechnikov with bandwidth $h = 0.004$.

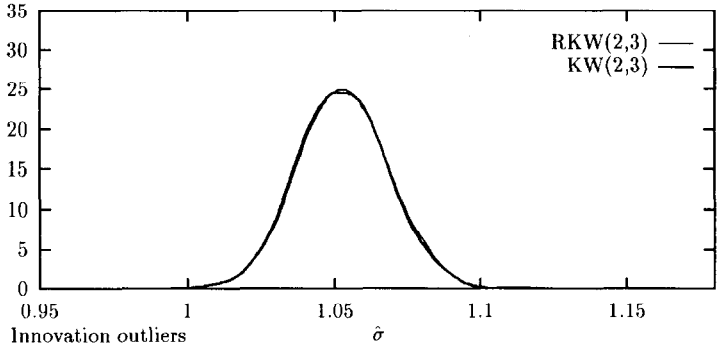

Fig. 8. Kernel estimate of the density of the recursive and off-line Krasker and Welsch scale estimate for data with innovation outliers. The kernel is Epanechnikov with bandwidth $h = 0.004$.

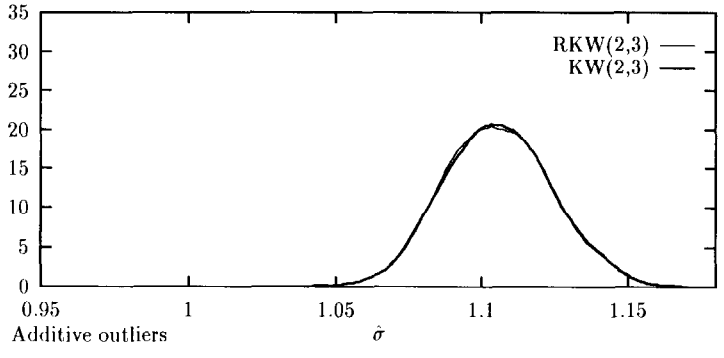

Fig. 9. Kernel estimate of the density of the recursive and off-line Krasker and Welsch scale estimate for data with additive outliers. The kernel is Epanechnikov with bandwidth $h = 0.005$.

bias will be small and the optimal c-value will not change too much. It might be appropriate to search for methods with smaller bias, for instance, a $\psi(u)$-function being zero for large values of u could be used in (6), but it is clear that when the model parameters are biased this will also be the case for the scale parameter estimate.

Table 5
Summary of the recursively obtained estimates at $t = 3000$ with second order model

Outliers/ Algorithm	$\theta = 1.20$						$\theta = -0.52$					
	$\overline{SSD}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.95}$	$\overline{SSD}$	$\overline{\hat{\theta}(t)}$	$s_{\hat{\theta}(t)}$	$\hat{\theta}(t)_{0.05}$	$\hat{\theta}(t)_{0.50}$	$\hat{\theta}(t)_{0.95}$
None												
RLS	0.292	1.200	0.016	1.173	1.200	1.224	0.293	−0.520	0.016	−0.545	−0.521	−0.493
RHU(2)	0.298	1.200	0.016	1.173	1.200	1.225	0.298	−0.520	0.016	−0.544	−0.521	−0.493
RKW(2, 5)	0.330	1.199	0.017	1.171	1.200	1.227	0.330	−0.520	0.016	−0.546	−0.521	−0.491
RMO(2)	0.452	1.199	0.019	1.167	1.199	1.232	0.468	−0.520	0.020	−0.551	−0.521	−0.489
Innovation												
RLS	0.289	1.200	0.016	1.173	1.200	1.225	0.295	−0.520	0.016	−0.545	−0.521	−0.493
RHU(2)	0.263	1.200	0.015	1.174	1.200	1.223	0.270	−0.520	0.015	−0.543	−0.521	−0.494
RKW(2, 5)	0.295	1.200	0.016	1.172	1.200	1.225	0.296	−0.520	0.016	−0.545	−0.521	−0.492
RMO(2)	0.391	1.199	0.018	1.168	1.200	1.228	0.405	−0.520	0.018	−0.550	−0.520	−0.488
Additive												
RLS	71.276	0.935	0.033	0.881	0.934	0.990	55.084	−0.287	0.029	−0.337	−0.287	−0.242
RHU(2)	43.156	0.995	0.029	0.947	0.996	1.043	36.346	−0.332	0.027	−0.377	−0.331	−0.290
RKW(2, 5)	13.554	1.086	0.022	1.049	1.086	1.121	12.145	−0.412	0.022	−0.449	−0.412	−0.376
RMO(2)	7.529	1.120	0.030	1.069	1.121	1.164	8.245	−0.436	0.031	−0.484	−0.437	−0.384

Table 5 shows the results obtained, when estimating the parameters of the second-order model. When innovation outliers are exciting the system, the RMO(2) algorithm gives a more flat distribution of the parameter estimates and a more disturbed parameter estimate trajectory than the rest of the algorithms. The differences are, however, small, and all of the algorithms are obviously working well. In the additive outlier case, however, the low value of the cut-off limit in RMO(2) forces the algorithm to expel of that large a part of the data that what remains is enough to make a good performance. The RLS estimates are brought far away from the true values of the parameters and hence, as is well known, RLS is a dangerous algorithm when applied to data with additive outliers. Irrespective of the nature of the outliers, the RKW algorithm performs satisfactorily.

5. Conclusion

The successful application of parameter estimation algorithms to real data demands that the algorithms are robust in order to reduce errors in the estimates originating from outliers of different kinds. The traditionally formulated methods for robust estimation only consider the off-line situation. However, for several purposes, e.g., adaptive forecasting and control or efficient treatment of large amounts of data, on-line estimation is desirable.

This paper describes a method for obtaining recursive robust on-line estimation algorithms. It is used to derive two recursive robust algorithms based on the minimization of robustified criteria, the Huber and the Krasker and Welsch criteria.

A simulation study is carried out to investigate the performance of the estimators, when innovation and additive outliers are considered. Among the conclusions, which can be drawn from the simulation results, is that the RLS algorithm is not suitable for parameter estimation when the outliers are additive, but when extended with a facility for treating the detected outliers as missing observations, it can handle also the additive outlier case. Furthermore, the recursive Huber method is sensitive to additive outliers, whereas the recursive Krasker and Welsch algorithm implies a considerable reduction of the bias which is due to the presence of additive outliers.

It should be noted that the RMO algorithm that cuts away parts of the data possibly can get stuck in situations where all data are considered outliers and the estimates do not change. This cannot happen for any of the other algorithms.

Acknowledgement

This work was supported by the Nordic Council of Ministers Research Program on District Heating, the Danish Energy Research Program 1989 and the

Swedish National Board for Technical Development under contracts 88-00784P, 88-02060P and 84-03554P. This research support is gratefully acknowledged.

References

Allende, H. and S. Heiler, Recursive generalized M-estimates for autoregressive moving-average models, *J. Time Ser. Anal.* , **13** (1992) 1–18.

Campbell, K., Recursive estimation of M-estimates for the parameters of a finite autoregressive process, *Ann. Statist.*, **10** (1982) 442–453.

Cipra, T., Robust exponential smoothing, *J. Forecasting*, **11** (1992) 57–69.

Denby, L. and R.D. Martin, Robust estimation of the first-order autoregressive parameter, *J. Amer. Statist. Assoc.*, **74** (1979) 140–146.

Englund, J-E., U. Holst and D. Ruppert, Recursive M-estimators for location and scale for dependent sequences, *Scand. J. Statist.*, **15** (1988) 147–159.

Englund, J-E., U. Holst and D. Ruppert, Recursive M-estimators for stationary, strong mixing processes - A representation theorem and asymptotic distributions, *Stochast. Process. Appl.*, **31** (1989) 203–222.

Englund, J-E., Recursive versions of the algorithm by Krasker and Welsch, *Sequential Anal.*, **10** (1991) 211–234.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseuw and W.A. Stahel, *Robust statistics* (Wiley, New York, 1986).

Huber, P.J., Robust estimation of a location parameter, *Annals Math. Statist.*, **35** (1964) 73–101.

Huber, P.J., Robust regression. Asymptotics, conjectures and Monte Carlo, *Anal. Statist.* **5** (1973) 799–821.

Huber, P.J., *Robust statistics* (Wiley, New York, 1981).

IMSL Math/Library, *Fortran subroutines for mathematical applications* (IMSL, Inc., Houston, 1987).

IMSL Stat/Library, *Fortran subsroutines for statistical analysis* (IMSL, Inc., Houston, 1987).

Kuh, E. and A. Samarov, Robust recursive estimation and detection of shifts in regression, *CompStat. 1986, Rome* (1986) 217–222.

Krasker, W.S. and R.E. Welsch, Efficient bounded-influence regression estimates, *J. Amer. Statist. Assoc.*, 77 (1982) 595–604.

Ljung, L., *System identification, theory for the user* (Prentice Hall, Englewood Cliffs, 1987).

Ljung, L. and T. Söderström, *Theory and practice of recursive identification* (MIT Press, Cambridge, London, 1983).

Martin, R.D. and C.J. Masreliez, Robust estimation via stochastic approximation, *IEEE Transactions on Information Theory*, **It-21** (1975) 263–271.

Martin, R.D. and V.J. Yohai, Robustness in time series and estimating ARMA models, in: Hannan, Krishnaiah, Rao (Eds), *Handbook of Statistics*, **5** (1985) 119–155.

Masrelicz, C.J., Approximate non-Gaussian filtering with state and observation relations, *IEEE Transactions on Automatic Control*, **AC-20** (1975) 107–110.

Poljak, B.T. and J.Z. Tsypkin, Robust identification, *Automatic*, **16** (1980) 53–63.

Poulson, N.K. and J. Holst, Robust self tuning controllers in nonstationary situations, *Ricerche di Automatica*, **13** (1982) 197–217.

Sejling, K., Adaptive prediction models for district heating systems, (in Danish: Adaptive prognosemodeller for fjernvarmesystemer), Master's thesis (The Institute of Mathematical Statistics and Operations Research, The Technical University of Denmark, Lyngby, Denmark, 1987).

Söderström, T. and P. Stoica, *System identification* (Prentice Hall International, New York, 1989).

West, M., Robust sequential approximate bayesian estimation, *J. Royal Statist. Soc.* B, **43** (1981) 157–166.