Mixture Based Outlier Filtration

Pavla Pecherková[a,b], Ivan Nagy[a,b]
[a] Faculty of Transportation Sciences CTU,
Na Florenci 25, 110 00 Prague 1, nagy@fd.cvut.cz
[b] Institute of Information Theory and Automation AV ČR,
P.O.B. 18, 18208 Prague 8

27th July 2005

Abstract

Success/failure of adaptive control algorithms—especially those designed using Linear Quadratic Gaussian criterion–depends on the quality of the process data used for model identification. One of the most harmful types of process data corruptions are outliers, i.e. 'wrong data' lying far away from the range of real data. The presence of outliers in the data negatively affects estimation of dynamics of the system. This effect is magnified when the outliers are grouped into blocks. In this paper, we propose an algorithm for outlier detection and removal. It is based on modelling of the corrupted data by a two-component probabilistic mixture. The first component of the mixture models uncorrupted process data, while the second one models outliers. When the outlier component is detected to be active, a prediction from the uncorrupted data component is computed and used as reconstruction of the observed data. The resulting reconstruction filter is compared to standard methods on simulated and real data. The filter exhibits excellent properties, especially in the case of block of outliers.

Keywords

Data filtration, system modelling, mixture models, Bayesian estimation, prediction.

This research was partly supported by grants MŠMT ČR 1M6798555601, MDČR 1F43A/003/120.

1 Introduction

Adaptive control systems typically work in feedback regime, hence the quality of control heavily depends on the quality of measurements of the process data. However, the measured data are often corrupted by various disturbances caused by uncertain elements of the process, such as measurement noise, malfunctions of measuring devices, etc. These disturbances can negatively influence performance of the resulting automatic system. Therefore, the task of data pre-processing (filtration) is of great importance in adaptive control, e.g. [1] or [2]. One of the most dangerous corruptions of the measured data is represented by the outliers. The outlier is an incorrect measurement of the process, which is significantly different from the real process data. Two type of outliers are distinguished: i) *isolated outliers*; which are caused by an isolated failure of the measurement; ii) *block of outliers*; which are caused by temporary breakdown of a measuring device. Detection of the former type is relatively easy. However, detection of the latter is more challenging since the block of outliers may have some characteristics of uncorrupted data. In this paper, we propose to model the corrupted data by a probabilistic mixture of dynamic (i.e. autoregressive) models [3, 4, 5]. The model is identified using Bayesian approach [6, 7, 8, 9, 10].

Aim and outline of the solution The *task addressed* in this paper is detection of outliers and reconstruction of the measured data. The proposed *solution of this task* is based on modelling of the corrupted data by a mixture model composed from two components. The first component models the uncorrupted data, the second one models the outliers. Detected outliers are replaced by predictions from the first component.

2 Principle of mixture model estimation

The Quasi-Bayes algorithm for recursive estimation of mixture model parameters was developed recently [11]. The algorithm is designed for estimation of mixture model parameters with components in the form of linear regression models. The mixing weights of the components are considered to be unknown, and their estimates are also provided by the algorithm.

Mixture models The mixture model is described as a conditional probability density

$$f(d_t \mid d(t-1), \theta, \alpha) = \sum_{i=1}^{\mathring{c}} \alpha_i f_i(d_t \mid \varphi_{t-1}, \theta_i) \tag{1}$$

where

$f(\cdot \mid \cdot)$ denotes conditional probability density function (pdf),

d is modelled (and filtered) variable; d_t is actual value at time t, φ_{t-1} is a vector of historical data on witch d_t depends,

$\theta = [\theta_1, \theta_2, \ldots, \theta_{\mathring{c}}]$ are parameters of individual components,

$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_{\mathring{c}}]$ is a vector of components weights,

$\mathring{c}$ is number of components.

The main advantage of this model is the ability to describe a system with finite amount of different states, even if relations between the states are very complex.

Bayes rule for mixture models Direct application of the well known Bayes rule,

$$f(\theta, \alpha \mid d(t)) \propto f(d_t \mid d(t-1), \theta, \alpha) f(\theta, \alpha \mid d(t-1)), \tag{2}$$

to the mixture model (1) yields intractable posterior distribution. Specifically, application of the Bayes rule (2) to the mixture model (1) yields posterior distribution in the form of a mixture with c_t components. Hence, complexity of the posterior grows with time, t, which is prohibitive for on-line processing.

Model approximation To solve the above problem, an approximate Bayesian estimation is used. It is achieved in three steps: **(i)** *introducing* a random variable c_t that indicates the active component at time t, **(ii)** *reformulation* of the model of the active component into a product form:

$$f_{c_t}(d_t \mid \varphi_{t-1}, \theta, \alpha) = \prod_{i=1}^{\mathring{c}} f_i(d_t \mid \varphi_{t-1}, \theta_i)^{\delta(i-c_t)}, \tag{3}$$

and **(iii)** *approximating* the Kronecker delta function $\delta(i - c_t)$ in (3) by its conditional mean value

$$E[\delta(i-c_t)|d(t)] = \sum_{i=1}^{\mathring{c}} \delta(i-c_t) f(c_t|d(t)) = \Pr(c_t = i|d(t)) = w_{i,t}, \tag{4}$$

where $\Pr(\cdot)$ denotes probability. Evaluation of the weight for linear regression models is available, [6] or [12].

Effect of the approximation The mean value (4) is a vector of probabilities, $w_{i,t}$, of individual components. Thus, at each time instant, the statistics of all components are updated by the observed data. Contribution of the observed data to each component is given by the estimated weight (4). For components from exponential family [6], the estimation is equivalent to the weighted least squares technique.

Initiation of the estimation The Bayesian estimation (2) updates the parameter description—represented by conditional pdf $f(\Theta, \alpha|d(t))$ — using the observed data, d_t, for all times $t = 1, 2, \ldots, \mathring{d}$, where $\mathring{d}$ is the number of available data. The recursion starts at $t = 1$ with pdf $f(\Theta, \alpha|d(0))$ which is

called the *prior pdf*. This pdf reflects our prior knowledge about the parameters Θ and α. Moreover, the prior can also be used to ensure that the estimated model have certain advantageous features [13].

Approximate estimation algorithm

The algorithm for approximate estimation of mixture model parameters with exponential family components is outlined in the following scheme:

A. Initial off-line part

- Choose the number of components of the mixture model and their structure.
- Set initial statistics of parameters.

B. On-line time loop

1. Measure the current data.
2. Compute probabilistic weights of all components using (4). The component with maximum weight is called the *active* component.
3. Update parameter statistics for each component.

C. Concluding off-line part

- Compute point estimates of the parameters from their statistics (if they are needed).

3 Mixture-based Outlier Filtration

The process of Bayesian mixture estimation, indicated above, is adapted for outlier detection and reconstruction.

Idea of the filter The main idea is to model the observed data by a probabilistic mixture with two components: 1) *the data component*; which models uncorrupted data, and 2) *the outlier component*; which models the outliers.

Initiation of the filter The initial description of the components is formalized by the prior pdf. The prior variance of the data component is chosen from prior analysis of the filtered data and thanks to forgetting [14], the variance is not allowed to change much. The prior variance of the outlier component is chosen significantly larger that that of the data component. Moreover, it is left relatively free, to be able to "catch" all that does not belong to the uncorrupted signal, i.e. the outliers.

Naturally, to better modelling of uncorrupted data allows better separation of the data from the corruptions. Dynamic models describe the variable in dependence on its historical values while static description is without it. Our experience with data modelling [15] suggest that even those data that are almost static deserve to be described by dynamic models to achieve high quality of the description. Therefore, the data component was chosen as first order regression model. The structure of the outlier component is relatively loose and is chosen as static, i.e. zero order regression model. Its only task is to "cover" all possible errors, mainly outliers.

Operation of the filter As described in the previous paragraph, the estimation of mixture model is based on weighting the data with respect to individual components. Using the estimated weights the active component can be detected. This mechanism is used for the outlier detection as follows:

1. if the dominant weight belongs to the data component no action is taken, and
2. if the dominant weight belongs to the outlier component, the actual data item is considered to be an outlier and the observed value is substituted by a simulated realization of the data component.

The problem occurs if in the following step the current data is not an outlier. Then the dominant weight belongs to the data component and it would be influenced by the old data item, which now is the outlier, through its regression vector. So, this value going to the regression vector of the data component must be substituted by the filtered value, too.

Algorithm of the filtration

The filter can be summarized by the following modification of the above mixture estimation algorithm.

A. Initial off-line part

- Set initial statistics of parameters for two components mixture model,
  - first component with small data covariance (the data component),
  - second component with large data covariance (the outlier component).
- Set forgetting coefficient.

B. On-line time loop

1. Measure the current data.
2. Compute probabilistic weights of both components with respect to the measured data.
3. Choose the component with the larger weight.
4. If the chosen component is that of data, go to 6.
5. If the chosen component is that of outliers,
   - generate data prediction using the data component,
   - use the predicted value as reconstruction of the observed data,
   - use the predicted value in the (future) regression vector of the data component.
6. Re-evaluate parameter statistics for each component separately.

4 Experiments

In this section, we test the proposed algorithm on real data. A sample of data from traffic miniregion in the center of Prague has been chosen for experiments. The data is formed by intensities of traffic flow measured in a single point of the miniregion. The noise, corrupting the data, is represented mainly by standard irregularities of the traffic (interactions between neighbouring control lights, accidental accumulations of cars, small accidents etc.).

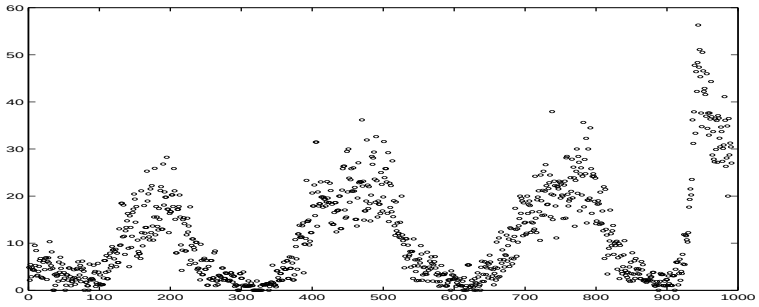

FIGURE 1. Uncorrupted transportation data.

The data sample is composed of 1000 data measured with period of 5 minutes. It involves data for approximately 3,5 days. The maxima of the intensity reflect the traffic load during each day. The noise causes dissimilarities of the courses for individual days. Different daily courses (visible at the beginning of the fourth day) are caused by different type of days, like weekdays and weekends.

The outlier are not frequent disturbances, but they are very important due to their devastating effects on model estimation. They are caused either

by accidental breakdowns of detectors or by their failure for several periods of measurements. Especially the latter ones are very difficult to distinguish automatically from the normal signal.

In order to test the filter, the data were artificially corrupted by various types of outliers. Basically, singular and block of outliers are used in all experiments. Then, various outlier amplitudes are tested—big, medium, small—and their combination in one data sample. For all experiments, results of the proposed filter are compared to those obtained using standard filters. These filters are based on a fixed-length window, moving along the current time, and evaluating some data characteristics for comparison with the current data measurement to detect an outlier. These characteristics are either mean value or median computed over the window. These characteristics are computed either equally for all data or via a kind of forgetting algorithm. A description of such filters can be found e.g. in [16, 17, 18, 19]. A lot of preliminary experiments was performed to compare the suggested mixture filter to the standard ones. All of them gave comparable results for isolated outliers but almost all standard filters were quite unsuitable for filtering of the block ones. Typically, the standard filters failed to detect the block of outliers. Of all those standard competitors, two were selected as the only ones that can be compared to the proposed mixture filter.

The standard filter No. 1 was designed for detecting block of outliers [16]. After detecting the borders of a block outlier, it models the data before and after the outlier with a simple regression model and substitutes the outlying values by a combination of predictions from both these models. The standard filter No. 2 is a median filter with window size 200 (time periods) and without forgetting. For demonstration of filtering results in this paper, only these two of all standard filters are used.

Example 1: All big outliers The fist experiment follows a typical scenario, i.e. outliers with big amplitudes. The level of the outliers is about 5000, which is approximately 100 times the level of the uncorrupted data. The mixture filter completely detects all outliers and leaves normal data without any change. The filtered variable is plotted in figure 2.

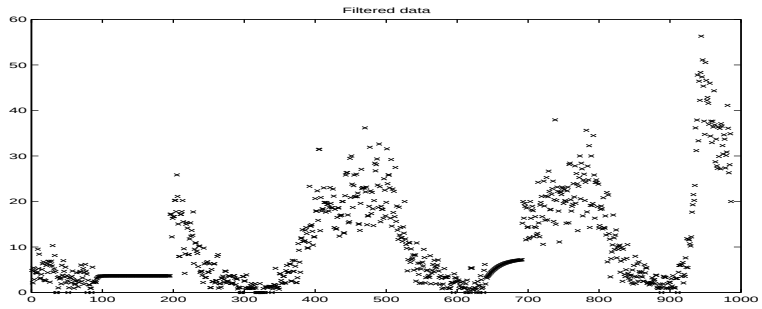

FIGURE 2. Filtered data.

The filtering gives practically identical data (cf. figure 1), up to 20 isolated outliers and two short blocks (first 100-200 items and second 650-700 items) where groups of outliers were located. All substitutions for outliers are in an apropriate range. For evaluation of the results in other than visual way and making use of the fact, that the outliers were introduced artificially, the corrupted data are compared to their predictions from a model estimated on the basis of filtered data sample. This quality evaluation is done through the prediction error (PE) coefficient which is square root of sum of squares of prediction error divided by variance of data. The results for the suggested mixture filter and the two chosen standard filters are in the following table.

TABLE 1: PE coefficients for all big outliers.

filter	*PE coefficient*
The mixture filter	0.49
The standard filter No. 1	0.72
The standard filter No. 2	4.80

REMARK: *The results of PE coefficient for the other standard filters were from 8 to 170. The big difference is caused by the fact, that the standard filters are not able to detect the blocks of outliers.*

Example 2: All small outliers Outlier is a value lying "far" out of the range of the corrupted data. *What happens if "far" is not so far as in the previous experiment?* Now, outliers of an amplitude about 5 times of the uncorrupted data amplitude are tested. The composition of data and outliers is the same. The results are summarized in the following table:

TABLE 2: PE coefficients for all small outliers.

filter	*PE coefficient*
The mixture filter	0.50
The standard filter No. 1	1.52
The standard filter No. 2	1.36

Once again, the mixture filter outperforms the others. The absolute values of differences of PE are smaller that in the previous experiment because the outliers are smaller and the failure to detect them results in smaller contribution to the PE.

Example 3: First big and then small outliers This last case is the most difficult, because the filter could "calibrate" the size of the outliers according to the first suspicious data and miss all that is smaller then its pattern. *Will the filter be able to recognize smaller outliers that follow the bigger ones?* The results are summarized in the following table:

TABLE 3: PE coefficients for first big and subsequent small outliers.

filter	*PE coefficient*
The mixture filter	0.49
The standard filter No. 1	1.52
The standard filter No. 2	1.36

Also in this example, which is the most difficult for the mixture filter, the results are stable and superior to the other filters.

5 Conclusions

A new type of filter for detection of outliers and data reconstruction has been described and demonstrated on a serial of examples. The filter is based on modelling of the data by a mixture model with two components: one for modelling of the uncorrupted data, and the second one for modelling of the outliers. The main advantage of the filter is its adility to detect groups of outliers which was demonstrated in simulation. This type of outliers arise from temporary breakdowns of measuring devices which are rather frequent in transportation system. In all simulated examples, both single and block of outliers were correctly detected and the data were reconstructed by reasonable values, generated by the uncorrupted data component. The comparison of the results with standard filters proved that reconstruction of block of outliers is a difficult task. Typically, standard filters were not be able to substitute the whole block of outliers. The best standard filters usually missed several outliers from the block before they "realized" that an outlier occur. The proposed mixture-based filter detects the outliers more correctly, and thus outperforms the standard filters in all simulated experiments.

References

[1] F. Zhao and TY Leong, "A data preprocessing framework for supporting probability-learning in dynamics decison modeling in medicine", *J AM MED INFORM ASSN*, vol. suppl. S2000, pp. 933–937, 2000.

[2] S. Dzerovski D. Gamberger, N. Lavrac, "Noise detection and elimination in data processing. experiments in medical domains", *APPL ARTIF INTELL*, vol. 14, no. 2, pp. 205–223, 2000.

[3] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixtures*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1985, ISBN 0 471 90763 4.

[4] S. Richardson and P.J. Green, "On bayesian analysis of mixtures with an unknown number of components, with discussion", *Journal of the Royal Statistical Society, Series B*, vol. 59, no. 4, pp. 731–792, 1997.

[5] G. J. McLachlan, *Finite Mixture Models*, Wiley, New York, 1999.

[6] M. Kárný, I. Nagy, and J. Novovičová, "Mixed-data multi-modelling for fault detection and isolation", *Adaptive control and signal processing*, , no. 1, pp. 61–83, 2002.

[7] M. Kárný, "Probabilistic support of operators", *ERCIM News*, , no. 40, pp. 25–26, 2000.

[8] P. Ettler, M. Kárný, and I. Nagy, "Employing information hidden in industrial process data", in *Preprints of Symposium Intelligent Systems for Industry*, Paisley, UK, 2001, pp. 1814–1817, Academic Press.

[9] M. Kárný, P. Nedoma, I. Nagy, and M. Valečková, "Initial description of multi-modal dynamic models", in *Artificial Neural Nets and Genetic Algorithms. Proceedings*, V. Kurková, R. Neruda, M. Kárný, and N. C. Steele, Eds., Wien, April 2001, pp. 398–401, Springer.

[10] I. Nagy, P. Nedoma, and M. Kárný, "Factorized EM algorithm for mixture estimation", in *Artificial Neural Nets and Genetic Algorithm. Proceedings*, V. Kurková, R. Netruda, M. Kárný, and N. C. Steele, Eds., Wien, April 2001, pp. 402–405, Springer.

[11] M. Kárný, J. Kadlec, and E. L. Sutanto, "Quasi-Bayes estimation applied to normal mixture", in *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*, J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, Eds., Praha, September 1998, pp. 77–82, ÚTIA AV ČR.

[12] I. Nagy, M. Kárný, P. Nedoma, and Š. Voráčová, "Bayesian estimation of traffic lane state", *International Journal of Adaptive Control and Signal Processing*, vol. 17, no. 1, pp. 51–65, 2003.

[13] M. Kárný, N. Khailova, J. Böhm, and P. Nedoma, "Quantification of prior information revised", *International Journal of Adaptive Control and Signal Processing*, vol. 15, no. 1, pp. 65–84, 2001.

[14] R. Kulhavý and M. B. Zarrop, "On general concept of forgetting", *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.

[15] I. Nagy, "Estimation of real data with dynamic mixtures", Tech. Rep., research report No. 2066, ÚTIA AV ÈR, Prague, 2002.

[16] L. Tesař and A. Quinn, "Detection and removal of outliers from multidimensional AR processes", in *Proceedings of Irish Signal and Systems Conference*, Maynooth, Ireland, August 2001.

[17] Sung-Jea Ko and Yong Hoon Lee, "Center weighted median filters and their applocations to image enhancement", *IEEE Transactions on Circuits and Systems*, vol. 38, no. 9, pp. 984–993, September 1991.

[18] L. Tesař and A. Quinn, "Method for artefact detection and supressing using alpha-stable distributions", in *Proceedings of ICANNGA Conference*, Prague, Czech Republic, March 2001.

[19] T. Cipra, "Dynamic credibility with outliers", *Applications of Mathematics*, vol. 41, no. 2, pp. 149–159, 1996.