Mixture Based Outlier Filtration

Pavla Němcová[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 CR,
P.O.B. 18, 18208 Prague 8

April 22, 2003

Abstract

Success/unsuccess of adaptive control algorithms, especially those based on Linear Quadratic Gaussian design, depends on the quality of process data used. One of the most harmful types of process data corruptions, are outliers, i.e. "wrong data" lying far from the range of real data and bringing totally wrong information about the process dynamics. These data, when grouped into blocks, can completely destroy estimation and consequentially the whole adaptive control. This paper proposes an algorithm for outlier detection and filtration. It is based on modelling of corrupted data by two-component mixture. The first component models pure process data, while the second one models outliers. Whenever during the filtration, the outlier component is declared as active, a prediction from the pure data component is computed and generated as filtered data item. Comparison of the suggested filter with other methods, tested on artificial and real data, illustrates the power of the proposed algorithm. It exhibits excellent properties, especially in the case of grouped outliers.

Keywords

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

This research was partly supported by grants GAČR 102/03/0049, AV ČR S1075102, VZ MŠMT 210000023 and COST Action 276.

1 Introduction

Automation is an inevitable tool when dealing with complex systems. Adaptive control systems are mostly work in feedback and the control quality heavily depends on a quality of the measured process data. The used process data are corrupted by various disturbances caused by uncertain elements of the process, measurement noise, malfunctions of measuring devices etc. These signal corruptions often completely devalue the performance of the resulting automatic system. This is why problems of data pre-filtration are of great importance e.g. [1] or [2] and need special attention. One of the most dangerous corruptions of measured data are represented by outliers. They are wrong data, bringing zero information about the process, with values far from the range of real process data. Two type of outliers are distinguished: i) caused by entirely wrong measurements which result in singular outliers; ii) caused by by total breakdown of a measuring device. Unlike easily detected single outliers, the second type is more difficult because it produces grouped outliers, which can be easily mistaken for big, but normal data.

The approach proposed in the paper is based on Bayesian identification [3] of mixture models [4, 5, 6], specifically on its approximate version capable of estimating dynamical mixtures [7, 8, 9, 10, 11].

Aim and outline of the solution

The *task solved* in this paper is detection of outliers in measured data and their correction. The *solution of this task* is based on modelling of filtered data with a mixture model consisting of two components. One of them models pure data, the second one describes outliers. The detected outliers are substituted by predictions from the pure data component.

2 Principle of mixture model estimation

The Bayesian approach to recursive mixture model estimation has been developed and introduced like quasi-Bayes algorithm recently [12]. It works with a model expressed as a mixture of weighted linear components, described as a set of linear regression models. The weights of components are stationary probabilities of the components.

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,

$\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_{\mathring{c}}]$ are probabilities of components weight,

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

The main advantage of such model is that it is able to describe a system with a 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 mixture model (1) leads to unfeasible computations. The reason is that repetitive utilization of the Bayesian update (2), which is represented by a sum, produces products of sums. Thus, the structure of evaluated statistics blows up with the increasing length of data sample.

Model approximation

To solve the above problem, an approximation of the mixture model is used. It consists in three steps: **(i)** *introducing* a random variable c_t that indicate the active component at the time instant t, **(ii)** formal *rewriting* the model of the active component into a product form

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

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. This computation is realizable with the models at disposal. For more details see [7] or [13].

Effect of the approximation

The mean value computed in (4) is a vector of probabilities $w_{i,t}$ of individual components. Thus, at each time instant, instead of looking for a single valued point estimate of the true component c_t, all components are taken into account. The statistics of all components are updated with the actual data item weighted by the corresponding probability weight. For components from exponential class [7], the estimation leads to the weighted least squares technique.

Initiation of the estimation
The relation (2) describes the process of Bayesian estimation. It incorporates the information carried by data into the parameter description representing by conditional pdf $f(\Theta, \alpha|d(t))$ for time instants $t = 1, 2, \ldots, \mathring{d}$, where $\mathring{d}$ is number of measured data items. The recursion starts in $t = 1$ with pdf $f(\Theta, \alpha|d(0))$ which is called prior pdf. This pdf reflects our prior knowledge about the parameters Θ and α. On the other way, it can also be used to force the estimated model some features we want to be preserved during the process of estimation. For more information about this, see [14].

3 Algorithm of approximated estimation

The algorithm of estimation for exponential class components of the mixture model can be sketched in the following scheme:

A. Initial off-line part

- Choose number and form of mixture components.
- Set initial statistics of parameter estimation.
- For data component, get prior data for its regression vector.

B. On-line time loop

1. Measure current data item.
2. Compute probabilistic weights of all components with respect to the measured data item; *the label of component with maximum weight can be considered a current point estimate of the label of active state.*
3. Update parameter statistics for each component separately, using data weighted the corresponding probability weights.

C. Concluding off-line part

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

4 Principle of the filtration

The process of Bayesian mixture estimation, indicated above, is used for outlier filtration.

Idea of the filter
The main idea is to use a mixture with two components as the filter. One

of them (the data component) models pure data and the other (the outlier component) describes outliers.

Initiation of the filter
The initial definition of the components can be done through an initiation of the mixture. The data component is pre-set with relatively small data variance derived from prior analysis of the filtered data and it is not allowed to change much. The outlier component is pre-set with rather large data variance and it is left relatively free, to be able to "catch" all that does not belong to the pure signal – mainly the outliers.

Naturally, to distinguish the useful data from the rest of erroneous signals, it is necessary to model them well. Dynamic models describe the variable in dependence on its historical values while static description is without it. According to our experience with data mixture modelling [15], even those data that are almost static, deserve to be described by dynamic models, to achieve high quality of the description. From this fact it follows, that the data component should be dynamic – the first order regression model seems to be fully sufficient. The structure of the outlier component is relatively loose and is chosen as static. 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. From this, the point estimate of the active component can be constructed. Thus, it is possible to recognize wether the actual data item is an outlier or not. If the dominant weight belongs to the outlier component, the actual data item is an outlier and at the output of the filter it is substituted by a value generated like a simulated realization of the data component. At this moment, the weight of the data component is small, so the outlier influences it in a negligible way. The problem occurs if in the following step the data item 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.

5 Algorithm of the filtration

The work of the filter can be summarized in the following algorithm which is just a modification of the above algorithm for mixture estimation.

A. Initial off-line part

- Set initial statistics of two component mixture,

- first component with small data covariance (data component),
- second component with large data covariance (outlier component).

• For data component, get prior data for its regression vector.

B. On-line time loop

1. Measure current data item.
2. Compute probabilistic weights of both components with respect to the measured data item.
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 item from the data component,
   - use the generated value as the filtered data item,
   - replace the generated value into the regression vector of the data component.
6. Recompute parameter statistics for each component separately, using data weighted by the corresponding probability weight.

C. Concluding off-line part

• The data sample is filtered.

6 Experiments

This section describes testing of the proposed algorithm of filtering on real data. A sample of data from traffic miniregion in the center of Prague has been chosen for experiments. They are 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.).

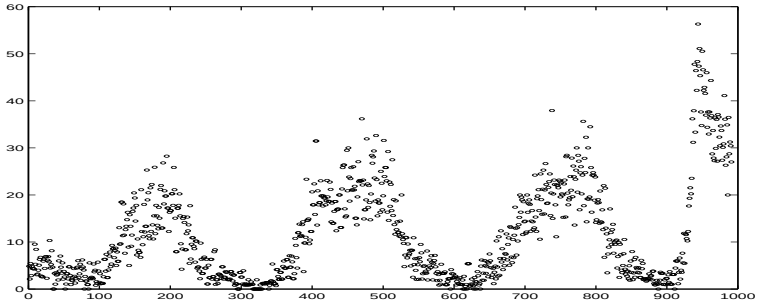

FIGURE 1. Pure transportation data.

The data sample is 1000 items long and it contains data measured each 5 minutes. It involves data for approximately 3,5 days, which can be clearly seen form the periodicity of the signal. The maxima of the intensity reflect the traffic load during a day. The noise mentioned 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.

Not so frequent, but very important disturbances, due to their devastating effect, are outliers. 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.

To test the filter ability, the data without outliers, artificially corrupted by various types of outliers, used. Basically, singular and block outliers are used in all experiments. Then, various outlier amplitudes are used – big, medium, small – and their combination in one data sample. Results of all examples are compared to those obtained with standard filters. These filters are based on a window, moving along the current time, and giving some data characteristics for comparison with the newly measured data to decide whether it is an outlier or not. These characteristics mostly are either mean value or median computed over the window. The characteristics are computed either equally for all data or a kind of forgetting is applied. A description of such filters can be found e.g. in [16, 17, 18, 19, 20, 21]. A lot of preliminary experiments were performed to compare the suggested mixture filter to the standard ones. All of them gave comparable results for singular outliers but almost all standard filters were quite unsuitable for filtering of the block ones. They mostly consider them normal data and copied them. 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 is a special one for detecting block outliers. 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 was chosen as a standard one. It uses outliers with big amplitudes. The level of the outliers is about 5000, which is approximately 100 times the level of the pure data. The filter completely substitutes outliers and leaves normal data without any change. The filtered variable is plotted in figure 2.

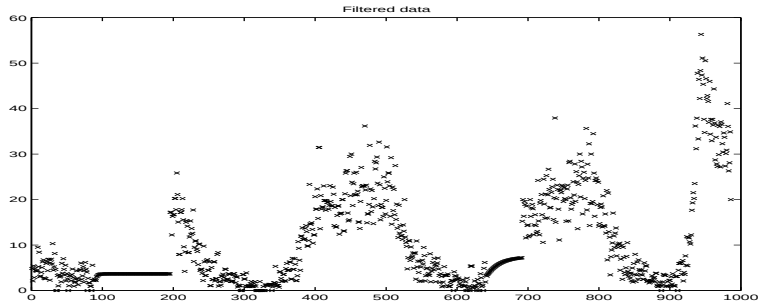

FIGURE 2. Filtered data.

The filtering gives practically identical data (cf. figure 1), up to 20 singular 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 a proper range. For evaluation of the results in other than visual way and making use of the fact, that the outliers were introduced artificially, the pure 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 recognize the blocks of outliers and they copy them.*

Example 2: All small outliers

Outlier is a value lying "far" out of the range of the pure data. *What happens if "far" is not so far as in the previous experiment?* Now, outliers of an amplitude about 5 times of the pure data amplitude are tested. The composition of data and outliers is the same. The results are

TABLE 2: PE coefficients for all small outliers.

filter	*PE coefficient*
The proposed filter	0.50
The best standard filter	1.52
The second best standard filter	1.36

The proposed filter wins again. The differences are not so big because even if the filters dos not remove the whole block outlier and consequently the outliers are predicted, the prediction error is not so big.

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 again in the table.

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

filter	*PE coefficient*
The proposed filter	0.49
The best standard filter	1.52
The second best standard filter	1.36

Also in this example, which is most difficult for suggested filter, the results are stable and best.

7 Conclusions

A new type of filter for detection and adaptation of outliers has been described and demonstrated on a serial of examples. The filter is based on modelling of the filtered signal by a mixture model with two components – one for pure data and the second for outliers. The experiments prove that the results of filtration are very good, even in the case of block outliers. This type of outliers arise from temporary breakdowns of measuring devices which are rather frequent in transportation. The results of performed experiments exhibit high quality of filtration. In all cases demonstrated, both single and block outliers were completely detected and substituted by reasonable values, generated from the data pure component. The comparison of the results with classical filters proved that block outliers are difficult for filtration. Mostly, those filters were not be able to substitute the whole block of outliers. The best classical filters usually copied several outlier values from the block before they "realized" that it is an outlier. And this is the reason why the suggested mixture based filter was better in all performed 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] V. Peterka, "Bayesian approach to system identification", in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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. Kůrková, R. Neruda, M. Kárný, and N. C. Steele, Eds., Wien, April 2001, pp. 398–401, Springer.

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

[12] 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.

[13] 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.

[14] 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.

[15] I. Nagy, "Estimation of real data with dynamic mixtures", Tech. Rep., research report No. 2066, TIA AV R, 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.

[20] M. Tanaka and T. Katayama, "A robust identification of a linear system with missing observations and outliers by the EM algorithm", *Transactions of the Institute of Systems, Control and Information Engineering*, vol. 1, no. 4, pp. 117–129, September 1988.

[21] S.J. Godsill, *The Restoration of Degraded Audio Signals*, PhD thesis, University of Cambridge, Department of Engineering, December 1993.