IDENTIFICATION AND PREDICTION OF MULTIPLE SHORT RECORDS BY DYNAMIC BAYESIAN MIXTURES

Pavel Ettler
COMPUREG Plzeň, s.r.o.
P.O.Box 334, 306 34 Plzeň, Czech Republic

Miroslav Kárný
Institute of Information Theory and Automation
P.O.Box 18, 182 08 Praha 8, Czech Republic

Keywords: System identification, prediction, probabilistic mixtures, decision support.

Abstract: A short data record is not suitable for proper identification of system model which is necessary for reliable data prediction. The idea consists in utilization of multiple similar short data records for identification of a dynamic Bayesian mixture. The mixture is used for prediction according to one of three methods described. Simulated and real data examples illustrate the methods.

1 INTRODUCTION

Applicable model-based control or decision support rely on system model identified from available data. Problem arises when data records are too short for standard identification procedure, especially for higher order models. Examples of such situation can be met in many areas including such diverse disciplines like analyzing treatment data in medicine or pass scheduling for metal rolling. Typically, there exists a set of data records each consisting of several samples mostly for several data channels. Such similar multiple records can be grouped according to a specific rule and processed in a way described in the following.

Section 2 sketches basics about Bayesian mixtures used for identification and prediction. The idea is demonstrated on a simple deterministic case in Section 3. Examples for noisy and multi-dimensional data are displayed in Section 4 while Section 5 concerns real data. Conclusions 6 summarize results and outline the future work.

2 EMPLOYING MIXTURES

External behavior of dynamic stochastic systems is the most generally described by a probability density function (pdf, denoted by f) relating the current system output y_t to the current system input u_t and past observed history of data $d(t-1) = (d_1, \ldots, d_{t-1})$, $d_t = (y_t, u_t)$. Such a model is rarely available directly. Instead, its version $f(y_t|u_t, d(t-1), \Theta)$ parameterized by unknown parameter Θ is assumed. Among various parameterized models, the prominent role is assigned to *finite, normal probabilistic mixtures*

$$f(y_t|u_t, d(t-1), \Theta) = \sum_{c=1}^{n_c} \alpha_c \mathcal{N}_{y_t}(\theta_c \psi_{c;t}, r_c), \quad (1)$$

where $\mathcal{N}_y(\hat{y}, r)$ is normal pdf given by the mean $\hat{y}$ and covariance matrix r; θ_c are regression coefficients of c-th normal pdf, called *component*; $\psi_{c;t}$ is regression vector formed in a known way from $u_t; d(t-1)$ and $\alpha_c \geq 0$ is *component weight* such that $\sum_{c=1}^{n_c} \alpha_c = 1$. The unknown parameter Θ is represented by the collection r_c, θ_c, α_c. The prominence of mixtures comes from their (asymptotic) universal approximation property: loosely speaking they are able to model any stochastic dynamic system (Haykin, 1994; Kárný et al., 2005).

Identification of mixtures is hard but relatively well elaborated task (Titterington et al., 1985; Kárný et al., 2005) and for asymptotically valid version with constant component weights can be taken as practically solved task. For instance, the extensive Matlab toolbox Mixtools (Nedoma et al., 2002) contains the relevant implementations of such an identification, which estimates also number of components and structures of respective regression vectors. The projection-based methodology, proposed in (Andrýsek, 2004) seems to be the best procedure available.

Ettler P. and Kárný M. (2005).
IDENTIFICATION AND PREDICTION OF MULTIPLE SHORT RECORDS BY DYNAMIC BAYESIAN MIXTURES.
In *Proceedings of the Second International Conference on Informatics in Control, Automation and Robotics - Signal Processing, Systems Modeling and Control*, pages 66-70
DOI: 10.5220/0001175400660070

The system model $f(y_t|u_t, d(t-1))$ obtained after "excluding" unknown parameters via identification is essentially predictor of the output y_t. Its performance depends weakly on overestimation of the structure of respective regressors but it is significantly influenced by the assumption that the component weights are time invariant. The assumption allows independent jumps between active (the best describing) components irrespectively of $u_t, d(t-1)$. This condition is met in some applications but in the considered technical ones is unrealistic: usually, the system is described just by a subset (often with a single term) of components for some period of time. Under this situation, the output prediction based on the whole mixture is poor. This problem can be overcome by detecting and utilizing the active components for time periods in question.

3 BASIC IDEA

Particular short data record, which is to be processed contains too few samples for valuable identification. The basic idea consists in rearrangement of data and their identification by a dynamic Bayesian mixture. The mixture or its selected components are then used for prediction.

3.1 Rearrangement of data

To illustrate the method, let us simulate a simple example of $n_r = 10$ one-dimensional data records each consisting of $n_d = 5$ samples generated by the model

$$y_k = a_1 y_{k-1}, \tag{2}$$

where $a_1 = 0.6$, $y_1 = 75$ and $k = 2, \ldots, 5$ is the discrete time index.

Data can be depicted by a mesh plot shown on the upper graph of Fig. 1. For the sake of identification particular records can be merged into a single vector with $l = n_r \cdot n_d$ items as shown on the lower graph of the same figure. Then, the overall sample index is $t = 1, \ldots, l$.

3.2 Identification – deterministic case

Mixtools package (Nedoma et al., 2002) was used for mixture identification. For the given deterministic example, the result came up to expectation exactly. The mixture is composed of two components, one corresponding to the model dynamics (2) and another modelling transitions among records.

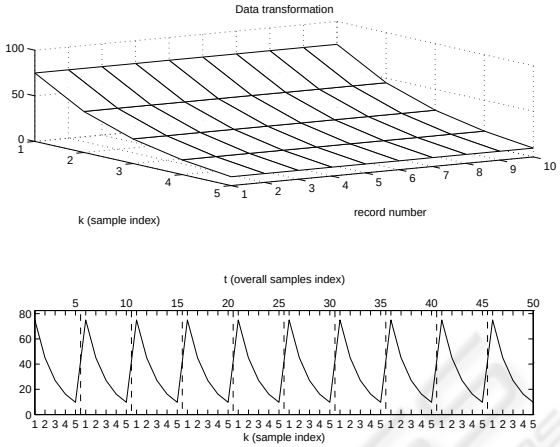

Figure 1: Data rearrangement. Short data records generated by a simple model shown on the upper mesh plot are merged into the single vector shown on the lower graph.

3.3 Prediction and evaluation criterion

The mixture was identified in order to get valuable prediction. One-step-ahead prediction is considered for the sake of simplicity. For an m-order model the prediction is accomplished for $(n_d - m)$ time instants for a single data record ($y_{c;k}$ means prediction by c-th component):

$$y_{p;k} = \sum_{c=1}^{n_c} \alpha_c y_{c;k}, \quad k = m+1, \ldots, n_d \tag{3}$$

Predictions are treated as merged original data forming a vector $y_p(t),\ t = 1, \ldots, l$. To evaluate prediction quality the following modified quadratic criterion E_s is used (subscript s stands for *selected* instants of t for which predictions are evaluated):

$$E_s = \frac{1}{l} \sum_{t=1}^{l} (y_t - y_{p;t})^2. \tag{4}$$

Fig. 2 shows the original data and predictions for 3 randomly chosen succeeding records. The whole mixture, ie. both components for this case were used for prediction on the upper graph. It is obvious that the prediction is poor ($E_s = 52.9$). The lower graph shows predictions calculated from the selected component. For this deterministic case the component matches the model (2) exactly and thus the prediction is perfect ($E_s = 0$).

3.4 Component selection

Criterion for selection of components to be used for prediction is crucial for the mentioned principle.

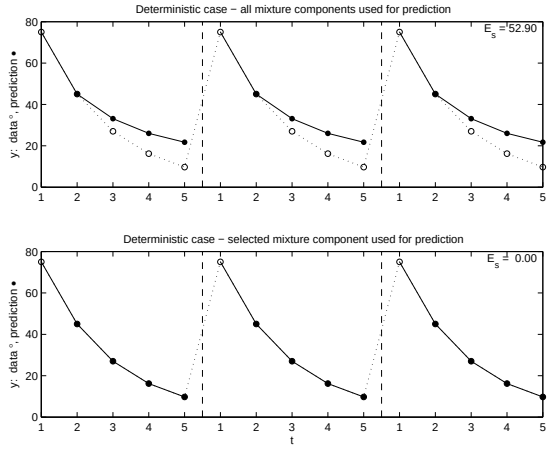

Figure 2: Deterministic case: whole mixture vs. selected component. Three data records put together and plotted by the dotted line (···o···). Prediction plotted by the solid line (—•—) was omitted for starting points of records for the 1st order model. For the upper graph whole mixture was used for prediction while the selected component was used for prediction shown on the lower graph.

Method A The simplest possibility is to engage just the component the weight α_c of which was identified as the maximal one. It means practically to set the weight to one for that component and to zero for the others and than to use the mixture for prediction.

Method B Another possibility consists in evaluation of prediction error criterion for all possible combinations of components to be involved. For most realistic cases the method results in utilizing several components instead of one. This can simply reflect uncertainty of measurement or indicate that data records should be split into two or more groups to allow approximation by a simpler mixture.

3.5 Extending data record

Method C A rather different approach consists in extending the merged data by an additional channel x_t, which indicates transitions among particular records:

$$x_t = \begin{cases} 1 & \text{for the last sample in the record} \\ 0 & \text{otherwise} \end{cases}$$

The situation is illustrated by two upper plots on Fig. 3.

In this case components preserve their identified weights α_c, ie. the whole mixture is used for prediction. Zero elements of $x(t)$, which is now included in the regression vectors $\psi_{c;t}$ (1) eliminate influence of "transient" components on computation of $y_p(t)$. On the other side, the ones in $x(t)$ enable to predict transitions accordingly. The lower plot on Fig. 3 shows the result for the deterministic case. The prediction is not exact ($E_s = 1.52$) but the problem of component selection was avoided.

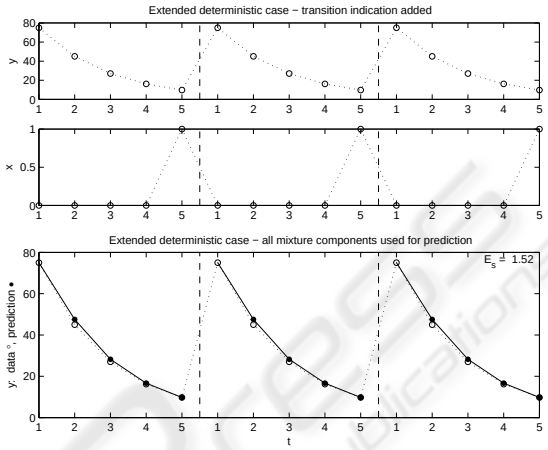

Figure 3: Extended deterministic case: additional data channel indicates transition among records. Two data (···o···) channels plotted on upper graphs. Whole mixture was used for prediction (—•—) shown on the lower graph.

4 SIMULATED EXAMPLES

4.1 Adding noise

Increased model order and introduction of noise make the simulation more realistic:

$$y_k = a_1 y_{k-1} + a_2 y_{k-2} + c_N e_{Nk}, \qquad (5)$$

where $a_1 = 0.6$, $a_2 = 0.1$, $c_N = 4$ are parameters of the model,
$k = 2, \ldots, 5$ and e_N is the output of a random number generator with normal distribution $\mathcal{N}(0, 1)$. Starting points of records are given by

$$y_1 = y_0 + c_U e_U, \qquad (6)$$

where $y_0 = 50$, $c_U = 50$ and e_U is the output of a random number generator with uniform distribution in the interval $< 0, 1 >$.

The identification was accomplished firstly on original merged data $d(t)$ where $d_t = y_t$ and on the extended data where $d_t = (y_t, x_t)$ afterwards to allow comparison of predictions shown on Fig. 4. For the upper plot, the original data were used for identification and the single most important component was used for prediction (method A). The middle plot uses the same data and multiple components (2 of 8) selected according to the method B. For the lower plot,

the extended data were utilized and the whole mixture used for prediction (method C). Values of E_s are very similar for all three methods for this case.

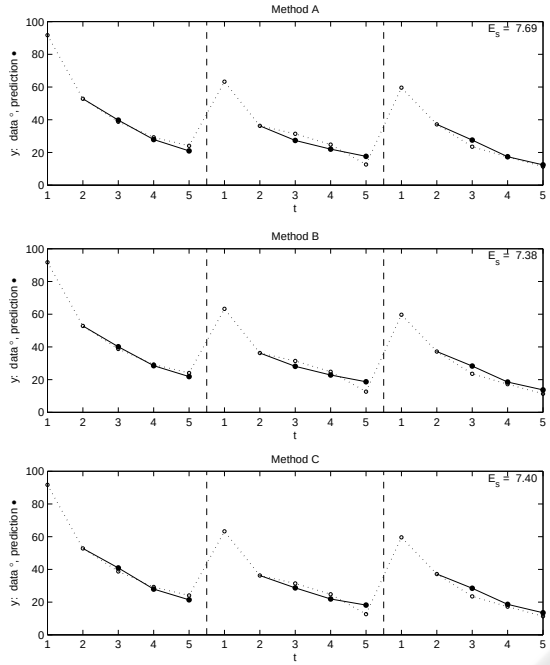

Figure 4: Comparison of prediction methods for noisy data.

4.2 Multiple dimensions

Multiple dimensions and increased uncertainty make the identification more difficult. Let us consider the model

$$Y_k = AY_{k-1} + c_N e_{N;k}, \qquad (7)$$

where $Y_k = \begin{bmatrix} y_{1;k} \\ y_{2;k} \end{bmatrix}$, $A = \begin{bmatrix} 0.6 & 0.1 \\ 0.2 & -0.8 \end{bmatrix}$ and $c_N = 4$.

Merged data to be identified consist of a $2 \times l$ matrix for methods A and B and $3 \times l$ matrix for method C. Fig. 5 compares the three methods of prediction. It can be seen that method B is becoming favourable for increasing uncertainty being involved.

5 REAL DATA EXAMPLE

A subset of records from a hot reversing rolling mill was selected for a real data example. The mill processes metal bars or slabs in several passes to produce thick strips. Thickness is not measured automatically on the given mill, which makes pass scheduling a non-trivial task.

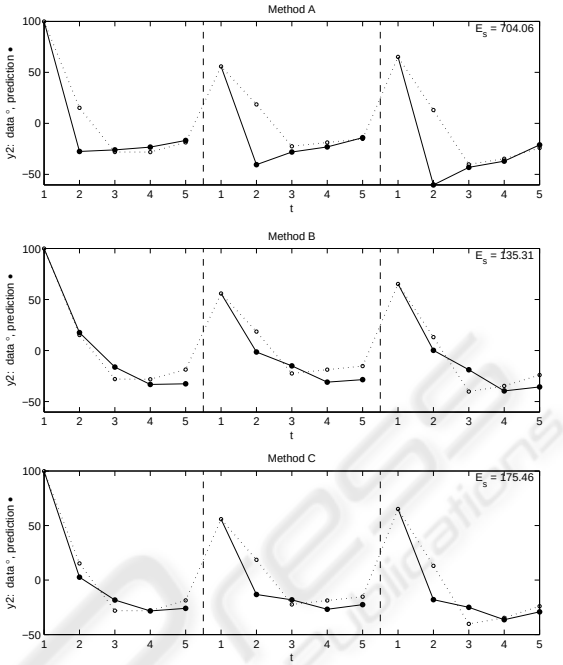

Figure 5: Comparison of prediction methods for multidimensional noisy data for channel y_2.

Three data channels – working roll position, rolling force and electric current of the roll drive were selected for the example. Characteristic values (means of selected parts of passes) were evaluated for 5 succeeding passes to produce 3×5 data matrix for a single record. Available records for a specific material were merged (Fig. 6) and used for identification.

Fig. 7 shows predictions for three real data channels. The mixture was identified for the second-order model. Method B turned out to be far most successful for this case (3 of 7 components were utilized). This confirmed the trend indicated by previous examples. Direct comparison of values of the criterion (4) would be misleading for this case because of dissimilar ranges and units used for particular data channels. Therefore Fig. 8 shows histograms of prediction errors recalculated to the percentage of range of the corresponding data channel.

6 CONCLUSIONS

Utilization of Bayesian dynamic mixtures for identification and subsequent prediction of multiple short data records was described. The principle was shown on a simple deterministic case. Two methods of prediction differing in number of mixture components to be utilized and the third method relying on extension of data were introduced and demonstrated on noisy

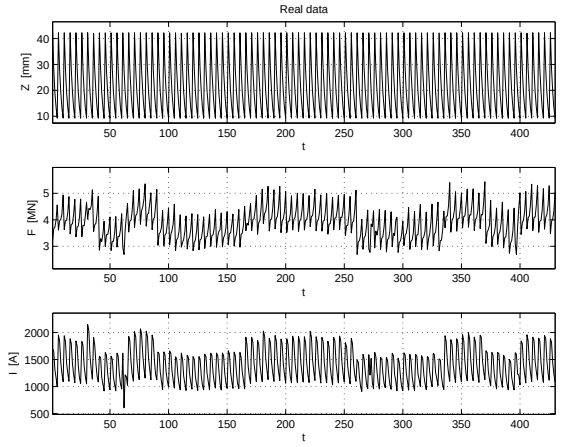

Figure 6: Real data from a hot reversing rolling mill. 86 records of characteristic values for final 5 passes were merged. Three data channels were selected: Z - position of the working roll, F - rolling force and I - electric current of the main mill drive.

and multidimensional data respectively. A simplified set of data records from a hot reversing rolling mill was used for a real data example.

Experiments showed that the mixture used for prediction should be composed by more than one component (method B). The algorithm for components selection will be more elaborated.

Further research will be focussed on utilization of the idea for real multidimensional short data records. Results should help to advance applications of the Bayesian decision support.

ACKNOWLEDGEMENTS

The work was accomplished within the research centre DAR, supported by the grant 1M6798555601 of the Czech Ministry of Education.

REFERENCES

Andrýsek, J. (2004). Approximate recursive Bayesian estimation of dynamic probabilistic mixtures. In Andrýsek, J., Kárný, M., and Kracík, J., editors, *Multiple Participant Decision Making*, pages 39–54. Advanced Knowledge International, Magill, Adelaide.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Macmillan College Publishing Company, New York.

Kárný, M., Böhm, J., Guy, T., Jirsa, L., Nagy, I., Nedoma, P., and Tesař, L. (2005). *Optimized Bayesian Dynamic Advising: Theory and Algorithms*. Springer, London. to appear.

Nedoma, P., Böhm, J., Guy, T. V., Jirsa, L., Kárný, M., Nagy, I., Tesař, L., and Andrýsek, J. (2002). Mixtools: User's Guide. Technical Report 2060, ÚTIA AV ČR, Praha.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixtures*. John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore. ISBN 0 471 90763 4.

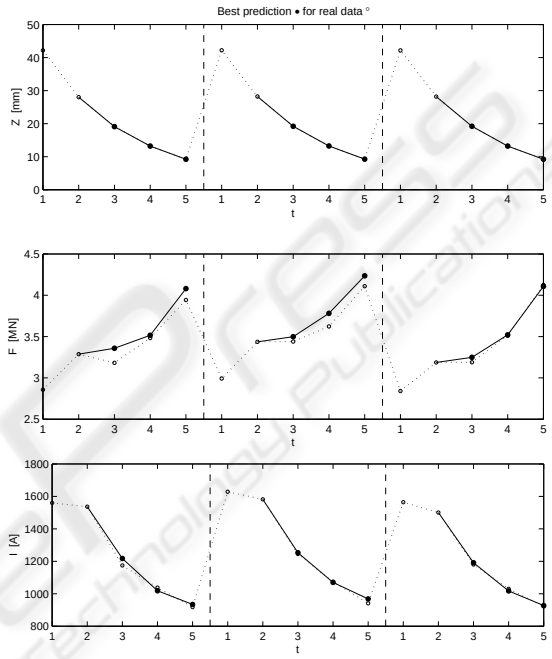

Figure 7: Prediction of real data made according to the method B. Data taken from a hot reversing rolling mill.

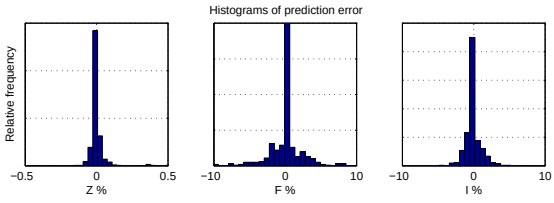

Figure 8: Histograms of percentage prediction error for three real data channels.