BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **7**, No 3

Sofia • 2007

Modifications of Estimation Procedures and Validation Criterion, Applied for Market System Identification

Alexander Efremov, Assen Todorov

Technical University of Sofia, 1000 Sofia

E-mails: alefremov@tu-sofia.bg assent@tu-sofia.bg

Abstract: *The problems arising during the automatic market system identification are discussed. A common case in the systems considered is the lack of input-output data related to products that are not on the market for a given period of time. Another possible reason for that is due to database errors or incorrect data collection, which may cause also incorrect records. To account for the uncertainty in the data, caused by the explained cases, modifications of the standard estimators are proposed. In order to avoid numerical problems, suitable realizations of the estimation procedures are used. Besides this, a modification of the validation criterion is utilized to assess the accuracy of the obtained models. A Monte Carlo simulation is run to determine the benefit from the use of modified estimators and the validation criterion.*

Keywords: *Automatic identification, market system, MIMO dynamic demand model.*

1. Introduction

Trying to improve the adequateness of the demand models, there is an increased tendency to account for the dynamic aspect of the market. For instance a dynamic demand model for cigarettes is investigated in [1]. Also a dynamic model is used in [5] for the future demand determination of different groups of products (food, alcoholic drink and tobacco, clothing and footwear, energy, etc).

A procedure for identification of multiple input multiple output (MIMO) dynamic demand models was designed. The approach is general enough since it is

automatically applied for different datasets, related to supermarket or hypermarket chains. The generalization of the procedure is connected with the solution of the arising problems without manual adjustments or taking expert decisions.

The problems encountered, which appear when automatic system identification cycle (ASIC) is applied for generation of MIMO representations, are outlined below. Also the suggested solutions are discussed.

1.1. Number of model parameters

The retailers actions are the inputs of the identified system. They are the units price, discount and the feature ads and displays information. The market reaction is the products sales. When the total number of products is tremendous, it is very difficult or even impossible to perform the identification. To decrease the inputs and outputs dimensions, the products are divided into categories and for each of them a MIMO model is estimated. In some cases the number of products in a category can be great (for example a category may contain few hundred products) and also the number of the processes observed can be quite big for an appropriate model determination. In a previous works [6, 7], a cross effects determination procedure (CED) was designed, to provide an appropriate problem dimension. It determines a subset of only significant retailers actions affecting the sales of every product.

1.2. Datasets quality

Another problem connected with the multivariate market identification is related to the quality of the datasets. For such kind of systems, the application of real experiments for the purpose of model determination can raise significantly the economic losses and it is also a very time consuming activity. Normally the data is collected weekly, which is the reason for the short observation intervals. Hence it is necessary to use the available data, but not to assign active experiments. The existing datasets can contain unrealistic values, uninformative observations or missing records.

The unrealistic values are due to data collection, such as barcode reading errors, time samples that are not exactly in one week, unaccounted promotions, etc. To avoid numerical problems, if some prices and sales are negative or zero, they are assumed as errors and replaced by the last appropriate records. Also a shaving procedure [9] is applied that corrects the data using linear interpolation. A problem can occur if the negative or zero sales are not wrong records (for example if the returned products by the customers are more than the sold ones). Another mistake is possible, if part of the data surrounding the wrong values corresponds to a promotion case. Thus the shaving procedure may replace the incorrect data with not precise values.

A common situation is when some products are not on the market for some period of time. This is related to a lack of records in the datasets that is an obstacle for MIMO identification. The multivariate structure is fixed and it takes into account the products appearing on the market during the period observed. So the identification procedure needs full input-output data. Also there is an additional effect, connected with the absence of products. It is connected with the increase or

decrease of the product sales if the correlated product is absent (Fig. 1).

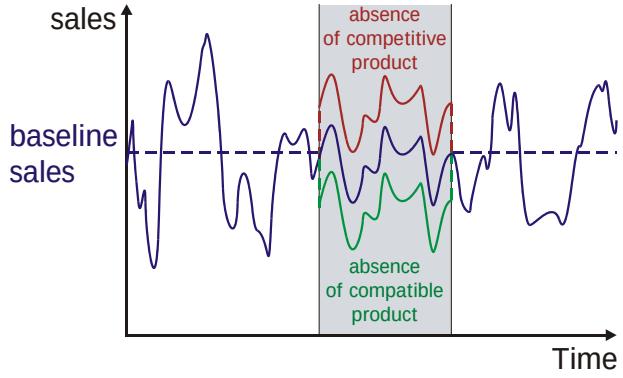

Fig. 1. The effects of missing competitive and compatible products

Usually the promotion case is run not so frequently and normally continues for a few weeks. Thus when the product is not on the market, we assume that the input-output data corresponds to the regular case. A problem can appear if a missing product is compatible or competitive with other products, which is the normal case, especially when the units belong to one category.

It is possible the lack of data to be caused by database errors. Then another problem is expected, if a promotion is run during the period of time with missing data. Hence the market reaction would be vastly different compared to the assumed non-promotional case.

There are other examples for mistakes that can be made during the datasets correction. In ASIC the problems with the datasets quality are accounted by different preprocessing techniques and estimators modifications. In order to decrease the effect of these problems, the datasets are preprocessed, so the irregular or missing observations are revealed, isolated and the datasets are automatically modified. After this process, an error remains, which (as we explained) can be significant. Thus in addition we developed a set of modified estimators and their numerically stable realizations.

1.3. Accounting for the feature ads and displays

The feature ads and displays are added in the demand model using two sets of flags. All flags related to the advertisements are zeroes for the regular case, but if a given ad is used, then the corresponding flag becomes one. The display type is taken into account in the same manner. Some retailers distinguish many types of advertisements and displays. By accounting for this kind of information, the input dimension can grow drastically, which leads to a huge number of model parameters. This is the reason to split up the inputs into two parts. The first part contains the significant prices and discounts and the second part consists of all the flags. To decrease the parameters all inputs from the first part are related to the output by dynamic terms and the remaining inputs are connected with the product sales by static relations.

2. Implemented estimators in ASIC

The current identification approach is oriented towards dynamic regression representations, for it is a continuation of ASIC [6], based on SISO regression models. The extension of the considerations for MIMO structures aimed to assess the ability of the multivariate dynamic models in presenting the market behaviour. Both approaches are applied in practice. The analysis made for Retail Analytics Ltd. company [10] shows the significant improvement of the MIMO dynamic representation.

A possible way to obtain a MIMO model of given dynamics is to use one MIMO regression model, but in this case a problem connected with the choice of the model structure may occur. On the other hand the number of the model parameters is an important quantity especially for the multivariate system identification. In particular, when the number of the inputs and outputs is enormous, some of the structure parameters can vastly exceed their optimal values. To avoid this, MIMO system can be decomposed to l multiple input single output (MISO) subsystems, where l is the output dimension. Each subsystem represents the relation between a given system output and all significant input processes. In this approach the freedom in the model structure selection is greater than in the first approach. This is the reason the above decomposition to be implemented in ASIC. The following estimators are used:

– Generalized Least Squares (GLS);
– Extended Matrix Least Squares (EMLS);
– Weighted Least Squares (WLS);
– Instrumental Variable (IV);
– Prediction Error (PE) applied for MISO Auto-Regressive Moving Average models with eXogenous input (MISO-ARMAX);
– PE applied for MISO Output Error (MISO-OE) model;
– PE applied for MISO Box-Jenkins (MISO-BJ) model.

2.1. Numerically stable estimators

The considerations below are related to a MISO demand model connected with a product from a concrete product category. The following notations are used: θ – a vector containing the model parameters; Φ – a data matrix with rows equal to the regression vector at the corresponding time instant; y – a vector containing the product sales for the observation interval. LS and PE are presented in the paper. The other estimators are modified in a similar way. According to the above notations, the LS criterion can be written as

$$\min_{\theta} J_{\theta} = \min_{\theta} \| \Phi \theta - y \|_2^2 .$$

The estimate of the parameter vector that minimizes J_θ is

(1) $$\hat{\theta} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} y .$$

The numerically reliable solutions of the non-recursive methods implemented in ASIC are based on the singular value decomposition (SVD) [8] applied to the data matrix Φ. The economy size SVD is

(2) $$\Phi = U\Sigma V^{\mathrm{T}} .$$

Since U and V are orthogonal matrices and Σ is square, the LS solution can be presented as

$$\hat{\theta} = V\Sigma^{-1}U^{\mathrm{T}} y .$$

If Φ is singular or close to singular, a reliable solution can be obtained by neglecting part of Σ. Let for instance r be the number of the significant singular values of Φ (greater than a certain low limit $\sigma_{\min} > 0$) that are the first r diagonal elements of Σ. And let also the matrix $\Sigma_r \in \mathsf{R}^{r\times r}$ be diagonal with elements equal to these r singular values. Hence, accounting (2), the data matrix can be approximated as

(3) $$\Phi \approx U_1\Sigma_r V_1^{\mathrm{T}} ,$$

with U_1 and V_1 – parts of U and V. Thus the modified LS solution becomes

(4) $$\hat{\theta} = V_1\Sigma_r^{-1}U_1^{\mathrm{T}} y .$$

With this change regularization is introduced as the singular values are restricted from below.

The PE methods are realized in ASIC again by the use of SVD. The PE approach is non-linear, because the cost function depends on the prediction error $e_{p,k}$. Generally $e_{p,k}$ is non-linear in θ and J_θ cannot be minimized analytically. To determine $\hat{\theta}$ as a solution of PE in ASIC is used the Gauss-Newton method. The updating rule is

(5) $$\hat{\theta}^{h+1} = \hat{\theta}^{h} - \mu^h H_{\hat{\theta}^h}^{-1} J'_{\hat{\theta}^h} .$$

Here by $J'_{\hat{\theta}^h} = \Psi_\theta e_p$ the gradient of J_θ is denoted at the current iteration and by $H_{\hat{\theta}^h} = \Psi_\theta \Psi_\theta^T$ – an approximation of its Hessian. The matrix $\Psi_\theta \in \mathsf{R}^{\dim\theta\times(N-n)}$ has columns equal to the gradients of $e_{p,k}$ with respect to θ and $e_p \in \mathsf{R}^{N-n}$ contains the residuals values. A stable estimation of θ can be obtained by using SVD applied to Ψ_θ. In accordance to the considerations made for (4), by introducing a regularization, we get

$$\Psi_\theta = U_1\Sigma_r V_1^{\mathrm{T}}$$

and (5) becomes

$$\hat{\theta}^{h+1} = \hat{\theta}^{h} - \mu^{h} U_1 \Sigma_r^{-1} V_1^{\mathrm{T}} e_p \,. \tag{6}$$

The presented ideas related to both linear and non-linear approaches are used to obtain all numerically stable estimators.

2.1. Estimators modifications accounting the presence of incorrect or missing data

Sometimes the apriori data is not enough to obtain adequate, full datasets. As it was above mentioned, there are cases where the remaining uncertainty after introducing or correcting some observations cannot be neglected. Therefore, during the data preprocessing the time instants, in which the data is absent or unrealistic, are accounted for. The collected information is necessary for the generation of a weighting matrix, used in the modifications presented below.

The idea of the modifications accounting incorrect or missing data is based on WLS, where the diagonal matrix $W \in \mathsf{R}^{N-n \times N-n}$ is used to assign different weights for the residuals in the cost function. The criterion corresponding to a given MISO model can be written as

$$\min_{\theta} J_{\theta} = \min_{\theta} \| y - \Phi\theta \|_W^2 \cdot$$

The problem solution is

$$\hat{\theta} = (\Phi^{\mathrm{T}} W \Phi)^{-1} \Phi^{\mathrm{T}} W y \,.$$

The modification is deducted and applied to all estimators. Here the weights $w_{kk} \in [0, 1]$ depend on the number of available data in the initial dataset at each time instant. Let the significant retailers actions with respect to a given product for the k-th week are collected in the vector u_k and the market reaction (the concrete product sales) for the same week is denoted by y_k. We made the following assumptions about the weights

$$w_{kk} \begin{cases} = 1, \text{ if } u_{k-1} \text{ and } y_{k-1} \text{ are available and correct,} \\ \in (0, 1), \text{ if part of } u_{k-1} \text{ and } y_{k-1} \text{ are available and correct,} \\ = 0, \text{ if } y_k \text{ is unavailable.} \end{cases}$$

A suitable way to compute the weights is to decrease them proportionally to the number of the absent or incorrect input data. If the sales y_k are not available, then $w_{kk} = 0$, because it is not possible to determine the corresponding residual. As the system and the regression models are dynamic, the system and the model output depend on the past input-output data. Thus the weights can be formed in such a way to account this relation. The above modification applied with the numerically stable solution (4) is

$$\hat{\theta} = V_{L,1} \sum\nolimits_{L,r}^{-1} U_{L,1}^{\mathrm{T}} L y. \tag{7}$$

Here $L = W^{1/2}$ and the matrices $V_{L,1}$, $\Sigma_{L,r}$ and $U_{L,1}$ are determined by SVD, applied to $L\Phi$.

The weighted numerically stable solution of the PE methods leads to the following updating rule

$$\hat{\theta}^{h+1} = \hat{\theta}^{h} - \mu^{h} U_{L,1} \Sigma_{L,r}^{-1} V_{L,1}^{\mathrm{T}} L e_p . \quad (8)$$

SVD is applied to $\Psi_\theta L$ and the solution is regularized by the technique explained in the previous subsection. In this paper with the term "weighted" used for the estimators, we denote the procedures that take into account the presence of wrong and missing data.

2.3. Modification of the validation criterion

To measure the accuracy of the predicted sales, we propose a weighted variant of the Variance Accounted For (WVAF) that is

$$\mathrm{WVAF} = \max(0,\ 1 - \mathrm{var}(Le)/\mathrm{var}(Ly)) \times 100\% . \quad (9)$$

This criterion is most sensitive to the observations if 100% from the records is available. But if a part of the processes is absent, the corresponding weights decrease. Here $e \in \mathrm{R}^{N-n}$ is the difference between the measured and the predicted sales. With this criterion we account the presence of errors or missing data in the datasets at the level of validation by using the matrix L defined above (when $L = I$, then WVAF = VAF).

3. Test, results and conclusions

3.1. Test description

In the previous works, real datasets were used to determine the performances of the designed procedures. To assess the benefit of applying the modified estimators and the validation criterion, we designed a Monte Carlo (MC) simulation. It is similar to the testing procedure applied during the design of the cross effects determination procedure, used in the first variant of ASIC. In Fig. 2 the structure of one simulation cycle is shown. The first stage is the input-output data generation using a model with parameters θ. Also an uncertainty is added in the data. Subsequently part of the data is replaced by baseline values, determined in accordance with the generated processes. In this way we simulate the case with unrealistic and missing input records, which are corrected by the preprocessing procedures. If a competitive or compatible product is removed from the set of significant cross related products, a bias in the current product sales appears. To account this effect, the generated sales are shifted in addition. The second stage is the model parameters estimation. To perform this and the last stage, the datasets are divided into two parts. The first part is used for model determination and the second one – for validation. The two sets of

numerically stable estimators mentioned in Section 2 are used. The first one is based on (4) and (6). The other set contains the weighted estimators based on (7) and (8). The remaining stage is the models validation, where both VAF and $WVAF$ are computed.

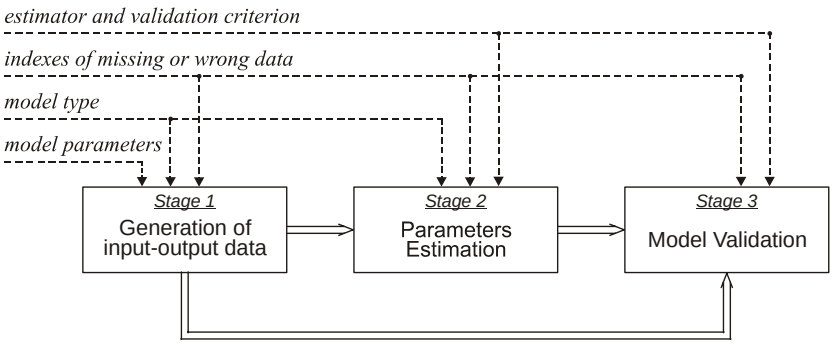

Fig. 2. Structure of the testing procedure

The MC simulation is run once for each model type. The three stages are run $M = 100$ times for a given MC. At each cycle $m = 6$ inputs and $l = 1$ output processes are formed using some concrete MISO parametric regression models. The first 3 inputs are related to the output by static and the other 3 – by dynamic relations. Additional output disturbance v_k with $\sigma_v = 0{,}3$ is generated to account the uncertainties in the market systems. 20% of the data are assumed to be unavailable or wrong. The biases in the output are $\pm 10\%$ from $y_{\max}$ with a random sign ("+" for the lack of a competitive and "–" for the lack of a compatible product). The data for the model estimation is $2/3$ part of the datasets. The model type and the structure are the same as the original structural parameters. As it was above mentioned, two models are formed for every dataset – one of them is obtained by a non-weighted estimator and the other one, by a weighted estimator. Finally, the validation criterions (VAF and WVAF) are applied to assess the models proximity to the simulated dynamics.

3.2. Results and conclusions

The results from Monte Carlo simulations are obtained for different parametric regression models. The average values of the validation criterions for both cases

$$\overline{\mathrm{VAF}} = \frac{1}{M}\sum_{i=1}^{M}\mathrm{VAF}_i \text{ and } \overline{\mathrm{WVAF}} = \frac{1}{M}\sum_{i=1}^{M}\mathrm{WVAF}_i$$

are shown in Table 1. The following conclusions can be made.

WVAF criterion is a more precise assessment of the estimators applicability than VAF. It is also seen that the models obtained by the estimators that account the presence of wrong or absent data are more accurate than the models generated by the non-weighted estimators. In both cases the reason is that the weighted criterion

and methods account by a higher weight the time instants with input-output records that correspond to the preliminarily chosen MISO structure. If a part of the processes, accounted by the model, is not available for some reasons, the sensitivity of WVAF and of the modified estimators decreases.

Table 1. MC simulation results

Methods	Criterion – models	Non-weighted estimators		Weighted estimators	
		$\overline{\text{VAF}}$, %	$\overline{\text{WVAF}}$, %	$\overline{\text{VAF}}$, %	$\overline{\text{WVAF}}$, %
BLS	MISO–ARX	85.96	87.40	95.16	98.57
WLS	MISO–ARX	86.19	87.64	95.17	98.57
IV	MISO–ARX	75.00	76.80	93.81	97.14
GLS	MISO–ARARX	81.29	83.75	93.37	96.41
EMLS	MISO–ARARMAX	85.67	87.14	95.15	98.56
ELS	MISO–ARMAX	85.98	87.65	89.06	94.13
PE	MISO–ARMAX	88.87	92.56	91.34	98.66
PE	MISO–BJ	89.47	93.22	92.81	94.28
PE	MISO–OE	81.69	82.99	87.53	91.58
Average $\overline{\text{VAF}}$ and $\overline{\text{WVAF}}$		84.46	86.57	92.60	96.43

A disadvantage of the presented MIMO identification procedure is that the computation time for generation of all models is enormous. To decrease the computational burden, recursive estimators for the MISO regression representations similar to [7] have been developed. The relations between the products that are not accounted in the multivariate structure lead to uncertainty, which is usually time-varying. For that case a variant of CED was used to assess the adequateness of a given MISO model structure. If it becomes inappropriate for the current market behaviour, a new structure is determined and ASIC is applied only to the corresponding model, but not to all MISO models. The estimated parameters are assumed as initial for the corresponding recursive estimator.

The system investigated is subject to different kinds of uncertainties caused by the competitors activities, weather conditions, products that are not accounted in the MIMO structure, etc. An appropriate way to predict the system behaviour, accounting the presence of these uncertainties, is to use Kalman filter (KF). It should be combined with the recursive MIMO identification, which will update the model. The market system is non-linear, so if a posteriori information is available (for instance restrictions upon the observed processes, or relations between some processes) then an extended KF (EKF) [3] can be applied.

When the input and output dimensions grow, the number of parameters of the regression models increases faster than the number of the state space representations. KF and EKF are also developed for state space representations. For this reason ASIC, which determines state space models [2, 4] can be developed.

References

1. Baltagi, B. H., Dan Levin. Cigarette taxation: raising revenues and reducing consumption. – Structural Change and Economic Dynamics, Vol. **3,** 1992, No 2, 321-335.

2. Bernal, D., B. Gunes. Performance of an Observer State-Space Identification in the Presence of Mild Nonlinearities. – In: Proc. of the American Control Conference, Vol. **2**, 2000, 986-990.
3. Bishop, G., G. Welch. An Introduction to the Kalman Filter. – In: Siggraph – Course 8, 2001.
4. Bos, R., X. Bombois, P.M.J. Van den Hof. On Model Selection for State Estimation for Nonlinear Systems. – In: Proc. of the 6th IFAC Symposium on Dynamics and Control of Process Systems (DYCOPS), Cambridge, MA, 5-7 July 2004.
5. Deschamps, Philippe J. Exact Small-Sample Inference in Stationary, Fully Regular, Dynamic Demand Models. – Journal of Econometrics, **97**, 2000, 51-91.
6. Efremov, A. An Automatic Identification Cycle for Dynamic Demand Model Generation. – International Conference on Computer Systems and Technologies, 2005, V.5-1-6. **http://anp.tu-sofia.bg/aefremov/index.htm.**
7. Efremov, A. Recursive Estimation of Dynamic Time-Varying Demand Models. – In: International Conference on Computer Systems and Technologies, 2006, V.7-1-6. **http://anp.tu-sofia.bg/aefremov/index.htm.**
8. Elden, L. Numerical Linear Algebra and Applications in Data Mining. Preliminary Version. Department of Mathematics, Linkoping University, Sweden, December 20, 2005.
9. Verhaegen, M., Vincent Verdult. Filtering and System Identification: An Introduction to Using Matlab Software. Software Manual for the Course sc4040 (et4094) Edition, Delft University of Technology, 2003.
10. **http://retail-analytics.com/**