Austrian Journal of Statistics
April 2017, Volume 46, 47–56.
http://www.ajs.or.at/
doi:10.17713/ajs.v46i3-4.670

Statistical Estimation and Classification Algorithms for Regime-Switching VAR Model with Exogenous Variables

Vladimir Malugin
Belarusian State University

Alexander Novopoltsev
Belarusian State University

Abstract

We consider a vector autoregression model with exogenous variables and Markov-switching regimes to describe complex systems with cyclic changes of states. To estimate and forecast the states, we propose EM and discriminant analysis algorithms based on non-classified and classified data samples accordingly. The accuracy of the algorithms is examined by means of computer simulation experiments.

Keywords: regime-switching models, vector autoregression models with exogenous variables, EM algorithm, discriminant analysis algorithm, dynamic programming approach, probability of misclassification.

1. Introduction

Regime-switching models are a convenient tool for the analysis of complex systems with cyclic changes of states (Hamilton 2008). Most studies are devoted to Markov-switching vector autoregression model (MS-VAR) (Krolzig 1997). If the regimes are independent or there is a high uncertainty regarding the classes of states, then the models with independent-switching regimes may be more preferable. The autoregression and regression models of such type were entirely studied in Malugin and Kharin (1986) and Malugin (2014). The object of the study is a vector autoregression model with Markov-switching states including exogenous variables (MS-VARX), thus allowing a multivariate linear regression ones (Malugin 2014).

2. Models and tasks of research

Let a complex system at time t be characterized by a random observation vector defined on the probability space $(\Omega, \mathbf{F}, \mathbf{P})$, where Ω is a space of elementary objects $\omega \in \Omega$; $\mathbf{P}$ is a probability measure: $\mathbf{P}(A) = \mathbf{P}\{\omega \in A\}$, $A \in \mathbf{F}$. Let $\{\Omega_0, \ldots, \Omega_{L-1}\}$ be a decomposition of Ω into a finite number of non-empty disjoint subsets, such that: $\Omega_l \in \mathbf{F}, \mathbf{P}\{\Omega_l\} = \mathbf{P}\{\omega \in \Omega_l\} > 0, \bigcup_{l \in S(L)} \Omega_l = \Omega, S(L) = \{0, \ldots, L-1\}$. These subsets are the classes of states of a complex system, and L is the number of classes.

A random vector $y_t = (x'_t, z'_t)' \in R^n$ can be partitioned into subvectors of endogenous

variables $x_t = (x_{tj}) \in \Re^N$ and deterministic exogenous variables (regressors) $z_t = (z_{tk}) \in \mathfrak{Z} \subset \Re^M$. It is assumed that, in general, the time series is described by a model RS-VARX$(p)(p \geq 1)$:

$$x_t = \sum_{i=1}^{p} A_{d(t),i} x_{t-i} + B_{d(t)} z_t + \eta_{d(t),t}, \quad t = 1, \ldots, T, \tag{1}$$

where $x_{1-p}, \ldots, x_0 \in \Re^N$ are a set of the given initial values; $\eta_{d(t),t} \in \Re^N$ is a random disturbances or innovation process; and $d(t) \in S(L) = \{0, \ldots, L-1\}$ is a state of a system at time t.

Model (1) satisfies the following assumptions:

M.1. Segmented-stationary condition: for each class of states $l \in S(L)$ matrices of autoregression coefficients $\{A_{l,i}\}(i = 1, ..., p)$ satisfy the stationarity condition for VAR(p) model (Lutkepohl 2005);

M.2. Disturbance assumptions: disturbances $\{\eta_{l,r}\}\,(t, s = 1, ..., T,\ l \in S(L))$ are independent Gaussian random vectors with parameters: $\mathbf{E}\{\eta_{l,t}\} = 0_N \in \Re^N, \mathbf{E}\{\eta_{l,t}\eta'_{l,s}\} = \delta_{t,s}\Sigma_l$, where $\delta_{r,s}$ — the Kronecker delta.

M.3. Structural heterogeneity conditions: for matrices of autoregression and regression coefficients: $A_l \neq A_k$ and (or) $B_l \neq B_k \forall k \neq l,\ k, l \in S(L)$.

We consider a model with L $(2 \leq L < s + 1)$ classes of states: where $s \geq 1$ — number of state switching points $1 < \tau_1 < \ldots < \tau_s < T$. Concerning the sequence of states $d(t) \equiv d_t \in S(L)(t = 1, ..., T)$ there are two types of assumptions:

d1. $d_t\,(t = 1, ..., T)$ — independent identically distributed random variables with probability distribution $\mathbf{P}\{d_t = l\} = \pi_l > 0\,(l \in S(L))\,, \sum_{l \in S(L)} \pi_l = 1;\ \mathbf{P}\{d_t = l\} = \pi_l > 0\,(l \in S(L)) \sum_{l \in S(L)} \pi_l = 1;$

d2. $d_t\,(t = 1, ..., T)$ — homogeneous ergodic Markov chain (GCM) with the distribution determined by the vector of probability of the initial state π and matrix one-step transition probabilities P:

$$\pi = (\pi_l),\ \pi_l = \mathbf{P}\{d_1 = l\} > 0\ \ (l \in S(L))\,, \sum_{l \in S(L)} \pi_l = 1;$$
$$P = (p_{kl}), p_{kl} = P\{d_{t+1} = l | d_t = k\} \geq 0 (k, l \in S(L)), \sum_{l \in S(L)} p_{kl} = 1, k \in S(L). \tag{2}$$

Under the conditions of d1 and d2, we deal with the models IS-VARX and MS-VARX respectively. Model (1) includes a number of special cases: model of multivariate linear regression RS-MLR, if $p = 0$, $M \geq 1$ (Malugin 2014); model RS-VAR without exogenous variables, if $p > 0$, $M = 0$ (Krolzig 1997).

The true values of model parameters $\{A_l, B_l, \Sigma_l\,(l \in S(L)), \pi\, P$ and the moments of switching state $\{\tau_i\}(i = 1, ..., s)$ are unknown. There is either classified or a non-classified sample of observations $(\bar{X}, \bar{Z})$ $(\bar{X} = (x_t) \in \Re^{NT}$, $\bar{Z} = (z_t) \in \mathfrak{Z}^T \subseteq \Re^T)$, so that the vector of states $\bar{d} = (d_t) \in S^T(L)$ is either known or unknown. We presented two statistical classification algorithms for MS-VARX model in these cases: EM algorithm for joint parameters and vector of states estimation for non-classified sample and discriminant analysis algorithm in the case of classified sample for classification of out-of-sample observations. For IS-MLR and IS-VARX models the listed tasks are solved in Malugin (2014).

3. Splitting of mixtures described by MS-VARX

Representations for the model parameters. Model (1) under the assumptions M.1-M.3, d.1 and d.2 can be represented in the regression form

$$x_t = \Pi_{d(t)} u_t + \eta_{d(t),t}, \tag{3}$$

where $\Pi_{d(t)} = (A_{d(t),1}, \ldots, A_{d(t),p}, B_{d(t)})$ is the block $N \times (pN + M)$ — matrix of parameters; $u_t = (x'_{t-1}, \ldots, x'_{t-p}, z_t)' \in \Re^{Np+M}$ — the stacked vector of predetermined variables formed from lagged endogenous and exogenous variables with values known at time t.

In this case we use a sample of observations $(\bar{X}, \bar{U})$, where $\bar{X} = (x'_1, \ldots, x'_T)' \in \Re^{NT}$ — the values of the endogenous variables, which correspond to the values $\bar{U} = (u'_1, \ldots, u'_T)' \in \Re^{NpT} \times \mathfrak{Z}^T \subseteq \Re^{(Np+M)T}$ of predefined variables. For the model (3) we will also denote:

$\theta_l \in \Re^m (m = N \times (pN + M) + N(N + 1)/2)$ — stacked vector of parameters for the class $l \in S(L)$ consisting of independent elements of matrices $\{\Pi_l, \Sigma_l\}$ $(l \in S(L))$;

$\phi \in \Re^q (q = Lm + (L-1)(L+1))$ — parameters of a mixture model, including $\{\theta_l\}$ and π, P, $\hat{\phi} \in \Re^q$ — statistical estimate of $\phi \in \Re^q$;

$D = (d_1, \ldots, d_T)' \in S^T(L)$ — the state vector for the period under observation;

$\tilde{\gamma}_{l,t} = \mathbf{P}\{d_t = l | \bar{X}, \bar{U}; \tilde{\phi}\}$ — posterior probabilities of the class $l \in S(L)$ at the moment t;

$\tilde{\xi}_{kl,t} = \mathbf{P}\{d_{t+1} = l | d_t = k; \bar{X}, \bar{U}; \tilde{\phi}\}$ — posterior probability of transition from class $k \in S(L)$ to class $l \in S(L)$ at the moment $t (t = 1, \ldots, T-1)$.

If the model (3) satisfies the assumptions M.1-M.3, then the random vector $x_t \in \Re^N$ under the given values $u_t \in \Re^{Np+M}$ and $d_t = l (l \in S(L))$ has conditional normal distribution with density

$$p_X(x, u, \theta_l) = (2\pi)^{-\frac{N}{2}} |\Sigma_l|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \Pi_l u)' \Sigma_l^{-1} (x - \Pi_l u)\right\}, \; x \in \Re^N, u \in \Re^{Np+M}. \quad (4)$$

The likelihood function for parameters ϕ under the fixed state vector $D \in S^T(L)$ and assumptions (4) and d.2 is of a form:

$$L(\phi; \bar{X}, \bar{U}, D) = \pi_{d_1} p_X(x_1; u_1, \theta_{d_1}) \prod_{t=2}^{T} p_{d_{t-1}, d_t} p_X(x_t; u_t, \theta_{d_t}). \quad (5)$$

Let $\Lambda(\phi, \tilde{\phi})$ be the conditional expectation of the log-likelihood function $l(\phi; \bar{X}, \bar{U}, D) = \ln L(\phi; \bar{X}, \bar{U}, D)$ induced by the distribution $P\{D | \bar{X}, \bar{Z}; \tilde{\phi}\}$ of the random vector D given the fixed sample $(\bar{X}, \bar{U})$ and initial value $\tilde{\phi}$ of the parameter vector, i.e.

$$\begin{aligned} \Lambda(\phi, \tilde{\phi}) &= E_{\tilde{\phi}}\{l(\phi; \bar{X}, \bar{U}, D) | \bar{X}, \bar{U}; \tilde{\phi}\} = \\ &= \sum_{l \in S(L)} \tilde{\gamma}_{l,1} \ln \pi_l + \sum_{t=2}^{T} \sum_{k \in S(L)} \sum_{l \in S(L)} \tilde{\xi}_{kl,t} \ln p_{kl} + \sum_{t=1}^{T} \sum_{l \in S(L)} \tilde{\gamma}_{l,t} \ln p_X(x_t; u_t, \tilde{\theta}_l) = \\ &= Q_1(\{\pi_l\}) + Q_2(\{p_{kl}\}) + Q_3(\{\theta_l\}). \end{aligned} \quad (6)$$

In accordance with a general approach (Malugin 2014; Bilmes 1998) we obtain an analytical representation for the unknown characteristics. In the considered case we have conditional normal distribution for vector of endogenous variables with the density $p_X(x; u, \theta_l)$ for the given vector of predetermined (lagged or exogenous) variables $u_t = (x'_{t-1}, \ldots, x'_{t-p}, z_t)' \in \Re^{Np+M}$. Formulas for the posterior probabilities $\{\tilde{\gamma}_{l,t}\}, \{\tilde{\xi}_{l,t}\}$ are based on the density $p_X(x; u, \theta_l)$ and followed from the Lemma 1.

Lemma 1. *For fixed values of the parameters $\{\tilde{\theta}_l\}$, $\tilde{\pi}$, $\tilde{P}$ of the model (3) posterior probabilities $\tilde{\gamma}_{t,l}$, $\tilde{\xi}_{kl,t}$ for the sample (X, Z) are of a form:*

$$\tilde{\gamma}_{l,t} = \frac{\tilde{\alpha}_{l,t} \tilde{\beta}_{l,t}}{\sum\limits_{k \in S(L)} \tilde{\alpha}_{k,t} \tilde{\beta}_{k,t}}, \quad l \in S(L), \; t = 1, \ldots, T; \quad (7)$$

$$\tilde{\xi}_{kl,t} = \frac{\tilde{\alpha}_{k,t} \tilde{p}_{kl} p_X(x_{t+1}; u_{t+1}, \tilde{\theta}_l) \tilde{\beta}_{l,t+1}}{\sum\limits_{r \in S(L)} \sum\limits_{s \in S(L)} \tilde{\alpha}_{r,t} \tilde{p}_{rs} p_X(x_{t+1}; u_{t+1}, \tilde{\theta}_s) \tilde{\beta}_{s,t+1}}, \quad k, l \in S(L), \; t = 1, \ldots, T-1; \quad (8)$$

$$\tilde{\alpha}_{l,1} = \tilde{\pi}_l p_X(x_1; u_1, \tilde{\theta}_l),\ \tilde{\alpha}_{l,t} = \left(\sum_{k \in S(L)} \tilde{\alpha}_{k,t-1} \tilde{p}_{kl} \right) p_X(x_t; u_t, \tilde{\theta}_l), \quad t = 2, \ldots, T; \tag{9}$$

$$\tilde{\beta}_{l,T} \equiv 1,\ \tilde{\beta}_{l,t} = \sum_{k \in S(L)} \tilde{p}_{lk} p_X(x_{t+1}; u_{t+1}, \tilde{\theta}_k) \tilde{\beta}_{k,t+1}, \quad t = T-1, T-2, \ldots, 1. \tag{10}$$

The proof of the Lemma 1 is based on the method from Bilmes (1998) for Gaussian Mixture with Markov regime switching.

The representation for estimate $\hat{\phi} \in \Re^q$ is obtained by maximization of the conditional expectation of the log-likelihood function (6) for some given initial value $\tilde{\phi} \in \Re^q$, that is:

$$\hat{\phi} = \arg \max_{\phi \in \Re^q} \Lambda(\phi, \tilde{\phi}) = \arg \max_{\phi \in \Re^q} E_{\tilde{\phi}} \{ l(\phi; \bar{X}, \bar{U}, D) | \bar{X}, \bar{U}; \tilde{\phi} \}, \tag{11}$$

Theorem 1. *If model MS-VARX (3) satisfies the assumptions M.1-M.3, d.2, the estimates* $\{\hat{\Pi}_l, \hat{\Sigma}_l\}$ $(l \in S(L))$, $\hat{\pi}$, $\hat{P}$ *on a sample* $(\bar{X}, \bar{U})$ *are the solution of equation (11) for a given* $\tilde{\phi} \in \Re^q$*:*

$$\hat{\pi}_l = \tilde{\gamma}_{l,1}, \quad \hat{p}_{kl} = \sum_{t=2}^{T} \tilde{\xi}_{kl,t} \left(\sum_{t=2}^{T} \tilde{\gamma}_{k,t-1} \right)^{-1}, \hat{\Pi}_l = \sum_{t=1}^{T} \tilde{\gamma}_{l,t} x_t u'_t \left(\sum_{t=1}^{T} \tilde{\gamma}_{l,t} u_t u'_t \right)^{-1}, \tag{12}$$

$$\hat{\Sigma}_l = \sum_{t=1}^{T} \tilde{\gamma}_{l,t} (x_t - \hat{\Pi}_l z_t)(x_t - \hat{\Pi}_l z_t)' \left(\sum_{t=1}^{T} \tilde{\gamma}_{l,t} \right)^{-1}, \tag{13}$$

where posterior probabilities $\{\tilde{\gamma}_{l,t}\}, \{\tilde{\xi}_{kl,t}\}$ *are described by the formulas (7)–(10).*

Proof. Three terms Q_1, Q_2 and Q_3 in the formula (6) depend on the various parameter sets. Therefore, the optimization problem for $\Lambda(\phi, \tilde{\phi})$ can be partitioned into three independent optimization problem for continuous in the parameters functions where a posterior probabilities $\{\tilde{\gamma}_{l,t}\}, \{\tilde{\xi}_{kl,t}\}$ are given. To maximize the functions Q_1, Q_2 with equality constrained we use the method of Lagrange multipliers. Maximizing the function Q_3 of the form

$$Q_3(\{\theta_l\}) = \sum_{l \in S(L)} \sum_{t=1}^{T} \tilde{\gamma}_{l,t} \left(-\tfrac{N}{2} \ln(2\pi) - \tfrac{1}{2} \ln |\Sigma_l| - \tfrac{1}{2} (x_t - \Pi_l u_t)' \Sigma_l^{-1} (x_t - \Pi_l u_t) \right)$$

is carried out separately on matrices Π_l and Σ_l by calculating the derivatives and using properties of matrices operations (Anderson 1984). □

Corollary. Using the known block structure for matrices $\hat{\Pi}_l$, we can get the estimates $\{\hat{A}_{l,1}, \ldots, \hat{A}_{l,p}, \hat{B}_l\}$ $(l \in S(L))$.

EM-algorithm for MS-VARX. For joint estimation of all parameters $\phi \in \Re^q$ and state vector $D \in S^T(L)$ the EM MS-VARX-algorithm (*Expectation-Maximization algorithm for MS-VARX*) is addressed. EM MS-VARX-algorithm belongs to the family of Baum – Welch algorithms of splitting of a mixture of multivariate distributions, controlled by a hidden Markov chain (Bilmes 1998).

The algorithm includes the following steps (superscript k in brackets indicates the iteration number).

Preliminary step, that includes: 1) setting the initial values of the parameters: $\phi^{(0)} \equiv \tilde{\phi}$; or $D^{(0)} = (d_t^{(0)})(t = 1, ..., T)$; 2) setting the parameters defining the accuracy rate of the objective function calculation ε $(0 < \varepsilon << 1)$ and the maximum number of iterations $\bar{k}$.

For iteration k $(k = 1, 2, \ldots)$:

Step E. Calculation of $\{\tilde{\gamma}_{l,t},\ \tilde{\xi}_{kl,t}\}$ by the formulas (7)–(10) assuming $\tilde{\phi} \equiv \phi^{(k-1)}$. Estimation of $D^{(k)} = (d_t^{(k)}) \in S^T(L)$ $(t = 1, \ldots, T\}$ by the decision rule of the maximum a posteriori probability of the class:

$$d_t^{(k)} = \arg\max_{l \in S(L)} \left\{ \gamma_{l,t}^{(k)} \right\}, \quad t = 1, \ldots, T. \tag{14}$$

Step M. Computation of the parameter estimates $\{\Pi_l^{(k)}, \Sigma_l^{(k)}\}$ $(l \in S(L)), \pi^{(k)}, P^{(k)}$by the formulas (12), (13) with using the probability $\gamma_{l,t}^{(k-1)}$ and $\xi_{ij,t}^{(k-1)}$ calculated in the Step E.

Checking two stop conditions (Bilmes 1998): 1) $k = \bar{k}$; 2) $l_X^{(k)} \geq l_X^{(k-1)}$ and $\left(l_X^{(k)} - l_X^{(2)}\right) < (1+\varepsilon)\left(l_X^{(k-1)} - l_X^{(2)}\right)$, where $l_X^{(k)} = \ln P\{\bar{X}, \bar{U}|\phi^{(k)}\} = \ln\left(\sum_{l \in S(L)} \alpha_{l,T}^{(k)}\right)$. If one of the conditions is satisfied, we set: $\hat{\phi} = \phi^{(k)}$, $\hat{D} = D^{(k)}$,$\hat{l}_X = l_X^{(k)}$, $\hat{\gamma}_{l,t} = \gamma_{l,t}^{(k)}$, $(l \in S(L),\ t = 1, \ldots, T)$. In this case, the algorithm terminates, otherwise the algorithm proceeds to Step E.

Convergence problems for this type of algorithms are investigated in numerous studies, particularly in Krolzig (1997); Malugin (2014). The convergence of the algorithm ensures the consistence of the resulting parameters estimates $\hat{\phi}$, $\hat{\pi}$, $\hat{P}$ as well as the consistence of the classification rule (13).

4. Discriminant analysis of the MS-VARX

The decision classification rule of multivariate autoregression observations $(\bar{X}, \bar{U})$ described by the MS-VARX model in general case can be defined as:$\hat{D} = (\hat{d}_t) = D(\bar{X}, \bar{U})$, $\hat{d}_t = \hat{d}_t(\bar{X}, \bar{U}) \in S(L)$, $t = 1, \ldots, T$. The accuracy of classification for this rule is characterized by the probability of misclassification:

$$r = r(D(\bar{X}, \bar{U})) = P\{||\hat{D} - D^0|| \neq 0\}, \quad ||D - D|| = \sum_{t=1}^{T} (1 - \delta_{\hat{d}t, d_t^0}), \tag{15}$$

where $D^0 = (d_t^0)$ and $\hat{D} = (\hat{d}_t)$ are the true state vector and its estimate respectively.

Assume first all parameters of the MS-VARX (3) to be known. Describe an optimal classification rule, called *Bayesian decision rules* (BDR) (Malugin 2014; Kharin 1996), which minimizes the probability of misclassification (15). Bayesian decision rules of pointwise and groupwise classification of multivariate observations described by IS-VARX and IS-MLR models, have been proposed and studied in Malugin (2014). In the considered case of MS-VARX model we addressing the groupwise classification decision rule. A similar problem in the case of a parametric family of continuous probability distributions was considered in Kharin (1996). To formulate the decision rule we will use the log-likelihood function, which for some fixed vector D according to (5) simplifies to:

$$l(\phi; \bar{X}, \bar{U}, D) = \ln(L(\phi; \bar{X}, \bar{U}, D)) = \ln \pi_{d_1} + \sum_{t=2}^{T} \ln p_{d_{t-1}, d_t} + \sum_{t=1}^{T} \ln p_X(x_t; u_t, \theta_{d_t}). \tag{16}$$

Lemma 2. *If model MS-VARX (3) satisfies the assumptions of M.1-M.3, d.2 and the staked vector of parameters $\phi \in \Re^q$ is known, BDR of groupwise classification is determined by the condition*

$$\hat{D} \equiv \hat{D}(\bar{X}_1^T, \bar{U}_1^T) = \arg\max_{D \in S^T(L)} l(\phi; \bar{X}_1^T, \bar{U}_1^T, D), \tag{17}$$

where $(\bar{X}_1^T, \bar{U}_1^T)$ $(\bar{X}_1^T = (x'_1, ..., x'_T)' \in \Re^{NT}$, $\bar{U}_1^T = (u'_1, \ldots, u'_T{}' \in \Re^{NpT} \times \mathfrak{Z}^T \subseteq \Re^{(Np+M)T})$ is a sample of observations to be classified.

Proof. It is known (Kharin 1996) that the decision rule of the form (17) for arbitrary family of parametric continuous distributions minimizes a probability of error classification. Such decision rules are known as Bayesian decision rules. Under the conditions of the Lemma 2 the vector of endogenous variables $x_t \in \Re^N$ corresponding to fixed values $u_t \in \Re^{Np+M}$ and

$d_t = l(l \in S(L))$ has conditional Gaussian distribution with density (4) which belong to the mentioned above family of parametric continuous distributions. □

To solve the integer optimization task (17) for some fixed continuous vector $\phi \in \Re^q (q = Lm + (L-1)(L+1))$ we will use the dynamic programming method (Kharin 1996; Bellman and Dreyfus 1962). Its implementation requires a special representation of the log-likelihood function $l(\phi; \bar{X}, \bar{U}, D)$ through the so-called Bellman functions.

Theorem 2. *Under the conditions of Lemma 2, the BDR of groupwise classification of sample* $(\bar{X}_1^T, \bar{U}_1^T)$ *is implemented using the dynamic programming method in accordance with the following relationships:*

$$\hat{d}_T = \arg\max_{k \in S(L)} F_T(k),\ \hat{d}_t = \arg\max_{k \in S(L)} \left(f_t(k, \hat{d}_{t+1}) + F_t(k) \right), \quad t = T-1, T-2, \ldots, 1, \tag{18}$$

$$F_1(l) \equiv 0,\ F_{t+1}(l) = \max_{k \in S(L)} \left(f_t(k,l) + F_t(k) \right),\ l \in S(L), \quad t = 1, \ldots, T-1, \tag{19}$$

where $\{F_t(k)\}$ *are Bellman functions and* $\{f_t(k,l)\}$ *are described by formulas*

$$f_t(k,l) = \delta_{t,1}(\ln \pi_k + \ln p_X(x_1; u_1, \theta_k)) + \ln p_{kl} + \ln p_X(x_{t+1}; u_{t+1}, \theta_l), k, l \in S(L), \tag{20}$$

δ_{t1} — *Kronecker symbol,* $t = 1, \ldots, T-1$.

Proof. In conditions of Lemma 2 the formulas (18)–(20) are obtained by means of equivalent transformation of function $l(\phi; \bar{X}, \bar{U}, D)$. Indeed, on the basis of (16), (17) and (20) we obtain:

$$\hat{D} = \arg\max_{D \in S^T(L)} \sum_{t=1}^{T-1} f_t(d_t, d_{t+1}). \tag{21}$$

It is known that a dynamic programming procedure includes the following two stages which use formulas (19) and (18) respectively:

1) recursive calculation of Bellman functions $\{F_t(l)\}$ $(l \in S(L), t = 1, ..., T-1)$ by the formulas

$$F_{t+1}(l) = \max_{k \in S(L)} (f_t(k,l) + F_t(k)),\ F_1(l) \equiv 0;$$

2) calculation of vector $\hat{D}$ components in the reverse order:

$$\hat{d}_t = \arg\max_{k \in S(L)} \left(f_t(k, \hat{d}_{t+1}) + F_t(k) \right)\ (t = T-1, T-2, ..., 1),$$
$$\hat{d}_T = \arg\max_{k \in S(L)} F_T(k).$$

□

Since parameters $\{\theta_l\}$ $(l \in S(L))$, π, P are unknown, we need to use their estimates obtained from some sample of classified observations. To get such a sample as to find the estimates $\{\hat{\theta}_l\}$ $(l \in S(L))$, $\hat{\pi}$, $\hat{P}$ it is suggested to apply the proposed above EM MS-VARX algorithm. Thus, the following statement is true.

Corollary. *If* $\{\hat{\theta}_l\}$ $(l \in S(L)), \hat{\pi}, \hat{P}$ *are consistent estimates of parameter for model (3), then using them in (15)–(17) instead of unknown values of parameters we obtain a consistent "plug-in" Bayesian decision rule.*

The "plug-in" BDR of group classification can be used to forecast future states of complex system for a given horizon $h \geq 1$ using new out-of-sample observations $(\bar{X}_{T+1}^{T+h}, \bar{U}_{T+1}^{T+h})$, where $\bar{X}_{T+1}^{T+h} = (x'_{T+1}, \ldots, x'_{T+h})' \in \Re^{Nh}$, $\bar{U}_{T+1}^{T+h} = (u'_{T+1}, \ldots, u'_{T+h})' \subseteq \Re^{(Np+M)h}$.

5. Performance evaluations

Description of test models and examples. We consider the model MS-VARX in the form (1) or (3) under the assumptions M.1–M.3, d.2 with cyclic changes in the matrix of

regression coefficients. The aim of experiments is to evaluate the accuracy of classification and prediction for the proposed decision rules. We use the following notation for the proposed classification algorithms for MS-VARX: BDR — Bayesian decision rule of groupwise classification algorithms; EBDR — estimated ("plug-in") BDR algorithms; EM — EM MS-VARX algorithms.

General description of the test models: $L = 2$, $N = 2$, $M = 3$; $A_1 = A_0$, $\Sigma_0 = \Sigma_1 = \Sigma$; $T \in \{100, 200, 500, 1000, 2000\}$, $h = 100$. The exogenous vector $z_t = (z_{tj}) \in \Re^M$ has a uniform distribution in $\mathfrak{Z} = a^M$, $a = [1, 10]$ with mean value $\tilde{z} = \mathbf{E}\{z\} = (5.5, 5.5, 5.5)'$. Interclass Mahalanobis distance defined for the mean value of the vector of exogenous variables is denoted by $\Delta(\tilde{z})$.

Parameter values for various experiments

$$\Sigma_0 = \Sigma_1 = \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}; \quad B_0 = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 0 & 3 \end{pmatrix}, \quad B_1 = B_0 + H;$$

$$B.1.\ H = \begin{pmatrix} 0 & 0 & 0 \\ -0.5 & 0 & 0 \end{pmatrix}, \quad B.2.\ H = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 1 \end{pmatrix}, \quad B.3.\ H = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & -1 \end{pmatrix};$$

$$\pi_0 = \pi_1 = 0.5, \quad P = \begin{pmatrix} 1-\omega & \omega \\ \omega & 1-\omega \end{pmatrix} \quad (0 < \omega < 0.5).$$

Characteristics of classification and estimation accuracy. The matrix $H = B_1 - B_0$ in the case $A_1 = A_0$, $\Sigma_0 = \Sigma_1 = \Sigma$ determines the degree of distinctiveness of classes, caused by structural changes in the matrix of regression coefficients. The probability of misclassification under the model assumptions is calculated according to the formulas (Malugin 2014; Kharin 1996):

$$r(\tilde{z}) = \pi_0 r_0(\tilde{z}) + \pi_1 r_1(\tilde{z}), \quad r_l(\tilde{z}) = \Phi\left(-\frac{\Delta(\tilde{z})}{2} - (-1)^l \frac{h}{\Delta(\tilde{z})}\right), \quad h = \ln\frac{\pi_0}{\pi_1}\ (l \in \{0, 1\}),$$

where $\Phi(\cdot)$ — the function of standard normal distribution, $\Delta(\tilde{z})$ — interclass Mahalanobis distance at point $\tilde{z}$.

The probability of misclassification is calculated by averaging the classification results of $K = 100$ random samples for each set of parameters using the formulas $\hat{r} = K^{-1}\sum_{i=1}^{K} \hat{r}_i$, $\hat{r}_i = 1 - T^{-1}\sum_{t=1}^{T} \delta_{\hat{d}_t^i, d_t^0}$ where $D^0 = (d_t^0)$, $\hat{D}^i = (\hat{d}_t^i)$ — true state vector and its estimate respectively for the i-th sample.

The accuracy of the parameter estimates is determined by the characteristics $\delta_\theta = ||\hat{\theta} - \theta||$, $\delta_P = ||\hat{P} - P||$, where $||\cdot||$ is the Euclidean norm of the matrix and vector.

Analysis of the results of experiments.

Case 1. The impact of differences in matrix of regression and autoregression coefficients for different classes. Parameters value (set 1): variants B.1–B.3 for the matrix of regression coefficients, $A_1 = A_2 = \mathrm{O}_{N\times N}$, $\omega = 0.2$. The estimates of accuracy measures for these experiments are presented in Table 1.

Table 1: The impact of structural changes in regression coefficients.

Variants of matrix B	Accuracy of classification and estimation algorithms					
	$\Delta(\tilde{z})$	$\hat{r}_{BDR}$	$\hat{r}_{EM}$	$\hat{r}^h_{EBDR}$	δ_θ	δ_P
B.1	1.23	0.198	0.294	0.34	0.265	0.28
B.2	2.46	0.097	0.1	0.109	0.191	0.073
B.3	4.919	0.017	0.02	0.018	0.166	0.059

Parameters values (set 2): variant B.3 for matrix of regression coefficients, $\omega = 0.2$,

$$A.1.\ A_1 = A_0 = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}; \quad A.2.\ A_0 = A_1 = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6 \end{pmatrix} \quad (the\ same\ matrices);$$

$$A.3.\ A_0 = -A_1 = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6 \end{pmatrix}; \quad A.4.\ A_0 = -A_1 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix} \quad (different\ matrices).$$

The estimates of accuracy measures for these experiments are presented in Table 2.

Table 2: The impact of structural changes in autoregression coefficients.

T	A.1		A.2		A.3		A.4	
	$\hat{r}_{BDA}$	$\hat{r}_{EM}$	$\hat{r}_{BDA}$	$\hat{r}_{EM}$	$\hat{r}_{BDA}$	$\hat{r}_{EM}$	$\hat{r}_{BDA}$	$\hat{r}_{EM}$
100	0.0077	0.0787	0.0077	0.0588	0.0013	0.0015	0.0001	0.0049
200	0.0074	0.0128	0.0074	0.0082	0.0012	0.0013	0.0002	0.0002

Conclusion 1. The accuracy of classification depends on the number of parameters, subject to structural changes and the severity of structural changes; the presence of structural changes in the matrices of autoregression coefficients leads to a decrease in the probability of misclassification (compare the values of $\hat{r}_{BDA}$ and $\hat{r}_{EM}$ for the cases A.2. ($A_0 = A_1$) and A.4. ($A_0 = -A_1$) in Table 2).

Case 2. The impact of training sample size T on the accuracy of the algorithms. Parameters value: variant B.2 for the matrix of regression coefficients; $T \in \{100, 200, 500, 1000, 2000\}$, forecast horizon $h = 100$. The dependence of the accuracy of classification and prediction on the size of training sample is illustrated in Figure 1.

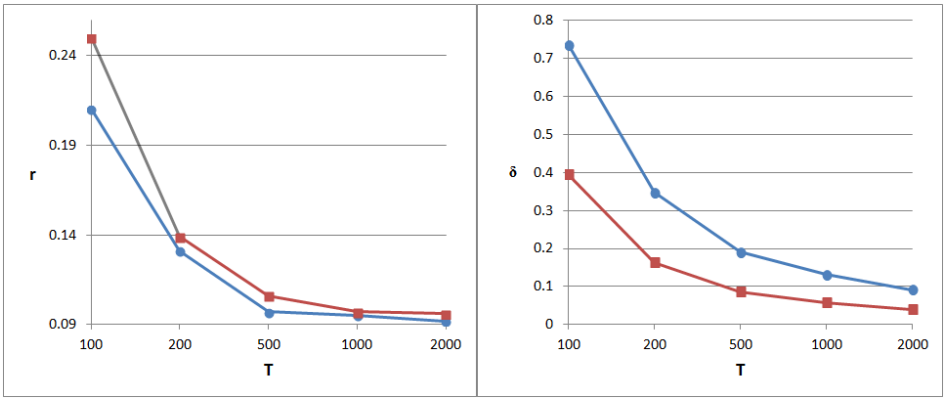

Figure 1: The dependence of the accuracy of algorithms on the size of training sample: left — $\hat{r}_{EM}$ (circles) and $\hat{r}^h_{EDA}$ (squares); right — δ_θ (circles) and δ_P (squares)

Conclusion 2. There is observed an expected rise in the accuracy of the classification and estimation algorithms with increasing interclass distance and volume of observations (Figures 1, 2, Table 1);

Case 3. The effect of uncertainty regarding the class of state on the efficiency of the EM MS-VARX algorithm. Parameters value: under conditions of Case 2 the uncertainty of state is described by the parameters $\omega \in \{0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5\}$, $T = 100$. The value $\omega = 0.1$ corresponds to the high degree of certainty, the value $\omega = 0.5$ corresponds to the highest degree of uncertainty. The dependence of the accuracy of classification and estimation on the parameter ω is illustrated in Figure 2.

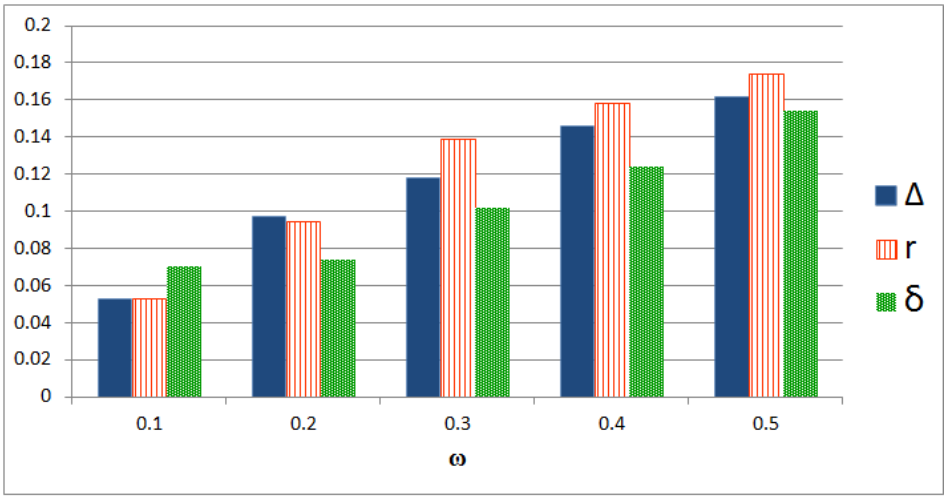

Figure 2: The effect of uncertainty regarding the class of the EM MS-VARX algorithm (columns from left to the right): interclass distance $\Delta(\tilde{z})$; estimate of the probability of misclassification $\hat{r}_{EM}$; characteristics of parameters estimation accuracy δ_θ

Conclusion 3. The increasing degree of uncertainty regarding the state of the system have the following effects for the EM MS-VARX and EDBR algorithms: interclass distance decreases and the probability of misclassification falls significantly (compare the values of $\hat{r}_{EM}$ for the cases $\omega = 0.1$ and $\omega = 0.5$ in Figure 2). This indicates the feasibility of using in these cases the IS MS-VARX algorithm (Malugin 2014) for independent classes of states.

References

Anderson T (1984). *An Introduction to Multivariate Statistical Analysis.* Wiley.

Bellman R, Dreyfus S (1962). *Applied Dynamic Programming.* Princeton University Press Princeton, New Jersey.

Bilmes J (1998). *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* Int. Computer Science Institute, Berkeley CA.

Hamilton J (2008). "Regime-switching Models." *New Palgrave Dictionary of Economics*, pp. 1755–1804.

Kharin Y (1996). *Robustness in Statistical Pattern Recognition.* Dordrecht, Boston, London Kluwer Academic Publishers.

Krolzig H (1997). *Markov Switching Vector Autoregressions. Modelling Statistical Inference and Application to Business Cycle Analysis.* Berlin, Springer-Verlag.

Lutkepohl H (2005). *New Introduction to Multiple Time Series Analysis.* Berlin, Springer-Verlag.

Malugin V (2014). *Methods of Analysis of Multivariate Econometric Models with Heterogeneous Structure.* Minsk, Belarusian State University.

Malugin V, Kharin Y (1986). "On Optimal Classification or Random Observations Different in Regression Equations." *Automation and Remote Control*, (7), 61–69.

Affiliation:

Vladimir Malugin
Department of Mathematical Modeling and Data Analysis
Belarusian State University
220030 Minsk, Belarus
E-mail: malugin@bsu.by
URL: http://fpmi.bsu.by/en/main.aspx?guid=24341

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 46
April 2017

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2016-11-15
Accepted: 2017-02-02