Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Robust Kalman filter and smoother for errors-in-variables model with observation outliers ⋆

Jaafar ALMutawa *

* *Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals, PO Box 158, Dhahran 31261, Saudi Arabia(e-mail:jaafarm@kfupm.edu.sa).*

Abstract: In this paper, we propose a robust Kalman filter and smoother for the errors-in-variables (EIV) state space model subject to observation noise with outliers. We introduce the EIV problem with outliers and then we present the minimum covariance determinant (MCD) estimator which is highly robust estimator to detect outliers. As a result, a new statistical test to check the existence of outliers which is based on the Kalman filter and smoother has been formulated. Since the MCD is a combinatorial optimization problem the randomized algorithm has been proposed in order to achieve the optimal estimate. However, the uniform sampling method has a high computational cost and may lead to biased estimate, therefore we apply the sub-sampling method. A Monte Carlo simulation result shows the efficiency of the proposed algorithm.

Keywords: Errors-in-variables model, minimum covariance determinant, Kalman filter and smoother, outliers, random search algorithm, sub-sampling method.

1. INTRODUCTION

A basic numerical routine for the classical EIV Kalman filter Diversi et al. [2005], Markovsky et al. [2005] and smoother computes the conditional expectation which is a least squares (LS) estimate. Since the LS method is rather sensitive to outliers (non Gaussian disturbances), so is the Kalman filter and smoother. Moreover, it is well known in real applications that most practical data contain outliers with a low probability, so that a standard Gaussian assumption for observation noises might fail. Following Rousseew Rousseeuw [1984], we define the outliers to be the observations which deviate from the pattern set of the majority of the data. There are many reasons for the occurrence of outliers, e.g. misplaces decimal points, recording or transmission errors, expectational phenomena such as earthquakes or strikes, or members of different population slipping in the sample etc.

Several algorithms have been proposed to deal with outliers in the output data Bai [2003], Proietti [2003], Masereliez et al. [1977], Meinhold et al. [1989], Fruhwirth [1997], however, there are some cases where the input data are observed quantities subject to random variability. Thus, there is no reason why gross errors would only occur in the response data. In a certain sense it is more likely to have outliers in the observed input data. As a technique for coping with this problem, Rousseeuw Rousseeuw [1984] suggested the MCD estimator and Rousseeuw et al. [2004,?] presented the fast MCD algorithm to compute the multivariate linear regression model. Another approach for the MCD estimator that is based on the covariance matrix of the residuals instead of the multivariate location and scatter has been proposed by Agullo et al. [2007]. Furthermore, the influence function and the efficiency of the MCD scatter estimator has been studied in Croux et al. [1999]. The MCD problem for the time series models, e.g. AR and ARMA models has discussed in Maronna et al. [2006]. However, for the EIV state space model where the outliers acts in the observed input data to the best of our knowledge, there is no paper that has been published in this area.

In this paper, we consider a filtering and smoothing problem in the presence of observation outliers with the aid of the MCD procedure. It is well known that the MCD is a highly robust estimator and its objective is to find a subset from the observation data with cardinality greater than half of the observed data and whose covariance matrix has minimum determinant. The random search algorithm Bai [2003] has been proposed to solve the MCD problem. However, the high computational complexity makes the MCD estimator impractical and may lead to bias estimate for the EIV state space model. Hence, we propose the sub-sampling method Heagerty [2000] which keeps the structure of the original data, decrease the computation time and is less sensitive to outliers. Another feature of the proposed algorithm is that the algorithm can be applied even if there is no outlier in the observed data. A minor contribution of the paper is that we derive the Kalman smoother for the EIV state space model which is required for the new statistics.

⋆ This work was supported by King Fahd University of Petroleum and Minerals.

This note is organized as follows. Section 2, gives the errors-in-variables problem in the presence of outliers, and introduces the MCD estimator for the EIV state space model. In section 3, we proposed the randomized algorithm as a method to solve the MCD problem and discuss the disadvantages of the algorithm. Section 4, is dedicated to the Kalman filter and smoother with outliers and propose the sub-sampling method. The Monte Carlo simulation is reported in section 5 and Appendix A is devoted to Kalman filter and smoother without outliers and proof of the proposition.

2. ERRORS-IN-VARIABLES MODEL

As depicted in Fig. 1, consider the errors-in-variables state space model described by

$$\begin{bmatrix} x(t+1) \\ \hat{y}(t) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x(t) \\ \hat{u}(t) \end{bmatrix} + \begin{bmatrix} w(t) \\ 0 \end{bmatrix}, \qquad (1)$$

where $x(t) \in \mathbb{R}^n$, $\hat{u}(t) \in \mathbb{R}^m$ and $\hat{y}(t) \in \mathbb{R}^p$ are unknown state, true input and output vectors respectively. Furthermore, $w(t)$ is the white Gaussian noise acting on the state whose mean is zero and has a covariance Σ_w. It should be noted that the output noise has been excluded here for the seek of simplicity, however it can be added and our technique can be easily generalized. The measured

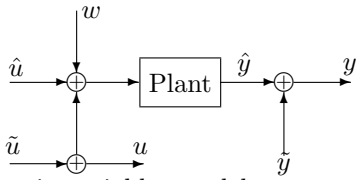

Fig. 1. Errors-in-variables model

input-output signals $u(t)$ and $y(t)$ are modelled as

$$u(t) = \hat{u}(t) + \tilde{u}(t), \qquad (2)$$
$$y(t) = \hat{y}(t) + \tilde{y}(t), \qquad (3)$$

where $\tilde{u}(t) \in \mathbb{R}^n$ and $\tilde{y}(t) \in \mathbb{R}^p$ are non-Gaussian white noises with zero mean and finite positive definite covariance matrices $\Sigma_{\tilde{u}}$ and $\Sigma_{\tilde{y}}$, respectively;

$$\mathbb{E}\left\{ \begin{bmatrix} \tilde{u}(t) \\ \tilde{y}(t) \end{bmatrix} \begin{bmatrix} \tilde{u}^{\mathrm{T}}(i) & \tilde{y}^{\mathrm{T}}(i) \end{bmatrix} \right\} = \begin{bmatrix} \Sigma_{\tilde{u}} & \Sigma_{\widetilde{uy}} \\ \Sigma_{\widetilde{uy}}^{\mathrm{T}} & \Sigma_{\tilde{y}} \end{bmatrix} \delta(t,i), \qquad (4)$$

where $\delta(t,i)$ denotes the Kronecker delta function. We will assume in the sequel, that $\tilde{u}(t)$ and $\tilde{y}(t)$ are uncorrelated with $w(t)$. Furthermore, the input and output noises $\tilde{u}(t)$ and $\tilde{y}(t)$ contain outliers with a low probability, therefore we write

$$\tilde{u}(t) = (I_m - \phi(t))\tilde{u}^n(t) + \phi(t)\tilde{u}^o(t),$$
$$\tilde{y}(t) = (I_p - \gamma(t))\tilde{y}^n(t) + \gamma(t)\tilde{y}^o(t),$$

where I_s is the $s \times s$ identity matrix for $s = m$ or $s = p$, $\psi(t) = \mathrm{diag}\{\psi_{t,i}\} = \mathrm{diag}\{\psi_{t,1}, \cdots, \psi_{t,s}\}$ and $\psi_{t,i} = 0$ or $\psi_{t,i} = 1$ for all i and where $\psi(t) = \gamma(t)$ or $\psi(t) = \phi(t)$. Moreover, $\mathrm{Prob}\{\psi_{t,i} = 1\}$ is small, i.e. the minority of the observed data are outliers. The noises $\{\tilde{u}^n(t), \tilde{u}^o(t), \tilde{y}^n(t), \tilde{y}^o(t)\}$ are Gaussian white noises with

$$\tilde{u}^n(t) \in \mathrm{N}(0, \Sigma_{\tilde{u}}^n), \qquad \tilde{u}^o(t) \in \mathrm{N}(0, \Sigma_{\tilde{u}}^o), \qquad (5)$$
$$\tilde{y}^n(t) \in \mathrm{N}(0, \Sigma_{\tilde{u}}^n), \qquad \tilde{y}^o(t) \in \mathrm{N}(0, \Sigma_{\tilde{y}}^o), \qquad (6)$$

where $\{\Sigma_{\tilde{u}}^n, \Sigma_{\tilde{u}}^o, \Sigma_{\tilde{y}}^n, \Sigma_{\tilde{u}}^o\}$ are positive definite covariance matrices. Furthermore, $\Sigma_{\tilde{u}}^o(i,i)$ and $\Sigma_{\tilde{y}}^o(i,i)$ are much larger than $\Sigma_{\tilde{u}}^n(i,i)$ and $\Sigma_{\tilde{y}}^n(i,i)$ respectively. Then, the problem of interest is to find a robust Kalman filter and smoother estimate $\hat{u}^*(t), \hat{y}^*(t)$ and $\hat{x}(t)$ for the input-output data $\hat{u}(t), \hat{y}(t)$ and the state vector $x(t)$ given that the observed input-output data are contaminated with outliers. The fact that we account for the possibility that the input signal is not exactly known and it may contain outliers, makes the problem difficult, and is often referred to as an outlier-errors-in-variables (OEIV) problem Maronna et al. [2006].

2.1 Minimum covariance determinant for the EIV models

The MCD technique has been introduced by Rousseeuw [1984] to detect the outliers for the high dimensional data set. In order to define the MCD for EIV state space model, consider a data set $\Omega(N) = \left\{ \omega(i) = \begin{bmatrix} u(i) \\ y(i) \end{bmatrix} : i = 1, \cdots, N \right\}$, and let $\mathcal{S} = \{S \subseteq \{1, \cdots, N\} : \#S = M\}$ [1] be the collection of all subsets with cardinality M from the set $\{1, \cdots, N\}$, where $[N/2] \leq M \leq N$ [2]. If the variable M equals to N, then we do not have any outlier. Moreover, the smallest possible value for M is $\frac{N}{2}$, because if more than half of the data were outliers, it would be unclear which data were from the main distribution and which were outliers. For any $S \in \mathcal{S}$, let $\Omega(S) = \left\{ \omega(i) = \begin{bmatrix} u(i) \\ y(i) \end{bmatrix} : i \in S \right\}$, and define the covariance as $\mathrm{cov}(S) = \frac{1}{M} \sum_{i \in S} (\omega(i) - T_S)(\omega(i) - T_S)^{\mathrm{T}}$ where $T_S = \frac{1}{M} \sum_{i \in S} \begin{bmatrix} \bar{u}(i) \\ \bar{y}(i) \end{bmatrix}$ and where $\bar{u}(i)$ and $\bar{y}(i)$ are the estimates based on the observations in $\Omega(S)$ to be obtained in section 4. The MCD estimator consist of two steps; the first step is to

$$J(S) = \text{Minimize } \det(\mathrm{cov}(S)), \qquad (7)$$

i.e. the MCD searches for a subset $S \in \mathcal{S}$ of size M whose covariance matrix has the smallest determinant. It is clear that the variables in the objective function (7) are the subset S and the estimates $\bar{y}(i)$ and $\bar{u}(i)$. The second step is to detect outliers by using the squared Mahalanobis distance $d(i)^2 = (\omega(i) - T_S)^{\mathrm{T}} \mathrm{cov}(S)^{-1} (\omega(i) - T_S)$, where T_S and $\mathrm{cov}(S)$, are computed by using the observed data in $\Omega(S)$ only. Furthermore consider the null hypothesis

$$H_0(t) : \omega(t) \text{ is not an outlier,}$$

against the alternative hypothesis

$$H_1(t) : \omega(t) \text{ is an outlier.}$$

Since $\mathrm{cov}(S)^{-1/2}(\omega(i) - T_S)$ has a standard Gaussian distribution function, therefore the squared Mahalanobis distance $d(t)^2$ has χ^2 distribution and a decision rule $\delta(t)$ can be found as

$$\delta(t) := \begin{cases} H_0(t) & \text{if } d(t) \leq \sqrt{\chi^2_{p+m}} \\ H_1(t) & \text{if } d(t) > \sqrt{\chi^2_{p+m}}, \end{cases}$$

where $p+m$ is the degrees of freedom of the χ^2 distribution. It is clear that if the observation $\omega(t)$ does not belong to the best subset S, then the Mahalanobis distance is greater

[1] # := cardinality of the subset S.
[2] $[\cdot]$ is the greatest integer number.

than the rejection point $\sqrt{\chi^2_{p+m,\nu}}$, so that $\delta(t)$ declared as $H_1(t)$.

In most cases, it is not feasible to generate all possible subsets provided that N is large due to computational cost. In the next section, we will generate finite number of subsets which will lead to a feasible solution that will converge with probability one to the true solution by using the randomized algorithm.

3. THE RANDOM SEARCH ALGORITHM

It is obvious that the minimum value for the objective function $J(S)$ in (7) can be found by searching for a subset $S \in \mathcal{S}$ that minimizes the value of $\det\mathrm{cov}(S)$. In fact there are $\binom{N}{M}$ subsets in $\mathcal{S}$, so that finding a subset that minimizes the value of the objective function is a very difficult combinatorial problem. However, we can easily calculate the value of $\det\mathrm{cov}(S)$, for each subset $S \in \mathcal{S}$ and then sort $\det\mathrm{cov}(S)$ in increasing order, i.e.

$$\det\mathrm{cov}(S)_{[1]} = \min_{S\in\mathcal{S}} \det\mathrm{cov}(S) \le \cdots \le \det\mathrm{cov}(S)_{[\binom{N}{M}]} = \max_{S\in\mathcal{S}} \det\mathrm{cov}(S).$$

Now we think of $S \in \mathcal{S}$ as a random variable that is uniformly distributed, and hence $\det\mathrm{cov}(S)$ is also a random variable depending on S. Let $F(\det\mathrm{cov}(S))$ denote the unknown cumulative distribution function of $\det\mathrm{cov}(S)$ for $i = 1, \cdots, L$ be L, independently generated samples of $S \in \mathcal{S}$. Furthermore, let $\bar{S} \in \{S\}_{r=1}^{L}$ be such that $\det\mathrm{cov}(\bar{S}) = \min_{1\le r\le L} \det\mathrm{cov}(S_r)$. We can derive the following theorem by using the result of Bai Bai [2003]. The theorem finds the sample size L so that $\det\mathrm{cov}(\bar{S})$ converges to the true solution with probability close to one.

Theorem 1. For the EIV model (1), the following (i) $\sim$ (ii) hold:

(i) For all $0 < F\left(\min_{S\in\mathcal{S}} \det\mathrm{cov}(S)\right) < \epsilon < 1$ and for all $0 < \delta < 1$, if

$$L \ge \frac{\ln(1/\delta)}{\ln(1/(1-\epsilon))}, \tag{8}$$

then

$$\mathrm{Prob}\left\{ F\left(\min_{1\le r\le L} \det\mathrm{cov}(S_r) \right) \le \epsilon \right\} \ge 1-\delta.$$

(ii) Let $\mathcal{S}_r$ for $r = 1, \cdots, k$ be k-th disjoint sub-samples such that $\cup_{r=1}^{k}\mathcal{S}_r = \{1, \cdots, N\}$ and run the randomized algorithm in each sub-sample. Then the overall probability that the confidence statement are simultaneously true is $1 - \sum_{i=1}^{k} \delta_i$.

Proof:

(ii) Assume that $S_i \subset \mathcal{S}_r$ and let $\mathcal{E}_g = F\left(\min_{1\le i\le L, S_i\subset\mathcal{S}_g} \det\mathrm{cov}(S_i)\right) \le \epsilon$, $(g = 1, \cdots, k)$ be the gth statement corresponds to the subset S_i, and assume that the gth statement $\mathcal{E}_g$, $(g = 1, \cdots, k)$ is correct, i.e.

$$\mathrm{Prob}[\mathcal{E}_g] \ge 1 - \delta_g,$$

and let $\bar{\mathcal{E}}_g$ be the complementary event of $\mathcal{E}_g$, then the overall probability that the statements are simultaneously true

$$\begin{aligned}\mathrm{Prob}[\cap\mathcal{E}_g] &= 1 - \mathrm{Prob}[\overline{\cap_g \mathcal{E}_g}] = 1 - \mathrm{Prob}[\cup_g \bar{\mathcal{E}}_g] \\ &\ge 1 - \sum_{g=1}^{k} \mathrm{Prob}[\bar{\mathcal{E}}_g] = 1 - \sum_{g=1}^{k} \delta_g,\end{aligned}$$

if $\delta_g = \delta$ for $g = 1, \cdots, k$. Then

$$\mathrm{Prob}[\cap\mathcal{E}_g] \ge 1 - k\delta.$$

Theorem 1 means that, whenever we generate L independent random subsets $S_L = \{S_i\}_{i=1}^{L}$ and compute the covariance for each subset $S_i \in S_L$, a subset $\bar{S} \in S_L$ with minimum covariance determinant $J(S)$ will improve our estimate. However, it may be noted that in the worst case this improvement is not considerable comparing to the LS estimate by using all observed data. In fact, if the number of the observed data is very large, then the probability of finding a subset $S \in \mathcal{S}$ with cardinality equal to M that does not contain any outlier approaches zero, i.e.

$$P_{\mathcal{I}} = \frac{\binom{\mathcal{I}}{M}}{\binom{N}{M}} = \frac{\mathcal{I}!(N-M)!}{(\mathcal{I}-M)!N!} = \prod_{j=0}^{M-1} \frac{\mathcal{I}-j}{N-j}, \tag{9}$$

where $\mathcal{I}$ stands for the number of clean data. According to (9) the random search algorithm can be improved by taking S with small cardinality and by finding the smallest M relative Mahalanobis distances d_i. This will increase the probability of finding a subset S_i from $\mathcal{S}$ that does not contain any outliers.

Part 2 of Theorem 1, means that even though we can attach a probability of $(1-\delta)$ to each separate sub-sample, the overall probability that the statements are simultaneously true is greater than $(1-k\delta)$. More discussion about the sub-samples will be given in section 4.1. It should be noted, that this result is known in statistics as the Bonferroni correction Seber [1977].

At this stage, we will derive the optimal estimate for the true input-output and the associated error covariances for the OEIV model using the Kalman filter and smoother.

4. KALMAN FILTER FOR THE ERRORS-IN-VARIABLES MODEL WITH OUTLIERS

Substituting (2) and (3) into (1) yields

$$\begin{bmatrix} x(t+1) \\ y(t) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + \begin{bmatrix} n_x(t) \\ n_y(t) \end{bmatrix}, \tag{10}$$

where $n_x(t) = -B\tilde{u}(t) + w(t)$ and $n_y(t) = -D\tilde{u}(t) + \tilde{y}(t)$. Let $z(t) = y(t) - Du(t)$, then (10) can be written as

$$\begin{bmatrix} x(t+1) \\ z(t) \end{bmatrix} = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + \begin{bmatrix} n_x(t) \\ n_y(t) \end{bmatrix}. \tag{11}$$

In addition, let $\mathcal{Z}(t) = \{z(0), \cdots, z(t)\}$, $\Phi(t) = \{\phi(0), \cdots, \phi(t)\}$ and $\Gamma(t) = \{\gamma(0), \cdots, \gamma(t)\}$ and define [3]

[3] The Kalman filter and smoother without outliers is given in Appendix A.

$$x(t \mid t) \equiv \mathbb{E}[x(t) \mid \mathcal{Z}(t), \Phi(t), \Gamma(t)], \quad (12)$$
$$x(t+1 \mid t) \equiv \mathbb{E}[x(t+1) \mid \mathcal{Z}(t), \Phi(t), \Gamma(t)], \quad (13)$$
$$y(t+1 \mid t) \equiv \mathbb{E}[y(t+1) \mid \mathcal{Z}(t), \Phi(t), \Gamma(t)], \quad (14)$$
$$P(t \mid t) \equiv \mathbb{E}[(x(t) - x(t \mid t))(x(t) - x(t \mid t))^{\mathrm{T}} \mid \mathcal{Z}(t), \Phi(t), \Gamma(t)], \quad (15)$$
$$P(t+1 \mid t) \equiv \mathbb{E}[(x(t+1) - x(t+1 \mid t))(x(t+1) - x(t+1 \mid t))^{\mathrm{T}} \mid \mathcal{Z}(t), \Phi(t), \Gamma(t)], \quad (16)$$

then the Kalman filter is given by

$$z(t+1 \mid t) = Cx(t+1 \mid t), \quad (17)$$
$$x(t+1 \mid t) = Ax(t \mid t) + Bu(t), \quad (18)$$

and we could compute the covariance of the errors as

$$\mathbb{E}\{(z(t+1) - z(t+1 \mid t))(x(t+1) - x(t+1 \mid t))^{\mathrm{T}}\} = CP(t+1 \mid t), \quad (19)$$

and

$$\mathbb{E}\{(z(t+1) - z(t+1 \mid t))(z(t+1) - z(t+1 \mid t))^{\mathrm{T}}\} = CP(t+1 \mid t)C^{\mathrm{T}} + (I_p - \gamma(t))\Sigma_{\tilde{y}}^n + \gamma(t)\Sigma_{\tilde{y}}^o + D[(I_m - \phi(t))\Sigma_{\tilde{u}}^n + \phi(t)\Sigma_{\tilde{u}}^o]D^{\mathrm{T}} - D\Sigma_{\underset{\sim}{uy}} - \Sigma_{\underset{\sim}{uy}}^{\mathrm{T}} D^{\mathrm{T}}, \quad (20)$$

where

$$P(t+1 \mid t) = \mathbb{E}[(x(t+1) - x(t+1 \mid t))(x(t+1) - x(t+1 \mid t))^{\mathrm{T}}] = AP_{t|t}A^{\mathrm{T}} + \Sigma_w + B(I_m - \phi(t))\Sigma_{\tilde{u}}^n B^{\mathrm{T}} + B\phi(t)\Sigma_{\tilde{u}}^o B^{\mathrm{T}}. \quad (21)$$

The optimal Kalman filter estimate for the state $x(t)$ is given by

$$x(t+1 \mid t+1) = x(t+1 \mid t) + P(t+1 \mid t)C^{\mathrm{T}}\Sigma_\epsilon(t)^{-1}\epsilon(t), \quad (22)$$

while $\epsilon(t)$ and $\Sigma_\epsilon(t)$ denote the innovation of $z(t)$ and its covariance matrix given by

$$\epsilon(t) = z(t) - Cx(t \mid t-1) = Cx(t) + n_y(t) - Cx(t \mid t-1) \quad (23)$$

$$\Sigma_\epsilon(t) = \mathbb{E}[\epsilon(t)\epsilon(t)^{\mathrm{T}}] = CP(t \mid t-1)C^{\mathrm{T}} + (I_p - \gamma(t))\Sigma_{\tilde{y}}^n + \gamma(t)\Sigma_{\tilde{y}}^o + D[(I_m - \phi(t))\Sigma_{\tilde{u}}^n + \phi(t)\Sigma_{\tilde{u}}^o]D^{\mathrm{T}} - D\Sigma_{\underset{\sim}{uy}} - \Sigma_{\underset{\sim}{uy}}^{\mathrm{T}} D^{\mathrm{T}}. \quad (24)$$

The optimal smooth estimates $\hat{u}(t \mid N)$, $\hat{y}(t \mid N)$ of $\hat{u}(t)$, $\hat{y}(t)$ that can be obtained from $\{u(0), y(0), \cdots, u(N), y(N)\}$, under constraints (1)-(3) are given by

$$\hat{u}(t \mid N) = u(t) - \tilde{u}(t \mid N) = u(t) - \mathbb{E}\{\tilde{u}(t) \mid z(0), \cdots, z(N)\}, \quad (25)$$
$$\hat{y}(t \mid N) = y(t) - \tilde{y}(t \mid N) = y(t) - \mathbb{E}\{\tilde{y}(t) \mid z(0), \cdots, z(N)\}, \quad (26)$$

where $\tilde{u}(t \mid N) = \mathbb{E}\{\tilde{u}(t) \mid z(0), \cdots, z(N)\}$ and $\tilde{y}(t \mid N) = \mathbb{E}\{\tilde{y}(t) \mid z(0), \cdots, z(N)\}$ are the optimal estimate for $\tilde{u}(t)$ and $\tilde{y}(t)$ respectively. To compute $\tilde{u}(t \mid N)$ and $\tilde{y}(t \mid N)$ we replace $z(t)$ by its innovation

$$\tilde{u}(t \mid N) = \mathbb{E}[\tilde{u}(t) \mid z(0), \cdots, z(t), \epsilon(t+1), \cdots, \epsilon(N)] = \mathbb{E}[\tilde{u}(t) \mid z(0), \cdots, z(t)] + \mathbb{E}[\tilde{u}(t) \mid \epsilon(t+1), \cdots, \epsilon(N)] = \tilde{u}(t \mid t) + \sum_{s=t+1}^{N} \mathrm{cov}\{\tilde{u}(t), \epsilon(s)\}\Sigma_\epsilon(s)^{-1}\epsilon(s) \quad (27)$$

$$\tilde{y}(t \mid N) = \mathbb{E}[\tilde{y}(t) \mid z(0), \cdots, z(t), \epsilon(t+1), \cdots, \epsilon(N)] = \mathbb{E}[\tilde{y}(t) \mid z(0), \cdots, z(t)] + \mathbb{E}[\tilde{y}(t) \mid \epsilon(t+1), \cdots, \epsilon(N)] = \tilde{y}(t \mid t) + \sum_{s=t+1}^{N} \mathrm{cov}\{\tilde{y}(t), \epsilon(s)\}\Sigma_\epsilon(s)^{-1}\epsilon(s), \quad (28)$$

where $\tilde{u}(t \mid t)$ and $\tilde{y}(t \mid t)$ are given in Appendix A. Now the covariances can be found as follows

$$\mathrm{cov}\{\tilde{u}(t), \epsilon(s)\} = [\Sigma_{\tilde{u}}(K(t)D - B)^{\mathrm{T}} - \Sigma_{\underset{\sim}{uy}} K^{\mathrm{T}}(t)]L(s-1, t)^{\mathrm{T}} C^{\mathrm{T}} \quad (29)$$

$$\mathrm{cov}\{\tilde{y}(t), \epsilon(s)\} = [\Sigma_{\underset{\sim}{uy}}^{\mathrm{T}}(K(t)D - B)^{\mathrm{T}} - \Sigma_{\tilde{y}} K^{\mathrm{T}}(t)]L(s-1, t)^{\mathrm{T}} C^{\mathrm{T}}, \quad (30)$$

where $\Sigma_{\tilde{u}} = (I_m - \phi(t))\Sigma_{\tilde{u}}^n + \phi(t)\Sigma_{\tilde{u}}^o$ and $\Sigma_{\tilde{y}} = (I_p - \gamma(t))\Sigma_{\tilde{y}}^n + \gamma(t)\Sigma_{\tilde{y}}^o$ and $\Sigma_{\underset{\sim}{uy}} = (I_m - \phi(t))\Sigma_{\underset{\sim}{uy}}^n (I_p - \gamma(t)) + \phi(t)\Sigma_{\underset{\sim}{uy}}^o \gamma(t)$. The definition of $L(s-1, t)$ and the proof of (29) and (30) are given in Proposition 4 (in Appendix A).

Proposition 2. Let π_t be a random integer number from 1 to N, and formulate the set $S = \{\pi_t : t = 1, \cdots, M\} \in \mathcal{S}$, $\omega_{\pi_t} = \begin{bmatrix} u(\pi_t) \\ y(\pi_t) \end{bmatrix}$, and $T_S = \frac{1}{M}\sum_{\pi_t \in S} \begin{bmatrix} \hat{u}(\pi_t \mid S) \\ \hat{y}(\pi_t \mid S) \end{bmatrix}$, where $\hat{u}(\pi_t \mid S)$ and $\hat{y}(\pi_t \mid S)$ can be calculated as in (25) and (26). Then the MCD cost function can be written as

$$J(S) = \min_S \det \frac{1}{M}\sum_{i \in S}(\omega_i - T_S)(\omega_i - T_S)^{\mathrm{T}}. \quad (31)$$

It should be noted that if t is included in the subset S, then $\phi(t)$ and $\gamma(t)$ will be the identity matrices, otherwise they are the zero matrices. In Proposition 2, if we apply the uniform sampling method then we will lose the structure of the original data and consequently the estimate will be biased. Therefore, we apply another sampling method which is called sub-sampling method Heagerty [2000].

4.1 Sub-sampling method

Instead of generating a random subsets from the observed input-output data we generate blocks of contiguous observations of fixed dimension b. That is, we divide the last $(N-n)$ observations into k subsets, where each subset contains the first initial data $(\omega(1), \cdots, \omega(n))$ and a set of $[(N-n)/k]$ contiguous observations. In other words, the subsets can be described as $S_r^{(b+n)} = \{\omega(1), \cdots, \omega(n), \omega(n+1+(r-1)b), \cdots, \omega(n+br)\}$, where $r = 1, \cdots, k$. Then we perform an exhaustive search of all possible blocks and choose the one which gives the minimum value for the objective function. It should be noted that, if $(N-n)/k$ is an integer then we have exactly k subsets. In general there are $k+1$ subsets, where the first k of size $n + [(N-n)/k]$, and the last of size $N - [(N-n)/k]k$. For the seek of simplicity and without loss of generality we assume that b is an integer where $b = (N-n)/k$.

Furthermore, if the number of the subsets k is large, then the probability of having at least a clean subset of data which does not contain any outlier will increase. However, if k is large, then the cardinality of each subset will be small, and consequently the estimate of the parameters can be unstable. P. Heagerty and T. Lumley Heagerty [2000] suggest that $b \approx \sqrt{N}$ to ensure a balance between the statistical properties of the estimated parameters and the robustness of the method.

Theorem 3. Let $\mid S_1^{(b+n)} \mid = M$. Put

$$T_1 = \frac{1}{M} \sum_{i \in S_1^{(b+n)}} \begin{bmatrix} \hat{u}(i) \\ \hat{y}(i) \end{bmatrix}$$

$$\mathrm{cov}(S_1^{(b+n)}) = \frac{1}{M} \sum_{i \in S_1^{(b+n)}} (\begin{bmatrix} u(i) \\ y(i) \end{bmatrix} - T_1)(\begin{bmatrix} u(i) \\ y(i) \end{bmatrix} - T_1)^{\mathrm{T}}$$

if $\det(\mathrm{cov}(S_1^{(b+n)})) \neq 0$, define

$$d_1(i) = \sqrt{\left\{ \begin{bmatrix} u(i) \\ y(i) \end{bmatrix} - T_1 \right\}^{\mathrm{T}} (\mathrm{cov}(S_1^{(b+n)}))^{-1} \left\{ \begin{bmatrix} u(i) \\ y(i) \end{bmatrix} - T_1 \right\}},$$

for $i = 1, \cdots, N$ and take $S_2^{(b+n)}$ such that $\{d_1(i); i \in S_2^{(b+n)}\} = \{(d_1)_{1:N}, \cdots, (d_1)_{M:N}\}$. Compute T_2 and $\mathrm{cov}(S_2^{(b+n)})$ based on $S_2^{(b+n)}$. Then

$$\det(\mathrm{cov}(S_2^{(b+n)})) \leq \det(\mathrm{cov}(S_1^{(b+n)}))$$

with equality if and only if $T_2 = T_1$ and $\mathrm{cov}(S_2^{(b+n)}) = \mathrm{cov}(S_1^{(b+n)})$.

The proof of Theorem 3 is a direct application of Theorem 1 in Rousseeuw et al. [2004]. It should noted that constructing a new subset $S_2^{(b+n)}$ from $S_1^{(b+n)}$ is called C-step where C stands for "concentration" because the new subset $S_2^{(b+n)}$ gives a lower value for the objective than $S_1^{(b+n)}$ does.

Random search algorithm:

Let $\cup_{i=1}^{k} S_i^{(b+n)} = \{1, 2, \cdots, N\}$,

- **Step 1:** Generate all subsamples of $S_i^{(b+n)}$, and for each sub-sample $S_i^{(b+n)}$, calculate $\det \mathrm{cov}(S_i^{(b+n)})$ and consequently find $\det \mathrm{cov}(\bar{S}) = \min_{S^{(b+n)} \in \mathcal{S}} \det \mathrm{cov}(S_i^{(b+n)})$.
- **Step 2:** Compute d_i, for $i = 1, \cdots, N$ and detect outliers using Chi-square distribution. Then, put $S_1 = \{\pi_i : i = 1, \cdots, M\}$.
- **Step 3:** Repeat step 1 to step 2, until convergent.

5. NUMERICAL EXAMPLE

The following example is a slightly modified version of the one used in Diversi et al. [2005]. Where the numerical simulation is performed on two inputs two outputs time-invariant system with $N = 500$ described by the following matrices:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -0.3 & 0.4 & -0.2 \\ -0.1 & 0.2 & 0.4 \end{bmatrix}, \quad B = \begin{bmatrix} 0.8 & -0.8 \\ 0.17 & -0.37 \\ 1.09 & 1.1 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1.7 & 1.5 \\ 0.51 & -1 \end{bmatrix}.$$

The noise free input sequence $\hat{u}(\cdot)$ is a zero mean, unit variance Gaussian process, and a sample of the unmeasurable output data $\hat{y}(\cdot)$ is shown in Fig. 2.

Furthermore, the noise sequences $w(t), \tilde{y}(t)$ and $\tilde{u}(t)$ are characterized as follows

$$w(t) \sim \mathrm{N}(0_3, 0.1 \times I_3),$$
$$\tilde{u}(t) \sim (I_2 - \phi(t))\mathrm{N}(0_2, 0.1 \times I_2) + \phi(t)\mathrm{N}(0, 2 \times I_2),$$
$$\tilde{y}(t) \sim (I_2 - \gamma(t))\mathrm{N}(0_2, 0.1 \times I_2) + \gamma(t)\mathrm{N}(0_2, 2 \times I_2),$$

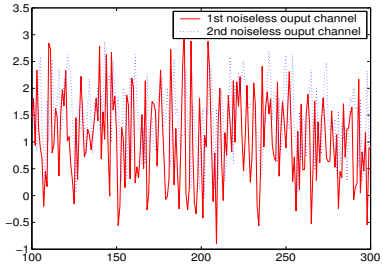

Fig. 2. Un-observed output data $\hat{y}$

where $\mathrm{Prob}(\phi(t) = I_2) = \mathrm{Prob}(\gamma(t) = I_2) = 1\%$. A sample for realizations of the noises $\tilde{y}(t)$ and $\tilde{u}(t)$ are shown in Fig. 3. Moreover, the initial state $x(0)$ is a random vector and

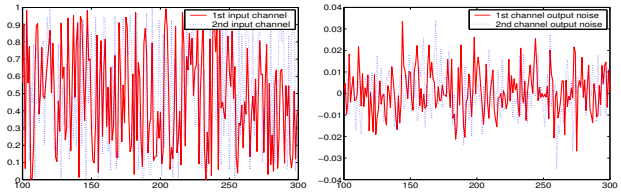

Fig. 3. A sample process noise $\tilde{u}$ and $\tilde{y}$

has been initialized as $x(0 \mid -1) = 0$ and $P(0 \mid -1) = I_n$.

In order to detect the outliers we generate all possible subsets S_r^{b+n} for $r = 1, \cdots N - (b + n) + 1$. Then for each subset S_r^{b+n}, we compute the MCD objective function and thus choose the best subset that gives the minimum value for the objective function. Hence the best subset has been used to detect the outliers in all observed data and consequently, we apply the C-step as in the Theorem 3. The Mahalanobis distance in first attempt to detect the outliers, i.e. before the C-step is shown Fig. 4.

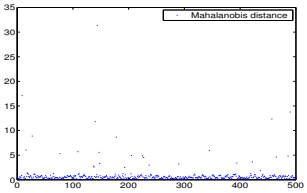

Fig. 4. Mahalanobis distance

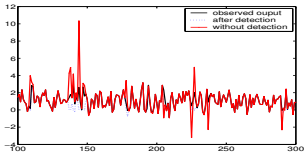

Fig. 5. The estimate

In Fig. 5, we show the noiseless output un-observed data, the estimate without detecting the outliers and after detecting the outliers. We see that the Kalman filter algorithm based on the MCD estimator significantly improve the estimate obtained by the EIV Kalman filter Diversi et al. [2005]. We noted that the proposed algorithm can deal with clean data as well as outlier problem.

6. CONCLUSION

In this paper, we have studied the Kalman filter and smoother for the Error-In-Variables state space models

with outliers. The outliers have been detected using highly robust estimator called minimum covariance determinant which requires the Kalman filter and smoother to be computed. In order to achieve the robust solution of the MCD problem, the random search algorithm has been proposed. However, applying the uniform sampling method to the randomized algorithm leads to complex calculation and biased estimate. Thus, we applied the subsampling method in order to keep the same dependence structure as the original data. The subsampling method leads to unbiased estimate and decrease the complexity issue of calculations. The proposed algorithm is highly robust to the effect of outliers.

REFERENCES

R. Diversi, R. Guidorzi and U. Soverini: Kalman filtering in extended noise environments; *IEEE Trans. Automatic Control*, Vol. 50, No. 9, pp. 1396-1402 (2005)

E. Bai: A random least-trimmed-squares identification algorithm; *Automatica*, Vol. 39, No. 9, pp. 1651-1659 (2003)

T. Proietti: Leave-k-out diagnostics in state-space models, *J. of Time Series Analysis*, Vol. 24, No. 2, pp. 221-236 (2003).

P. J. Rousseeuw, S. Van Aelst, K. Van Driessen and J. Agulló: Robust multivariate regression; *Technometrics*, Vol. 46, No. 3, pp. 293–305 (2004)

P. J. Rousseeuw and K. Van Driessen: A fast algorithm for the minimum covariance determinant estimator; *Technometrics*, Vol. 41, No. 3, pp. 212-223 (1999)

P. Rousseeuw: Least median of squares regression; *J. American Statistical Assoc.*, Vol. 79, pp. 871-880 (1984)

Ivan Markovsky and Bart De Moor: Linear dynamic filtering with noisy input and output; *Automatica*, Vol. 41, No. 1, pp. 167-171 (2005)

P. Heagerty and T. Lumley: Window subsampling of estimating functions with application to regression models; *J. American Statistical Assoc.*, Vol. 95, pp. 197-211 (2000)

R. Maronna, D. Martin and V. Yohai: Robust statistics: Theory and Methods; *John Wiley & sons* (2006)

C. Croux and G. Haesboeck: Influnece function and efficiency of the minimum covariance determinant scatter matrix estimator; *J. of Multivariate Analysis*, Vol. 71, pp. 161-190 (1999)

J. Agullo, C. Croux and S. Van Aelst: The Multivariate lest trimmed squares estimator; *J. of Multivariate Analysis*, in press (2007)

C. Masereliez and R. Martin: Robust Bayesian estimation for the linear model and robustifying the Kalman filter; *IEEE Trans. Automatic Control*, Vol. 22, No. 3, pp. 361-271 (1977)

R. Meinhold and N. Singpurwalla: Robsutification of Kalman filter models; *J. American Statistical Assoc.*, Vol. 84, N0. 406, pp. 479-486 (1989)

R. Frühwirth: Truck fitting with non-Gaussian noise; *Comp. Phys. Comm.*, Vol. 100, No. 1, pp. 479-486 (1997)

G.A.F. Seber: Linear regression analysis; *John Wiley & sons* (1977)

Appendix A. KALMAN FILTER AND SMOOTHER WITHOUT OUTLIERS

The Kalman filter is given by

$$z(t+1 \mid t) = Cx(t+1 \mid t) \tag{A.1}$$

$$x(t+1 \mid t) = Ax(t \mid t-1) + Bu(t) + K(t)\epsilon(t) \tag{A.2}$$

$$K(t) = [AP(t \mid t-1)C^{\mathrm{T}} + S(t)]\Sigma_\epsilon(t)^{-1} \tag{A.3}$$

$$P(t+1 \mid t) = AP(t \mid t-1)A^{\mathrm{T}} + Q(t) - [AP(t \mid t-1)C^{\mathrm{T}} + S(t)]\Sigma_\epsilon(t)^{-1}[AP(t \mid t-1)C^{\mathrm{T}} + S(t)]^{\mathrm{T}} \tag{A.4}$$

and the Kalman smoother for $t = N, N-1, \cdots, 1$ is given by

$$x(t-1 \mid N) = x(t-1 \mid t-1) + J(t-1)[x(t \mid N) - x(t \mid t-1)] \tag{A.5}$$

$$P(t-1 \mid N) = P(t-1 \mid t-1) + J(t-1)[P(t \mid N) - P(t \mid t-1)] \times J(t-1)^{\mathrm{T}} \tag{A.6}$$

$$J(t-1) = P(t-1 \mid t-1)AP(t \mid t-1)^{-1} \tag{A.7}$$

$$\tilde{u}(t \mid t) = [\Sigma_{\underset{\sim}{uy}}(t) - \Sigma_{\tilde{u}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t) \tag{A.8}$$

$$\tilde{y}(t \mid t) = [\Sigma_{\tilde{y}} - \Sigma^{\mathrm{T}}_{\underset{\sim}{uy}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t) \tag{A.9}$$

By using (A.8) and (A.9), the minimal variance estimates of $\hat{y}(t)$ and $\hat{u}(t)$ can be written in the form

$$\hat{u}(t \mid t) = u(t) - [\Sigma_{\underset{\sim}{uy}} - \Sigma_{\tilde{u}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t) \tag{A.10}$$

$$\hat{y}(t \mid t) = y(t) - [\Sigma_{\tilde{y}} - \Sigma^{\mathrm{T}}_{\underset{\sim}{uy}}D^{\mathrm{T}}]\Sigma_\epsilon(t)^{-1}\epsilon(t) \tag{A.11}$$

Proposition 4. For $1 \le t \le s$, the followings hold

$$\mathrm{cov}\{\tilde{u}(t), \epsilon(s)\} = [\Sigma_{\tilde{u}}(t)(K(t)D - B)^{\mathrm{T}} - \Sigma_{\underset{\sim}{uy}}(t)K^{\mathrm{T}}(t)] \times L(s-1,t)^{\mathrm{T}}C^{\mathrm{T}} \tag{A.12}$$

$$\mathrm{cov}\{\tilde{y}(t), \epsilon(s)\} = [\Sigma^{\mathrm{T}}_{\underset{\sim}{uy}}(t)(K(t)D - B)^{\mathrm{T}} - \Sigma_{\tilde{y}}(t)K^{\mathrm{T}}(t)] \times L(s-1,t)^{\mathrm{T}}C^{\mathrm{T}}, \tag{A.13}$$

where $L(s-1,t) = L(s-1)\cdots L(t)$ and $L(s-1) = A - K(s-1)C$.

Proof:

$$\begin{aligned}
&x(s+1) - x(s+1 \mid s)\\
&= A\left(x(s) - x(s \mid s-1)\right) + n_x(s) - K(s)\epsilon(s)\\
&= \left(A - K(s)C\right)\left(x(s) - x(s \mid s-1)\right) + n_x(s) - K(s)n_y(s)\\
&= L(s)\left(x(s) - x(s \mid s-1)\right) + n_x(s) - K(s)n_y(s),
\end{aligned}$$

$$\begin{aligned}
&\mathrm{cov}\{\tilde{u}(t), \epsilon(s)\} = \mathrm{cov}\{\tilde{u}(t), z(s) - Cx(s \mid s-1)\}\\
&\quad = \mathrm{cov}\{\tilde{u}(t), C\left(x(s) - x(s \mid s-1)\right) + n_y(s)\}\\
&\quad = \mathrm{cov}\{\tilde{u}(t), C(L(s-1)(x(s-1) - x(s-1 \mid s-2))\\
&\qquad\qquad + n_x(s-1) - K(s-1)n_y(s-1)) + n_y(s)\}\\
&\quad = \mathrm{cov}\{\tilde{u}(t), n_x(t) - K(t)n_y(t)\}L(s-1,t)^{\mathrm{T}}C^{\mathrm{T}}\\
&\quad = [\Sigma_{\tilde{u}}(K(t)D - B)^{\mathrm{T}} - \Sigma_{\underset{\sim}{uy}}K^{\mathrm{T}}(t)]L(s-1,t)^{\mathrm{T}}C^{\mathrm{T}}
\end{aligned}$$

$$\begin{aligned}
&\mathrm{cov}\{\tilde{y}(t), \epsilon(s)\} = \mathrm{cov}\{\tilde{y}(t), z(s) - Cx(s \mid s-1)\}\\
&\quad = \mathrm{cov}\{\tilde{y}(t), C\tilde{x}(s \mid s-1) + n_y(s)\}\\
&\quad = \mathrm{cov}\{\tilde{y}(t), n_x(t) - K(t)n_y(t)\}L(s,t)^{\mathrm{T}}C^{\mathrm{T}}\\
&\quad = [\Sigma^{\mathrm{T}}_{\underset{\sim}{uy}}(K(t)D - B)^{\mathrm{T}} - \Sigma_{\tilde{y}}K^{\mathrm{T}}(t)]L(s-1,t)^{\mathrm{T}}C^{\mathrm{T}},
\end{aligned}$$