A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 255, 2011

Marcin Hławka[*], *Maciej Kawecki*[**]

MODEL SELECTION CRITERIA FOR REDUCED RANK MULTIVARIATE TIME SERIES WITH APPLICATION IN IDENTIFICATION OF PERIODIC COMPONENTS

Abstract. The main focus of this paper is to present an application of different model selection criteria used for multivariate reduced rank time series. We consider one of the most commonly used reduced-rank models, Regularized Reduced Rank Vector Autoregression (RRRVAR(p,r,λ)). In our study, the most popular model selection criteria are included. The criteria are divided into two groups: simultaneous selection and two-step selection criteria, accordingly.

We applied RRRVAR model in the task of an identification of periodic components in high-dimensional, short and often highly-correlated multivariate time series. Additionally, we proposed a universal and well-parameterized simulation framework which allows to mimic almost any scenario that may occur in real experimental settings. Moreover, efficiency of all methods is compared using well-known time-course Spellman (1998) microarray data, used to find cell-cycle regulated yeast genes.

Key words: multivariate time series, periodicity, microarray experiment, spectral analysis, regularization, reduced rank model, model selection criteria.

I. INTRODUCTION

Recent progress in high-throughput technologies in molecular biology resulted in a vast amount of experimental data on the one hand, and brought many analytical challenges, on the other. In particular, time-course microarray assays have been found useful in answering a broad spectrum of biological problems (see e.g. Andersson (2006)). One of the scientific problems solved with the aid of time-course microarray experiments is to find genes which are cell-cycle regulated (Spellman (1998)). From a statistical perspective, this is equivalent to finding periodically expressed genes, i.e. periodic components of multivariate microarray times series.

In this paper, we address the problem of identification of periodic components in high dimensional, short and often highly-correlated multivariate time series. As mentioned above, this problem is strongly motivated by time-

[*] MSc., Institute of Mathematics and Computer Science, Wroclaw University of Technology.
[**] MSc., Institute of Mathematics and Computer Science, Wroclaw University of Technology.

course microarray experiments in molecular biology, however it does occur in other areas, e.g. economic or financial multivariate time series, where periodicity of components has to be taken into account. The task is often exacerbated by the so-called large p, small n problem (p<<n), large number of components and only short time horizon. In typical microarray assay, we observed several thousands of components and at most few tens of measurements (i.e. time points). Under these circumstances, one cannot expect traditional statistical models or tests (used to identify periodic components) to perform well (see i.e. Wichert (2004)). In this case, it is necessary to develop new, more efficient methods.

In this paper we will assume that time series $Y(t)$ can be represented as follow: for each $i = 1,2,\ldots,M$ each variable $Y_i(t)$ of a multivariate time series of length T is defined as:

$$Y_i = a_{i,0} + a_{i,1} f_i(t + \tau_i) + Z_i(t), \tag{1}$$

where f_i is a periodic function, $a_{i,0}$ (where $a_{i,0} \in \Re$) is a mean of the i-th component, $a_{i,1}$ (where $a_{i,1} \geq 0$) is an amplitude of a periodic component, τ_i is a phase shift ($\tau_i \in [0,T]$) and $Z_i(t)$ is a white noise with variance σ^2. We note, that $Z_i(t)$ may be correlated.

With such assumptions our goal is to proceed a binary classification, that means assign for each component value 0 – nonperiodic component or 1, which means periodic component. We can obtain it by determining appropriate scoring for each component and then using scoring classification methodology. In next paragraph, we described different techniques to determine scoring of periodicity.

II. DIFFERENT APPROACHES IN FINDINING PERIODIC COMPONENTS

Methods developed recently we can classify into three groups. This classification takes into the account the way of consideration of correlation of error matrix and using additional prior knowledge:

1. Univariate methods (not using prior knowledge) - analyses every component separately (independently from others). We assume that the data suits model (1) with additional constraint $cor(Z_i(t), Z_j(t)) = 0$ for $i \neq j$.

- classical spectral analysis (Wichert(2004))
- Bayesian approach (Andersson(2006))

2. Multivariate methods (not using prior knowledge) - we have taken into the account correlation between components

- combined test for regulation and periodicity (Lichtenberg(2005))
- RRRVAR model (Zagdanski(2008))

3. Methods using prior knowledge (both univariate and multivariate).

- B-spline (Lichtenberg(2005))
- correlation test (Spellman(1998))

In next section, we describe in details new methods based on fitting RRRVAR model.

III. FITTING RRRVAR MODEL

Regularized Reduced Rank Vector Autoregressive model (RRRVAR) is based on very famous and well known VAR model. Multidimensional time series Y_t is called Vector Autoregression of order p (VAR(p)), if:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} + Z_t , \tag{2}$$

where: $Y_t = (Y_{1t}, Y_{2t} \ldots, Y_{Kt})^T$, Φ_i stand for $K \times K$ coefficient matrices and $Z_t = (Z_{1t}, Z_{2t} \ldots, Z_{Kt})^T$ is a noise vector.

One of the simplest method of reducing VAR model complexity is a reduction of rank of each coefficient matrix Φ_i to r instead of K, which is particularly important when we analyze time series with many components. After reducing rank and imposing normalization condition (Velu, 1986) we have to estimate only $r[K(p+1)-r]$ coefficients. Such simplified model is called Reduced Rank Vector Autoregression (RRVAR) of order p and rank r. and is defined as (Velu, 1986): $Y_t = ABX_t + Z_t$, (for $t = 1,2,\ldots T$), where: $X_t = (Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})$ is a vector $Kp \times 1$, A denotes matrix of size $K \times r$, $B = (B_1, B_2 \ldots, B_p)$ is matrix $r \times Kp$, B_i matrices $r \times K$. Note that the matrices $\Phi_i = AB_i$ defined in (2) (for $i = 1,2,\ldots, p$) have rank $r \le \min(K,T)$. Next, we will focus on a fundamental task of choice parameters of the model. We consider two main categories of methods using to select parameters *p* and *r*.

First type of methods in two steps selects parameters of model RRVAR. In the first step, we assume that our time series comes from VAR model and using classical method we determine autoregressive order (i.e. AIC, SC, HQ, FPE, CV and others). In this case, when we have already known the parameter *p,* we choose the rank of coefficient matrices r by statistical tests: Bartlett test or Bartlett-Lawley test. Starting with the null hypothesis of $r = 1$, a sequence of

tests is performed. If the null hypothesis is rejected, r is augmented by one and the test is repeated (Camba-Mendez(2003)).

For the second type of methods, using criteria we choose simultaneously both the autoregressive order and the rank of matrix. To this type of techniques belong criteria based on the likelihood function (Lutkepohl(2005)) and the prediction error (Lutkepohl(2005)).

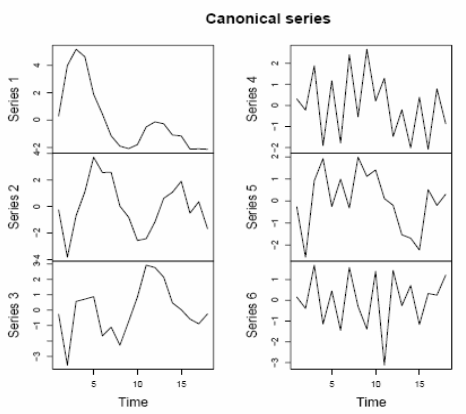
Now, let us consider the problem of estimation coefficient matrix in RRVAR model. The population of coefficient matrices A and B of rank r are given: $\hat{A}(p,r) = \hat{\Omega}^{-1/2}\hat{V}$ and $\hat{B}(p,r) = \hat{V}^T\hat{\Omega}^{1/2}\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}{}^{-1}$, where: $\hat{V} = (\hat{V}_1,\ldots\hat{V}_r)$ and $\hat{V}_i$ is the eigenvector corresponding to the i-th largest eigenvalue of matrix $W = \hat{\Omega}^{1/2}\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}{}^{-1}\hat{\Sigma}_{XY}\hat{\Omega}^{1/2}$. Matrix $\Sigma_{XX} = Cov(X_t, X_t)$ has the diagonal submatrix $\Gamma = \Gamma(0) = E(X_t X_t)^T$. In our problem (i.e. short time series with many components) the estimator of matrix Σ_{XX} is usually not of a full rank (so not invertible). Zagdanski and Kustra (2008) propose use of penalized version, i.e. just replace $\hat{\Gamma}$ by $\hat{\Gamma} = \hat{\Gamma} + \lambda I$, where λ is a penalty coefficient. In this way, we obtain RRRVAR model. The parameter λ is not the most important, so we propose a selection of a such minimal parameter, which allows to invert matrix Σ_{XX} numerically. In this way, the problem of fitting the best model is reduced to the fitting appropriate RRVAR(p,r) model. Considered model is strictly associated with canonical analysis of time series. This connection caused, that RRVAR model is so useful. When the rank of matrix $C = AB$ is equal to r, then the $K - r$ of the smallest eigenvalues is equal to zero and for corresponding eigenvectors we have $\omega_j(t) = l_j Y_t = 0$ ($l_j = \Sigma_{ZZ}^{-1/2} V_j$). On the other hand, the r of the first canonical series assigns the main patterns in the analyzed multidimensional time series. The analysis of the canonical series allowed us also to identify and determine periodic trends. By using canonical weights we could affirm which components are seasonal. This leads us to the following procedure generating scoring of periodicity: test, which canonical time series are periodic (get r series), then choose canonical weights of these series (obtain matrix $T \times r$) and compute maximum of the absolute value of weights for each row, receiving „scoring of periodicity".

IV. CASY STUDY – YEAST CELL DIVISION

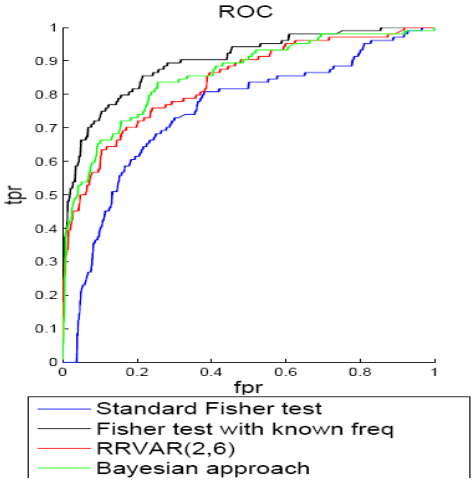
The methods described in the previous sections, we have applied in the problem of identification of periodic components. We have used a benchmarking time series for cell-cycle division of the yeast (Lichtenberg(2005)). By taking into the consideration a computational complexity, we have chosen randomly

subset of 1500 genes, including 104 genes, which were classified as a periodic in the previous experiments (Spellman(1998)).

During simulations we obtained, that the most efficient criterion is the method based on a cross-validation. Next, we have chosen the optimal parameter $\lambda = 4$. Finally we have got the RRRVAR(2,6) model. In Picture (1) we present 6 canonical time series. Statistical test (Fisher G-test) proves, that the series with numbers 2 and 3 are significantly periodic. Moreover, the analysis of the canonical series has yielded additional knowledge about a frequency, which equals to 2 and a lack of the other trends. This is an extra a-prior information, which can be used while applying different methods, described in the second section of this paper.



Picture 1: The first 6 canonical series for Spellman dataset



Picture 2: Comparison of ROC curves for selected methods

Basing on the periodic canonical profiles, we have got the scoring of periodicity, which efficiency was examined by Receiver Operating Curves (ROC). Picture (2) shows a comparison between RRRVAR method and 3 different techniques proposed in the literature. We can say that, RRRVAR is a method, which does not use any of a-prior knowledge and is comparable to other methods.

V. DISCUSION

RRRVAR can be used as a promising method in finding significant patterns in short time series highly correlated with many components. RRRVAR can be also used as a tool in time series clustering. There are many further topics to be developed, i.e. research into choosing optimal parameter lambda or modified simultaneous selection criteria of parameters lambda, p and r. We can see a necessity of developing procedure, which would combine RRRVAR methods with other methods and finally would get a "combined scoring of periodicity".

REFERENCES

Anderson T. W. (2002), Canonical Correlation Analysis and Reduced Rank Regression in Autoregressive Models, *The Annals of Statistics*.

Andersson C. R., Isaksson A., Gustafsson M. G. (2006), Bayesian detection of periodic mRNA time profiles without use of training examples, *Bioinformatics*, Vol. 7, pp. 7–63.

Camba-Mendez G., Kapetanios G., Smith R. J., Weale M. R. (2003), Test of Rank in Reduced Rank Regression Models, *Journal of Business & Economic Statistics*, Vol. 21.

Lichtenberg U. (2005), Comparison of computational methods for the identification of cell cycle-regulated genes, *Bioinformatics*, Vol. 21, pp. 1164–1171.

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

Spellman P. T. et.al. (1998), Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization, *Molecular Biology of the Cell*, Vol. 9, pp. 3273–3297.

Velu R. P., Reinsel G. C., Wichern D. W. (1986), Reduced Rank Models for Multiple Time Series, *Biometrika*, Vol.73, pp. 105–118.

Wichert S., Fokianos K., Strimmer K. (2004), Indentifying periodically expressed transcripts in microarray time series, *Bioinformatics*, Vol. 20, pp. 5–20.

Zagdanski A., Kustra R. (2008), Exploration of High-dimensional Time Series Using Regularized Reduced Rank Approach: Application in Time-course Microarray. Proceedings of IADIS European Conference on Data Mining (ECDM2008), part of the IADIS Multi Conference on Computer Science and Information Systems 2008, Amsterdam

Marcin Hławka, *Maciej Kawecki*

KRYTERIA WYBORU MODELU O ZREDUKOWANYM RZĘDZIE W WIELOWYMIAROWYCH SZEREGACH CZASOWYCH Z ZASTOSOWANIEM W METODACH IDENTYFIKACJI SKŁADOWYCH OKRESOWYCH

W pracy jest przedstawione zastosowanie kryteriów wyboru modelu dla wektorowego modelu autoregresji o zredukowanym rzędzie (Reduced Rank Vector Autoregression (RRVAR(p,r)). W analizie uwzględniono najbardziej popularne kryteria wyboru modelu, podzielone na dwie grupy: kryteria równoczesnego wyboru oraz tzw. kryteria dwukrokowe.

Model RRVAR został użyty w zagadnieniu identyfikacji składowych okresowych dla wielowymiarowych szeregów czasowych, zawierających dużą liczbę, zazwyczaj istotnie skorelowanych składowych, obserwowanych w krótkim horyzoncie czasowym. Przedstawione zostaną rezultaty porównujące efektywność metody opartej na dopasowaniu wektorowego modelu autoregresji o zredukowanym rzędzie z tradycyjnymi jednowymiarowymi metodami. Wykorzystano bazę rzeczywistych danych mikromacierzowych Spellman'a (1998), służącą do identyfikacji genów drożdży, związanych z cyklem podziału komórki.