A SIMPLE SEQUENTIAL OUTLIER DETECTION WITH SEVERAL RESIDUALS

Ji Won Yoon

Cyber defense department, Korea University, Republic of Korea (South Korea)

ABSTRACT

Outlier detection schemes have been used to identify the unwanted noise and this helps us to obtain underlying valuable signals and predicting the next state of the systems/signals. However, there are few researches on sequential outlier detection in time series although a lot of outlier detection algorithms are developed in off-line systems. In this paper, we focus on the sequential (on-line) outlier detection schemes, that are based on the *'delete-replace'* approach. We also demonstrate that three different types of residuals can be used to design the outlier detection scheme to achieve accurate sequential estimation: *marginal residual, conditional residual*, and *contribution*.

Index Terms— Outlier detection, Marginal residual, Conditional residual, Contribution

1. INTRODUCTION

In research fields related to defense science and technology, one of the interesting topics is methods for obtaining accurate underlying valuable signals and predicting the next state of the systems/signals using signal processing techniques that are effective in unwanted environments. If the linear Gaussian assumption is valid for the signal, the Kalman filter is often used for the prediction of the systems/signals. However, the signals can be corrupted by malicious jammers or an unwanted external influence. In this case, it is better to remove the corrupted observations than to consider them under Gaussian uncertainty, since handling noise is rather difficult, and Gaussian assumption may not be realistic. From this point of view, the corrupted signals can be regarded as outliers, which are commonly called novelty, anomaly, or abnormal signals, depending on the research field. Therefore, we need to combine outlier detection algorithms and sequential prediction techniques to handle such unwanted outliers.

Anomaly detection is one of the key problems in signal processing, statistics, and machine learning [1, 2, 3, 4, 5, 6, 7]. It is often variously called: outlier detection in the machine learning field, anomaly detection in the signal processing and measurement diagnostics in the field of statistics. In general, anomaly detection is one of the most difficult issues since a new pattern that has not been found or trained has to be detected. In the literature, there are several types of outliers [1]: the innovation outlier (IO), the additive outlier (AO), a level change/level shift outlier (LC/LSO), transient change (TC) and variance change (VC) outliers. In this paper, we focus on the subsequently multiple additive outlier (AO), innovation outlier (IO), and level shift outlier (LSO). Other types of outliers will be considered in future work.

This research is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A1012797).

2. TECHNICAL BACKGROUND

2.1. Several residuals in linear mixed model

Anomaly detection in statistics is studied mostly in the generalized linear (mixed) model and generalized least squares (GLS) with a vector of observations $\mathbf{Y}$: $\mathbf{Y} = \mathbf{B}\mathbf{X} + \epsilon$, where $\epsilon \sim MVN(\cdot; \mathbf{0}, \mathbf{V})$ and $\mathbf{X}$ is an unknown parameter to be fitted and estimated. Specialization of $\mathbf{V}$ allows us to address a large range of situations. When ordinary least squares (OLS) suffices, $\mathbf{V} = \sigma^2\mathbf{I}$ provides the greatest specialization. In this model there are, in general, two different types of residual with two different aspects of the deviation of the data from the fitted model [8, 9]. One of the types of residual is the *marginal/classical residual*, which is the difference between observation and the fitted value in 'estimation'. The other type of residual is the *conditional residual*, which is the difference between observation and the predicted value in 'prediction'. In the linear model, it is known that

$$\hat{\mathbf{X}} = (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}Y$$

and that

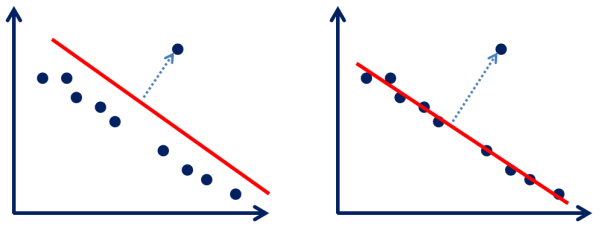
$$var(\hat{\mathbf{X}}) = (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}.$$

Hence, the marginal/classical residual $\hat{\epsilon} = \mathbf{Y} - \mathbf{B}\hat{\mathbf{X}}$ can be represented by the form of $\mathbf{Y}$ as $\hat{\epsilon} = \mathbf{V}\mathbf{Q}\mathbf{Y}$ where $\mathbf{Q} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{B}\hat{\mathbf{X}}\mathbf{Y}^{-1}$.

Now, let us consider the conditional residual which is based on the prediction model. Suppose that $\mathbf{Y}$ can be partitioned into two subsets $\mathbf{Y}_a$, and $\mathbf{Y}_b$, where $\mathbf{Y}_a \cup \mathbf{Y}_b = \mathbf{Y}$ and $\mathbf{Y}_a \cap \mathbf{Y}_b = \{\}$. We can estimate the *conditional residual* $\tilde{\epsilon}_{(a)}$ [8] by $\tilde{\epsilon}_{(a)} = \mathbf{Y}_a - \tilde{\mathbf{Y}}_a(\mathbf{Y}_b)$, which is the difference between $\mathbf{Y}_a$ and $\tilde{\mathbf{Y}}_a(\mathbf{Y}_b)$.

Figure 1 shows two different types of residual for a particular data point, *marginal* and *conditional*. The red solid

978-0-9928626-3-3/15/$31.00 ©2015 IEEE



(a) Marginal residual $\hat{\epsilon}_i$ (b) Conditional residual $\tilde{\epsilon}_{(i)}$

Fig. 1. Two types of residual

lines represent the regression of the fitted model, and therefore the *marginal* data points $\hat{\mathbf{Y}}_i$ and the *conditional* data point $\tilde{\mathbf{Y}}_{(i)}$ lie on the lines. That is, while the marginal residual is obtained from the regression with full data, the conditional residual is obtained from the regression with deleted subsets. The residuals via different aspects of the regression may have their own interesting characteristics which have their practical advantages and disadvantages. For instance, if there is a singleton outlier at a large distance, as an outlier as shown in figure 1, we can easily identify it by using the conditional rather than the marginal residual. However, the conditional residual does not always outperform the marginal residual, since the computation can suffer from over-fitting because of the small number of subsets used for the conditional residuals. Therefore, we need to consider both conditional and marginal residuals. Haslett et al. [10, 11] showed the relations between marginal and conditional residuals in order to consider both residuals. According to the leave-A-out conditional residual $\tilde{\epsilon}_{(J)}$, we have

$$\mathbf{D}_J^{-1}\tilde{\epsilon}_{(J)} = (\mathbf{V}^{-1}\hat{\epsilon})_J \tag{1}$$

where $\mathbf{D}_J = var(\tilde{\epsilon_{(J)}}) = (\mathbf{Q}_{JJ})^{-1}$. If $J = \{i\}$, $d_i^{-1}\tilde{\epsilon}_{(i)} = (\mathbf{V}^{-1}\hat{\epsilon})_i$ where $d_i = var(\tilde{\epsilon}_{(i)}) = q_{ii}^{-1}$ [11]. From equation (1), the special lack of statistics may be written as $C_J = \sum_i C_{J_i}$, where $C_{J_i} = \tilde{\epsilon}_{(J_i)}^T \mathbf{D}_{J_i}^{-1}\hat{\epsilon}_{J_i}$. This is named a *contribution*. Haslett and Hayes [11] also derived the expectation and variance of the contribution C_J by using an approximated distribution of the contribution. The contribution $C_J = \hat{\epsilon}_J^T \mathbf{D}_J^{-1}\tilde{\epsilon}_{(J)}$ is related to the distribution of the difference of two independent $\chi^2_{k_J}$ random variables for a subset of size k_J and can be modeled by

$$C_J = [(\gamma_J + \phi_J)/2k_J]V_1 - [(\gamma_J - \phi_J)/2k_J]V_2$$

where V_1 and V_2 are independent $\chi^2_{k_J}$. Given this approximation of the distribution, the expectation and variance are written as

$$\begin{aligned} E[C_J] &= \sum_{i\in J} E[C_i] = \sum_i \phi_i = \phi_J \text{ and} \\ var(C_J) &= \frac{1}{k}(\gamma_J^2 + \phi_J^2), \end{aligned}$$

where ϕ_i is the i-th diagonal element of $\hat{\mathbf{V}}\mathbf{Q}$. In addition, $\gamma_J = k_J^{-1} tr\{(\mathbf{G}^{1/2}\mathbf{D}^{-1}\mathbf{G}^{1/2})^{1/2}\}$, where $\mathbf{G} = var(\hat{e}_J)$.

2.2. Simple outlier detection

In general, given the mean and variance of the random variables, we can calculate in what way a particular data is outlying. Assume that we have a univariate/multivariate random variable z with a mean μ and a variable Σ from a certain target distribution. In this case, we use the Mahalanobis distance to transform the multivariate to a nonnegative univariate by $\alpha(z) = \sqrt{(z-\mu)^t\Sigma^{-1}(z-\mu)} \geq 0$, as shown in [12]. With this Mahalanobis distance, we have a truncated normal distribution $p(\alpha(z)|\mu,\Sigma) = \frac{2}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\alpha(z)^2\right\}$ and its cumulative distribution is $p(\alpha(z) \leq f|\mu,\Sigma) = \frac{1}{\sqrt{2}} erf\left(\frac{\alpha(z)}{\sqrt{2}}\right)$ where $erf(\cdot)$ is an error function.

2.3. Bayesian State Space Model in Time series

In this study, the Bayesian sequential estimation framework for the time series state space model for the sequential prediction technique is considered for performing long term prediction. The well-known state space model consists of several parameter variables: k dimensional observations $\mathbf{y}_t$, the hidden states $\mathbf{x}_t$, and a set of the time-invariant control parameter θ. In the Bayesian sequential estimation framework, we can obtain the posterior distribution for the prediction by recursively estimating the following equations: $p(\mathbf{x}_t|\mathbf{y}_{1:t-1},\theta)$ for the prediction step, and $p(\mathbf{x}_t|\mathbf{y}_{1:t},\theta)$ for the filtering step. In addition, the marginal likelihood is often useful for achieving an efficient prediction of the future observation by $p(\mathbf{y}_t|\mathbf{y}_{1:t-1},\theta) = \int p(\mathbf{y}_t|\mathbf{x}_t,\theta)p(\mathbf{x}_t|\mathbf{y}_{1:t-1},\theta)d\theta$.

In this paper, we consider primarily a simple but a well-known linear Gaussian state space model: $\mathbf{y}_t = \mathbf{B}_t\mathbf{x}_t + \epsilon_t$ and $\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \rho_t$ where ϵ_t is an assumed Gaussian noise with $\epsilon_t \sim \mathcal{N}(\mathbf{0}_{k\times 1}, \sigma^2\mathbf{I}_{k\times k})$ and ρ_t is an uncertain noise of the coefficients and $\rho_t \sim \mathcal{N}(\mathbf{0}_{2k\times 1}, g\mathbf{R})$. Here we have $\mathbf{B}_t = [\mathbf{I}_{k\times k}, \mathbf{0}_{k\times k}]$, $\mathbf{A} = \begin{bmatrix} \mathbf{I}_{k\times k}, & \Delta\, t\mathbf{I}_{k\times k} \\ \mathbf{0}_{k\times k}, & \mathbf{I}_{k\times k} \end{bmatrix}$, $\mathbf{R} = \mathbf{G}\mathbf{G}^T$, and $\mathbf{G} = \left[\frac{\Delta\, t^2}{2}\mathbf{I}_{k\times k}, \Delta_t\mathbf{I}_{k\times k}\right]^T$. Given this model and its model parameters $\theta = (\sigma, g)$, we have the following steps for estimation in linear dynamics:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t},\theta) = \mathcal{N}(\mathbf{x}_t; \mu_{t|t}, \Sigma_{t|t})$$

for filtering,

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1},\theta) = \mathcal{N}(\mathbf{x}_t; \mu_{t|t-1}, \Sigma_{t|t-1})$$

for prediction,

$$p(\mathbf{y}_t^*|\mathbf{y}_{1:t},\theta) = \mathcal{N}(\mathbf{y}_t^*; \mathbf{m}_{t|t}, \mathbf{M}_{t|t})$$

for marginal filtering, and

$$p(\mathbf{y}_t^*|\mathbf{y}_{1:t-1},\theta) = \mathcal{N}(\mathbf{y}_t^*; \mathbf{m}_{t|t-1}, \mathbf{M}_{t|t-1})$$

for marginal prediction where $\mathbf{m}_{t|t} = \mathbf{B}_t\mu_{t|t}$ and $\mathbf{M}_{t|t} = \mathbf{B}_t^T\Sigma_{t|t}\mathbf{B}_t + \sigma^2\mathbf{I}_{d\times\, d}$ and $\mathbf{m}_{t|t-1} = \mathbf{B}_t\mu_{t|t-1}$ and $\mathbf{M}_{t|t-1} = \mathbf{B}_t^T\Sigma_{t|t-1}\mathbf{B}_t + \sigma^2\mathbf{I}_{d\times\, d}$.

3. PROPOSED APPROACH

3.1. Outlier detection in time series

Since the linear Gaussian state space model can be reinterpreted as a generalized linear model, the prediction and filtering steps of the state space model correspond to the prediction and estimation of the linear mixed model. Therefore, we can obtain the above three residuals for the outlier detection using marginal prediction and marginal filtering distributions.

- **Marginal residual**: the marginal residual is the difference between observation and marginalized filtering, $\hat{\epsilon}_t = \mathbf{y}_t - \mathbf{m}_{t|t}$, where $E[\hat{\epsilon}_t] = \mathbf{0}$ and $var(\hat{\epsilon}_t) = \mathbf{M}_{t|t}$.
- **Conditional residual**: the conditional residual is the difference between observation and marginalized prediction, $\tilde{\epsilon}_{(t)} = \mathbf{y}_t - \mathbf{m}_{t|t-1}$ where $E[\tilde{\epsilon}_{(t)}] = \mathbf{0}$ and $var(\tilde{\epsilon}_{(t)}) = \mathbf{M}_{t|t-1}$.
- **Contribution**: $C_t = \tilde{\epsilon}_{(t)}^T\mathbf{D}_t^{-1}\hat{\epsilon}_t$ where $E[C_t] = \sum_{i\in\{1,2,\cdots,k\}}\phi_i$ and $var(C_t) = \frac{1}{k}(\gamma_t^2 + \phi_t^2)$

where ϕ_i is the i-th diagonal element of $\hat{\mathbf{V}}\mathbf{Q}$ and $\gamma_t = k_t^{-1}tr\{(\mathbf{G}^{1/2}\mathbf{D}^{-1}\mathbf{G}^{1/2})^{1/2}\}$. In this model, we have a slightly different $\mathbf{Q}$, since we have a hierarchical model in the state space model, while the linear mixed model has a single layer. Recall $\hat{\epsilon}_t = \mathbf{y}_t - \mathbf{m}_{t|t} = \mathbf{y}_t - \mathbf{B}_t\mu_{t|t} = \mathbf{V}\mathbf{Q}\mathbf{y}_t$. Therefore, we have

$$\mathbf{Q} = \mathbf{V}^{-1}(\mathbf{y}_t - \mathbf{B}_t\mu_{t|t})\mathbf{y}_t. \tag{2}$$

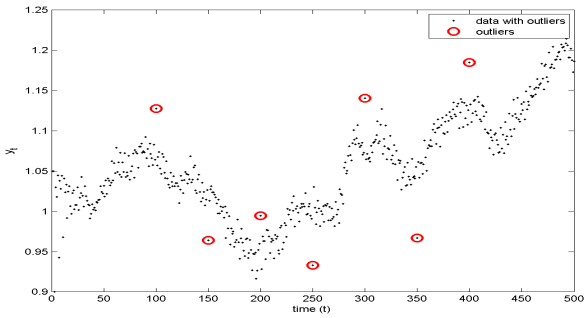
Finally, we have the detailed algorithm 1 for residual-based outlier detection for the state space model.

4. EXPERIMENTAL RESULTS

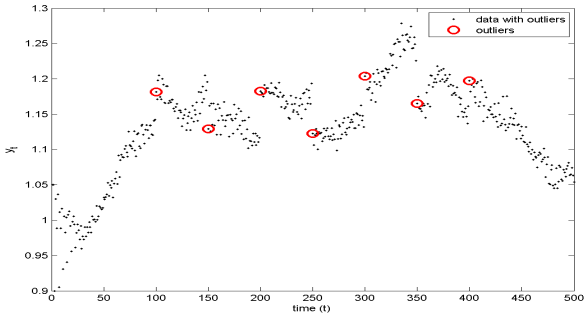
Figure 2(a) shows a synthetic data set with additive outliers (AO) with a fixed $\theta = (\sigma, g) = (10^{-1}, 10^2)$. We generated 1000 random trajectories ($T = 500$) using linear Gaussian space model. Here, the AOs were added with a variance τ varying from 0.1 to 10. Figure 3(a) and (b) show the results obtained from the filtering operation for full trajectories and the region of interest (ROI), where the data points are the outliers beyond the trajectory. Figure 3(c) and (d) show the results of the prediction operation for the full trajectory and ROI. Each result represents the area under curve of the root mean square error (RMSE) of the estimated trajectories when outlier detection techniques were and were not used. Black bars represent the performance of filtering algorithms without outlier detection. Red, green, and blue bars represent

Algorithm 1 $outputs = ProbNotOutlier(inputs)$

Input: $\mathbf{y}_t, \mathbf{m}_{t|t}, \mathbf{M}_{t|t}, \mathbf{m}_{t|t-1}, \mathbf{M}_{t|t-1}, \sigma^2, k$
Output: s_t (the probability that the t-th measurement is not an outlier)
$\hat{\mathbf{V}} = \sigma^2\mathbf{I}$
$\hat{\epsilon}_t = \mathbf{y}_t - \mathbf{m}_{t|t}$ and $\tilde{\epsilon}_{(t)} = \mathbf{y}_t - \mathbf{m}_{t|t-1}$.
$\mathbf{G} = \mathbf{M}_{t|t}$ and $\mathbf{D} = \mathbf{M}_{t|t-1}$.
$\mathbf{Q} = \hat{\mathbf{V}}^{-1}(\mathbf{y}_t - \mathbf{m}_{t|t})\mathbf{y}_t$
$C_t = \tilde{\epsilon}_{(t)}^T\mathbf{D}_t^{-1}\hat{\epsilon}_t$.
$\gamma_t = \frac{1}{k}tr\{(\mathbf{G}^{1/2}\mathbf{D}^{-1}\mathbf{G}^{1/2})^{1/2}\}$
$\phi_t = \sum_{i\in t}\phi_i \leftarrow$ sum of diagonal elements of $\hat{\mathbf{V}}\mathbf{Q}$.
$m^* = E[C_t] = \phi_t$ and $M^* = \frac{1}{k}(\gamma_t^2 + \phi_t^2)$
if Marginal residual is used **then**
  $\alpha(\hat{\epsilon}_t) = \sqrt{\hat{\epsilon}_t^T\mathbf{M}_{t|t}^{-1}\hat{\epsilon}_t}$
else if Conditional residual is used **then**
  $\alpha(\tilde{\epsilon}_{(t)}) = \sqrt{\tilde{\epsilon}_{(t)}^T\mathbf{M}_{t|t-1}^{-1}\tilde{\epsilon}_{(t)}}$
else if Contribution i used **then**
  $\alpha(C_t) = \sqrt{(C_t - m^*)^T M^{*-1}(C_t - m^*)}$
end if
$s_t = 1 - \frac{1}{\sqrt{2}}erf\left(\frac{\alpha(z)}{\sqrt{2}}\right)$ for $z \in \{\hat{\epsilon}_t, \tilde{\epsilon}_{(t)}, C_t\}$



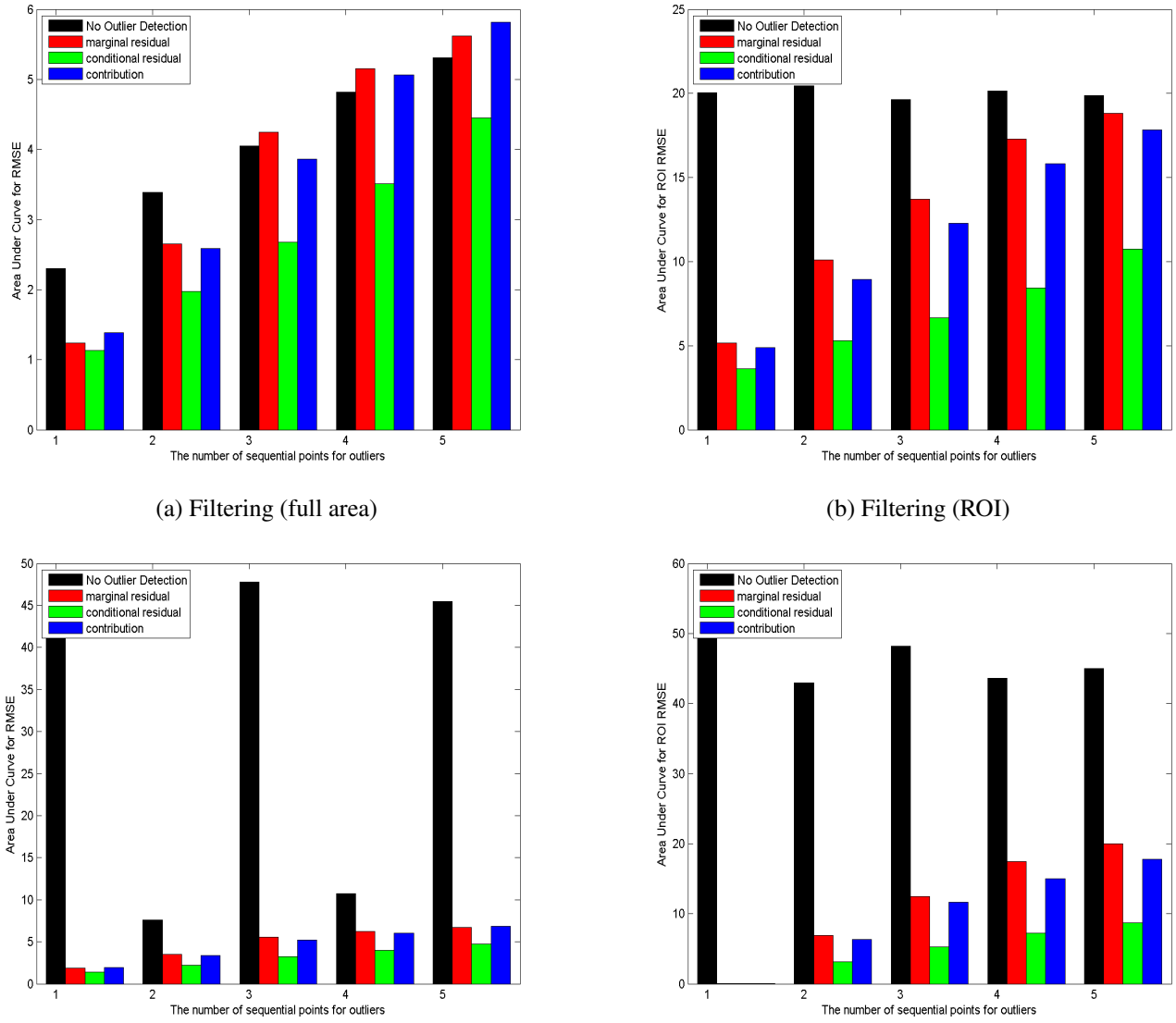
(a) Additive outlier (AO)



(b) Innovative/level shift outlier (IO/LSO)

Fig. 2. Synthetic datasets with outliers

the performance of filtering algorithms with outlier detection based on marginal residual, conditional residual, and



(a) Filtering (full area)

(b) Filtering (ROI)

(c) Prediction (full area)

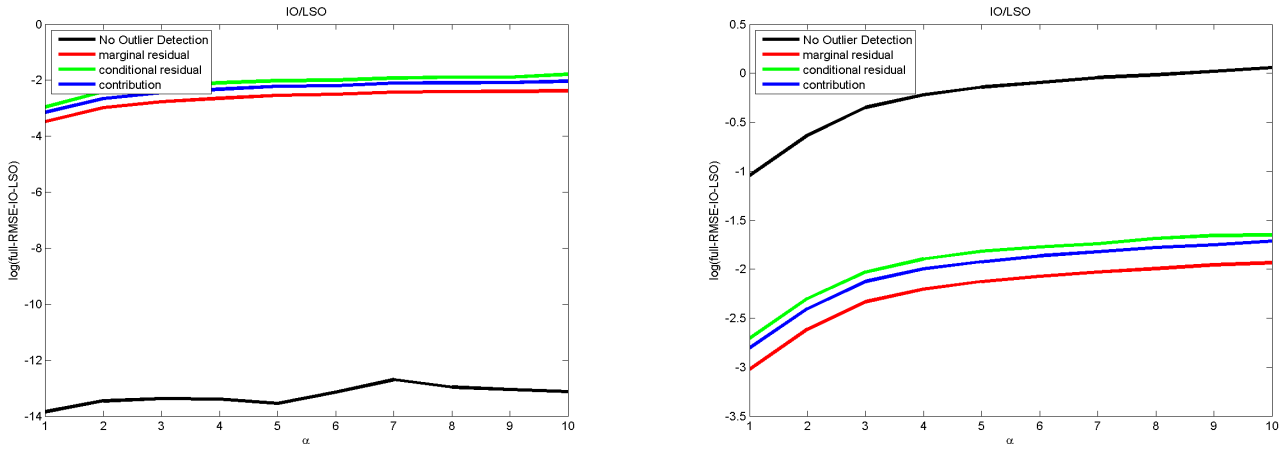
(d) Prediction (ROI)

Fig. 3. The x-axis represents the number of subsequent outliers and the y-axis represents the area under curves of the RMSE for AO: no outlier detection (black), marginal residual-based outlier detection (red), conditional residual-based outlier detection (green) and contribution-based outlier detection (blue)

contribution, respectively. In addition, the x-axis of figure 3 represents the number of subsequent outliers. As can be seen in the figure, the prediction performance of each algorithm is displayed as 'black (no outlier detection)<red (marginal residual)<blue(contribution)<green(conditional residual)'. That these results would be obtained is obvious, since removing such outliers can lead to a more accurate estimate. However, the results are rather different from those produced in the case of innovation outlier and level shift outliers. As shown in figure 4, the contribution-based approach is effective for the prediction operation for both AOs and IO/LSOs while marginal and conditional residual-based approaches are effective only for either AOs or IO/LSOs.

5. CONCLUSION

We proposed a simple Bayesian sequential estimation scheme in which outlier detection techniques based on three different residuals: marginal residual, conditional residual and contribution are applied. We demonstrated that the marginal resid-



(a) Filtering

(b) Prediction

Fig. 4. The x-axis represents the standard deviation of the outliers and the y-axis shows the log of the RMSE with innovation and level shift outliers for full trajectory (IO/LSO)

uals are highly useful for prediction when IOs and LSOs are present, but do not provide an accurate estimation for trajectories with AOs, which can be handled effectively by using conditional residuals. From this point of view, the contribution, which hybridizes both marginal and conditional residuals, is recommended for detecting abnormal signals (outliers/anomalies) in the case of several types of outliers: IO, LSO, and AO.

6. REFERENCES

[1] R. S. Tsay, "Outliers, level shifts, and variance changes in time series," *Journal of Forecasting*, vol. 7, no. 1, pp. 1–20, 1988.

[2] V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004.

[3] R. Baragona and F. Battaglia, "Outliers Detection in Multivariate Time Series by Independent Component Analysis," *Neural Computation*, vol. 19, no. 7, pp. 1962–1984, July 2007.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, July 2009.

[5] Q. Huang and P. P. C. Lee, "Ld-sketch: A distributed sketching design for accurate and scalable anomaly detection in network data streams," in *2014 IEEE Conference on Computer Communications, INFOCOM 2014, Toronto, Canada, April 27 - May 2, 2014*, 2014, pp. 1420–1428.

[6] K. Cohen and Q. Zhao, "Active Hypothesis Testing for Anomaly Detection," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1432–1450, 2015.

[7] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal sequential outlier hypothesis testing," in *48th Asilomar Conference on Signals, Systems and Computers, ACSSC 2014, Pacific Grove, CA, USA, November 2-5, 2014*, 2014, pp. 281–285.

[8] J. Haslett and K. Hayes, "Residuals for the linear model with general covariance structure," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 1, pp. 201–215, 1998.

[9] K. Hayes and J. Haslett, "Simplifying General Least Squares," *The American Statistician*, vol. 53, no. 4, pp. pp. 376–381, 1999.

[10] J. Haslett and D. Dillane, "Application of 'Delete = Replace' to Deletion Diagnostics for Variance Component Estimation in the Linear Mixed Model," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 66, no. 1, pp. 131–143, 2004.

[11] J. Haslett, "A Simple Derivation of Deletion Diagnostic Results for the General Linear Model with Correlated Errors," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 603–609, 1999.

[12] S. J. Roberts, "Novelty Detection using Extreme Value Statistics," in *IEE Proceedings on Vision, Image and Signal Processing*, 1999, pp. 124–129.