Communications for Statistical Applications and Methods
2018, Vol. 25, No. 1, 99–107

https://doi.org/10.29220/CSAM.2018.25.1.099
Print ISSN 2287-7843 / Online ISSN 2383-4757

Robustizing Kalman filters with the M-estimating functions

Ro Jin Pak[1,a]

[a]Department of Applied Statistics, Dankook University, Korea

Abstract

This article considers a robust Kalman filter from the M-estimation point of view. Pak (*Journal of the Korean Statistical Society*, **27**, 507–514, 1998) proposed a particular M-estimating function which has the data-based shaping constants. The Kalman filter with the proposed M-estimating function is considered. The structure and the estimating algorithm of the Kalman filter accompanying the M-estimating function are mentioned. Kalman filter estimates by the proposed M-estimating function are shown to be well behaved even when data are contaminated.

Keywords: Kalman filter, M-estimation, redescending function, robust estimation

1. Introduction

The Kalman filter, named after Rudolf E. Kalman (1960), has been an important algorithm in the fields of control theory and time series analysis. The Kalman filter leads the estimators of the signal or the state of an state-space model. Its usefulness and versatility have prevailed with adequate performance when observations are made online. Recently, it has become a major tool for artificial intelligence and robotics utilizing big data (Thrun, 2002) as well as for space time series analysis (Lee and Kim, 2010; Lee, *et al.*, 2011).

The Kalman filter uses the mean squared error (MSE) as a criterion of optimality but this criterion exaggerates the magnitude of errors. Therefore, the estimates of the signals are heavily influenced by abnormal observations or outliers. There have been numerous attempts to make the Kalman filter robust against abnormal observations based on M-estimation methodology (Ruckdeschel, 2000; Gandhi and Mili, 2010).

In order to use M-estimating functions, we need the values of shaping constants such as ‘cut-off constant’, ‘bending constant’, or ‘tuning constant’. In the previous works by Ruckdeschel (2000) and by Gandhi and Mili (2010), these constants are usually fixed or predetermined while updating the estimates for the signals themselves. However, the M-estimating function proposed by Pak (1998) is actually based on data-driven shaping constants, so that those constants can be easily updated as an observation comes in or as the iteration continues. As a result, it is found out that the Kalman filter with the proposed M-estimating function performs very stably under contaminated situations.

It should be noted that this article is not actually talking about M-estimation on the Kalman filter, rather it uses the M-estimating function themselves to handle unusual observations. In order to carry out the Kalman filtering stably with unusual observations, this article concerns to utilize the M-estimating functions to treat those observations. Robustly estimating the Kalman filter is another difficult problem that needs to be solved.

[1] Department of Applied Statistics, Dankook University, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 16890, Korea. E-mail: rjpak@dankook.ac.kr

Published 31 January 2018 / journal homepage: http://csam.or.kr

2. Kalman filter

This section is based on the references by Brockwell and Davis (1986) and by Kay (1993). Assume that the $M \times 1$ signal vector $\mathbf{x}[n]$ follows the vector state-vector observation model (or the state-space model):

$$\mathbf{s}[n] = \mathbf{A}\mathbf{s}[n-1] + \mathbf{B}\mathbf{u}[n], \quad n \geq 0, \tag{2.1}$$

where $\mathbf{A}$, $\mathbf{B}$ are known $p \times p$ and $p \times r$ matrices, $\mathbf{u}[n]$ is vector white Gaussian noise (WGN) with $\mathbf{u}[n] \sim N(\mathbf{0}, \mathbf{Q})$, $\mathbf{s}[-1] \sim N(\mu_{\mathbf{s}}, \mathbf{C}_{\mathbf{s}})$, and $\mathbf{s}[-1]$ is independent of the $\mathbf{u}[n]$'s. The observations are modeled as

$$\mathbf{x}[n] = \mathbf{H}[n]\mathbf{s}[n] + \mathbf{w}[n], \tag{2.2}$$

where $\mathbf{H}[n]$ is a known $M \times p$ matrix, and $\mathbf{x}[n]$ is an $M \times 1$ observation vector, and $\mathbf{w}[n]$ is a $M \times 1$ observation noise sequence. The $\mathbf{w}[n]$'s are independent of each other and of $\mathbf{u}[n]$ and $\mathbf{s}[-1]$, and $\mathbf{w}[n] \sim N(\mathbf{0}, \mathbf{C}[n])$.

The estimator of $\mathbf{s}[n]$

$$\hat{\mathbf{s}}[n|n] = E\left(\mathbf{s}[n]|\mathbf{x}[0], \mathbf{x}[1], \ldots, \mathbf{x}[n]\right)$$

can be sequentially obtained in the following manner:

- Step 0. Initialization:

$$\hat{\mathbf{s}}[-1|-1] = \mu_{\mathbf{s}} \quad \text{and} \quad \mathbf{M}[-1|-1] = \mathbf{C}_{\mathbf{s}}.$$

- Step 1. Prediction:

$$\hat{\mathbf{s}}[n|n-1] = \mathbf{A}\hat{\mathbf{s}}[n-1|n-1]. \tag{2.3}$$

- Step 2. Mean squared error matrix for prediction:

$$\mathbf{M}[n|n-1] = \mathbf{A}\mathbf{M}[n-1|n-1]\mathbf{A}^T + \mathbf{B}\mathbf{Q}\mathbf{B}^T. \tag{2.4}$$

- Step 3. Kalman gain:

$$\mathbf{K}[n] = \mathbf{M}[n|n-1]\mathbf{H}^T[n]\left(\mathbf{C}[n] + \mathbf{H}[n]\mathbf{M}[n|n-1]\mathbf{H}^T[n]\right)^{-1}. \tag{2.5}$$

- Step 4. Correction:

$$\hat{\mathbf{s}}[n|n] = \hat{\mathbf{s}}[n|n-1] + \mathbf{K}[n]\left(\mathbf{x}[n] - \mathbf{H}[n]\hat{\mathbf{s}}[n|n-1]\right). \tag{2.6}$$

- Step 5. Mean squared error matrix for correction:

$$\mathbf{M}[n|n] = \left(\mathbf{I} - \mathbf{K}[n]\mathbf{H}[n]\right)\mathbf{M}[n|n-1]. \tag{2.7}$$

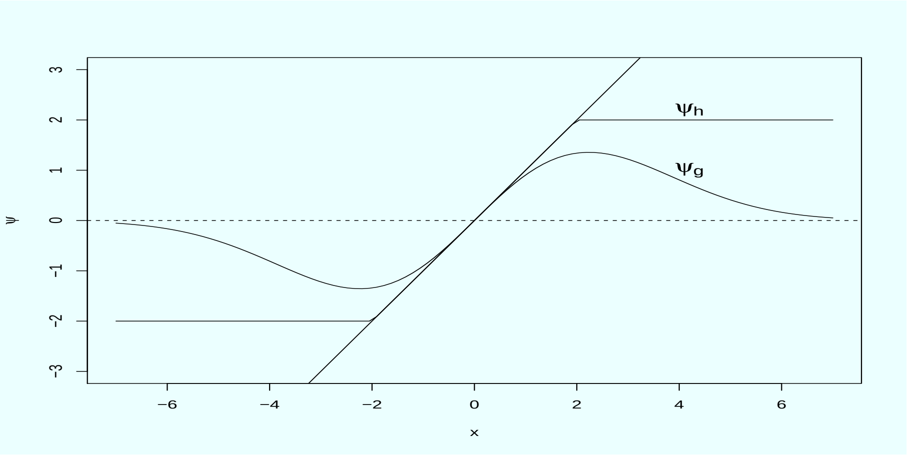


Figure 1: *Examples of ψ_h and ψ_g.*

3. M-estimation

M-estimation is a representative statistical method to estimate parameters robustly. The content in this section are based mainly on the books by Huber and Ronchetti (2009) and by Hampel *et al.* (2011). The M-estimation has roots on minimization problems such as the least squares estimation. Huber (1964) proposed the estimating method as a minimization problem for a parameter θ such as

$$\hat{\theta} = \arg\min_{\theta} \left(\sum_{i=1}^{n} \rho(X_i, \theta) \right), \tag{3.1}$$

when the random samples $\{X_1, \ldots, X_n\}$ from a density $f(x, \theta)$ are given. If the $\rho(r)$ is r^2, the M-estimation is nothing but the least squares estimation and if the $\rho(r)$ is $-\log f(r)$, the M-estimation is equivalent to the maximum likelihood estimation.

The minimization problem is actually turned to solve

$$\sum_{i=1}^{n} \frac{d}{d\theta} \rho(X_i, \theta) = \sum_{i=1}^{n} \psi(X_i, \theta) = 0, \quad \psi = \rho'. \tag{3.2}$$

For example, a representative ψ-function, which was proposed by Huber (1964), is

$$\psi_h(r) = \begin{cases} r, & |r| < b, \\ c \cdot \mathrm{sgn}(r), & |r| \geq b, \end{cases}$$

for a given cutoff constant b. The Huber's M-estimating function was designed to bound the influence of outlying observations (Figure 1).

However, Pak (1998) introduced a new type of M-estimating function as

$$\psi_g(r) = r \exp\left\{ -\frac{r^2}{2\,(h^2 + \sigma^2)} \right\},$$

where the h is a bandwidth of a density estimator, and the σ is a standard deviation (Figure 1). The above ψ_g-function is a redescending and differentiable everywhere unlike the other ψ-functions. In practice, it is needed to replace σ by the sample standard deviation or by a robust estimator like the median absolute deviance (MAD). The h can be replaced by the optimal bandwidth proposed by Scott (2009), Silverman (1986), or Sheather and Jones (1991). As h and σ approach to 0, a $\psi_g(r)$ then becomes just r, which produces the least squares estimator (LSE) or in some cases the maximum likelihood estimator (MLE). The asymptotic properties about the estimator with a ψ_g are the same as those of the LSE or the MLE. Details can be found in Pak (1998) but we briefly explain below how to get the above ψ_g-function.

Let $g_\theta(x) \in L_2$ be a family of probability densities indexed by θ and let $p(x) \in L_2$ be a density estimator for $g_\theta(x)$ such as

$$p(x) = \frac{1}{n}\sum \frac{1}{h}K\left(\frac{x - X_i}{h}\right),$$

where $K(\cdot)$ is a kernel density and h is a window width (or bandwidth). The estimator $\hat{\theta}$ which minimize the L_2 distance,

$$\int (p(x) - g_\theta(x))^2 dx,$$

is called the minimum L_2 distance estimator. Minimizing the above L_2 distance to get an estimator is equivalent to solve the equation,

$$\int (p(x) - g_\theta(x))\nabla_\theta g_\theta(x) dx = 0,$$

which is also equivalent to solve

$$\int p(x)\nabla_\theta g_\theta(x) dx = 0 \tag{3.3}$$

because $\int g_\theta(x)\nabla g_\theta(x)dx = (1/2)\nabla_\theta \int g_\theta^2(x)dx = 0$.

The equation (3.3) can be rewritten as

$$\sum \nabla_\theta \int \frac{1}{h}K\left(\frac{x - X_i}{h}\right) g_\theta(x) dx = 0,$$

and then an M-estimating function ψ is defined as

$$\psi(X_i, \theta) = \nabla_\theta \int \frac{1}{h}K\left(\frac{x - X_i}{h}\right) g_\theta(x) dx. \tag{3.4}$$

For example, if g_θ is $N(\mu, \sigma)$ and $K(t) = (1/\sqrt{2\pi})\exp\{-t^2/2\}$ (Gaussian kernel) then $\int h^{-1}K\{h^{-1}(x - X_i)\}g_\theta(x)dx$ becomes $N(\mu, h^2 + \sigma^2)$. After dropping unnecessary coefficients, we have

$$\psi_g(r) = r\exp\left\{-\frac{r^2}{2(h^2 + \sigma^2)}\right\},$$

where the g stands for the Gaussian. The $\psi_g(r)$ can be thought as $(r)\times$ (weighting factor), which controls the magnitude of r exponentially. We will use this $\psi_g(r)$-function in robustizing the Kalman filter.

4. Robust scalar Kalman filter

For simplicity, we assume that the signals follow the scalar Gauss-Markov signal model (or state model);

$$s[n] = as[n-1] + u[n], \quad n \le 0, \tag{4.1}$$

where a is a scalar coefficient, and $u[n]$ is a Gaussian noise with a variance σ_u^2, $s[-1] \sim N(\mu_s, \sigma_s^2)$, and $s[-1]$ is independent of $u[n]$ for all $n \ge 0$. Also, assume that the observations follow the scalar observation model;

$$x[n] = s[n] + w[n], \tag{4.2}$$

where $w[n]$ is zero mean Gaussian noise with a variance σ_w^2.

We can summarize the algorithm for the scalar state-scalar observation Kalman filter according to the following steps.

$$\begin{aligned}
\text{Prediction} &: \hat{s}[n|n-1] = a\hat{s}[n-1|n-1] \\
\text{Prediction MSE} &: M[n|n-1] = a^2 M[n-1|n-1] + \sigma_n^2 \\
\text{Kalman gain} &: K[n] = \frac{M[n|n-1]}{\sigma_n^2 + M[n|n-1]} \\
\text{Correction} &: \hat{s}[n|n] = a\hat{s}[n|n-1] + K[n](x[n] - \hat{s}[n|n-1]) \\
\text{MSE} &: M[n|n] = (1 - K[n])M[n|n-1]
\end{aligned}$$

The robust version of the Kalman filter can be proposed by replacing (4.3) by

$$\hat{s}[n|n] = \hat{s}[n|n-1] + K[n]\psi_g(x[n] - \hat{s}[n|n-1])$$

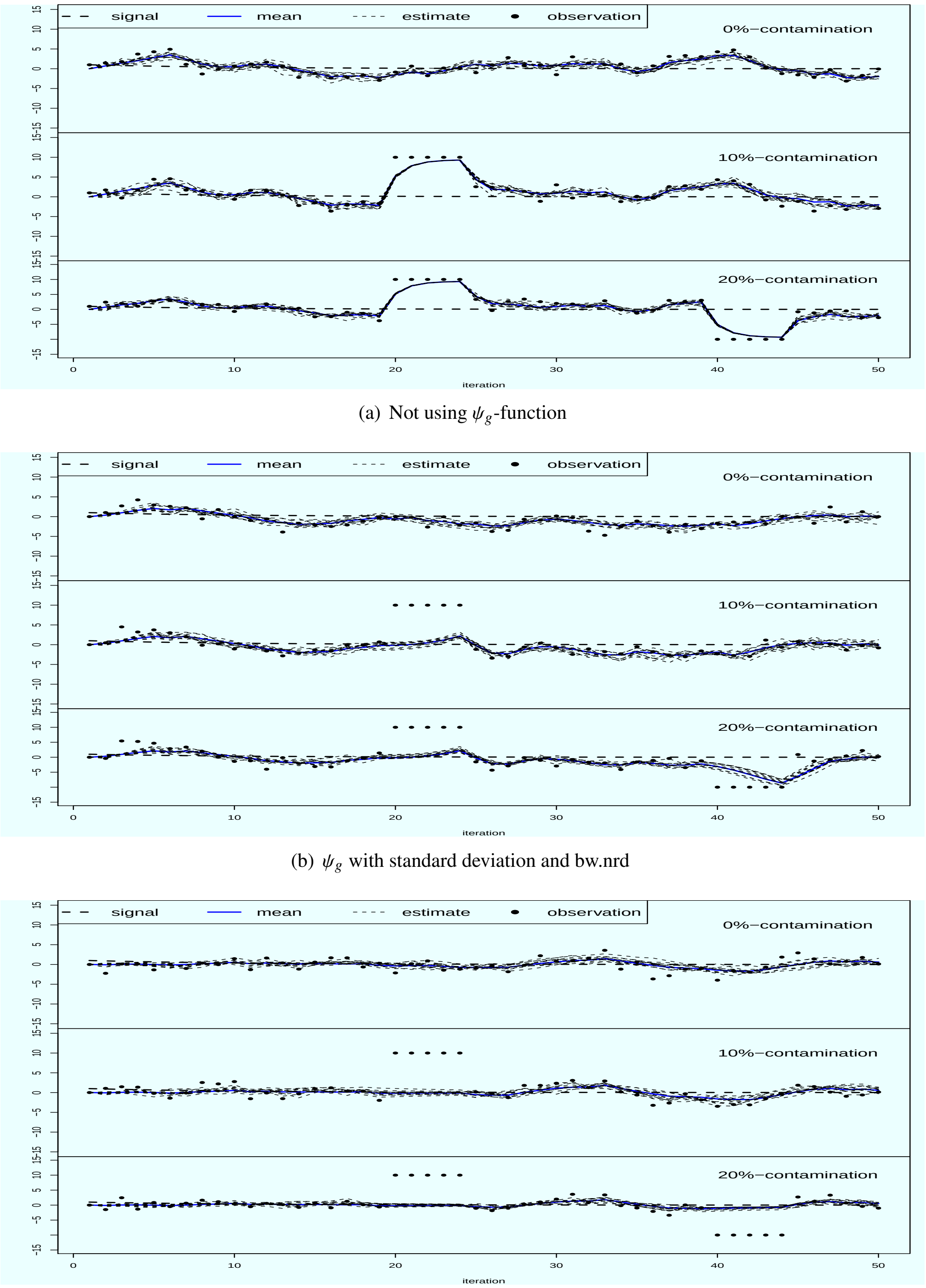
in order to bound the influence of the one-stop prediction error. The signal estimate is then

$$\hat{s}[n|n] = a\hat{s}[n-1|n-1] + K[n]\psi_g(x[n] - \hat{s}[n|n-1]).$$

5. Data analysis

Suppose that the data are from $x[n] = s[n] + w[n]$, $w[n] \sim N(0, 1)$ and $N(0, 4)$ and the signal model is an autoregressive model of order 1, AR(1), $s[n] = 0.9s[n-1] + u[n]$ for $n \ge 1$, where $u[n] \sim N(0, 1)$ and $s[1] = 1$. Five hundreds sets of fifty observations are generated based on the above model. In order to verify the robustness of the signal estimates, we simulated the contaminated data sets with the 10% and 20% of the observations replaced by the number 5 or -5 on purpose. We estimate the signal by the Kalman filter algorithm: (1) without using an M-estimating function, (2) with the ψ_g along with $\hat{\sigma}_n$ and the bandwidth by Scott (2009) (bw.nrd), and (3) with the ψ_g along with MAD and the bandwidth by Sheather and Jones (1991) (bw.SJ). An initial signal estimate is assumed to be 0.

The true signal (- - -), the ten examples of the signal estimates ($\cdots$), the mean of the all signal estimates (—) and an sample of observations (●) are plotted in Figure 2. We can observe that the ψ_g with MAD and bw.SJ produces stable and robust signal estimates in Figure 2(c) and Figure 3(c) while the other signal estimates in Figure 2(a) and (b) are heading toward the outlying observations which are either 5 or -5. When we use ψ_g with MAD and bw.SJ, the minimum squared errors of the estimates from the true signals are actually smaller than the other cases (Table 1). However, when the error variances are large, for example σ_w^2 is 4 in this case, the minimum squared errors tend to be relatively larger than when σ_w^2 is 1 (Table 2). In fact, the M-estimating function ψ_g is designed to handle a location parameter so that when the variance is large, the performance is relatively poor.

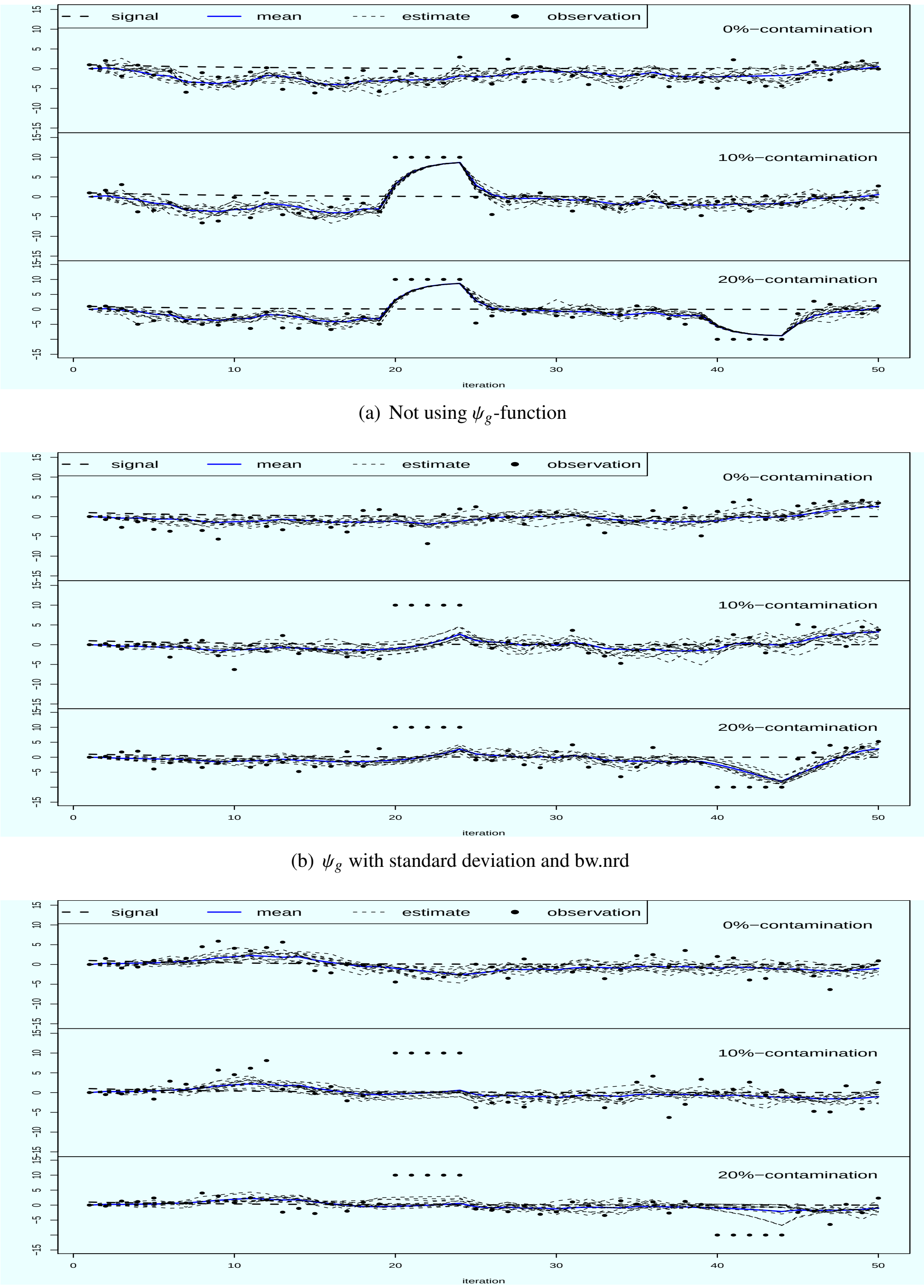


(a) Not using ψ_g-function

(b) ψ_g with standard deviation and bw.nrd

(c) ψ_g with MAD and bw.SJ

Figure 2: *Simulation results when the error is* $N(0,1)$. *bw.nrd = bandwidth by Scott (2009); bw.SJ = bandwidth by Sheather and Jones (1991); MAD = median absolute deviance.*



(a) Not using ψ_g-function

(b) ψ_g with standard deviation and bw.nrd

(c) ψ_g with MAD and bw.SJ

Figure 3: *Simulation results when the error is* $N(0,4)$. *bw.nrd = bandwidth by Scott (2009); bw.SJ = bandwidth by Sheather and Jones (1991); MAD = median absolute deviance.*

Table 1: minimum squared error statistics when $\sigma_e^2 = 1$

Contamination	Min	1st quarter	Median	Mean	3rd quarter	Max
Without ψ_g-function						
0%	4.422	5.563	5.946	5.989	6.365	8.050
10%	9.913	10.843	11.175	11.211	11.564	12.969
20%	16.810	18.170	18.540	18.550	18.930	20.180
With ψ_g-function; σ = sample standard deviation, h = bw.nrd						
0%	1.015	1.780	2.060	2.058	2.314	3.162
10%	1.438	2.294	2.564	2.585	2.859	3.942
20%	1.615	2.558	2.892	2.946	3.274	4.995
With ψ_g-function; σ = MAD, h = bw.SJ						
0%	0.573	1.440	1.720	1.736	2.009	3.280
10%	0.679	1.491	1.779	1.775	2.017	3.335
20%	0.301	1.047	1.261	1.2897	1.526	2.794

bw.nrd = bandwidth by Scott (2009); MAD = median absolute deviance; bw.SJ = bandwidth by Sheather and Jones (1991).

Table 2: minimum squared error statistics when $\sigma_e^2 = 4$

Contamination	Min	1st quarter	Median	Mean	3rd quarter	Max
Without ψ_g-function						
0%	4.422	5.563	5.946	5.989	6.365	8.050
10%	9.913	10.843	11.175	11.211	11.564	12.969
20%	16.810	18.170	18.540	18.550	18.930	20.180
With ψ_g-function; σ = sample standard deviation, h = bw.nrd						
0%	0.811	1.846	2.335	2.358	2.793	4.775
10%	1.227	2.444	2.900	2.977	3.432	6.523
20%	2.549	5.180	6.647	6.754	8.242	12.537
With ψ_g-function; σ = MAD, h = bw.SJ						
0%	0.540	1.872	2.396	2.436	2.954	5.614
10%	0.508	1.399	1.843	1.945	2.288	7.145
20%	0.537	1.381	1.865	2.374	2.773	10.664

bw.nrd = bandwidth by Scott (2009); MAD = median absolute deviance; bw.SJ = bandwidth by Sheather and Jones (1991).

6. Summary

We have demonstrated to run the Kalman filter with a special M-estimating function as well as indicated that the estimated signal can cope with unusual observations. This article utilized the M-estimating functions to treat those observations in order to conduct the Kalman filtering stably with unusual observations. The robustly estimating the Kalman filter is another difficult problem that needs to be solved. In this article, only the scalar Kalman filter has been treated, though an idea how to extend the proposed methodology to the multivariate situation, but has to be fully studied in the future.

References

Brockwell PJ and Davis RA (1986). *Time Series: Theory and Methods*, Springer-Verlag, New York

Gandhi MA and Mili L (2010). Robust Kalman filter based on a generalized maximum-likelihood-type estimator, *IEEE Transactions on Signal Processing*, **58**, 2509–2520.

Hampel FR, Ronchetti EM, Rousseeuw PJ, and Stahel WA (2011). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

Huber PJ (1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics*, **35**, 73–101.

Huber PJ and Ronchetti EM (2009). *Robust Statistics* (2nd ed), John Wiley, Chichester.

Kalman RE (1960). A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, **82**, 35–45.

Kay SM (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, NJ.

Lee SD, Han EH, and Kim DK (2011). Kalman-Filter estimation and prediction for a spatial time series model, *Communications for Statistical Applications and Methods*, **18**, 79–87.

Lee SD and Kim DK (2010). The comparison of imputation methods in space time series data with missing values, *Communications for Statistical Applications and Methods*, **17**, 263–273.

Pak RJ (1998). M-estimation function induced from minimum l2 distance estimation, *Journal of the Korean Statistical Society*, **27**, 507–514.

Ruckdeschel P (2000). *Robust Kalman Filtering* (discussion paper), Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes.

Scott DW (2009). *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York.

Sheather SJ and Jones MC (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B (Methodological)*, **53**, 683–690.

Silverman BW (1986). *Density Estimation for Statistics and Data Analysis*, CRC Press, New York.

Thrun S (2002). Particle filters in robotics. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, 511–518.

Received October 27, 2017; Revised November 25, 2017; Accepted December 2, 2017