KYBERNETIKA — VOLUME 47 (2011), NUMBER 2, PAGES 165–178

EXPONENTIAL SMOOTHING FOR TIME SERIES WITH OUTLIERS

Tomáš Cipra and Tomáš Hanzák

Recursive time series methods are very popular due to their numerical simplicity. Their theoretical background is usually based on Kalman filtering in state space models (mostly in dynamic linear systems). However, in time series practice one must face frequently to outlying values (outliers), which require applying special methods of robust statistics. In the paper a simple robustification of Kalman filter is suggested using a simple truncation of the recursive residuals. Then this concept is applied mainly to various types of exponential smoothing (recursive estimation in Box–Jenkins models with outliers is also mentioned). The methods are demonstrated using simulated data.

Keywords: exponential smoothing, Kalman filter, outliers, robust smoothing and forecasting

Classification: 62M10, 62M20, 90A20, 60G35

1. INTRODUCTION

Kalman filter represents a theoretical framework for various recursive methods in time series, i. e. for recursive estimating, smoothing and forecasting. In particular, all types of exponential smoothing can be derived using this concept, see e. g. [1, 3, 8, 9]. If there are outliers in an analyzed time series one should respect this fact: (1) it is possible to identify and then to remove these outlying observations and treat the remaining data as a time series with missing observations, see e. g. [6] or (2) one can robustify classical statistical methods to make them insensitive (robust) against outliers (e. g. to apply medians instead of means). The latter approach is usually more simple and comfortable from the numerical point of view, and therefore various robust modifications of Kalman filter have been suggested in literature, see e. g. [5].

In this paper we try to robustify the classical Kalman filter (i. e. Kalman filter in a simple linear state space model with scalar observations under the assumption of normality) using a simple truncation of the recursive residuals (i. e. a truncation of the recursive prediction errors). The corresponding robust Kalman filter is introduced in Section 2. Several recursive scale estimators are also discussed here. Various types of robustified exponential smoothing procedures together with a robustified recursive estimation procedure in autoregressive models are presented in Section 3, see e. g. [4, 5, 7, 10]. Finally, the methods are demonstrated and compared

using simulated data in Section 4 (the implementation details and the methods used for the comparison are also presented here). Section 5 brings the summary of the paper.

2. ROBUST KALMAN FILTER

2.1. Classical Kalman filter

Let's consider a simple discrete time linear state space model of the form [1]

$$\mathbf{S}_{t+1} = \mathbf{A}\mathbf{S}_t + \mathbf{a}_{t+1}\,, \quad \mathbf{a}_t \sim iid\ N_n(0,\ \mathbf{R}_2)\,, \tag{1}$$

$$y_t = \mathbf{h}_t'\mathbf{S}_t + \varepsilon_t\,, \quad \varepsilon_t \sim iid\ N(0,\ r_1^2)\,, \tag{2}$$

where $\mathbf{S}_t$ is the n-dimensional state vector of the system with a fixed initial value $\mathbf{S}_0$, y_t is the one-dimensional observation process, the observation noise $\{\varepsilon_t\}$ and the n-dimensional innovation process $\{\mathbf{a}_t\}$ are mutually independent, $\mathbf{A}$ is a fixed $n \times n$ matrix of parameters, $\mathbf{h}_t$ is n-dimensional vector of parameters varying in time and $\mathbf{R}_2$ and $r_1^2 > 0$ describe the variance-covariance structure of $\mathbf{a}_t$ and ε_t, see e. g. [1].

The updating equations referred to as Kalman filter (or Kalman–Bucy filter) are

$$\hat{\mathbf{S}}_{t+1|t} = \mathbf{A}\hat{\mathbf{S}}_{t|t}\,, \tag{3}$$

$$\mathbf{P}_{t+1|t} = \mathbf{A}\mathbf{P}_{t|t}\mathbf{A}' + \mathbf{R}_2\,, \tag{4}$$

$$\hat{\mathbf{S}}_{t+1|t+1} = \hat{\mathbf{S}}_{t+1|t} + \mathbf{k}_{t+1}\left(y_{t+1} - \mathbf{h}_{t+1}'\hat{\mathbf{S}}_{t+1|t}\right)\,, \tag{5}$$

$$\mathbf{P}_{t+1|t+1} = \mathbf{P}_{t+1|t} - \mathbf{k}_{t+1}\mathbf{h}_{t+1}'\mathbf{P}_{t+1|t}\,, \tag{6}$$

$$\mathbf{k}_{t+1} = \frac{\mathbf{P}_{t+1|t}\mathbf{h}_{t+1}}{\mathbf{h}_{t+1}'\mathbf{P}_{t+1|t}\mathbf{h}_{t+1} + r_1^2}\,, \tag{7}$$

where $\hat{\mathbf{S}}_{r|s}$ is an estimate of $\mathbf{S}_r$ based on the observations of y up to time s, $\mathbf{P}_{r|s}$ is its estimation error covariance matrix and $\mathbf{k}_{t+1}$ is called *gain* vector.

Since the one-step-ahead prediction of y_{t+1} from time t is naturally

$$\hat{y}_{t+1|t} = \mathbf{h}_{t+1}'\hat{\mathbf{S}}_{t+1|t}\,, \tag{8}$$

see (2), one can rewrite (5) to the form

$$\hat{\mathbf{S}}_{t+1|t+1} = \hat{\mathbf{S}}_{t+1|t} + \mathbf{k}_{t+1}(y_{t+1} - \hat{y}_{t+1|t}) = \hat{\mathbf{S}}_{t+1|t} + \mathbf{k}_{t+1}e_{t+1}\,, \tag{9}$$

where

$$e_{t+1} = y_{t+1} - \hat{y}_{t+1|t} \tag{10}$$

are the corresponding prediction errors. These errors can be normalized to have unit variances:

$$\tilde{e}_{t+1} = \frac{e_{t+1}}{\sigma(e_{t+1})} = \frac{e_{t+1}}{\sqrt{\mathbf{h}_{t+1}'\mathbf{P}_{t+1|t}\mathbf{h}_{t+1} + r_1^2}}\,. \tag{11}$$

The estimates and predictions delivered by this Kalman filter are optimal in the MSE sense.

[1] Matrices and vectors are printed in bold. Vectors are always column vectors.

2.2. Robust Kalman filter

When an outlier is present in the observation y_{t+1} at time $t+1$ then the prediction error e_{t+1} on the right hand side of the recursive formula (5) or (9) estimating the state $\mathbf{S}_{t+1}$ is distorted, and one should adjust it in order to robustify the filter. The natural way how to achieve this aim is to apply error truncation of the form

$$\hat{\mathbf{S}}^{robust}_{t+1|t+1} = \hat{\mathbf{S}}_{t+1|t} + w_{t+1}(e_{t+1})\,\mathbf{k}_{t+1}\,e_{t+1}\,, \tag{12}$$

where $0 < w_{t+1}(\cdot) \leq 1$ is a robustifying weight function of the form

$$w_{t+1}(x) = \begin{cases} 1\,, & |x| \leq \dfrac{u_{t+1}}{\sqrt{\mathbf{k}'_{t+1}\mathbf{W}'_{t+1}\mathbf{W}_{t+1}\mathbf{k}_{t+1}}} \\ \dfrac{1}{|x|}\dfrac{u_{t+1}}{\sqrt{\mathbf{k}'_{t+1}\mathbf{W}'_{t+1}\mathbf{W}_{t+1}\mathbf{k}_{t+1}}}\,, & |x| > \dfrac{u_{t+1}}{\sqrt{\mathbf{k}'_{t+1}\mathbf{W}'_{t+1}\mathbf{W}_{t+1}\mathbf{k}_{t+1}}} \end{cases} \tag{13}$$

with a suitable truncation value $u_{t+1} > 0$. In the literature on robust statistics w_{t+1} is called Huber's weight function. $\mathbf{W}_{t+1}$ is a diagonal $n \times n$ weighting matrix with positive elements on its diagonal which solves the problem of uncomparable units of individual components of $\mathbf{k}_{t+1}$ and $\mathbf{S}_{t+1}$. It is obvious that then Kalman filter is less sensitive to outliers in y_t since the influence of large prediction errors in (5) or (9) is reduced, see (12) and (13).

However, it is not difficult to show rigorously that the choice of the weight function (13) leads to the optimal estimate of the state $\mathbf{S}_{t+1}$ in the $\mathbf{W}_{t+1}$-weighted MSE sense under the additional robustifying condition

$$\left\| \mathbf{W}_{t+1} \left(\hat{\mathbf{S}}^{robust}_{t+1|t+1} - \hat{\mathbf{S}}_{t+1|t} \right) \right\| \leq u_{t+1} \tag{14}$$

(we restrict the magnitude of $\hat{\mathbf{S}}$ update). Obviously $\hat{\mathbf{S}}^{robust}_{t+1|t+1}$ defined by (12) together with (13) fulfills the condition (14). Further one can decompose the minimized $\mathbf{W}_{t+1}$-weighted MSE criterion as

$$\begin{aligned} &\mathsf{E} \left\| \mathbf{W}_{t+1} \left(\hat{\mathbf{S}}^{robust}_{t+1|t+1} - \mathbf{S}_{t+1} \right) \right\|^2 \\ &\quad = \mathsf{E} \left\| \mathbf{W}_{t+1} \left(\hat{\mathbf{S}}^{robust}_{t+1|t+1} - \hat{\mathbf{S}}_{t+1|t+1} \right) \right\|^2 + \mathsf{E} \left\| \mathbf{W}_{t+1} \left(\hat{\mathbf{S}}_{t+1|t+1} - \mathbf{S}_{t+1} \right) \right\|^2 \end{aligned} \tag{15}$$

due to orthogonality

$$\begin{aligned} &\mathsf{E} \left\{ \left(\hat{\mathbf{S}}^{robust}_{t+1|t+1} - \hat{\mathbf{S}}_{t+1|t+1} \right)' \mathbf{W}'_{t+1}\mathbf{W}_{t+1} \left(\hat{\mathbf{S}}_{t+1|t+1} - \mathbf{S}_{t+1} \right) \right\} \\ &= \mathsf{E} \left\{ \mathsf{E} \left[\left(\hat{\mathbf{S}}^{robust}_{t+1|t+1} - \hat{\mathbf{S}}_{t+1|t+1} \right)' \mathbf{W}'_{t+1}\mathbf{W}_{t+1} \left(\hat{\mathbf{S}}_{t+1|t+1} - \mathbf{S}_{t+1} \right) \middle| y_1, \ldots, y_{t+1} \right] \right\} \\ &= \mathsf{E} \left\{ \left(\hat{\mathbf{S}}^{robust}_{t+1|t+1} - \hat{\mathbf{S}}_{t+1|t+1} \right)' \mathbf{W}'_{t+1}\mathbf{W}_{t+1} \mathsf{E} \left(\hat{\mathbf{S}}_{t+1|t+1} - \mathbf{S}_{t+1} \middle| y_1, \ldots, y_{t+1} \right) \right\} \\ &= \mathsf{E} \left\{ \left(\hat{\mathbf{S}}^{robust}_{t+1|t+1} - \hat{\mathbf{S}}_{t+1|t+1} \right)' \mathbf{W}'_{t+1}\mathbf{W}_{t+1} \left(\hat{\mathbf{S}}_{t+1|t+1} - \hat{\mathbf{S}}_{t+1|t+1} \right) \right\} = 0\,. \end{aligned} \tag{16}$$

Substituting (12) into (15) we get

$$\mathsf{E}\left\|\mathbf{W}_{t+1}\left(\hat{\mathbf{S}}^{robust}_{t+1|t+1}-\mathbf{S}_{t+1}\right)\right\|^2$$
$$=\mathsf{E}\left\{\left[1-w_{t+1}\left(e_{t+1}\right)\right]^2\left\|\mathbf{W}_{t+1}\mathbf{k}_{t+1}e_{t+1}\right\|^2\right\}+\mathsf{E}\left\|\mathbf{W}_{t+1}\left(\hat{\mathbf{S}}_{t+1|t+1}-\mathbf{S}_{t+1}\right)\right\|^2. \quad (17)$$

As the second summand does not depend on the choice of $w_{t+1}(\cdot)$, this really shows the optimality of the function (13): to minimize (17) one takes the value of $w_{t+1}\left(e_{t+1}\right)$ as close to 1 as possible under the restriction given by (14).

In practical situations, vector $\mathbf{k}_t$ is used constant (equal to the steady state solution of the filter) and the same may hold for the weighting matrix $\mathbf{W}_{t+1}$. So the only practically relevant problem remains how to choose the truncation value u_{t+1} in (13). The approach suggested in this paper makes use of the assumption of normality of the residuals in (1) and (2) (if outliers are ignored). In particular, it is

$$\tilde{e}_{t+1}=\frac{e_{t+1}}{\sigma(e_{t+1})}=\frac{e_{t+1}}{\sqrt{\mathbf{h}'_{t+1}\mathbf{P}_{t+1|t}\mathbf{h}_{t+1}+r_1^2}}\sim N(0,1)\,. \quad (18)$$

Using the symbol $u_{1-p/2}$ as the normal $(1-p/2)$-quantile, then the outliers should be identified for

$$\frac{|e_{t+1}|}{\sqrt{\mathbf{h}'_{t+1}\mathbf{P}_{t+1|t}\mathbf{h}_{t+1}+r_1^2}}>u_{1-p/2} \quad (19)$$

since this inequality occurs with a small probability p in the situation without outliers (e. g. $u_{0.975}\doteq 1.96$ for the probability $p=5\,\%$). Therefore the choice of the constant u_{t+1} in (13) should be such that

$$\frac{u_{t+1}}{\sqrt{\mathbf{k}'_{t+1}\mathbf{W}'_{t+1}\mathbf{W}_{t+1}\mathbf{k}_{t+1}\left(\mathbf{h}'_{t+1}\mathbf{P}_{t+1|t}\mathbf{h}_{t+1}+r_1^2\right)}}=u_{1-p/2}\,. \quad (20)$$

Finally it is convenient to rewrite (12) and (13) as

$$\hat{\mathbf{S}}^{robust}_{t+1|t+1}=\hat{\mathbf{S}}_{t+1|t}+\sigma(e_{t+1})\;\psi(\tilde{e}_{t+1})\;\mathbf{k}_{t+1}\,, \quad (21)$$

where $\sigma(e_{t+1})$ and $\tilde{e}_{t+1}$ are defined in (11) and the truncation function $\psi(\cdot)$ is defined as

$$\psi(x)=\begin{cases} x\,, & |x|\le u_{1-p/2}\\ \operatorname{sign}(x)\cdot u_{1-p/2}\,, & |x|>u_{1-p/2}\,. \end{cases} \quad (22)$$

One can recapitulate that our robustification of Kalman filter obviously consists in replacing the original recursive formula (5) or (9) by the new one (21) together with (22) introducing the prediction error truncation. The remaining formulas stay unchanged.

The approach described in this section can be generalized to robustify Kalman filter with m-dimensional vector observations $\{\mathbf{y}_t\}$, i. e. for m-dimensional time series. In such a case the analogy of the truncation function (22) can be based on the square root of the quantile $\chi^2_{1-p}(m)$ of the distribution $\chi^2(m)$.

2.3. Recursive scale estimation

In practice the normalized prediction error (11) can be estimated by various approaches that differ by their technical sophistication. All the approaches presented here rely on a recursive scale estimator $s_t \approx \sigma(y_{t+1} - \hat{y}_{t+1|t}) = \sigma(e_{t+1})$ with an initial value s_0 and a preset smoothing constant $\nu \in (0, 1)$. The normalized prediction error is then estimated as $\tilde{e}_{t+1} \approx e_{t+1}/s_t$.

The first approach is based on L_1-norm:

$$s_t = \nu \cdot 1.2533 \cdot |e_t| + (1 - \nu) s_{t-1} \,. \tag{23}$$

The factor $\sqrt{\pi/2} \doteq 1.2533$ is used to make the scale estimation unbiased for normally distributed errors.

The second approach is based directly on L_2-norm (therefore no normalizing factor is needed) and it improves the previous one since it takes into account the identified outliers:

$$s_t^2 = \nu \cdot [s_{t-1}\, \psi(\tilde{e}_t)]^2 + (1 - \nu) \cdot s_{t-1}^2 \,, \tag{24}$$

$$\tilde{e}_t = e_t / s_{t-1} \,. \tag{25}$$

Value of s_t^2 is in fact (for $t \gg 0$) an exponentially weighted average of the squared truncated prediction errors up to time t. The truncation used here is the same as in Kalman filter itself, see (21) and (22). This approach reminds the volatility modeling in a GARCH$(1, 1)$ model. [1]

The third approach makes use of the so-called τ^2-scale estimator by [12], see also [7] or [10]:

$$s_t^2 = \nu \cdot s_{t-1}^2 \cdot \rho(\tilde{e}_t) + (1 - \nu) \cdot s_{t-1}^2 \,, \tag{26}$$

where the so-called *biweight* (or *bisquare*) ρ-function is given by

$$\rho(x) = \begin{cases} c_k \left\{ 1 - \left[1 - (x/k)^2 \right]^3 \right\} \,, & |x| \le k \\ c_k \,, & |x| > k \end{cases} \tag{27}$$

(the common values of the constants are $k = 2$ and $c_k = 2.52$).

3. SOME SPECIAL CASES

If one wants to apply the procedure from Section 2 numerically, a lot of technical problems and details must be solved and possible practical improvements must be taken into account. In any case, the resulting algorithm should remain recursive and should be as simple as possible from the numerical point of view, since just the simplicity supports applicability of methods of this type (it holds e. g. for the exponential smoothing).

[1] It would be possible to alternatively use the lagged scale estimate s_{t-1} to truncate the forecasting error, update the scale estimate to s_t based on this truncated error and then use this updated scale estimate to produce the final truncated error. But this would be equivalent just to replacing u in (22) by $u \left[\nu(u^2 - 1) + 1 \right]^{-1/2}$.

Fortunately, there are special cases of the general procedure from Section 2, where such aims may be achieved using acceptable approximations. Examples, which include various types of exponential smoothing (see 3.1–3.3) and recursive estimation of simple Box–Jenkins models (see 3.4), are given in this section.

3.1. Simple exponential smoothing

Various types of exponential smoothing are special cases of Kalman filter, see e. g. [1, 8]. For instance, using the state space model of the form

$$y_t = L_t + \varepsilon_t\,, \quad \varepsilon_t \sim \ iid\ N(0,\sigma^2)\,, \tag{28}$$

$$L_t = L_{t-1} + \eta_t\,, \quad \eta_t \sim iid\ N(0,\sigma^2\omega)\,, \tag{29}$$

one obtains (after stabilization $\mathbf{P}_{t|t} \to \mathbf{P}$) the following robust version of simple exponential smoothing (with the smoothing constant $\alpha \in (0,1)$ driven by ω):

$$\hat{y}_{t+1} = \hat{y}_t + \alpha\, s_t\, \psi\left(\tilde{e}_{t+1}\right)\,, \tag{30}$$

$$\hat{y}_{t+k|t} = \hat{y}_t\,, \quad k \geq 0\,, \tag{31}$$

where the prediction error is

$$e_{t+1} = y_{t+1} - \hat{y}_t\,, \tag{32}$$

the robustifying function $\psi(\cdot)$ is as in (22) and the normalized prediction error $\tilde{e}_{t+1}$ can be estimated by means of formulas (23) or (24) or (26) for a suitable smoothing constant $\nu \in (0,1)$.

The principle of the robust simple exponential smoothing is natural: an automatic reduction of the smoothing constant occurs when an outlier is identified. It resembles to various adaptive approaches to the exponential smoothing, e. g. [11] suggests

$$\hat{y}_{t+1} = \hat{y}_t + \alpha_{t+1} \cdot e_{t+1} \tag{33}$$

with

$$\alpha_{t+1} = \frac{1}{1 + \exp\left[b + c(y_{t+1} - \hat{y}_t)^2\right]} \tag{34}$$

for suitable parameters b and c. In our case, the adaptive adjustment reacts only to outliers.

3.2. Double exponential smoothing

The robust version of double exponential smoothing with a smoothing constant $\alpha \in (0,1)$ is

$$\hat{S}_{t+1} = \hat{S}_t + \ \hat{T}_t + \alpha \cdot s_t\, \psi\left(\tilde{e}_{t+1}\right)\,, \tag{35}$$

$$\hat{T}_{t+1} = \hat{T}_t + \ \alpha^2 \cdot s_t\, \psi\left(\tilde{e}_{t+1}\right)\,, \tag{36}$$

$$\hat{y}_{t+k|t} = \hat{S}_t + \frac{1-\alpha}{\alpha}\hat{T}_t + k \cdot \hat{T}_t\,, \quad k \geq 0\,, \tag{37}$$

where the prediction error is

$$e_{t+1} = y_{t+1} - \hat{S}_t - \frac{1}{\alpha}\,\hat{T}_t\,, \tag{38}$$

the robustifying function $\psi(\cdot)$ is as in (22) and the normalized prediction error $\tilde{e}_{t+1}$ can be estimated by means of (23) or (24) or (26) for a suitable smoothing constant $\nu \in (0,1)$.

3.3. Holt method

The robust version of Holt method with smoothing constants $\alpha, \gamma \in (0,1)$ is

$$\begin{aligned} \hat{S}_{t+1} &= \hat{S}_t + \hat{T}_t + \alpha \cdot s_t\, \psi\left(\tilde{e}_{t+1}\right)\,, & (39)\\ \hat{T}_{t+1} &= \hat{T}_t + \alpha \cdot \gamma \cdot s_t\, \psi\left(\tilde{e}_{t+1}\right)\,, & (40)\\ \hat{y}_{t+k|t} &= \hat{S}_t + k \cdot \hat{T}_t\,, \quad k \geq 0\,, & (41) \end{aligned}$$

where the prediction error is

$$e_{t+1} = y_{t+1} - \hat{S}_t - \hat{T}_t\,, \tag{42}$$

the robustifying function $\psi(\cdot)$ is as in (22) and the normalized prediction error $\tilde{e}_{t+1}$ can be estimated by means of (23) or (24) or (26) for a suitable smoothing constant $\nu \in (0,1)$.

Any other variant of Holt method can be robustified exactly in the same way. This relates to additive or multiplicative Holt–Winters method, Holt method with exponential, damped linear or damped exponential trend and all the combinations of trend and seasonality types, see e. g. [8] or [9].

3.4. Robust recursive estimation of AR(1) model

Our goal is to estimate the parameter φ in AR(1) model for the series y in a robust and recursive way. Using the state space model of the form, see e. g. [1],

$$\begin{aligned} y_t &= \varphi_t\, y_{t-1} + \varepsilon_t\,, \quad \varepsilon_t \sim iid\, N(0,\sigma^2)\,, & (43)\\ \varphi_{t+1} &= \varphi_t & (44) \end{aligned}$$

one obtains the robust Kalman filter in the form

$$\begin{aligned} P_{t+1} &= \frac{P_t}{1 + P_t\, y_t^2}\,, & (45)\\ \hat{\varphi}_{t+1} &= \hat{\varphi}_t + s_t \cdot \psi\left(\tilde{e}_{t+1}\right) \cdot P_{t+1} \cdot y_t\,, & (46) \end{aligned}$$

where we put $\hat{\varphi}_t = \hat{\varphi}_{t,t} = \hat{\varphi}_{t+1,t}$ and $P_t = P_{t,t}/\sigma^2 = P_{t+1,t}/\sigma^2$. The prediction error is

$$e_{t+1} = y_{t+1} - \hat{\varphi}_t \cdot y_t \tag{47}$$

and the robustifying function $\psi(\cdot)$ is as in (22) and the normalized prediction error $\tilde{e}_{t+1}$ can be estimated by means of (23) or (24) or (26) for a suitable smoothing constant $\nu \in (0,1)$. Other types of estimators could be also considered, see e. g. [2].

4. OTHER METHODS, IMPLEMENTATION AND SIMULATION STUDY

In this section we compare the suggested methods with other ones already published. We focus on a simple exponential smoothing for non-seasonal time series with locally constant trend and on a double exponential smoothing (or alternatively Holt method) for non-seasonal time series with locally linear trend. We believe that such a comparison is sufficient to evaluate the performance of different robustification approaches.

4.1. Methods for comparison

The methods for the comparison came from [4] and [7]. Simple and double exponential smoothing based on approximate discounted M-estimation of constant and linear trend respectively was suggested in [4]. However, [7] showed the numerical instability of double exponential smoothing formulas and proposed a different computational scheme for a theoretically equivalent method.

Let us briefly present these two methods. Both the simple and double exponential smoothing follow the same idea of discounted M-estimation provided by Iteratively Reweighted Least Squares (IRLS) algorithm in [4]. This is a favorite estimation technique transferring a general minimization problem (e. g. M-estimation) into the Weighted Least Squares (WLS) problem using the weights depending on the parameter's estimates from the previous iteration. The weights adjust the Least Squares (LS) criterion for the actual loss which is not of the LS type. The solution of the original problem is obtained after the convergence of the algorithm.

To keep the exponential smoothing methods recursive and computationally simple (with no need for multiple iterations), the IRLS algorithm is followed only approximately. In each iteration, instead of recalculation of all the weights, a new observation is included and its weight is assigned based on the trend fitted in the previous time step. The remaining weights are not recalculated, just discounted in time.

The double exponential smoothing fitting a linear trend $y_i \sim A+F\cdot i$ through time series $\{y_i\}$, $i = 1, 2, \ldots, t, t+1, \ldots$ by discounted M-estimation with a loss function ρ (with $\psi = \rho'$) and discount factor $\beta \in (0, 1)$ can be summarized as follows (we use the notation largely consistent with [7]):

$$\hat{y}_{t+k|t} = \hat{a}_t + \hat{F}_t \cdot (t+k)\,, \quad k = 0, 1, 2, \ldots\,, \tag{48}$$

$$\hat{a}_t = \frac{N_t^y - \hat{F}_t N_t^x}{N_t^c}\,, \quad \hat{F}_t = \frac{N_t^c N_t^{xy} - N_t^x N_t^y}{N_t^c N_t^{xx} - (N_t^x)^2}\,, \tag{49}$$

where the N-statistics are updated recursively as

$$N_{t+1}^c = \beta\, N_t^c + w_t\,, \tag{50}$$

$$N_{t+1}^y = \beta\, N_t^y + w_t\, y_t\,, \tag{51}$$

$$N_{t+1}^x = \beta\, N_t^x + w_t\, t\,, \tag{52}$$

$$N_{t+1}^{xx} = \beta\, N_t^{xx} + w_t\, t^2\,, \tag{53}$$

$$N_{t+1}^{xy} = \beta\, N_t^{xy} + w_t\, t\, y_t\,. \tag{54}$$

Here w_t is the weight assigned to observation y_t (classical non-robust method is obtained by taking $w_t \equiv 1$). It is

$$w_t = \frac{s_{t-1} \cdot \psi\left(\frac{y_t - \hat{y}_{t|t-1}}{s_{t-1}}\right)}{y_t - \hat{y}_{t|t-1}}\,, \tag{55}$$

where s_{t-1} is a scale estimate for the one-step-ahead forecasting error $e_t = y_t - \hat{y}_{t|t-1}$ (see 2.3 for particular scale estimators). For $e_t = 0$ we put $w_t = 1$ by definition.

Analogously the simple exponential smoothing is given by formulas

$$\hat{y}_{t+k|t} = \hat{a}_t\,, \quad k = 0, 1, 2, \ldots\,, \tag{56}$$

$$\hat{a}_t = N_t^y / N_t^c\,, \tag{57}$$

$$N_{t+1}^c = \beta\, N_t^c + w_t\,, \tag{58}$$

$$N_{t+1}^y = \beta\, N_t^y + w_t\, y_t \tag{59}$$

and the weight w_t again according to (55).

Both [4] and [7] use Huber ψ-function defined already in (22) as the truncation function to be used in our robust Kalman filter.

Holt method with two independent smoothing constants can't be derived as a solution to a certain discounted linear trend fitting. In [10] an approach to robust exponential smoothing is suggested (independently of our approach) which is in fact equivalent to that suggested here. The difference is that [10] interpret it as replacing outlying observations while we speak on truncating the forecasting errors.

4.2. Implementation details

To run all the methods suggested and presented in this paper one needs to specify starting values for the trend components (level and possibly slope) and a scale estimate s_0. For double exponential smoothing and Holt method this can be done by fitting a linear regression $y_t \sim \hat{a}_0 + \hat{F}_0 \cdot i$ through the initial m observations of the series (with let's say $m = 10$). Value of s_0 can then be calculated as a scale measure of the residuals of this fit.

Since we suppose that the series $\{y_i\}$ can contain outliers, it is desired that also these starting values are designed to be robust. We adopt the particular starting values from [7], based on *repeated median* regression and Median Absolute Deviation (MAD) scale estimation:

$$\hat{F}_0 = \operatorname*{med}_{i=1,\ldots,m} \left(\operatorname*{med}_{\substack{j=1,\ldots,m \\ j \neq i}} \frac{y_i - y_j}{i - j} \right)\,, \tag{60}$$

$$\hat{a}_0 = \operatorname*{med}_{i=1,\ldots,m} \left(y_i - \hat{F}_0 i \right)\,, \tag{61}$$

$$s_0 = 1.4826 \cdot \operatorname*{med}_{i=1,\ldots,m} \left| y_i - \hat{a}_0 - \hat{F}_0 i \right|\,, \tag{62}$$

where $\left[\Phi^{-1}(0.5)\right]^{-1} \doteq 1.4826$ is a normalizing factor to make the scale estimator unbiased for normally distributed residuals. For the simple exponential smoothing the starting values are obtained in a similar way (the repeated median becomes a simple median).

Of course one must finally choose the values of parameters α, γ (in the case of Holt method), p, ν and m to apply the methods. The meaning and importance of smoothing constant(s) α (and γ) is the same as in classical non-robust variants. Numerical searching technique can be employed to find their optimal choice.

The choice of the parameter p reflects the nature of outliers in the analyzed series and our preferences about the robustness of the method on one hand and efficiency on the other hand. It can be viewed as a frequency of false outlier detection when applied to normally distributed data, see (19). The value of $p = 5\,\%$ is a reasonable default or routine choice. Higher values p can be compensated by higher values of α and γ.

The parameter ν should be chosen depending on how quickly (if at all) the scale of forecasting errors changes. The starting period length m has lower impact on the results, especially for longer time series.

4.3. Simulation study

In the following simulation study, we compare these methods: simple exponential smoothing (30)–(32) with the similar method from [4], see (56)–(59), and Holt method (39)–(42) with double exponential smoothing from [7], see (48)–(54). In the simulation study these methods will be referred to as "Error truncation" and "M-estimation", respectively.

Always the "GARCH" scale estimator (24) and "biweight" scale estimator (26) are used. In addition to these two robust variants, always also the classical non-robust version of the method is tested (this can be achieved by applying $p \gg 0$). So for each of the two trend types we have six particular methods tested.

The simulation study itself is purposely designed in the same way as in [7] and [10]. In addition to their simulation, we test also the simple exponential smoothing methods. For this purpose we use the random walk plus noise model

$$y_t = L_t + \varepsilon_t\,, \tag{63}$$

$$L_t = L_{t-1} + \eta_t\,, \quad \eta_t \sim iid\ N(0, 0.1^2) \tag{64}$$

with ε_t and η_t mutually independent. For the locally linear trend methods we use the model

$$y_t = L_t + \varepsilon_t\,, \tag{65}$$

$$L_t = L_{t-1} + T_t + \eta_t\,, \quad \eta_t \sim iid\ N(0, 0.1^2)\,, \tag{66}$$

$$T_t = T_{t-1} + \theta_t\,, \quad \theta_t \sim iid\ N(0, 0.1^2)\,. \tag{67}$$

Innovation terms η_t and θ_t are mutually independent and also independent of the noise term ε_t. We initialize the models by $L_0 = T_0 = 0$ without loss of generality.

As in [7] and [10], we consider four different scenarios for the noise component ε_t which are described in Table 4.3. Non-contamination CD setting is used to be able

to evaluate the impact of outliers on both the non-robust and robust methods. In SO, the $N(0,1)$ observation error (or noise) ε_t is multiplied by 20 with probability of 5 %. In AO, the $N(0,1)$ error is shifted upward by 20 with probability of 5 %. FT setting uses a fat tailed Student t-distribution with 3 degrees of freedom for the observation errors (its variance is 3 and kurtosis is infinite).

Table 1. Contamination schemes for the noise component ε_t.

Scheme		Description
CD	Clean Data	$\varepsilon_t \sim iid\ N(0,1)$
SO	Symmetric Outliers	$\varepsilon_t \sim iid\ 0.95\,N(0,1) + 0.05\,N(0,20^2)$
AO	Asymmetric Outliers	$\varepsilon_t \sim iid\ 0.95\,N(0,1) + 0.05\,N(20,1)$
FT	Fat Tailed Errors	$\varepsilon_t \sim iid\ t_3$

Combined with the two types of trend (locally constant and locally linear), we have totally 8 different generating schemes. We simulate $N = 100\ 000$ time series from each of them. The level L_t in (64) and (66) is always common to all the four contamination schemes so that we reduce the unnecessary random impact on our results. Moreover, all the compared methods are applied to the same N time series.

The time series length is always 101; we construct the forecast for time 101 at time 100 and compare it with the actual value. We suppress the contamination in SO and AO at time 101. Mean Square Forecasting Error (MSFE) is evaluated for each method:

$$\mathrm{MSFE} = \frac{1}{N}\sum_{n=1}^{N} r_n^2\,, \tag{68}$$

where r_n is the forecasting error occurred in the nth time series.

As the parameters of the methods are concerned, we always use $p = 5\,\%$, $\nu = 0.1$ and $m = 10$. The smoothing constant(s) are used fixed for each trend type. It is $\alpha = 0.095$ for the simple exponential smoothing and $\alpha = 0.25$ for the double exponential smoothing. These values are optimal for the non-robust methods when applied to the time series generated by (63)–(67) together with CD scheme for ε_t. For Holt method, we use the combination $\alpha = 0.25\cdot(2 - 0.25) = 0.4375$ and $\gamma = 0.25/(2 - 0.25) \doteq 0.1429$ which makes the classical non-robust Holt method equivalent to the classical non-robust double exponential smoothing with $\alpha = 0.25$. In such a way one eliminates the advantage of Holt method consisting in its two independent smoothing constants. The optimal combination would be $\alpha = 0.37$ and $\gamma = 0.21$ which would lead to slightly better results for Holt method than with the previously stated combination used.

The resulted MSFE values for all 48 combinations of trend type, contamination scheme, method and scale estimation are reported in Tables 4.3 and 4.3. The non-robust method is always the best one (with lowest MSFE) for CD scheme since it does not loss efficiency by unnecessary robustness when no outliers are present. But the loss of efficiency occurred here by all the robust methods is negligible (for locally constant trend there is even no measurable difference in MSFE, see Table 4.3).

Table 2. MSFE values for locally constant trend.

Contamination	Method	Non-robust	GARCH	Biweight
CD	M-estimation	1.097	1.097	1.097
CD	Error truncation	1.097	1.098	1.097
SO	M-estimation	2.100	1.127	1.127
SO	Error truncation	2.100	1.125	1.126
AO	M-estimation	3.044	1.148	1.150
AO	Error truncation	3.044	1.145	1.146
FT	M-estimation	3.065	3.005	3.006
FT	Error truncation	3.065	3.004	3.004

Table 3. MSFE values for locally linear trend.

Contamination	Method	Non-robust	GARCH	Biweight
CD	M-estimation	1.604	1.611	1.609
CD	Error truncation	1.604	1.621	1.617
SO	M-estimation	9.646	1.964	1.977
SO	Error truncation	9.646	1.799	1.808
AO	M-estimation	10.310	2.241	2.248
AO	Error truncation	10.310	1.872	1.883
FT	M-estimation	4.325	3.820	3.829
FT	Error truncation	4.325	3.776	3.786

As expected, the non-robust methods applied to contaminated time series (especially in SO and AO case) give poor results. Most of this impact of outliers on prediction accuracy can be eliminated by using robust variants of the methods. The approach with error truncation gives generally better results than the M-estimation solved by approximate IRLS algorithm. This difference is only slight for the locally constant trend (see Table 4.3) and becomes significant for the locally linear trend (see Table 4.3).

For FT contamination scheme the difference between non-robust and robust methods is smaller than for SO or AO since the noise component of observation at time 101 is not "cleaned" (as it is done in SO and AO case).

The scale estimation choice seams to have very little impact on the results, i. e. the both variants give very similar results for both methods and all contamination schemes. In our opinion, GARCH-like scale estimator (24) should be preferred in practice due to its simplicity and consistency with the error-truncation philosophy of the method, compare (21) and (24). Moreover, the truncation function ψ used in (24) is parameterized by the same p as the truncation function ψ in the method itself.

5. CONCLUSIONS

The robustification approach consisting in prediction error truncation is easy to implement as a modification or extension of the classical non-robust methods. The idea of error truncation is intuitive and can be visualized transparently in the time series plot produced by the software. Theoretical justification based on the robust Kalman filter formulation is provided.

The robustness-efficiency trade off of the method can be easily balanced by tuning the value of parameter p. The necessary scale estimation can be based on GARCH-like tracking of the truncated errors variance, see (24). If outliers can occur even in the starting period of the series, we should not forget to set up the starting values of the recursive procedure also in a robust way, see (60)–(62).

The robust methods proved their satisfactory forecasting accuracy in the simulation study performed. They managed to overcome the presence of outliers while having negligible loss of efficiency when applied to "clean data". So it is worth to think about using the robust methods as the routine choice even if no prior evidence of outliers is available.

ACKNOWLEDGEMENT

The work is a part of the research project MSM0021620839 (Czech Republic).

(Received September 9, 2010)

REFERENCES

[1] B. Abraham and J. Ledolter: Statistical Methods for Forecasting. Wiley, New York 1983.

[2] J. Anděl and J. Zichová: A method for estimating parameter in nonnegative MA(1) models. Comm. Statist. Theory Methods *31* (2002), 2101–2111.

[3] T. Cipra: Some problems of exponential smoothing. Appl. Math. *34* (1989), 161–169.

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

[5] T. Cipra, and R. Romera: Robust Kalman filter and its applications in time series analysis. Kybernetika *27* (1991), 481–494.

[6] T. Cipra, J. Trujillo, and A. Rubio: Holt-Winters method with missing observations. Management Sci. *41* (1995), 174–178.

[7] C. Croux, S. Gelper, and R. Fried: Computational aspects of robust Holt-Winters smoothing based on M-estimation. Appl. Math. *53* (2008), 163–176.

[8] E. S. Gardner: Exponential smoothing: The state of the art. J. Forecasting *4* (1985), 1–28.

[9] E. S. Gardner: Exponential smoothing: The state of the art - Part II. Internat. J. Forecasting *22* (2006), 637–666.

[10] S. Gelper, R. Fried, and C. Croux: Robust forecasting with exponential and Holt-Winters smoothing. J. Forecasting *29* (2010), 285–300.

[11] J. W. Taylor: Smooth transition exponential smoothing. J. Forecasting *23* (2004), 385–404.

[12] V. Yohai and R. Zamar: High break down point estimates of regression by means of the minimization of an efficient scale. J. Amer. Statist. Assoc. *83* (1988), 406–413.

Tomáš Hanzák, Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics – Charles University, Sokolovská 83, 186 75 Praha 8. Czech Republic.
e-mail: hanzak@karlin.mff.cuni.cz

Tomáš Cipra, Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics – Charles University, Sokolovská 83, 186 75 Praha 8. Czech Republic.
e-mail: cipra@karlin.mff.cuni.cz