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

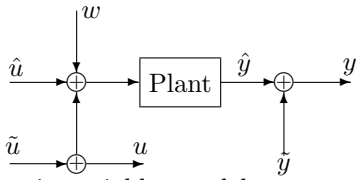

Fig. 1. Errors-in-variables model

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.