This collection of papers includes proceedings of the Tenth International Conference “Computer Data Analysis and Modeling: Theoretical and Applied Stochastics” organized by the Belarusian State University and held in September 2013 in Minsk. Papers are reviewed by qualified researchers from Belarus, Russia, Austria, Great Britain, Germany, Lithuania.

The papers are devoted to the topical problems: robust and multivariate data analysis; statistical analysis of time series and stochastic processes; probabilistic and statistical analysis of discrete data; computer simulation of stochastic systems; computer data analysis in applications; econometric and financial analysis and modeling; survey analysis and official statistics.

For specialists who work in the fields of mathematical statistics and its applications, computer data analysis, statistical modeling and statistical software development.

© BSU, 2013
PROGRAM COMMITTEE
HONORARY CHAIRS
S. Ablameyko, Rector of the Belarusian State University (Minsk, Belarus)
V. Zinovskiy, Chair of the National Statistical Committee (Minsk, Belarus)

CO-CHAIRS
Prof. S. Aivazian (Moscow, Russia)
Prof. P. Filzmoser (Vienna, Austria)
Prof. Yu. Kharin (Minsk, Belarus)

MEMBERS
W. Charemza (Leicester, United Kingdom) B. Lemeshko (Novosibirsk, Russia)
A. Christmann (Bayreuth, Germany) Yu. Mishura (Kiev, Ukraine)
D. Chibisov (Moscow, Russia) S. Morgenthaler (Lausanne, Switzerland)
C. Croux (Leuven, Belgium) V. Mukha (Minsk, Belarus)
K. Ducinskas (Klaipeda, Lithuania) H. Oja (Turku, Finland)
R. Dutter (Vienna, Austria) Yu. Pavlov (Petrozavodsk, Russia)
G. Dzemyda (Vilnius, Lithuania) G. Shevlyakov (St.-Petersburg, Russia)
A. Egorov (Minsk, Belarus) N. Troush (Minsk, Belarus)
H. Friedl (Graz, Austria) A. Tuzikov (Minsk, Belarus)
I. Gaishun (Minsk, Belarus) R. Viertl (Vienna, Austria)
K. Hron (Olomouz, Czech Republic) N. Yanev (Sofia, Bulgaria)
V. Krasnoproschin (Minsk, Belarus) A. Zubkov (Moscow, Russia)

LOCAL ORGANIZING COMMITTEE
CO-CHAIRS
P. Mandrik, Dean of the Faculty of Applied Mathematics and Informatics
Yu. Kharin, Director of the Research Institute for Applied Problems of Mathematics and Informatics

MEMBERS
V. Malugin – conference secretary, invitation letters, program
H. Aheyeva – visa support
I. Bodiagin – technical organizing of sessions
O. Golos – registration
A. Kapusta, M. Mitskevich – conference web site
A. Kharin – e-correspondence, reviewing process
A. Shilkin – publishing
M. Maltsew, E. Vechorko – transport service
E. Melnikova – finance, registration
E. Orlova, S. Staleuskaya, N. Shenets – social program
I. Pirshtuk, V. Voloshko – accommodation, arrival, departure
To the 25th Anniversary
of the Department
of Mathematical Modeling
and Data Analysis
of the Belarusian State University
GREETINGS
FROM THE PROGRAM COMMITTEE
CO-CHAIRS

S.A. Aivazian, P. Filzmoser, Yu.S. Kharin
Central Economics and Mathematics Institute of RAS, Moscow, RUSSIA
Vienna University of Technology, Vienna, AUSTRIA
Belarusian State University, Minsk, BELARUS

e-mail: aivazian@cemi.rssi.ru,
P.Filzmoser@tuwien.ac.at,
kharin@bsu.by

Dear Participants
of the International Conference on Computer Data Analysis and Modeling!

Dear Friends and Colleagues!

The history of our CDAM Conferences goes from the year 1988, when the first of them have been organized in a picturesque place on the east part of Minsk close to a small river. Since that time 25 years passed, and the Conference is already the 10th in the sequence.

Looking to the quantity of participants, to the quality of their papers in the Proceedings, we would like to tell that the Conference turned into the successful and well-known scientific forum.

This year the Conference focuses on the problems of theoretical and applied stochastics. The number of papers included into the Proceedings is 100, their depth and spectrum are quite impressive.

Let us thank all of you for your contribution to the CDAM Conference and wish you interesting lectures and talks, fruitful discussions, new friends and enjoyable time for CDAM 2013 scientific forum!

Minsk, September 10, 2013
The Tenth International Conference “Computer Data Analysis and Modeling: Theoretical and Applied Stochastics” (CDAM 2013) organized by the Belarusian State University and Vienna University of Technology on September 10-14, 2013, is devoted to the topical problems in computer data analysis and modeling. The methods to be discussed during the Conference work for development of the theory and are applied in different areas influenced by stochastics: decision making in economics, engineering, medicine, ecology and others. The previous CDAM Conferences were held in September 1988, December 1990, December 1992, September 1995, June 1998, September 2001, September 2004, September 2007 and September 2010 in Minsk.

The Proceedings of the CDAM 2013 Conference include two volumes containing 100 papers by more than 120 participants from more than 20 countries. The topics of the papers cover the following research directions: Robust and Multivariate Data Analysis; Statistical Analysis of Time Series and Stochastic Processes; Probabilistic and Statistical Analysis of Discrete Data; Computer Simulation of Stochastic Systems; Computer Data Analysis in Applications; Econometric and Financial Analysis and Modeling; Survey Analysis and Official Statistics.

The Organizing Committee of the CDAM 2013 Conference makes it acknowledgements to: Belarusian State University, Vienna University of Technology, Research Institute for Applied Problems of Mathematics and Informatics of the Belarusian State University, Faculty of Applied Mathematics and Informatics of the Belarusian State University, Belarusian National Science Foundation, Ministry of Education of the Republic of Belarus, BPS-Sberbank, BelSwissBank, Belarusian Science and Technology Association “Infopark” for the support in organizing of the Conference.

S.A. Aivazian,
P. Filzmoser,
Yu.S. Kharin
## CONTENTS

### PLENARY LECTURES

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aivazian S., Afanasiev M., Rudenko V.</td>
<td>Some specification aspects for three-factor models of a company’s production potential taking into account intellectual capital</td>
<td>11</td>
</tr>
<tr>
<td>Bernatavičienė J., Bazilevičius G., Dzemyda G., Medvedev V.</td>
<td>Comparative analysis of similarity measures for medical streaming data</td>
<td>17</td>
</tr>
<tr>
<td>Charemza W., Díaz C., Makarova S.</td>
<td>Too many skew normal distributions? The practitioneer’s perspective</td>
<td>21</td>
</tr>
<tr>
<td>Filzmoser P.</td>
<td>A projection–pursuit method for sparse robust PCA</td>
<td>32</td>
</tr>
<tr>
<td>Fišerová E., Kubáček L.</td>
<td>Some remarks on decomposition of partitioned multivariate models into two seemingly unrelated submodels</td>
<td>36</td>
</tr>
<tr>
<td>Francq C., Meintanis S.G.</td>
<td>Fourier–type estimation of the power GARCH model with stable–paretian innovations</td>
<td>42</td>
</tr>
<tr>
<td>Fried R., Elsaied H., Liboschik T.</td>
<td>On outliers and interventions in count time series following GLMs</td>
<td>50</td>
</tr>
<tr>
<td>Kharin A. Yu.</td>
<td>Sequential testing of parametric hypotheses: performance, robustness and application</td>
<td>58</td>
</tr>
<tr>
<td>Kharin Yu. S.</td>
<td>Long–memory discrete–valued time series models and methods</td>
<td>64</td>
</tr>
<tr>
<td>Kolchin A. V., Ionkina E. G.</td>
<td>On acquisition of nociceptive evoked potentials in rats cerebral cortex</td>
<td>72</td>
</tr>
<tr>
<td>Leri M.M., Pavlov Yu. L.</td>
<td>Power-law graphs robustness and forest fires</td>
<td>74</td>
</tr>
<tr>
<td>Malugin V.</td>
<td>Testing for structural heterogeneity in vector autoregressive model by means of statistical classification methods</td>
<td>78</td>
</tr>
<tr>
<td>Mishura Y. S., Ralchenko K. V., Shevchenko G. M.</td>
<td>Parameter estimation in the models with long-range dependence</td>
<td>83</td>
</tr>
<tr>
<td>Perrotta D., Riani M., Cerioli A., Torti F.</td>
<td>Scalability properties of a new forward search algorithm</td>
<td>87</td>
</tr>
<tr>
<td>Rudzikis R., Bakshaev A.</td>
<td>Goodness of fit tests based on kernel density estimators</td>
<td>89</td>
</tr>
</tbody>
</table>
SECTION 1
ROBUST AND MULTIVARIATE DATA ANALYSIS

Andrea K., Shevlyakov G.L., Smirnov P.O. Detection of outliers with boxplots ................................................................. 141
Chernov S.Yu., Kharin A.Yu. Error probabilities for sequential testing of simple hypotheses under functional distortions in the $L_2$-metric ........ 145
Doronin A.V. Semiparametric estimation in mixture models with varying mixing probabilities ......................................................... 149
Esaulov D.M. Application of residual empirical processes to robust linear hypotheses testing in autoregression ......................................................... 153
Kirlitsa V.P. On D- and A-optimal designs of experiments ........................................ 157
Lobach V.I. On robust Kalman filtering with using wavelet analysis .................. 161
Luz M.M., Moklyachuk M.P. Robust interpolation problem for stochastic sequences with stationary increments .............................................. 165
<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maltsew M.V.</td>
<td>On statistical estimation for Markov chain of conditional order</td>
</tr>
<tr>
<td>Mert C., Filzmoser P., Hron K.</td>
<td>Sparse principal balances for high-dimensional compositional data</td>
</tr>
<tr>
<td>Shiryaeva L.K.</td>
<td>The joint distribution of the standardized maximum and standardized minimum for a normal sample</td>
</tr>
<tr>
<td>Zalesky B.A.</td>
<td>Regression smoothing Kalman filter</td>
</tr>
<tr>
<td>Zhuk E.E.</td>
<td>Assignment of multivariate samples to the fixed classes by the maximum likelihood method and its risk</td>
</tr>
</tbody>
</table>

**SECTION 2**

**STATISTICAL ANALYSIS OF TIME SERIES AND STOCHASTIC PROCESSES**

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Badziahin I.A.</td>
<td>On parameters estimation of autoregressive time series under right censoring</td>
</tr>
<tr>
<td>Boldin M.V.</td>
<td>On qualitatively robust sign test in random walk model</td>
</tr>
<tr>
<td>Dubovetska I.I., Moklyachuk M.P.</td>
<td>Minimax extrapolation problem for periodically correlated processes</td>
</tr>
<tr>
<td>Dučinskas K., Karaliutė M., Šimkienė I.</td>
<td>Misclassification probability based on linear discriminant function for SAR error models</td>
</tr>
<tr>
<td>Egorov A.D.</td>
<td>Approximate formulas for expectation of functionals from solution to stochastic equation with random measure</td>
</tr>
<tr>
<td>Lobach S.V.</td>
<td>Prediction of time series based on state space model</td>
</tr>
<tr>
<td>Mukha V.S., Prishchepchik M.V.</td>
<td>Modeling of spatial queues system</td>
</tr>
<tr>
<td>Tomashyk V.V.</td>
<td>Properties of optimal stopping and exit times for diffusion processes and random walks</td>
</tr>
<tr>
<td>Valoshka V.A., Kharin Yu.S.</td>
<td>On comparison of classical and cepstrum–based forecasts</td>
</tr>
<tr>
<td>Yanovich L.A.</td>
<td>On formulas of linear interpolation for operators generated by solutions of stochastic differential equations</td>
</tr>
</tbody>
</table>
PLENARY LECTURES
SOME SPECIFICATION ASPECTS FOR THREE-FACTOR MODELS OF A COMPANY’S PRODUCTION POTENTIAL TAKING INTO ACCOUNT INTELLECTUAL CAPITAL

S. Aivazian¹, M. Afanasiev², V. Rudenko
Central Economics and Mathematics Institute of RAS
Moscow, RUSSIA

e-mail: ¹ aivazian@cemi.rssi.ru; ² afanasiev@cemi.rssi.ru

Abstract

We propose a general algorithm that gives a solution of some problems related to specification of 3-factor stochastic models of a company’s production potential that take into account main production factors. The proposed formal scheme is based on procedures of statistical hypothesis testing and provides the necessary means to choose a reasonable alternative within the analyzed class of models.

According to [1], a model of production function is a deterministic component of a production potential model. We consider the class of production potential models that is given by relation

\[ R = \beta_0 K^{\beta_2} L^{\beta_3} I^{\beta_3} e^{V-U}, \]

(1)

where \( R \) is the overall production of a company, \( K \) is the physical and financial capital input, \( L \) is the labor input, \( I \) is the intellectual capital input; the random variable (r.v.) \( V \) is normally distributed with parameters \( (0; \sigma_V^2) \) (i.e. \( V \in N(0; \sigma_V^2) \)), the r.v. \( U \) is distributed according to a truncated at zero normal law with a mean value \( \mu \) and a variance \( \sigma_U^2 \) (i.e. \( U \in N^+(\mu; \sigma_U^2) \)). The r.v.’s \( V \) and \( U \) are considered to be stochastically independent. The parameters \( \beta_0, \ldots, \beta_3 \) are subject to statistical estimation.

The main goal of the research is to describe a general algorithm of statistical hypothesis testing that provides answers to the following questions related to the model’s specification.

(a) Is it reasonable to consider the 3-factor model (1) under given methods of intellectual capital measurement (alternative is the standard 2-factor model)?

(b) In case the answer to the question (a) is positive: is there any inefficiency in use of production factors (alternative: \( \sigma_U^2 = 0 \))? 

(c) If the answer to the question (a) is positive and there’s inefficiency in use of inputs (\( \sigma_U^2 > 0 \)): is it reasonable to apply the model (1) with \( \mu = 0 \) (alternative: \( \mu \neq 0 \))? 

(d) Finally, in case the answers to the questions (a) and (b) are positive and there’s possibility to identify indicators \( z^{(1)}, z^{(2)}, \ldots, z^{(p)} \) that can influence efficiency in use of the main production factors: how one can proceed with statistical hypothesis testing related to the dependence character of the parameters \( \mu \) or \( \sigma_U^2 \) of those indicators?

In order to estimate intellectual capital (IC) one can use any known method that meets the requirements of the research. We use the following approach to find out whether it is reasonable to apply one or another way of IC measurement. If the
influence of IC on the overall production of a company is positive and statistically
significant (under positive and statistically significant estimated coefficients for capital
input K and labor input L) we say that the corresponding measurement method is
acceptable for practical purposes. Otherwise the considered way of IC measurement is
said to be inapplicable for the analysis.

Proceeding with econometric analysis of the model (1) we generally have the data
array \( E^2 = \{ R_i, K_i, L_i, I_i, z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(p)} \}_{i=1}^n \), where \( R_i \) is the total production
of the \( i \)th company, \( K_i, L_i, I_i \) are the values of the main production factors for the \( i \)th company; \( z_i^{(1)}, \ldots, z_i^{(p)} \) are the values of the measurable indicators that characterize
efficiency in use of the main production factors for the \( i \)th company; \( n \) is the number of
companies in the sample. If there’s no information about any efficiency indicators
the specification of the model (1) is carried out on the basis of the reduced data array
\( E^1 = \{ R_i, K_i, L_i, I_i \}_{i=1}^n \).

In order to formalize the problems given by items (a) - (d) we should consider the
following models:

\[
M_0: R = \beta_0 K^{\beta_1} L^{\beta_2} I^{\beta_3} \varepsilon^V, \quad \text{where } V \in N(0; \sigma_V^2).
\]

\[
M_1: R = \beta_0 K^{\beta_1} L^{\beta_2} I^{\beta_3} e^{-U}, \quad \text{where } V \in N(0; \sigma_V^2), \quad U \in N^+(0; \sigma_U^2).
\]

\[
M_2: R = \beta_0 K^{\beta_1} L^{\beta_2} I^{\beta_3} e^{-U}, \quad \text{where } V \in N(0; \sigma_V^2), \quad U \in N^+(\mu; \sigma_U^2).
\]

\[
M_3: R = \beta_0 K^{\beta_1} L^{\beta_2} I^{\beta_3} e^{-U}, \quad \text{where } V \in N(0; \sigma_V^2), \quad U \in N^+(0; \sigma_U^2(z)), \quad \ln \sigma_U^2(z) =
\]

\[
\theta_0 + \theta_1 z^{(1)} + \ldots + \theta_p z^{(p)},
\]

\[
M_4: R = \beta_0 K^{\beta_1} L^{\beta_2} I^{\beta_3} e^{-U}, \quad \text{where } V \in N(0; \sigma_V^2), \quad U \in N^+(\mu(z); \sigma_U^2(z)), \quad \mu(z) =
\]

\[
\delta_0 + \delta_1 z^{(1)} + \ldots + \delta_p z^{(p)}.
\]

To find out whether a method of IC measurement is acceptable for practical use
one should test the following hypothesis:

\[
H_0: \text{there exists } i(i \in \{1, 2, 3\}) \text{ such that: } \beta_i \leq 0 \text{ i.e. there exists a production}
\]

factor that is not statistically significant or that provides negative influence on the
total production;

\[
H_0^{(i)}: \beta_i > 0, \quad i = 1, 2, 3, \text{ i.e. all the considered production factors are statistically}
\]

significant and provide positive impact on the overall production.

Testing procedure of the hypothesis \( H_0 \) against the alternative \( H_0^{(i)} \) is based on the
fact that the statistics \( \hat{t}_i = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \) (where \( \hat{\beta}_i \) is an estimate of the coefficient \( \beta_i \) and \( s_{\hat{\beta}_i} \),
is an estimate of the standard deviation in \( \beta_i \) estimation) is distributed according to
\( t(n - k) \) - law.

To get an answer to the question (b) one should test the following statistical hypo-
thesis within the frame of the model \( M_1 \):

\[
H_1: \sigma_U^2 = 0 \text{ (no inefficiency in use of production factors) with the alternative}
\]

\[
H_1^A: \sigma_U^2 > 0 \text{ (inefficiency is observed)}.
\]

Testing of the hypothesis \( H_1 \) (against the alternative \( H_1^A \)) is done on the basis of
asymptotic characteristics of the likelihood ratio statistics (see corresponding results
in [3, 4, 6]).

The choice of a proper model between \( M_1 \) and \( M_2 \) is formalized be the hypothesis:

\[
H_{1,2}: \mu = 0 \text{ for the model } M_2 \text{ (inefficiency in the models } M_1 \text{ and } M_2 \text{ cannot be}
\]

distinguished),
\(H^4_{1,2} : \mu \neq 0\) for the model \(M_2\) (inefficiency in the models \(M_1\) and \(M_2\) can be distinguished).

The corrected Akaike criterion is used in testing of the hypothesis \(H_{1,2}\) against the alternative \(H^4_{1,2}\) (see [5], [2]).

To find out whether the indicators \(z^{(1)}, \ldots, z^{(p)}\) have a real effect on the variance \(\sigma^2_U\) in the model \(M_3\) we test the hypotheses:

- \(H^4_{3,1} : \forall j = 1, \ldots, p : \theta_j = 0\) (influence of all the efficiency indicators in the model \(M_3\) are not statistically significant),
- \(H^4_{3,1} : \exists j = 1, \ldots, p : \theta_j \neq 0\) (there exists at least one statistically significant efficiency factor in the model).

The problem is solved by means of general theory of linear hypotheses testing using the corresponding \(F\) statistics (see [2]).

To form the a posteriori set of efficiency indicators for the model \(M_3\) one should test the hypotheses:

- \(H^4_{3,2} : \exists j = 1, \ldots, p : \theta_j = 0\) (there are some (at least one) non-significant efficiency factors in the model \(M_3\)),
- \(H^4_{3,2} : \forall j = 1, \ldots, p : \theta_j \neq 0\) (all the efficiency indicators in the model \(M_3\) are statistically significant).

Testing procedure of the hypothesis \(H_{3,2}\) (against the alternative \(H^4_{3,2}\)) is based on the fact that the statistics \(\hat{s}_j = \hat{\theta}_j^2 / s^2_{\hat{\theta}_j}\) is distributed according to \(\chi^2(1)\)-law.

Analysis of the model \(M_4\) should clarify whether the efficiency indicators \(z^{(1)}, \ldots, z^{(p)}\) really influence the value \(\mu\) in the distribution \(N^+(\mu; \sigma^2_U)\). The following hypothesis should be tested:

- \(H^4_{4,1} : \forall j = 1, \ldots, p : \delta_j = 0\) (all the efficiency factors in the model \(M_4\) are not statistically significant) against the alternative

  - \(H^4_{4,1} : \exists j = 1, \ldots, p : \delta_j \neq 0\) (there’s at least one statistically significant efficiency indicator in the model \(M_4\)).

To form the a posteriori set of efficiency indicators for the model \(M_4\) one is advised to test the hypotheses:

- \(H^4_{4,2} : \exists j = 1, \ldots, p : \delta_j = 0\) (there are some (at least one) non-significant efficiency factors in the model \(M_4\)),
- \(H^4_{4,2} : \forall j = 1, \ldots, p : \delta_j \neq 0\) (all the efficiency indicators in the model \(M_4\) are statistically significant).

The dependence character between efficiency in use of the main production factors and the indicators \(z^{(1)}, \ldots, z^{(p)}\) can be clarified by testing the following hypotheses

- \(H^4_{2,3} : \mu \neq 0, \sigma^2_U = \text{const}\) (the variance of the component \(U\) should not be expressed via the efficiency indicators),
- \(H^4_{2,3} : \mu = 0, \sigma^2_U = e^{\theta_0 + \theta_1 z^{(1)} + \ldots + \theta_p z^{(p)}}\) (the variance of the component \(U\) should be decomposed by the efficiency indicators under assumption that the mathematical expectation \(\mu\) is equal to 0).

Finally, if one assumes that both parameters of the r.v. \(U\) (\(\mu\) and \(\sigma^2_U\)) might depend on the indicators \(z^{(1)}, \ldots, z^{(p)}\) we recommend to test the hypothesis:

- \(H^4_{3,4} : \mu = 0, \sigma^2_U = e^{\theta_0 + \theta_1 z^{(1)} + \ldots + \theta_p z^{(p)}}\) (the variance \(\sigma^2_U\) (but not the mathemat-
ical expectation $\mu$) should be decomposed by the efficiency indicators, against the alternative

$H_{3,2}^A: \quad \mu = \delta_0 + \delta_1 z^{(1)} + \ldots + \delta_p z^{(p)}, \quad \sigma^2_U = \text{const}$ (the mathematical expectation $\mu$ (but not the variance $\sigma^2_U$) should be decomposed by the efficiency indicators).

To describe the expanded general methodological algorithm that helps to choose a proper model for the class (1) and takes into account availability of the information regarding efficiency indicators we use the following notation:

- Input 1 — start of the analysis with the data array $E^1 = \{R_i, K_i, L_i, I_i\}_{i=1}^n$,
- $M_i$ — calculation of the estimates in the model $M_i$, 
- $M_{i+}$ — the estimates are successfully obtained, 
- $M_{i-}$ — the estimates are not obtained (due to the sample specifics, problems of non-identifiability etc.),
- $H_i$ — application of the hypothesis testing procedure, 
- $H_{i+}$ — the hypothesis is accepted, 
- $H_{i-}$ — the hypothesis is discarded in favor of the alternative $H_{i}^A$,
- $M_i$ — the model $M_i$ is finally chosen, 
- ICNS — conclusion that the used estimate of intellectual capital (IC) is not statistically significant,

- Input 2 — start of the analysis with the data array $E^2 = \{R_i, K_i, L_i, I_i, z_{i}^{(1)}, z_{i}^{(2)}, \ldots, z_{i}^{(p)}\}_{i=1}^n$,
- $\text{ENSF}(z_i)$ — exclusion of the $i$th non-significant efficiency indicator that has the maximum $p$-value in the testing of the hypotheses $H_{3,2}$ and $H_{4,2}$,
- $\text{ENF}(z_i)$ — exclusion of the $i$th efficiency indicator that has the maximum absolute correlation coefficient with intellectual capital indicator,
- Check — check whether there still exist non-excluded efficiency indicators, 
- Check+ — in the analyzed model there still exist non-excluded efficiency indicators, 
- Check- — in the analyzed model all the efficiency indicators are excluded.

As shown at figure 1, the algorithm starts with the model that has the biggest number of variables and provides the widest opportunities for analysis, i.e. with the
model $M_4$. In case there’s at least one statistically significant efficiency indicator in the models $M_4$ or $M_3$ one should choose one of these models as the final result (given that the basic principles of the provided methodology are not violated). If in both models $M_3$ and $M_4$ all the efficiency indicators are not statistically significant or at least one of production factors is not statistically significant the analysis is reduced to the models of production potential that do not take into account the dependence of efficiency in use of production factors of any additional indicators.

In [2] we consider the modeling of production potential for a sample of US companies that operate in the sector “Biotechnology and Drugs”. The sequence of procedures given below leads to a conclusion that one should use the 3-factor model $M_3$ to estimate the production potential:

$$
\begin{align*}
\begin{array}{c}
E^2; M_4; M_4^-; ECF(z^1); \text{Check}^+; M_4; M_4^-; ECF(z^2);
\text{Check}^-; E^2; M_3; M_3^+; H_0; H_0^-; H_3; H_3; H_3; H_3^+; H_3^-

ENSF(z^1); \text{Check}^+; M_3; M_3^+; H_0; H_0^-; H_3; H_3; H_3^+; H_3^-

H_3^2; E^1; M_2; M_2^+; \hat{M}_3
\end{array}
\end{align*}
$$

The detailed description of initial data and the results of calculations done according to the described scheme is given in [2].

**References**


Figure 1: General methodological algorithm of specification problem solution
Abstract

In this paper, we analyse the similarity measures for comparison of medical streaming data (MSD). In general, the real time streaming data can be interpreted as the multivariate time series. In medicine, the comparison of patient physiological MSD data can be useful in disease detection. Four similarity measures Correlation Coefficient, Frobenius norm, Principal Component Analysis similarity factor, Multidimensional Dynamic Time Warping are presented and the experiments of the comparison of these measures have been performed.

1 Introduction

Measuring the similarity between objects described by several features is very important in data mining and decision-making process. Moreover, values of features vary in time. In this case, we have some multivariate time series that characterizes the behaviour of particular object in time.

In this paper, we investigate the comparison problem of multivariate physiological time series. Different similarity measures are used and compared. A physiological time series is a series of some medical observations over a period of time. Such type of data can be collected using devices (or sensors) that collect personal medical features, such as heart rate, blood pressure, etc. When the observed space is multidimensional, the time series becomes multivariate [1].

In general, the problem is to detect events in real time streaming data that can be interpreted as the multivariate time series. Decisions are made in accordance on the previously detected and estimated streaming data. Therefore, the problem is to compare the new real time streaming data with the previously detected one. The comparison problem may be formulated as the search of most similar segment in the previously detected and estimated streaming data to the new real time streaming data.

Denote the previously detected and estimated streaming data and the new real time streaming data as

\[ X^a = \begin{pmatrix} x_{11}^a & \cdots & x_{1T_a}^a \\ \vdots & \ddots & \vdots \\ x_{n1}^a & \cdots & x_{nT_a}^a \end{pmatrix} \quad \text{and} \quad X^b = \begin{pmatrix} x_{11}^b & \cdots & x_{1T_b}^b \\ \vdots & \ddots & \vdots \\ x_{n1}^b & \cdots & x_{nT_b}^b \end{pmatrix}. \]

Here \( T_a \) and \( T_b \) are the numbers of observations, \( T_a > T_b \); \( n \) is the number of measured features. In a result, we need to find the optimal place of \( X^b \) on \( X^a \). The place is defined by some time moment \( T_* : 1 \leq T_* \leq T_a - T_b + 1 \). Similarity measures for multivariate time series are used as the criterion of optimality.
2 Similarity Measures for Multivariate Time Series

Different techniques and similarity measures are introduced and used for comparison of multivariate time series of different nature: [4], [8]. Four similarity measures for MTS are presented here.

The Frobenius norm is often used in matrix analysis [6]. This similarity measure is based on Euclidean distance. Frobenius norm of a matrix $X_b$ is defined by formula:

$$\|X_b\|_F = \sqrt{\sum_{p=1}^{n} \sum_{q=1}^{T_b} (x_{pq}^b)^2} = \sqrt{\text{tr}((X_b^b)^T X_b^b)},$$

here $\text{tr}$ is the sum of the elements on the diagonal of the square matrix. Frobenius norm is used to compare the similarity of two matrices $X_b$ and $X_c$. The similarity is defined by formula $\|X_b - X_c\|_F$. In our case, the matrix $X_c$ is a segment of length $T_b$ in the previously detected and estimated streaming data $X_a$. The best possible value of the Frobenius norm is 0.

Correlation coefficient between two matrices of the same size also can be used as similarity measure [2]:

$$r = \frac{\sum_{p=1}^{n} \sum_{q=1}^{T_b} (x_{pq}^b - \bar{X}_b^b)(x_{pq}^c - \bar{X}_c^c)}{\sqrt{\sum_{p=1}^{n} \sum_{q=1}^{T_b} (x_{pq}^b - \bar{X}_b^b)^2} \sum_{p=1}^{n} \sum_{q=1}^{T_b} (x_{pq}^c - \bar{X}_c^c)},$$

where $\bar{X}_b^b$ and $\bar{X}_c^c$ are the means of the $X_b$ and $X_c$, respectively. The best possible value of correlation coefficient is 1.

The third similarity measure for multivariate time series is Principal Component Analysis (PCA) similarity factor [5], [8]. PCA similarity factor is defined by formula: $S_{PCA}(X_b, X_c) = \text{tr}(L^T MM^T L)$, where $L$ and $M$ are matrices that contained the first $k$ principal components of $X_b$ and $X_c$. The best possible value of the PCA similarity factor is $k$. In our experiments $k = 1$.

Multidimensional Dynamic Time Warping (MDTW) is presented in [3]. Some distance matrix is defined: \(d(p, q) = \sum_{k=1}^{n} (x_{kp}^b - x_{kq}^c)^2, p, q = 1, \ldots, T_b\). Then the matrix $D$ of cumulative distances is calculated as in the traditional DTW algorithm [7]:

$$D(p, q) = \begin{cases} d(1, 1), & \text{if } p = 1, q = 1, \\ d(p, q) + D(p - 1, q), & \text{if } p = 2, \ldots, T_b, q = 1, \\ d(p, q) + D(p, q - 1), & \text{if } p = 1, q = 2, \ldots, T_b, \\ d(p, q) + \min \left\{ \begin{array}{ll} D(p - 1, q) & \text{if } p = 2, \ldots, T_b, q = 1, \\ D(p, q - 1), & \text{if } p = 1, q = 2, \ldots, T_b, \\ D(p - 1, q - 1) & \text{other cases.} \end{array} \right. \end{cases}$$

$(p, q)$ defines the pair of $p$th observation in $X_b$ and $q$th observation in $X_c$. Finally, the minimal path and distance along minimal path is obtained using the matrix $D$. The path must start at the beginning of each time series at $(1, 1)$ and finish at the end of both time series at $(T_b, T_b)$. See [7] for details. The best possible value of the MDTW is 0.
3 Comparative Analysis of Similarity Measures for Multivariate Time Series

Data from PhysioNet/Computing in Cardiology Challenge (http://www.physionet.org/challenge/2012/) is used for the experimental analysis. The records were collected in the Intensive Care Unit. In the experiments we used multivariate time series dataset of 5 patients of the same age, i.e. if to follow the notations of Section 1, we have 5 different matrices $X_i^a$, $i = 1, 5$, consisting of $T_a = 47$ observations of $n = 4$ features (non-invasive diastolic arterial blood pressure, non-invasive systolic arterial blood pressure, heart rate, temperature).

As the new real time streaming data $X^b$, we have chosen first ten observations from one more patient record from PhysioNet data base. For each similarity measure, the optimal place of $X^b$ on $X^a$ has been found. Then the values of remaining measures were computed for the same place of $X^b$ on $X^a$. The place of $X^b$ on $X^a$ may be denoted as follows: $i[T_{start};T_{end}]$, where $i$ is the order number of patient, $i = 1, 5$, $T_{start}$ and $T_{end}$ are start and end positions of $X^b$ on $X^a$. The results are presented in Table. The best-found places of $X^b$ on $X^a$ for different measures are given in bold.

<table>
<thead>
<tr>
<th>$i$</th>
<th>Place of $X^b$ on $X^a$</th>
<th>$|X^b|_F$</th>
<th>$r$</th>
<th>$S_{PCA}$</th>
<th>MDTW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1[6;12]</td>
<td>5.9039</td>
<td>0.7635</td>
<td>0.4986</td>
<td>39.6146</td>
</tr>
<tr>
<td></td>
<td>1[11;17]</td>
<td>5.3568</td>
<td>0.6873</td>
<td>0.4650</td>
<td>17.0764</td>
</tr>
<tr>
<td></td>
<td>1[4;10]</td>
<td>5.7248</td>
<td>0.6423</td>
<td>0.5902</td>
<td>51.0581</td>
</tr>
<tr>
<td>2</td>
<td>2[6;12]</td>
<td>3.4326</td>
<td>0.7935</td>
<td>0.4113</td>
<td>18.2572</td>
</tr>
<tr>
<td></td>
<td>2[31;37]</td>
<td>8.5973</td>
<td>-0.4569</td>
<td>0.8039</td>
<td>26.7820</td>
</tr>
<tr>
<td>3</td>
<td>3[37;43]</td>
<td>2.4704</td>
<td>0.9279</td>
<td>0.8029</td>
<td>29.7421</td>
</tr>
<tr>
<td></td>
<td>3[36;42]</td>
<td>2.8350</td>
<td>0.9260</td>
<td>0.8179</td>
<td>23.2644</td>
</tr>
<tr>
<td></td>
<td>3[41;47]</td>
<td>2.5146</td>
<td>0.9139</td>
<td>0.8205</td>
<td>23.5281</td>
</tr>
<tr>
<td>4</td>
<td>4[36;42]</td>
<td>6.5803</td>
<td>0.6427</td>
<td>0.7029</td>
<td>20.5356</td>
</tr>
<tr>
<td></td>
<td>4[11;17]</td>
<td>7.0432</td>
<td>0.3517</td>
<td>0.0058</td>
<td>17.5301</td>
</tr>
<tr>
<td></td>
<td>4[20;26]</td>
<td>6.8849</td>
<td>-0.1647</td>
<td>0.8860</td>
<td>24.5282</td>
</tr>
<tr>
<td>5</td>
<td>5[29;35]</td>
<td>8.9804</td>
<td>-0.9298</td>
<td>0.4815</td>
<td>60.3518</td>
</tr>
<tr>
<td></td>
<td>5[24;31]</td>
<td>8.6541</td>
<td>-0.7769</td>
<td>0.1075</td>
<td>24.3271</td>
</tr>
<tr>
<td></td>
<td>5[15;21]</td>
<td>7.4975</td>
<td>-0.4360</td>
<td>0.2291</td>
<td>27.4271</td>
</tr>
<tr>
<td></td>
<td>5[7;13]</td>
<td>9.8165</td>
<td>-0.0824</td>
<td>0.6498</td>
<td>64.0362</td>
</tr>
</tbody>
</table>

4 Conclusions

Four similarity measures for comparison of medical streaming data are analysed. The results of this paper may be considered as the preliminary ones. For more comprehensive conclusions, the wide statistical estimations must be done.
Application of the Frobenius norm, Correlation coefficient between matrices and Multidimensional dynamic time warping gives similar results - these measures indicate often the same optimal place of $X^b$ on $X^a$. PCA similarity factor gives much more different results as compared with the first three measures. This is advantage of application of measures of different nature.

Data are very specific for the individual patient. Therefore, in all the experiments, the values of all the measures are far from the best possible ones, given in Section 2. This may cause the problem of the reliability of the decision when the problem is to detect events in real time streaming data in accordance on the previously detected and estimated streaming data of various patients. Some threshold of measure values is necessary to be fixed for proper decision.

Acknowledgement. The research has been inspired and funded partly by joint stock company ”Algoritmų sistemos” and Agency for Science, Innovation and Technology (MITA), Lithuania.

References


The paper tackles the issue of possible misspecification in fitting skew normal distributions to empirical data. It is shown, through numerical experiments, that it is easy to choose a distribution which is different from this which actually generated the sample, if the minimum distance criterion is used. It is suggested that, in case of similar values of distance measures obtained for different distributions, the choice should be made on the grounds of parameters’ interpretation rather than the goodness of fit. This is supported by empirical evidence of fitting different skew normal distributions to the estimated monthly inflation uncertainties for Belarus, Poland, Russia and Ukraine.

KEYWORDS: skew normal distribution, simulated minimum distance estimators, inflation uncertainties, monetary policy in Eastern Europe.

JEL codes: C15, C46, E52, E37

ACKNOWLEDGEMENT: Financial support of the ESRC/ORA project RES-360-25-0003 Probabilistic Approach to Assessing Macroeconomic Uncertainties is gratefully acknowledged. This research used the ALICE High Performance Computing Facility at the University of Leicester. We are solely responsible for all remaining deficiencies.

1 Introduction

During the last decade a substantial development of the theory of skew normal distributions, that is distributions which contain normal distribution as the special symmetric case, can be observed. The first distribution of this kind was probably the so-called two-piece normal (or split normal) distribution, TPN, originated by [14] and developed further by [15]. It gained substantial popularity among the practitioners; in particular it has been widely used by economic forecasters for describing uncertainties of the probabilistic forecasts of inflation (for a current review see e.g. [16]). Further breakthrough was made by [1, 2], who developed an elegant theory of univariate, and then multivariate, skew normal distributions. These distributions have been recently subject of substantial generalisations. Most notable, the Balakrishnan skew normal distribution
has been proposed by [22], generalized Balakrishnan skew normal distribution, GBSN, by [26], and developed further by [17], [13], [10] and others.

These distributions, albeit fairly general and elegant, provide the potential user with three practical problems: (i) estimation, (ii) interpretation of the parameters and (iii) possible distributional misidentification. In this paper the problems (i) and (ii) are considered only indirectly. Regarding (i), the identification and numerical problems have been discussed in a number of papers, e.g. in [20], [18] and [5]. Problem (ii) can be tackled by developing skew normal distributions with parameters directly related to the particular theory or the phenomenon described. In particular, in [7] we have proposed a skew normal distribution, called weighted skew normal distribution, WSN, which parameters are directly interpretable in the context of macroeconomic density forecasting under inflation targeting. The current paper deals predominantly with (iii), that is the possibility of distributional misspecification. After overcoming (or skipping) problems (i) and (ii), a practitioner faces a dilemma of choosing from a plethora of different skew normal specifications. It seems to be natural that the researcher would choose that one which fits the best to the data. And here the old problem arises: is the distribution which fits to the data in the best way really the true one?

We tackle (iii) by putting three skew normal distributions mentioned above, that is TPN, GBSN and WSN, to the goodness of fit contest. In Section 2 we give brief description of the distributions we are considering. Section 3 explains general settings and estimation procedure. Section 4 presents the results of a Monte Carlo study evaluating the probabilities of choosing a wrongly specified skew normal distribution on the basis of its fit. Section 5 shows empirical results of estimation skew normal distributions, for the one-step ahead forecasts errors of monthly inflation in Belarus, Poland, Russia and Ukraine. Section 6 concludes.

2 Three skew normal distributions

There are three distributions which we consider in this paper: weighted skew normal, WSN (which we regard as the benchmark one), two-piece normal, TPN, and the [26] generalized Balakrishnan skew normal distribution, GBSN.

A random variable with WSN distribution is defined by [7] as:

\[ Z = X + \alpha \cdot Y \cdot I_{Y > \tau_{up}} + \beta \cdot Y \cdot I_{Y < \tau_{low}}, \]

where

\[ I_{Y > \tau_{up}} = \begin{cases} 1 & \text{if } Y > \tau_{up} \\ 0 & \text{otherwise} \end{cases}, \quad I_{Y < \tau_{low}} = \begin{cases} 1 & \text{if } Y < \tau_{low} \\ 0 & \text{otherwise} \end{cases}, \]

\[ (X, Y) \sim N \left( \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix} \right), \]

\[ \tau_{low} < \tau_{up}; \quad \alpha, \beta, \tau_{low}, \tau_{up}, \mu_X, \mu_Y \in \mathbb{R}, \sigma_X, \sigma_Y \in \mathbb{R}^+, \quad \text{and} \quad |\rho| \leq 1. \]

It is suggested that the parameters \( \alpha \) and \( \beta \) are, in the economic context, related to the strength of the 'corrective' monetary policy (it is sensible to restrict them as
non-positive under some additional assumptions considered below). The density and moment-generating functions and the main moments of WSN are given in [7].

A random variable with TPN distribution is defined by its pdf:

\[
f_{\text{TPN}}(t; \sigma_1, \sigma_2, \mu) = \begin{cases} A \cdot \exp\left(-\frac{(t - \mu)^2}{2\sigma_1^2}\right), & \text{if } t \leq \mu \\ A \cdot \exp\left(-\frac{(t - \mu)^2}{2\sigma_2^2}\right), & \text{if } t > \mu \end{cases}, \quad t \in \mathbb{R},
\]

where \(A = \sqrt{2/\pi} (\sigma_1 + \sigma_2)^{-1}\). Three parameters to be estimated are \(\sigma_1, \sigma_2, \mu\) \(\in \mathbb{R}^+\) and \(\mu \in \mathbb{R}\).

The third distribution considered here, the GBSN, is given by the following pdf:

\[
f_{\text{GBSN}}(t; n, m, \delta) = \frac{1}{C(n, m, \delta)} \left[\Phi(\delta t)\right]^n \left[1 - \Phi(\delta t)\right]^m \varphi(t), \quad t \in \mathbb{R},
\]

where \(C(n, m, \delta) = \sum_{i=0}^m \left(\begin{array}{c} m \\ i \end{array}\right) (-1)^i \int_{-\infty}^{\infty} \left[\Phi(\delta t)\right]^{n+i} \varphi(t) dt\), \(\Phi\) and \(\varphi\) are respectively the cdf and pdf of the standard normal distribution; \(n\) and \(m\) are non-negative integers and \(\delta \in \mathbb{R}\) are the parameters. The GBSN includes the Balakrishnan skew normal distribution for \(m = 0\), and the original Azzalini skew normal distribution for \(n = 1\) and \(m = 0\). Azzalini distribution is also a special case of the WSN for \(\alpha = -2\rho\), \(\tau_{up} = \beta = 0\), \(\sigma_X = \sigma_Y = 1\) and \(\mu_X = \mu_Y = 0\). All three distributions can be reduced to a standard normal: WSN for \(\alpha = \beta = \mu_X = 0\) and \(\sigma_X = 1\); TPN for \(\sigma_1 = \sigma_2 = 1\) and \(\mu = 0\); GBSN for \(n = 1\) and \(\delta = m = 0\) or \(n = m = 0\).

### 3 Estimation and general settings

As mentioned above, estimation of WSN, TPN and GBSN distributions by the maximum likelihood or the generalized method of moments is numerically awkward. This problem is particularly well discussed for the Azzalini distribution (see e.g. [3], [21]), and is evident also for all three families of distributions considered here. However, it is straightforward to derive random number generators for all three distributions. For WSN given by (1) it is described in [7], for TPN in [19] and for GBSN in [26].

With the use of these generators and inspired by [12] we have applied the simulated minimum distance estimators method (SMDE, see [6]), which consists of fitting the approximated by simulation density function to empirical histograms of data and applying a minimum distance criterion. The algorithm requires conducting an iterative grid search over the pre-defined range of admissible parameters.

The version of SMDE applied here can be defined as:

\[
\hat{\omega}_{n}^{\text{SMDE}} = \arg \min_{\omega \in \Omega} \left\{ \mu_w \left( d(g_n, f_{\omega}) \right)_{t=1}^T \right\},
\]

where \(f_{\omega}\) is the approximation of the pdf, \(f_w\), of a random variable obtained by generating \(t = 1, \ldots, T\) replications (drawings) from a distribution with parameters \(\omega\) \((\omega \in \Omega \subset \mathbb{R}^k\)), \(g_n\) denotes the density of empirical sample of size \(n\), \(\mu_w\) is an aggregation operator based on \(T\) replications, which deals with the problem of the ‘noisy’
Table 1: Mean, st. deviation and skewness of DGP’s

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>st. dev.</th>
<th>skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGP1</td>
<td>-0.363</td>
<td>1.069</td>
<td>-0.628</td>
</tr>
<tr>
<td>DGP2</td>
<td>-0.398</td>
<td>1.113</td>
<td>-0.695</td>
</tr>
<tr>
<td>DGP3</td>
<td>-0.207</td>
<td>0.925</td>
<td>-0.687</td>
</tr>
</tbody>
</table>

criterion function (median, in this case), and \(d(\cdot, \cdot)\) is the distance measure. The distance measures, \(MD\), applied here are that of the [8] power divergence disparities family given by:

\[
d(g_n, f_{t,\omega}) = \frac{1}{\lambda_{CR}(\lambda_{CR} + 1)} \sum_{i=1}^{m+1} g_n(i) \left[ \left( \frac{g_n(t)}{f_{t,\omega}(t)} \right)^{\lambda_{CR}} - 1 \right],
\]

where \(m\) denotes the number of cells in which data are organized. For \(\lambda_{CR} = 1\) formula (2) gives the Pearson \(\chi^2\) measure, for \(\lambda_{CR} = -1/2\) the Hellinger twice squared distance (\(HD\)) and for \(\lambda_{CR} = -2\) the Neyman \(\chi^2\) measure. For \(\lambda_{CR} \to 0\) and \(\lambda_{CR} \to -1\) the continuous limits of the right-hand side expression in (2) are respectively the likelihood disparity (\(LD\)) and the Kullback-Leibler divergence statistics. [8] advocate optimal setting \(\lambda_{CR} = 3/2^1\).

4 Fit of true and false models

As the main objective of this paper is to decide whether using the best fit criterion for selecting type of a skew normal model might lead to distributional misspecification, we have set up three data generating processes (DGP’s, or ‘true models’) and fitted all three models to data generated by each of them. The DGP’s are:

- **DGP 1**: WSN with \(\alpha = -2.0, \beta = -0.5, \mu_X = \mu_Y = 0, \sigma_X = \sigma_Y = \sigma = 1, \tau_{up} = -\tau_{low} = 1\) and \(\rho = 0.75\). As in this paper we intend to compare three-parameter distributions only, we are keeping \(\mu_X = \mu_Y = 0\) and \(\tau_{up} = -\tau_{low} = 1\) constant, so that we are hence left with three parameters to estimate: \(\alpha, \beta,\) and \(\sigma\).

- **DGP 2**: TPN with \(\sigma_1 = 1.5, \sigma_2 = 0.5, \mu = 0.4\).

- **DGP 3**: GBSN with \(n = 2, k = 1\) and \(\delta = -0.3\).

All three DGP’s have similar first three moments, as given in Table 1. For each DGP, and for sample sizes of 100, 150, 200, 250, 300, 350, 400, 450 and 500, there have been generated \(N_{repl} = 1,000\) replications. For each simulated sample we have fitted all three distributions using the SMDE method outlined in Section 3. We apply TPN,
GBSN and two variations of WSN. In the first variant, denoted by WSN(0) we keep the thresholds fixed as in the DGP 1, that is $\tau_{up} = -\tau_{low} = 1$. In the second variant, we made the thresholds dependent on $\sigma$ in such way that $\sigma \tau_{up} = -\sigma \tau_{low} = 1$. We denote this as WSN(1).

As a simple, nave, misspecification measure, we use the frequency of cases when $d_0(\zeta^i) > d_1(\zeta^i)$, where $d_0$ denotes the minimum distance measure computed for the estimated properly specified distribution in the $i$th replication $\zeta^i$, and $d_1$ denotes the minimum distance measure computed for the estimated misspecified distribution in the same sample. By the properly specified distribution we understand the distribution of the same type as used for generating the sample. The distance criterion used here is the Hellinger distance, $HD$ (results for other criteria are available on request; they do not differ much from these presented in this paper).

Another misspecification measure is based on the bootstrapping the ratios of two alternative distance measures obtained for the same sample. We have used methodologies developed originally for comparing variances: simple bootstrap and Efron bootstrap (see e.g. [24]).

The algorithm for simple bootstrap is the following:

1. Step 1: Draw $M$ pairs of $d_0(\zeta^k), d_0(\zeta^j), k, j = 1, \ldots, Nrepl, k \neq j$. $M$ should be large, e.g. 10,000;
2. Step 2: Compute the ratio of distance measures $r_h^0 = \frac{d_0(\zeta^k)}{d_0(\zeta^j)}$, $h = 1, 2, \ldots, M$;
3. Step 3: Compute the 95th quantile of the distribution of $r_h^0$ denoted as $q_{0.95}$;
4. Step 4: Check the simulated bootstrap criterion for the case where $d_0(\zeta^i) > d_1(\zeta^i)$ as: $$\frac{d_1(\zeta^i)}{d_0(\zeta^i)} > q_{0.95}.$$ 

The frequency of cases where the above inequality is fulfilled tells about the probability of undertaking the right decisions regarding the distribution by rejecting the wrong one. It approximates the probability of rejecting the null hypothesis that the distance measures for the true and false distributions are identical with the implicit alternative that the distribution on which $d_1(\zeta^i)$ is based is false. Efron bootstrap is similar, except that in Step 1 drawing is made from the set of all $d_0(\zeta^k), d_1(\zeta^k)$ rather than from $d_0(\zeta^k)$ alone. Results in this case are more robust, as the equality of the distance measures is explicit under the null.

Tables 2, 3 and 4 present respectively the nave misspecification measure and also these based on the simple and Efron bootstraps. Results for other criteria and for different sample sizes are available on request.

Tables 2-4 show that results of fitting WSN and TPN to data generated from GBSN behave differently to that fitted to data generated from WSN or TPN distributions. Let us first concentrate on evaluating the misspecification in case when data are generated by WSN and TPN; it is clearly difficult to distinguish between these two distributions. For the small sample size it is practically haphazard to find out which statistic is
smaller regardless of the data generating process. In particular, if data are generated from TPN, there is a virtually equal chance that WSN would fit better than the true TPN distribution. However, with the increase in sample size the frequencies of cases where the MD statistics for the ‘true’ distribution is smaller than for the ‘false’ one increase, suggesting the consistency of choice based on the MD criterion. This is confirmed by the bootstrap results. In Tables 3a, 3b and 4a, 4b frequencies of the rejection of the null that the MD statistics are identical increase with the increase in sample size. Nevertheless, the empirical power of the tests based on the MD statistics is, in absolute terms, not high. Even for samples of size 500 it is not reaching 20%. In another words, it is in practice problematic to distinguish between the WSN and TPN distributions.

Nevertheless, some differences between the fits given by WSN and TPN can be observed here. Generally TPN is more often falsely well approximated by WSN, particularly WSN(1), than WSN by TPN. Also, for middle-sized samples (150-350 observations) chances for proper identification of WSN against TPN by rejecting the null of identical MD statistics are visibly higher than otherwise, albeit still small in absolute terms. It is also worth noting that the differences between particular MD criteria, in terms of power, and frequencies of the false choice based on the minimum of competing statistics, are meaningless.

For data generated by WSN and TPN, the danger of misspecification by falsely fitting GBSN is visibly smaller. Except for small samples of data generated by WSN, MD statistics for GBSN are usually bigger than for two remaining distributions in this case than the corresponding WSN and TPN statistics, reducing the chance of
distributional misspecification. Also the empirical power of the MD ratio test rises relatively quickly with the increase in sample size exceeding, in some cases, 20% for large samples.

In contrast to WSN and TPN, data generated by GBSN exhibit different patterns. In terms of power of the bootstrap tests, they are also be easily confused with two other distributions, as the power of the MD ratio test is low. However, the power of the test is not visibly increasing with the increase of sample size, causing doubts regarding the consistency. On the positive side, the nave misspecification benchmark based on the differences between the MD statistics for the true and false distributions is less often false than in the case of data generated from WSN and TPN.

5 Empirical results: Assessing inflation uncertainty in Belarus, Poland, Russia and Ukraine

The distributions discussed above have been used for modelling short-run inflation variability uncertainties, approximated by one-step ahead forecast error. As it is discussed in [7], possible skewness of such uncertainties is caused by monetary policy asymmetries, characterised by the thresholds in the short-run (past-independent) inflation forecast and effectiveness of the anti-inflationary and output-stimulating policies. In the context of WSN, the thresholds are represented by τ_{up} and τ_{low}, and monetary policy effectiveness respectively by α (for the anti-inflationary policy) and β (for the output-stimulating policy).

The four East and Central European countries studied here for the period from January 1995 to December 2012, Belarus, Poland, Russia and Ukraine, represent different types and practices of monetary policy. Poland, for the period under investigation, conducted inflation targeting policy. For Belarus and Russia the targets have been less clear. For Belarus it was predominantly the currency stability, although recently a policy of inflation stabilisation has been announced. For Russia the target was, formally, inflation, for most of the period under study, but practically stabilisation of the exchange rate. For both Belarus and Russia which relies on exporting (in case of Belarus, re-exporting) natural resources, it lead to appreciation pressures and 'dirty float' inflationary effects. In Ukraine the targets and instruments have been usually multiple and loosely defined, with the emphasis on controlling bank liquidities, periods of nominal anchoring the currency to the US dollar, direct commercial banking supervision, etc. Russia and Ukraine and, to a lesser extent, Belarus, have been affected by the Russian currency crisis in 1998.

The data used here are on monthly inflation, not de-seasonalised, with 223 observations per country. After checking for the order of seasonal and non-seasonal integration by the Taylor (2003) test which takes into account the possibility of the presence of unit roots at frequencies other than tested, we have identified the variability

Table 5: Basic characteristics of empirical distributions of uncertainties

<table>
<thead>
<tr>
<th></th>
<th>Belarus</th>
<th>Poland</th>
<th>Russia</th>
<th>Ukraine</th>
</tr>
</thead>
<tbody>
<tr>
<td>std. dev.</td>
<td>0.0184</td>
<td>0.0035</td>
<td>0.0209</td>
<td>0.0098</td>
</tr>
<tr>
<td>skewness</td>
<td>0.2033</td>
<td>0.0479</td>
<td>6.1632</td>
<td>0.4141</td>
</tr>
<tr>
<td>BP (p-val)</td>
<td>1.000</td>
<td>0.7869</td>
<td>0.9932</td>
<td>0.2800</td>
</tr>
</tbody>
</table>

uncertainty as ut in the seasonal ARMA (SARMA) model:

$$\phi(B)\Phi(B^s)\Delta^\kappa \Delta^D y_t = \theta(B)\Theta(B^s)u_t,$$

where $B$ is the lag operator, $\Delta^\kappa = (1 - B)^\kappa$ is the regular difference operator, $\kappa$ is the order of integration of the regular part of $y_t$, $\Delta^D = (1 - B^s)^D$ is the seasonal difference operator for a seasonal $I(D)$ process, $\phi(B) = (1 - \phi_1 B - \cdots - \phi_p B^p)$ is the polynomial of order $p$ in the lag operator $B$ and similarly, the seasonal AR operator is defined as $\Phi(B^s) = (1 - \Phi_1 B^s - \cdots - \Phi_P B^{sP})$. Regular, $\theta(B)$, and seasonal, $\Theta(B^s)$, moving average polynomials are defined similarly with their orders denoted by $q$ and $Q$ respectively. The orders $p, P, q$ and $Q$ have been obtained using the [11] procedure which is based on an automatic lag selection criterion that minimises the Bayesian Information Criteria (BIC) of the residuals. The algorithm applied here is equivalent to the well-known TRAMO-SEATS and X-11 adjustment methods.

Basic characteristics of the estimates of uncertainties: standard deviations, coefficients of skewness and p-values of Box-Pierce (BP) portmanteau autocorrelation statistics are given in Table 5. Table 6 presents the empirical results. As in the previous section, for each distribution three parameters have been estimated by the SMDE. In case of WSN, the range of selection of other parameters: $\tau_{up}, \tau_{low}$ and $\rho$ has been sparse, with only few values searched. For this reason, we do not report standard errors for these parameters. For other, non-integer, parameters, standard errors are given in brackets below the estimates.

The distance measure criterion suggests the choice of different distributions for particular countries. For Belarus, the best fit is that of TPN, closely followed by GBSN, for Poland and Ukraine the best is WSN, albeit for Poland the difference with respect to TPN is slight. For Russia, the best fitted distribution is that of GBSN. These differences can be explained by different types of monetary policy in each country. It is worthwhile to note that there is no systematic relationship between the absolute level of inflation uncertainties, measured by standard deviations, and the type of best-fitted distribution.

For the countries where WSN fits best, that is Poland and Ukraine, differences in the parameters estimated reflects the changes in monetary policy. For Poland, where there are inflation target bands, the thresholds are non-existing and the estimates of $\alpha$ and $\beta$ are very close and to each other, reflecting the near symmetricity of the uncertainties, which are, in turn, related to the balance between anti-inflationary and output-stimulating policies imposed by the inflation target bands of the inflationary targeting. For Ukraine, the distance between the estimated thresholds is substantial and the estimated $\alpha$ is markedly lower than the estimated $\beta$. This indicates the prefer-
Table 6: Results of empirical estimation of skew normal distributions

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Belarus</th>
<th>Poland</th>
<th>Russia</th>
<th>Ukraine</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>-0.8169 (0.4526)</td>
<td>-3.997 (0.00854)</td>
<td>-3.624 (0.6461)</td>
<td>-3.729 (0.5502)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.7226 (0.7507)</td>
<td>-3.713 (0.09892)</td>
<td>-3.548 (0.397)</td>
<td>-1.765 (1.532)</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.03355 (0.4028)</td>
<td>0.001027 (0.000798)</td>
<td>1.875 (0.3758)</td>
<td>0.009011 (0.4804)</td>
</tr>
<tr>
<td>$\tau_{up}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\tau_{low}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>$MD$</td>
<td>46.47</td>
<td>8.693</td>
<td>55.35</td>
<td>4.611</td>
</tr>
</tbody>
</table>

For TPN:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Belarus</th>
<th>Poland</th>
<th>Russia</th>
<th>Ukraine</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_1$</td>
<td>0.0712 (0.3098)</td>
<td>0.0117 (0.0053)</td>
<td>0.3550 (0.0840)</td>
<td>0.0652 (0.1656)</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>0.1520 (0.3060)</td>
<td>0.0149 (0.0154)</td>
<td>0.1492 (0.3148)</td>
<td>0.0944 (0.2243)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0.2408 (1.5152)</td>
<td>-1.1874 (0.9721)</td>
<td>1.3737 (0.9685)</td>
<td>-1.9402 (0.1795)</td>
</tr>
<tr>
<td>$MD$</td>
<td>14.7413</td>
<td>9.1105</td>
<td>52.2228</td>
<td>17.7917</td>
</tr>
</tbody>
</table>

For GBSN:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Belarus</th>
<th>Poland</th>
<th>Russia</th>
<th>Ukraine</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>20</td>
<td>16</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>$m$</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$\delta$</td>
<td>-2.586 (0.03822)</td>
<td>1.908 (0.8716)</td>
<td>-2.558 (0.1176)</td>
<td>-2.558 (0.1176)</td>
</tr>
<tr>
<td>$MD$</td>
<td>18.76</td>
<td>27.90</td>
<td>15.56</td>
<td>15.56</td>
</tr>
</tbody>
</table>

ence (or better effectiveness) of the anti-inflationary policy over the output-stimulating one.

6 Conclusions

The general message from this paper is somewhat pessimistic. It might be difficult to tell one skew normal distribution from another on the basis of the best fit, especially if the sample size is not very large. As the number of potential skew normal candidates for fitting to data is substantial (especially in the light of the fact that there are other propositions in the literature not considered in this paper) it seems to be sensible to decide on the type of distribution not on the basis of the best fit but rather on the basis of interpretation of its parameters. In the context of inflation uncertainties in countries conducting consistent and reasonably tight monetary policy, the weighted skew normal distribution seems to be a sensible choice. For modelling other phenomena, different distributions can be of a better use.

The difficulty in deciding on the type of skew normal distribution are deepened by the fact that there are no operational statistics developed for testing the degree of disparities between distance measures (or other characteristics) of these distributions. The bootstrap procedure used in this paper offers some hope in this respect. However, further studies are needed here.
References


A PROJECTION-PURSUIT METHOD FOR SPARSE ROBUST PCA

P. Filzmoser
Vienna University of Technology, Vienna, AUSTRIA
e-mail: P.Filzmoser@tuwien.ac.at

Abstract
A method based on the idea of projection-pursuit is introduced for obtaining principal components that are robust and sparse [3]. The term robust refers to resistance against outlying observations, while the expression sparse means that some of the principal component loadings will be forced to zero values. In that way, the resulting method is insensitive to outliers in the data, and due to the sparse loading pattern, results are simpler to interpret.

1 Introduction
Principal component analysis (PCA) is a very popular statistical tool for the analysis of multivariate data. The main objective of PCA is dimension reduction: with reduced dimension one still wants to express much of the overall data information. Dimension reduction is in particular useful for high-dimensional data which are difficult or impossible to study visually. Principal components (PCs) are linear combinations of the original variables, and they are constructed in order to maximize the variance. The resulting coefficients of the linear combinations, the so-called loadings, are then used for the interpretation of the PCs.

The interpretation of PCs can be improved by different procedures. One possibility is to use rotation methods: the extracted PCs are rotated according to some criterion in order to simplify the pattern of the loadings matrix. Specifically, one wants to obtain small and large (absolute) loadings, but rather avoid intermediate values, which stand for an unclear or weak contribution of a variable to a PC. Here we will focus on a different strategy, namely on sparse estimation of the loadings matrix.

The concept of sparse parameter estimation was introduced by [7] with the method Lasso in the context of regression analysis. The idea is to use an $L_1$ penalty which shrinks some of the regression coefficients to zero. In the context of PCA, this sparsity concept was first considered by [5], who introduced the so-called SCOTLLASS method. For this method it is necessary to plug in an estimate of the covariance matrix. Although this would be straightforward to robustify by using a robustly estimated covariance matrix, it is unclear how this idea can be applied to data containing more variables than observations. This setting is particularly important for many applications, e.g. from biostatistics. The problem is that there is no method for robust covariance estimation for such situations. Also another proposal for robust PCA of [8] is not straightforward to robustify for the high-dimensional case, since it is formulated in a regression context.

Here we will use a projection-pursuit definition of PCA [1] for proposing a robust and sparse PCA method. Projection-pursuit PCA goes back to the original definition of
PCA: search for a direction such that the variance of the projected data is maximized. Similarly, subsequent directions are defined to maximize the variance of projected data, by imposing additional orthogonality constraints to previously found directions. Using robust variance estimators results in robust PCs, and imposing an $L_1$ penalty gives sparseness [3].

2 Method

Assume that we have given $n$ multivariate observations $x_1, \ldots, x_n \in \mathbb{R}^p$. Then the first sparse PC can be defined as

$$\tilde{a}_1 = \arg \max_{\|a\|=1} V(a'^x_1, \ldots, a'^x_n) - \lambda_1 \|a\|_1.$$  

(1)

Here, “V” is a variance estimator. For classical sparse PCA, this is the empirical variance, for robust sparse PCA this is a robust variance estimator, like the squared $Q_n$ estimator [6]. The tuning parameter $\lambda_1$ regulates the compromise between maximizing the variance and obtaining sparsity. The resulting vector $\tilde{a}_1$ is the first sparse PCA direction. Similarly, the $j$th sparse PCA direction ($1 < j \leq p$) is defined by

$$\tilde{a}_j = \arg \max_{\|a\|=1, a \perp \tilde{a}_1, \ldots, a \perp \tilde{a}_{j-1}} V(a'^x_1, \ldots, a'^x_n) - \lambda_j \|a\|_1.$$  

(2)

We require orthogonality to previously found directions. $\lambda_j$ is the sparsity parameter that can be different from the tuning parameters for other directions.

The optimization problems (1) and (2) are not trivial to solve. In [3] we describe an algorithm that is based on iterative grid searches in planes spanned by pairs of variables, see also [2]. An implementation in the R package pcaPP [4] makes its use very attractive for real applications.

3 Example

We consider a data set of 166 alcoholic fermentation mashses of different feedstock (rye, wheat and corn) that were analyzed with near infrared spectroscopy (NIR). In total, absorbance values for 235 different spectral values (variables) corresponding to wavelengths of 1115-2285 nm have been measured. The data set is available in the R package chemometrics as data frame NIR. Figure 2 (top) shows the mean-centered data as they are used for PCA: each curve corresponds to one observation. One can see different “streams” in the figure, corresponding to different feedstocks. Nevertheless, the data structure is quite homogeneous, without any obvious severe outliers. Therefore, we apply classical sparse PCA, by using the empirical variance for “V”. Here we only present the first two sparse PCs. The choice of the tuning parameters $\lambda_1$ and $\lambda_2$ is shown in the plots in Figure 1: These plots show the proportion of explained variance for the first (left) and second (right) PC when the tuning parameters are increased. The explained variance decreases, but at the same time the number of zeros in the
loading vectors gets larger. The latter are shown on the horizontal axes. The optimal tuning parameters are selected by maximizing the explained variance multiplied by the number of zero loadings. The dashed lines show the selected optima, which refer to the choice of the tuning parameters $\lambda_1$ and $\lambda_2$. With these tuning parameters, the final sparse PCs are computed, and the resulting first two loadings vectors are shown as the lines in Figure 2 (bottom). It can be seen that many of the loadings are zero, and this is essentially at positions where the original signals seem to include rather random noise and not structural information. The different PCs refer to different feedstocks.

Figure 1: Selection of the optimal tuning parameters $\lambda_1$ and $\lambda_2$ for sparse PCA.

References


Figure 2: Mean-centered NIR data (top) and first two sparse PCA loading vectors (bottom).


SOME REMARKS ON DECOMPOSITION OF PARTITIONED MULTIVARIATE MODELS INTO TWO SEEMINGLY UNRELATED SUBMODELS

E. Fišerová, L. Kubáček

Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University
Olomouc, CZECH REPUBLIC
e-mail: eva.fiserova@upol.cz

Abstract

Procedure on testing decomposition of partitioned multivariate models into two seemingly unrelated submodels in order to obtain more efficient estimators is presented. The multiresponses are independently normally distributed with the same covariance matrix. Moreover, the partitioned multivariate model is considered either with, or without an intercept. The elimination transformation of the intercept that leads to the same BLUEs of parameter matrices in multivariate models with and without intercept is stated. Due to the elimination transformation the same procedure for testing decomposition can be applied for models with and without the intercept.

Key Words: multivariate model, decomposition of model, elimination transformation, unbiased estimator

2010 Mathematics Subject Classification: Primary 62H12; Secondary 62J05, 62H15.

1 Introduction

Multivariate models describe relationships between multiresponse data and a set of predictors, see, e.g., [1, 3, 5, 8]. Multivariate approach has several advantages in comparison with series of univariate models. Multivariate models respect association between outcomes, and thus, in general, procedures are more efficient. Further, they can evaluate joint influence of predictors on all outcomes and avoid issue of multiple testing.

There are situations when a multivariate model can be decomposed to a series of simpler models, univariate or multivariate depending on issue. Obviously, if the multiresponses are not intercorrelated, the multivariate model can be equivalently expressed as series of independent univariate models. In this paper we deal with a special case of decomposition of a partitioned multivariate model with independent multiresponses with the same covariance matrix into two seemingly unrelated multivariate submodels [9] in order to obtain more efficient estimators. Namely, the multiresponse variables in the model are partitioned into two sets $Y^1$ and $Y^2$. Similarly, the set of predictors is partitioned into two sets $X_1$ and $X_2$. As an example, let us consider the nutrigenomic study in the mouse. Response variable might be expressions of chosen
genes \((Y^1_i)\) and concentrations of hepatic fatty acids \((Y^2_i)\) measured on subjects. The predictors might be genotype \((X_1)\) and type of diet \((X_2)\). The problem is to decide, roughly speaking, if it is possible to explain separately expressions of genes by genotype and hepatic fatty acids concentrations by diet or not. Fišerová and Kubáček (2012) proposed tests for verification significance of model decomposition under normality of random errors in the case when covariance matrix is known or completely unknown. Short overview is presented in Section 2. Here we focus on situation when the partitioned multivariate model includes also an intercept. As it will be shown in Section 3, the intercept can be eliminated by a suitable transformation such that the best linear unbiased estimators (BLUEs) of parameter matrices in partitioned models with and without intercept are the same. Therefore the results from [2] can be applied also for models with the intercept.

2 Tests for decomposition of partitioned model without intercept

Let us consider multivariate model in a partitioned form

\[
\begin{pmatrix}
Y^1 \\
Y^2
\end{pmatrix} = \begin{pmatrix}
X_1 \\
X_2
\end{pmatrix} \begin{pmatrix}
B_{11} \\
B_{21}
\end{pmatrix} + \begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{pmatrix},
\]

(1)

Here \(Y = (Y^1, Y^2)\) is a random matrix (observation matrix), \(X = (X_1, X_2)\) is a known design matrix, \(B_{11}, B_{12}, B_{21}, B_{22}\) are matrices of unknown parameters and \(\varepsilon = (\varepsilon_1, \varepsilon_2)\) is a random error matrix. We will denote a multiresponse as \(Y_i = (Y^1_i, Y^2_i)\). Further, we will assume that the matrix \(X\) is of full column rank, the multiresponses are independent with the same covariance matrix \(\Sigma\) which is positive definite and random errors are normally distributed. The covariance matrix \(\Sigma\) of the multiresponse \(Y_i\) is partitioned in the same way, i.e.,

\[
\text{var} \begin{pmatrix}
Y^1_i \\
Y^2_i
\end{pmatrix} = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}, \quad i = 1, 2, \ldots, n.
\]

(2)

Further, let us consider a system of two seemingly unrelated [9] multivariate submodels

\[
\begin{aligned}
Y^1_i &= X_1 B_1 + \varepsilon_1 \\
Y^2_i &= X_2 B_2 + \varepsilon_2
\end{aligned}
\]

(3)

with the covariance matrix \(\Sigma\) of the multiresponse \(Y_i\) in the form (2). The problem is to decide which of the models (1) and (3) should be chosen for modelling in order to obtain more efficient estimators. Note that models in (3) are seemingly unrelated because there is a link between them described by \(\text{cov}(Y^1_i, Y^2_i) = \Sigma_{12}\). If \(\Sigma_{12} = 0\), models in (3) are independent.

Obviously, if the parameter matrices \(B_{12}\) and \(B_{21}\) in model (1) are zeros, the parameter matrices \(B_{11}, B_{22}\) in model (1) and \(B_1, B_2\) in model (3), respectively, are the
same, however, the estimators in model (3) are more efficient due to fewer unknown parameters. Thus the problem with decomposition of model (1) into (3) leads to testing the hypothesis that “a system of two seemingly multivariate submodels (3) is a true model”, i.e., to test $B_{12} = 0$ and $B_{21} = 0$ simultaneously. If the covariance matrix $\Sigma$ is known, the test statistics are given as [2]

$$T_{21} = \text{Tr} \left[ (Y^1)'M_{X_1}X_2(X_2' M_{X_1}X_2)^{-1} X_2' M_{X_1}Y^1 \Sigma^{-1} \right] \sim \chi^2_{p_1k_2} \text{ under } B_{21} = 0;$$

$$T_{12} = \text{Tr} \left[ (Y^2)'M_{X_2}X_1(X_1' M_{X_2}X_1)^{-1} X_1' M_{X_2}Y^2 \Sigma^{-2} \right] \sim \chi^2_{p_2k_1} \text{ under } B_{12} = 0.$$ 

The symbol $\text{Tr}(\Sigma)$ denotes trace of the matrix $\Sigma$ and $M_{X_i} = I_n - X_i(X_i'X_i)^{-1}X_i'$, $i = 1, 2$.

If the covariance matrix $\Sigma$ of multivariate $Y_i$ is unknown, an unbiased estimator of $\Sigma$ in model (1) is [3]

$$\hat{\Sigma} = \left( \hat{\Sigma}_{11}, \hat{\Sigma}_{12} \right) = \frac{1}{n-k_1-k_2} \left( \begin{array}{c} (Y^1)' \\ (Y^2)' \end{array} \right) \left( \begin{array}{c} M_{X_1}X_2(Y^1, Y^2) \end{array} \right).$$  \hspace{1cm} (4)

With $n-k_1-k_2 \geq p_1+p_2$, $\hat{\Sigma}$ is nonsingular with probability 1. The test statistics are in the form [2]

$$F_{21} = - \left[ n-k_1 - \frac{p_1+k_2+1}{2} \right] \log \frac{\det \left[ (Y^1)'M_{(X_1,X_2)}Y^1 \right]}{\det \left[ (Y^1)'M_{(X_1,X_2)}Y^1 + \hat{B}_{21}' X_2 M_{X_1} X_2 \hat{B}_{21} \right]},$$

$$F_{12} = - \left[ n-k_2 - \frac{p_2+k_1+1}{2} \right] \log \frac{\det \left[ (Y^2)'M_{(X_1,X_2)}Y^2 \right]}{\det \left[ (Y^2)'M_{(X_1,X_2)}Y^2 + \hat{B}_{12}' X_1 M_{X_2} X_1 \hat{B}_{12} \right]},$$

where the BLUEs of $B_{12}$ and $B_{21}$ are given by

$$\hat{B}_{12} = (X_1'M_{X_2}X_1)^{-1}X_1'M_{X_2}Y^2, \quad \hat{B}_{21} = (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}Y^1.$$ 

Under $B_{21} = 0$, the statistic $F_{21}$ is asymptotically distributed as $\chi^2_{k_2p_1}$; under $B_{12} = 0$, $F_{12}$ is asymptotically distributed as $\chi^2_{k_1p_2}$.

To test the hypotheses $B_{21} = 0$ and $B_{12} = 0$ simultaneously, one can use, e.g., the Bonferroni correction in order to preserve type I error rate $\alpha$. More precisely, if

$$T_{21} \leq \chi^2_{p_1k_2}(1-\alpha/2) \quad \text{and} \quad T_{12} \leq \chi^2_{p_2k_1}(1-\alpha/2),$$

where $\chi^2_{p_1k_2}(1-\alpha/2)$ denotes the $(1-\alpha/2)$-quantile of a $\chi^2_{p_1k_2}$ distribution, neither of the hypotheses $B_{21} = 0$, $B_{12} = 0$ can be rejected on the significance level $\alpha$. Similarly we proceed with test statistics $F_{12}$ and $F_{21}$. If the decomposition is significant, the prediction of $Y^1$ conditional on $X_1$ is not improved also by regressing on $X_2$. However the predictors $X_2$ are necessary for calculation of the prediction of $Y^1$. Analogous conclusion holds for the prediction of $Y^2$.  

38
### 3 Partitioned model with an intercept

A partitioned form of multivariate model with an intercept can be written as

\[
\begin{pmatrix}
Y^1 \\
Y^2
\end{pmatrix}_{(n \times p_1) \ (n \times p_2)} = \begin{pmatrix}
1 \\
X_1 \\
X_2
\end{pmatrix}_{(n \times 1) \ (n \times k_1) \ (n \times k_2)} \begin{pmatrix}
b_1 \\
b_2
\end{pmatrix}_{(1 \times p_1) \ (1 \times p_2)} + \begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2
\end{pmatrix}_{(n \times p_1) \ (n \times p_2)}, \tag{5}
\]

where \(1\) is a vector of ones. In this case we will also assume that the design matrix \((1, X_1, X_2)\) is of full column rank, the multiresponses are independent with the same covariance matrix \(\Sigma\) which is positive definite and random errors are normally distributed. If the model includes also an intercept, then the question is, where the intercept should go in decomposed model, in \(X_1\), in \(X_2\), or in both \(X_1\) and \(X_2\). Naturally, all cases are possible and results depend on particular tasks. To avoid this situation we propose general solution based on a suitable transformation for elimination of the intercept. Such a transformation is stated in the following theorem and leads to the identical BLUEs of parameter matrices \(B_{11}, B_{12}, B_{21}\) and \(B_{22}\) in models with and without the intercept.

**Theorem 1.** The BLUEs of parameter matrices \(B_{11}, B_{12}, B_{21}\) and \(B_{22}\) in model (5) are the same as in the model

\[
(M_1 Y^1, M_2 Y^2) = (M_1 X_1, M_1 X_2) \begin{pmatrix}
B_{11} \\
B_{12}
\end{pmatrix} + (M_1 \varepsilon_1, M_1 \varepsilon_2). \tag{6}
\]

Here \(M_1 = I - 1 1'\).

**Proof.** For the sake of simplicity we proceed for univariate form of model (5). Let us denote

\[
\varepsilon = \begin{pmatrix}
\vec(\varepsilon_1) \\
\vec(\varepsilon_2)
\end{pmatrix}, \ Y = \begin{pmatrix}
\vec(Y^1) \\
\vec(Y^2)
\end{pmatrix}, \ \Sigma \otimes I = \begin{pmatrix}
\Sigma_{11} \otimes I, & \Sigma_{12} \otimes I \\
\Sigma_{21} \otimes I, & \Sigma_{22} \otimes I
\end{pmatrix},
\]

\[
A_1 = \begin{pmatrix}
I \otimes 1, & 0 \\
0, & I \otimes 1
\end{pmatrix}, \ A_2 = \begin{pmatrix}
I \otimes X_1, & 0, & 0 \\
0, & I \otimes X_2, & 0 \\
0, & I \otimes X_1, & 0
\end{pmatrix}, \ \beta_1 = (b_1', b_2'), \ \beta_2 = (\vec(B_{11})', \vec(B_{22})', \vec(B_{21})', \vec(B_{12})').
\]

Here, the symbol \(\vec(Y^1)\) denotes the column vector composed of the columns of \(Y^1\). The notation \(\otimes\) means the Kronecker multiplication of matrices [6]. Then the model (5) can be rewritten as

\[
Y = (A_1, A_2) \begin{pmatrix}
\beta_1 \\
\beta_2
\end{pmatrix} + \varepsilon. \tag{7}
\]

The model is regular (design matrices has full column rank and covariance matrix is positive definite) and thus the BLUEs of vector parameters \(\beta_1\) and \(\beta_2\) are

\[
\begin{pmatrix}
\tilde{\beta}_1 \\
\tilde{\beta}_2
\end{pmatrix} = \left(A_1' (\Sigma \otimes I)^{-1} A_1, \ A_1' (\Sigma \otimes I)^{-1} A_2\right)^{-1} \left(A_1' (\Sigma \otimes I)^{-1} Y\right),
\]

\[
\left(A_2' (\Sigma \otimes I)^{-1} A_1, \ A_2' (\Sigma \otimes I)^{-1} A_2\right)^{-1} \left(A_2' (\Sigma \otimes I)^{-1} Y\right).
\]
Using well known Rohde formula for inverse of a block matrix [7] we get
\[
\begin{pmatrix}
A'_1(\Sigma \otimes I)^{-1}A_1,
A'_2(\Sigma \otimes I)^{-1}A_2
\end{pmatrix}^{-1} = \begin{pmatrix}
Z_{11},
Z_{12}
\end{pmatrix},
\]
where
\[
Z_{21} = -(A'_1(\Sigma \otimes I)^{-1}A_1)^{-1}A'_1(\Sigma \otimes I)^{-1}A_2A'_2(M_{A_1}(\Sigma \otimes I)M_{A_1})^+A_2, \\
Z_{22} = [A'_2(M_{A_1}(\Sigma \otimes I)M_{A_1})^+A_2]^{-1}.
\]
Thus the BLUE of \( \beta_2 \) in the model (5), i.e., in the model with the intercept, is
\[
\hat{\beta}_2 = Z_{21}A'_1(\Sigma \otimes I)^{-1}Y + Z_{22}A'_2(\Sigma \otimes I)^{-1}Y \\
= [A'_2(M_{A_1}(\Sigma \otimes I)M_{A_1})^+A_2]^{-1}A'_2(M_{A_1}(\Sigma \otimes I)M_{A_1})^+Y.
\]
Since it holds that
\[
[M_{A_1}(\Sigma \otimes I)M_{A_1}]^+ = [M_{A_1}(\Sigma \otimes I)M_{A_1}]^+M_{A_1},
\]
the estimator \( \hat{\beta}_2 \) is the BLUE of \( \beta_2 \) in the model
\[
M_{A_1}Y = M_{A_1}A_2\beta_2 + M_{A_1}\epsilon \quad (8)
\]
as well. This model can be directly rewritten into multivariate form (6) using the relationship
\[
M_{A_1} = \begin{pmatrix}
I \otimes M_1, & 0 \\
0, & I \otimes M_1
\end{pmatrix}.
\]

According to Theorem 1, partitioned multivariate model (5) with the intercept can be transformed into partitioned multivariate model (6) without the intercept. The transformation preserves the BLUEs of parameter matrices \( B_{11}, B_{12}, B_{21}, \) and \( B_{22} \), therefore testing decomposition of a partitioned model with the intercept can be proceed similarly as for the model without the intercept. The process is the following. First we transform model (5) to (6). In model (6), the covariance matrix of the multiresponse \( M_1\Sigma j \), is \( M_1\Sigma M_1 \). If \( \Sigma \) is unknown, the covariance matrix \( M_1\Sigma M_1 \) can be unbiasedly estimated in model (6) by (4) using the substitution
\[
Y^j \to M_1Y^j, \quad X_j \to M_1X_j, \quad \Sigma \to M_1\Sigma M_1, \quad \kappa_j \to r(M_1X_j), \quad j = 1, 2.
\]
The symbol \( r(M_1X_j) \) denotes the rank of the matrix \( M_1X_j \). Next we test hypotheses \( B_{12} = 0 \) and \( B_{21} = 0 \) simultaneously applying tests statistics \( T_{12}, T_{21} \), or \( F_{12}, F_{21} \) using the same substitution together with \( p_j \to r(M_1Y^j) \), \( j = 1, 2 \). Naturally, test statistics have different degrees of freedom in this case. Specifically, for test statistics \( T_{12} \) and \( T_{21} \) it holds that
\[
T_{21} \sim \chi^2_{r(M_1Y^1)r(M_1X_2)}, \quad T_{12} \sim \chi^2_{r(M_1Y^2)r(M_1X_1)}.
\]
Similarly, test statistic \( F_{21} \) and \( F_{12} \) are asymptotically distributed as \( \chi^2_{r(M_1Y^1)r(M_1X_2)} \) and \( \chi^2_{r(M_1Y^2)r(M_1X_1)} \), respectively.
Conclusion

We shortly discuss the problem of decomposition of a partitioned multivariate model into two seemingly unrelated model in order to obtain more efficient estimators. In the case the case when the partitioned multivariate model includes also the intercept, the elimination transformation for the intercept that preserve the BLUEs of parameter matrices can be used. Therefore the same procedure for testing decomposition can be applied for models with and without the intercept.

Acknowledgements

The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic) and the Grant No. PrF-2013-013 of the Internal Grant Agency of the Palacký University in Olomouc.

References


FOURIER–TYPE ESTIMATION OF THE POWER GARCH MODEL WITH STABLE–PARETIAN INNOVATIONS

C. FRANCOQ, S.G. MEIANTANIS
CREST and University Lille 3, Villeneuve d’Ascq, FRANCE
e-mail: christian.francq@univ-lille3.fr

Department of Economics, National and Kapodistrian University of Athens
Athens, GREECE

and Centre for Business Mathematics and Informatics, North–West University
Potchefstroom, SOUTH AFRICA
e-mail: simosmei@econ.uoa.gr

Abstract

We consider the estimation of general power GARCH models with stable–Paretian innovations. Exploiting the simple structure of the conditional characteristic function of the observations driven by these models, we propose minimum distance estimation based on the empirical characteristic function of corresponding residuals. Consistency of the estimators is proved, and we obtain a singular asymptotic distribution which is concentrated on a hyperplane.

1 Introduction

Consider observations $y_t$ from the so–called power GARCH model defined by

$$
\begin{align*}
    y_t &= c_t \varepsilon_t \\
    c_t^\rho &= \mu + \sum_{j=1}^{p} b_j |y_{t-j}|^\rho + \sum_{j=1}^{q} \gamma_j c_{t-j}^\rho, \quad \forall t = 1, 2, \ldots,
\end{align*}
$$

(1.1)

where $\{\varepsilon_t\}$ is a sequence of i.i.d. random variables (with location zero and unit scale) independent of $\{c_t\}$, and $\rho, \mu, \{b_j\}_{j=1}^{p}$, and $\{\gamma_j\}_{j=1}^{q}$ denote unknown parameters. If the innovations $\{\varepsilon_t\}$ are standard normal and the power parameter $\rho$ is set equal to two we obtain the classical Gaussian GARCH model. From the time of Mandelbrot (1963) and Fama (1965) however there is strong evidence that the distribution of financial returns could be heavy–tailed and possibly asymmetric, and many authors advocated the use of the stable–Paretian (SP) distribution instead of the normal distribution in financial modelling. For more recent evidence of stable–Paretian behavior of financial assets the reader is referred to the papers of Mittnik and Rachev (1993), Koutourelis and Meintanis (1999), Liu and Borsen (1995a), Paolella (2001), Tsionas (2002), Akgül and Sayyan (2008), Tavares et al. (2008), Curto et al. (2009), and Xu et al. (2011), and the volumes by Adler et al. (1998), Rachev and Mittnik (2000), Rachev (2003) and Nolan, 2012).

---

1 This research was partially supported by Grant No. 11699 of the Special Account for Research Grants of the National and Kapodistrian University of Athens, and by the Agence Nationale de la Recherche (ANR) and the Economic and Social Research Council (ESRC) through the ORA Program (Project PRAM ANR-10-ORAR-008-01)
Therefore one of the popular generalizations of model (1.1) is to assume that $\{\varepsilon_t\}$ follow a stable-Paretian distribution. We shall call this model SP power GARCH (SP–PGARCH) model. The most convenient way to introduce SP distributions is by means of their characteristic function (CF). Specifically if we assume that $\varepsilon_t$ are zero–location SP random variables with unit scale, then their CF is given by

$$\varphi(u, \lambda) = e^{-|u|^{\alpha}(1-i\beta \text{sgn}(u) \tan(\pi \alpha/2))}, \alpha \neq 1,$$

$$= e^{-|u|^{\alpha}(1+i\beta \frac{2}{\pi} \text{sgn}(u) \log |u|)}, \alpha = 1,$$

where $\lambda = (\alpha, \beta)'$ is the vector of the SP parameter such that $0 < \alpha \leq 2$ and $-1 \leq \beta \leq 1$. We also used the notation $\text{sgn}(u) = 1$, $u > 0$, $\text{sgn}(0) = 0$, and $\text{sgn}(u) = -1$, $u < 0$.

Note that $\alpha$ is a shape parameter often referred to as the ‘tail index’ and that the SP law reduces to the Gaussian distribution at $\alpha = 2$. On the other hand $\beta$ measures skewness, and if $\beta = 0$ the corresponding SP law is symmetric. Besides the normal law, well known particular cases are the Cauchy distribution for $(\alpha, \beta) = (1, 0)$ and the Lévy distribution which corresponds to $(\alpha, \beta) = (1/2, 1)$. Several authors (see for instance Mittnik et al. 1999, Liu and Brorsen 1995b, and Bonato 2009) proposed maximum likelihood estimation of the SP–PGARCH model. However since the density of the SP law is generally not available in closed form various approximations are needed, and therefore likelihood methods may be characterized as computationally demanding.

In the next section we capitalize on the simplicity of eq. (1.2) and suggest an estimation procedure based on a comparison between the empirical CF (ECF) of the PGARCH residuals and the (theoretical) CF of the SP distribution. With the ECF-based estimators that have been proposed in the literature, the ECF depends on the observations and the CF depends on the unknown parameter. Our estimator is different because the ECF not only depends on the observations but also on the GARCH parameter, whereas the CF depends on the SP parameter.

## 2 CF estimation of the SP–PGARCH model

Consider the SP–PGARCH model whereby the observations $y_t$, $(t = 1, ..., T)$, are driven by equation (1.1) and the innovations $\varepsilon_t$ have CF given by (1.2). We assume the standard positivity conditions $\mu > 0$, $\{b_j \geq 0, 1 \leq j \leq p\}$ and $\{\gamma_j \geq 0, 1 \leq j \leq q\}$.

Denote by $\theta = (\rho, \mu, b_1, ..., b_p, \gamma_1, ..., \gamma_q)'$ the PGARCH parameter and by $\lambda = (\alpha, \beta)'$ the SP parameter. We suggest to estimate the parameter $\hat{\theta} = (\theta', \lambda')'$, by minimum distance between the CF and a suitable empirical counterpart. Specifically, given the observations $(y_1, ..., y_T)$ and fixed initial values $(y_0, ..., y_{1-p})$ and $(\tilde{c}_0, ..., \tilde{c}_{1-q})$, the estimation method is defined as

$$\hat{\theta}_T = \arg \min_{\theta \in \Xi} \tilde{\Delta}_T(\theta),$$

where $\Xi$ denotes the parameter space and

$$\tilde{\Delta}_T(\theta) = \int_{-\infty}^{\infty} |\tilde{\varphi}_T(u, \theta) - \varphi(u, \lambda)|^2 W(u)du,$$
with \( W(\cdot) \) a nonnegative weight function. In (2.2) \( \tilde{\varphi}_T(u, \theta) := \varphi_T \{ u; \tilde{e}_1(\theta), ..., \tilde{e}_T(\theta) \} \) is the empirical CF (ECF) defined by
\[
\varphi_T(u; x_1, ..., x_T) = \frac{1}{T} \sum_{t=1}^{T} e^{iux_t},
\]
and computed from the residuals \( \tilde{e}_t(\theta) = y_t/\tilde{c}_t(\theta) \), with \( \tilde{c}_t(\theta) \) being recursively defined for \( t \geq 1 \), by
\[
\tilde{c}_t^\rho(\theta) = \mu + \sum_{j=1}^{p} b_j |y_{t-j}|^\rho + \sum_{j=1}^{q} \gamma_j \tilde{c}_{t-j}^\rho(\theta).
\]
Note that the introduction of the weight function \( W(\cdot) \) in (2.2) is necessary in order to neutralize the periodic components in the ECF \( \tilde{\varphi}_T(u, \theta) \) and thus render the corresponding integral finite.


Note that the ECF, \( \tilde{\varphi}_T(u, \theta) \), involves the PGARCH parameter and the CF, \( \varphi(u, \lambda) \), involves the SP parameter. In the existing literature, the unknown parameter is only involved in the CF, the ECF being computed directly from the observations \( y_1, ..., y_T \).

A major difference with the estimator defined by (2.1)-(2.4) and the estimators considered in the above-mentioned references is that we consider an ECF of residuals instead of an ECF of observations. The reason why we can not use the standard approach in our framework is that, for a PGARCH model, there exists no closed form for the CF of a vector of the form \( (y_t, ..., y_{t-h}) \), \( h \geq 0 \).

### 3 Asymptotic properties

Now consider the asymptotic properties of the estimator (2.1) of the parameter of the SP–PGARCH\((p, q)\) model (1.1)-(1.2). Recall that the parameter vector is decomposed as \( \vartheta = (\vartheta', \lambda')' \) with \( \theta = (\rho, \mu, b_1, ..., b_p, \gamma_1, ..., \gamma_q)' \in \Theta \) and \( \lambda = (\alpha, \beta)' \in \Lambda \). The true parameter value is denoted by \( \vartheta_0 = (\vartheta_0', \lambda_0)' \) with \( \theta_0 = (\rho_0, \mu_0, b_{01}, ..., b_{0p}, \gamma_{01}, ..., \gamma_{0q}) \) and \( \lambda_0 = (\alpha_0, \beta_0) \). Following the seminal paper of Bougerol and Picard (1992), it is easy to see that the necessary and sufficient condition for strict stationarity of (1.1) takes the form \( \gamma(\theta_0) < 0 \), where \( \gamma(\theta_0) \) is the top-Lyapounov exponent of the model, as defined in Appendix A of Hamadeh and Zakoian (2011).

In addition to this strict stationarity condition and to standard identifiability conditions, the strong consistency requires the following assumptions:
\[
\vartheta_0 \in \Xi := \Theta \times \Lambda \text{ where } \Theta \text{ is a compact subset of } (0, \infty)^2 \times [0, \infty)^{p+q}
\]
and \( \Lambda \) is a compact subset of \( (1, 2) \times [-1, 1] \).
and \( W(\cdot) \) is strictly positive over \( \mathbb{R} \setminus \{0\} \), with \( \int |u|^p W(u)du < \infty \) for \( j = 0 \) and \( j = 1 \). Assumption (3.1) imposes standard positivity constraints on the PGARCH coefficients. It also puts restrictions on the value of the tail index \( \alpha_0 \). DuMouchel (1983) showed that, in the case of a sample of stable distribution, the asymptotic distribution of the MLE is not standard when \( \alpha_0 = 2 \). Note also that when \( \alpha_0 = 2 \), the coefficient \( \beta_0 \) is not identifiable. We impose \( \alpha_0 > 1 \) because we need \( E|\varepsilon_t| < \infty \).

It is also assumed that \( \sum_{j=1}^{q} \gamma_j < 1 \) for all \( \theta \in \Theta \). A stationary sequence \( \{c_t(\theta)\} \) can then be defined such that \( |c_t(\theta) - c_t(\tilde{\theta})| \to 0 \) almost surely as \( t \to \infty \). To show the identifiability of the conditional characteristic function, the following assumption is also needed:

\[
E \sup_{\theta \in \Theta} \left\{ \frac{c_1(\theta_0)}{c_1(\theta)} \right\}^{10} < \infty. \tag{3.2}
\]

For an ARCH\((p)\) model, we have

\[
\frac{c_t(\theta_0)}{c_t(\theta)} = \left( \frac{\mu_0 + \sum_{j=1}^{p} b_{0j} y_{t-j}^2}{\mu + \sum_{j=1}^{p} b_{j} y_{t-j}^2} \right)^{1/p} \leq \left( \frac{\mu_0}{\mu} + \sum_{j=1}^{p} \frac{b_{0j}}{b_j} \right)^{1/p}.
\]

Therefore, in the ARCH case, (3.2) is satisfied when \( \inf_{\theta \in \Theta} \min b_j > 0 \). In the general case, it can be shown that (3.2) is satisfied when \( \Theta \) is sufficiently small (see (5.15) and (5.16) in Hamadeh and Zakoïan (2011), referred to as HZ hereafter).

Under the previous assumptions, any sequence \( (\tilde{\theta}_T) \) of CF estimators satisfying (2.1) is strongly consistent:

\[
\tilde{\theta}_T \to \theta_0 \quad \text{almost surely, as } T \to \infty.
\]

To obtain the asymptotic normality, it is obviously necessary to assume that \( \theta_0 \) does not stand at the boundary of the parameter space \( \Xi \). We also need to introduce few additional notations. Let

\[
g_t(u, \theta) = e^{iu(\theta)} - \varphi(u, \lambda)
\]

and the vector of dimension \( d = p + q + 3 \)

\[
\Upsilon_t = \int_{-\infty}^{\infty} \text{Re} \left( g_t(u, \theta_0) E \frac{\partial g_t(u, \theta_0)}{\partial \theta} \right) W(u)du.
\]

It is then possible to show the existence of the matrices \( V = \text{Var} \, \Upsilon_1 \) and

\[
G = \int_{-\infty}^{\infty} \text{Re} \left( E \frac{\partial g_t(u, \theta_0)}{\partial \theta} E \frac{\partial g_t(u, \theta_0)}{\partial \theta'} \right) W(u)du.
\]

The matrices \( V \) and \( G \) are actually singular, and one can show that

\[
G \sqrt{T}(\tilde{\theta}_T - \theta_0) \text{converges in law to the } N(0, V) \text{ distribution as } T \to \infty.
\]
By showing that \( G = ABA' \) where

\[
A = \begin{pmatrix}
E \frac{1}{c_t} \frac{\partial c_t(\theta_0)}{\partial \theta} & 0 \\
0 & I_2
\end{pmatrix}
\]

has full rank 3 and \( B \) is an invertible \( 3 \times 3 \) matrix, we conclude that \( \sqrt{T} A' (\hat{\theta}_T - \theta_0) \) converges in law to the \( N(0, \Sigma) \) distribution, for some matrix \( \Sigma \). When \( W(\cdot) \) is even the latter matrix is of the form

\[
\Sigma = \begin{pmatrix}
0 & 0 \\
0 & S
\end{pmatrix}.
\]

It follows that the asymptotic distribution of \( \sqrt{T} (\hat{\theta}_T - \theta_0) \) is concentrated on the line

\[
\Delta_c = \left\{ x \in \mathbb{R}^{p+q+2} : x'E \frac{1}{c_t(\theta_0)} \frac{\partial c_t(\theta_0)}{\partial \theta} = c \right\}
\]

for some constant \( c \).

4 A numerical illustration

The aim of the simulation experiment presented in this section is to illustrate that the ECF is consistent, but has a non standard asymptotic distribution concentrated on a line, as stated in the previous section. We thus consider the following very simple version of the SP–PGARCH model

\[
y_t = c_t \epsilon_t, \quad c_t^2 = \mu_0 + b_0 y_{t-1}^2
\]

where \( \theta_0 = (0.5, 0.2, 1.6, 0) \), i.e., we have a SP-PARCH(1) model with symmetric SP innovations and tail index equal to 1.6. Moreover we assume that the value of \( \rho \) is known to be equal to 2. We estimated the four parameters by ECF over \( N = 1,000 \) independent simulations of length \( T = 20,000 \) of the process. As expected, because the sample size is large for each parameter, the estimated values were very close to the true value. Figure 1 displays the scatter plot of the 1,000 values of \((\hat{\mu} - \mu_0, \hat{b} - b_0)\). In accordance with the previous section, the points are concentrated along the red line, carried by the vector

\[
E \frac{1}{c_t(\theta_0)} \frac{\partial c_t(\theta_0)}{\partial \theta}.
\]

For comparison, we plotted the linear regression of \( \hat{b} - b_0 \) on \( \hat{\mu} - \mu_0 \) as a dotted line. This line is almost confused with the full red line.\(^2\)

\(^2\)Because there exists no explicit form for this vector, it has been evaluated on the basis of a simulation of length 50,000.
Figure 1: Empirical distribution of the ECF estimator over 1,000 independent simulations of length $T = 20,000$ of the SP-PARCH(1) model (4.1). The red line of the scatter plot corresponds to the direction of the vector $E c^{-1} \partial c_t(\theta_0)/\partial \theta$. 
References


ON OUTLIERS AND INTERVENTIONS IN COUNT TIME SERIES FOLLOWING GLMs

R. Fried, H. Elsaied, T. Liboschik
Department of Statistics, TU Dortmund University
44221 Dortmund, Germany
e-mail: {fried, saied, liboschik}@statistik.tu-dortmund.de

K. Fokianos, S. Kitromilidou
Department of Mathematics and Statistics, University of Cyprus
e-mail: {fokianos, kitromelidou.stella}@ucy.ac.cy

Abstract

We discuss the analysis of count time series following generalized linear models in the presence of outliers and intervention effects. Different modifications of such models are formulated which allow to incorporate, detect and to a certain degree distinguish extraordinary events (interventions) of different types in count time series retrospectively. An outlook on future extensions to the problem of online surveillance and robust parameter estimation is provided.

1 Introduction

Time series of counts are measured in various disciplines whenever a number of events is counted during certain time periods. Examples are the monthly number of car accidents in a region, the weekly number of new cases in epidemiology, the number of transactions at a stock market per minute in finance, or the number of photon arrivals per microsecond in a biological experiment. A natural modification of the popular autoregressive moving average (ARMA) models for continuous variables is based on the assumption that the observation $Y_t$ at time $t$ is generated by a generalized linear model (GLM) conditionally on the past, choosing an adequate distribution for count data like the Poisson and a link function $\eta(\cdot)$. This approach of time series following a GLM is pursued e.g. by Kedem and Fokianos (2002). Restricting ourselves to first order models, we consider time series $(Y_t : t \in \mathbb{N}_0)$ following a Poisson model

$$Y_t|\mathcal{F}_{t-1}^Y \sim \text{Pois}(\lambda_t),$$

$$\eta(\lambda_t) = \beta_0 + \beta_1 \eta(Y_{t-1} + c) + \alpha_1 \eta(\lambda_{t-1}), \ t \geq 1,$$

where $\mathcal{F}_{t-1}^Y$ stands for the $\sigma$-algebra created by $\{Y_{t-1}, \ldots, Y_0, \lambda_0\}$, while $\beta_0, \beta_1, \alpha_1$ are unknown parameters, and $c$ is a known constant. Models employing other distributions like the negative binomial could be treated similarly.

The natural choice for $\eta$ is the logarithm, and this is the reason for adding the constant $c$ to $Y_{t-1}$ in the term $\eta(Y_{t-1} + c)$, since we need to avoid difficulties arising from observations which are equal to 0. Following Fokianos and Tjøstheim (2011), who develop ergodicity conditions for a subclass of the arising log-linear models, we set $c = 1$. Another choice for $\eta$ which has received some attention is the identity,
\( \eta = id \), see e.g. Ferland, Latour and Oraichi (2006). In this case we can set \( c \) to 0. For ergodicity conditions for this model class see Fokianos, Rahbek and Tjøstheim (2009).

We briefly discuss possible interpretations of models like those given in (1) in the context of epidemiology, with \( Y_t \) denoting the number of new cases observed at time \( t \). For a fixed population size, the conditional mean \( \lambda_t \) measures the risk of a person to fall ill at time \( t \). Our model assumes that all effects on \( \lambda_t \) are linear after transformation to a suitable scale by \( \eta \). The term \( \eta(Y_{t-1} + c) \) in the second equation models the dependence of the transformed conditional mean \( \eta(\lambda_{t-1}) \) and thus of the observation \( Y_t \) on the previous value \( Y_{t-1} \), with \( \beta_1 \) measuring the strength of this dependence. A large number of cases \( Y_{t-1} \) at time \( t-1 \) can cause a large number of cases \( Y_t \) at time \( t \) because the risk of infection increases. The term \( \eta(\lambda_{t-1}) \) additionally describes that there can be periods of increased risk also because of certain weather conditions or expositions, for instance, and \( \alpha_1 \) measures the size of such dependencies.

Given a model as formulated in (1), a basic question is whether it properly describes all the observations of a given time series, or whether some observations have been influenced by extraordinary effects, which are called interventions in what follows. Outlier and intervention analysis for ARMA processes of continuous variables has been developed by Fox (1972), Box and Tiao (1975), Tsay (1986), Chang, Tiao and Chen (1988) and Chen and Liu (1993), among others. However, counts are positive and typically right-skewed, causing a need for especially designed models and procedures.

In the following sections, we review the intervention models proposed by Fokianos and Fried (2010, 2012) for time series which are Poisson conditionally on the past, with \( \eta \) being the identity and the log-link, respectively, and describe some extensions.

## 2 Models for Intervention Analysis

A possibility to introduce an extraordinary effect on a time series \((Y_t)\) generated by (1) is the assumption that from a time point \( \tau \) on the underlying conditional mean process is changed by adding terms \( \omega \delta^{t-\tau} I(t \geq \tau) \) to \( \eta(\lambda_t) \), so that instead of \((Y_t)\) we observe a contaminated process \((Z_t)\) generated from a model with contamination,

\[
Z_t | F_{t-1}^Z \sim \text{Pois}(\lambda^c_t), \\
\eta(\lambda^c_t) = \beta_0 + \beta_1 \eta(Z_{t-1} + c) + \alpha_1 \eta(\lambda^c_{t-1}) + \omega \delta^{t-\tau} I(t \geq \tau), \ t \geq 1.
\]

In obvious notation, \((\lambda^c_t)\) is the contaminated process of conditional means, which coincides with \((\lambda_t)\) until time \( \tau - 1 \) and then becomes affected, while \( F_{t-1}^Z \) denotes the \( \sigma \)-algebra representing the information on the past of the contaminated process and the initial values, analogous to \( F_{t-1}^Y \). The new parameter \( \omega \) determines the size of the effect, \( I(t \geq \tau) \) indicates whether \( t \geq \tau \) or not, and \( \delta \in [0,1] \) determines whether the effect is concentrated on time \( \tau \) (in case of \( \delta = 0 \)), causing a spiky outlier, whether the whole level is shifted from time \( \tau \) on (\( \delta = 1 \)), or whether a geometrically decaying transient shift with rate \( \delta \in (0,1) \) occurs. Note that even in case of \( \delta = 0 \) the whole future of the process is affected by an intervention, since its effect enters the dynamics both via \( Z_t \) and \( \eta(\lambda^c_t) \), \( t \geq \tau \). Continuing the explanations given above in the context
of epidemiology, an intervention according to (2) can be interpreted as an internal change of the data generating process. For some reason, the conditional mean of the process (the risk) changes in an unpredictable manner at time $\tau$, and this changes the observation for that time point, and also the observations thereafter.

Liboschik et al. (2013) explore another intervention model in case of the identity link. In their approach, an intervention affects the observation at time $\tau$, but not the underlying conditional mean. This can be understood as an external change, as the contaminated observation $Z_{\tau}$ equals the sum of the uncontaminated value $Y_{\tau}$ plus a random number $C_{\tau}$, which arises because of extraordinary reasons and enters the dynamics of the process in the same way as $Y_{\tau}$, while the underlying risk $\lambda_{\tau}$ initially is not affected. An example might be people being infected due to external reasons, e.g. on a journey. The modified intervention model with a general link function $\eta$ reads

$$Z_{t} | F_{t-1} \sim \text{Pois}(\lambda_{t}),$$

$$\eta(\lambda_{t}) = \eta(\lambda_{t-1} + c) + \alpha_{1}\eta(\lambda_{t-1}), \ t \geq 1.$$ (3)

The last two equations describing the conditional mean process can be summarized as

$$\eta(\lambda_{t}) = \beta_{0} + \beta_{1}\eta(Z_{t-1} + c) + \alpha_{1}(\eta(\lambda_{t-1}) - \omega\delta^{t-\tau}I(t - 1 \geq \tau)) + \omega\delta^{t-\tau}I(t \geq \tau).$$

This shows the difference to model (2) more clearly.

If the time point $\tau$ and the type of an intervention, i.e. the value of $\delta$, both are known, an intervention model as formulated in (2) or (3) can be fitted by maximizing the conditional likelihood iteratively, starting from suitable initial values. The existence of such a known intervention can be confirmed by comparing the test statistics of the corresponding score test to the upper percentiles of its asymptotical $\chi^2$-distribution, as described in the papers mentioned above. If only the time point $\tau$ is unknown, but the type is known, simulation experiments indicate that parametric bootstrap procedures work rather well: fit the model without intervention effects and calculate the score test statistics for all time points. Use the maximum of all score test statistics for all time points as the final test statistic. Then generate artificial time series without interventions from the fitted model and calculate the corresponding maximum score test statistic as well. Opt for an intervention at that time point which maximizes the score test statistic for the real data, if it is among the largest $100\alpha$-percent of all maximum score test statistics. If the type of the intervention is unknown as well, the maximum score test statistics can be calculated for each type given either model (2) or (3). The simulations suggest that preference should be given to level shifts ($\delta = 1$) if they turn out to be significant, since a level shift usually causes the test statistics for the other types of intervention effects also to become large, while the reverse effect is much less pronounced. Multiple interventions can be dealt with by estimating the effect of a detected intervention and subtracting it from the time series, before the cleaned data are analyzed with respect to further interventions.

Note that the above intervention models are not able to describe so called additive outliers representing e.g. pure measurement or reporting errors, i.e. the case where a
single observation is changed without any effects on the future of the process. Actually, such additive outliers are difficult to deal with by a frequentist approach, since we would need to condition on the unobserved value $Y_\tau$ instead of the contaminated $Z_\tau$. Fried et al. (2012) develop a Bayesian approach for additive outliers, applying Markov Chain Monte Carlo techniques. Their simulation results provide evidence that in this way it is possible to deal with additive outliers if there are several of them. A single or very few additive outliers pose difficulties to a Bayesian approach based on little informative prior distributions, since they do not provide enough information on that component of the underlying mixture distribution which causes the outliers.

Furthermore it should be noted that we implicitly assume intervention effects to be additive when using the identity link, and multiplicative on the original scale when using the log-link, since for simplicity we introduce the intervention effects in the same way as the dependencies on the past. Another assumption underlying the intervention models formulated above, and also the common outlier and intervention models which have been proposed for ARMA processes in the literature, is that the dynamics of the process does not change and follows the same model after an intervention as before it.

For an illustration we analyze an artificial time series of length $n = 200$ generated from model (2) with $\eta = id$, $\beta_0 = 3$, $\beta_1 = 0.4$, $\alpha_1 = 0.3$, an internal level shift of size $\omega_1 = 4$ at time $\tau_1 = 100$ and an internal spike of size $\omega_2 = 30$ at time $\tau_2 = 150$.

![Figure 1: Results obtained from fitting both intervention models to a time series with an internal level shift at time 100 and an internal spike at time 150.](image)

The results obtained from fitting both intervention models to these data are illustrated in Figure 1. The spike and the level shift are detected when using either of these two models, albeit with some differences between the estimated parameter values and outlier sizes, according to the different influences of such patterns on the time series. These findings confirm those of Liboschik et al. (2013): interventions can be detected successfully even if the wrong model is used. This is good news and also bad news: it is good news since it implies a certain robustness against model misspecification, but it makes a statement about the cause of an intervention effect and about its mechanism (internal / external) difficult. More work on model selection is needed for this.
3 Extensions and Outlook

3.1 Surveillance

The methods for detection of intervention effects in count time series mentioned above work retrospectively, i.e., we observe the whole time series before it is analyzed. An open problem so far is how these models can be used for surveillance, i.e. online detection of changes. This is an interesting problem not only in epidemiology, where we want to detect the outbreak of an epidemic with only short time delays. An intuitive approach is to compare an incoming observation $y_{n+1}$ to its 1-step prediction $\hat{\lambda}_{n+1}$, obtained by estimating the parameters of model (1) from the data observed until time point $n$, plugging in these estimates into the formula for $\eta(\lambda_{n+1})$ and applying the inverse transform $\eta^{-1}$. There is evidence of an extraordinary effect if $y_{n+1}$ is larger than the upper, say, 99% percentile of a Poisson distribution with parameter $\hat{\lambda}_{n+1}$.

An analysis of a single observation cannot tell us which type of intervention occurs, e.g. whether there is a spiky outlier or a level shift. For this we need to wait some more time points until further values $y_{n+2}, y_{n+3}, \ldots, y_{n+m}$ are observed, with a suitably chosen delay $m \in \mathbb{N}$. Instead of its 1-step ahead prediction, a comparison of $y_{n+h}$ to its $h$-step ahead prediction might be advantageous then, since the 1-step ahead prediction will strongly be affected by a level shift at time $n+1$ due to its use of $y_{n+1}, \ldots, y_{n+h-1}$. To the best of our knowledge, so far there are no simple formulae available for the conditional expectation of $Y_{n+h}$ given $\mathcal{F}_n^Y$ if $h \geq 2$, which is the natural candidate for $h$-step ahead prediction, so that we would need to rely on simulating the future given the fitted model, or use simple linear predictions instead, sticking the previous predictions $\hat{y}_{t+h-1} = \hat{\lambda}_{t+h-1}$ into the formula for $\eta(\hat{\lambda}_{t+h})$ for $h = 2, 3, \ldots, m$. However, note that the conditional distribution of $Y_{n+h}$ given $\mathcal{F}_n^Y$ is not Poisson for $h \geq 2$, so that there is need for more research on these models.

3.2 Robust estimation

Further open questions remain concerning the robust estimation of the model parameters in the presence of outliers and intervention effects. This is even more important because it is difficult to specify intervention effects correctly and because of the difficulties in dealing with a single or a few additive outliers outlined above.

M-estimators are a popular generalization of (conditional) maximum likelihood estimators which provide some robustness against outliers by replacing the log-likelihood or the score function by more robust alternatives. An M-estimator of a parameter $\theta$ can be defined as the solution of a score equation

$$\sum_{t=1}^{n} \psi(y_t, \hat{\theta}) = 0 .$$

Maximum likelihood estimation is derived by choosing $\psi(y, \theta)$ as the derivative of the log-density $\ln f_\theta(y)$ with respect to $\theta$, i.e. as the usual score function, while $\psi(y, \theta) = y - \theta$ corresponds to least squares and $\psi(y, \theta) = \text{sign}(y - \theta)$ to least absolute deviation.
estimation of location. The popular Huber M-estimator of the location parameter $\theta$ in a location-scale model with known (or preliminarily estimated) scale $\sigma$ is derived from

$$\psi(y, \theta) = \frac{y - \theta}{\sigma} I(-k\sigma \leq y - \theta \leq k\sigma) + k \text{ sign}(y - \theta) I(|y - \theta| > k\sigma),$$

where $k$ is a tuning constant which determines the efficiency and the robustness of the resulting estimator. For $k = 0$ we get least absolute deviations and for $k \to \infty$ we get least squares. The score function of the Huber M-estimator is monotone. This guarantees a unique solution which can easily be determined iteratively starting from any initial value. The score function of the Tukey M-estimator,

$$\psi(y, \theta) = \frac{y - \theta}{\sigma} \left( k^2 - \frac{(y - \theta)^2}{\sigma^2} \right)^2 I(-k\sigma \leq y - \theta \leq k\sigma),$$

however, is redescending to 0 as $y - \theta$ approaches $\pm k\sigma$. This leads to the possibility of multiple solutions of the defining score equations (4).

M-estimation of generalized linear models using the Huber $\psi$-function has been treated by Cantoni and Ronchetti (2001). However, in our basic model (1) we regress on previous observations and previous conditional means, and it is well known that monotone M-estimators like the Hubers need further modifications to become robust against outlying regressors. Cantoni and Ronchetti (2001) consider covariates following an elliptical distribution and use weights based on robustly estimated Mahalanobis distances to downweight observations with outlying regressors. This approach is not natural in our context, since we regress on previous observations, which are conditionally Poisson, or some transformation of them. Empirical work on model (2) with the log-link and $\alpha_1 = 0$, that is a model without feedback, indicates that in the cases of level shift and transient shift there are no significant differences between the classical maximum likelihood estimation and the approach based by Cantoni and Ronchetti (2001). This agrees with findings for Gaussian ARMA models, that maximum likelihood and least squares work rather well in case of outliers which conform to the dynamics of the process. In the case of additive outliers, the weighted approach through robust Mahalanobis distances was found to perform much better than the classical maximum likelihood estimation, especially as the number of outliers increases. In fact, some further empirical work on the feedback case ($\alpha_1 \neq 0$) indicates that the Cantoni and Ronchetti (2001) estimation approach performs better with weights.

Maronna, Martin and Yohai (2006) recommend Tukey’s $\psi$-function since its redescending behavior completely eliminates the influence of huge outliers and provides some robustness even in the case of outlying regressors. However, we need to use highly robust initial parameter estimates then, in order not to get trapped in a wrong solution when trying to solve (1) iteratively. This and the discreteness and strong asymmetries of Poisson models pose further problems which are not encountered in ordinary symmetric location-scale models. This will briefly be illustrated in the context of independent Poisson data in the following.

Cadigan and Chen (2001) investigate a modification of the Huber score function for the Poisson distribution. Under Poisson assumptions, the variance $\sigma^2$ equals the
mean $\theta$, so that we can replace $\sigma$ by $\sqrt{\theta}$ in the above score functions, see also Elsaied (2012). Furthermore, the expectation of $\psi(Y, \theta)$ has to be zero for getting asymptotically unbiased estimates. This can be accomplished by introducing a bias correction $a$ and replacing $(y - \theta)/\sigma$ by $(y - \theta)/\sqrt{\theta} - a$ in the above formulae. Given the need for a highly robust initial estimate when using the Tukey $\psi$-function, we might want to apply the median of the data, but this only works if it is not zero because of our scaling by $\sqrt{\theta}$, and it provides only a very rough estimate if the sample median is small. Elsaied (2012) proposes an adaptive estimate instead, combining the sample median with an estimate derived from the frequency of zero observations.

The asymptotical distribution of an M-estimator under suitable regularity conditions is $N(\theta, V_\psi(\theta))$, with the asymptotical variance $V_\psi(\theta) = E(\psi(Y, \theta)/B_\theta)^2$, where $B_\theta = \partial E\psi(Y, \theta)/\partial \theta$, see e.g. Maronna, Martin and Yohai (2006). The relative efficiency of an M-estimator as compared to the maximum likelihood estimator, which is the sample mean, under these conditions thus becomes $\theta/V_\psi(\theta)$, and is illustrated in Figure 2. Note that an estimator with a fixed tuning constant $k$ does not achieve a desirable high level of efficiency for all possible values of $\theta$. For further investigations in this respect and a first attempt to formulate robust M-estimators for model (1) with $p = 1$, $q \in \{0, 1\}$ and the identity link see Elsaied (2012).

Acknowledgement: The work of K. Fokianos and S. Kitromilidou is supported by Cyprus Research Promotion Foundation TEXNOLOGIA/THEPIS/0609(BE)/02.

References


SEQUENTIAL TESTING OF PARAMETRIC HYPOTHESES: PERFORMANCE, ROBUSTNESS AND APPLICATION

A.Yu. Kharin
Belarussian State University
Minsk, BELARUS
e-mail: KharinAY@bsu.by

Abstract

The problems of sequential testing of hypotheses on parameters of stochastic data being observed are considered. An approach to the calculation of the performance characteristics (error probabilities and expected sample number of observations to be used) is developed. Asymptotic expansions for the performance characteristics are derived under distortions of the hypothetical models for sequential tests from the proposed family of tests. The robust sequential tests are constructed under the minimax risk criterion. The results are used in medicine for the quickest in the mean accessibility detection of a parameter value set in the model with a dynamic parameter.

1 Introduction

The sequential approach [9] is intensively used in computer data analysis of stochastic data for hypotheses testing [8]. The performance characteristics of sequential tests are problematic to calculate with a given accuracy even for basic hypothetical models [7]. The sequential procedures are applied for the real data sets that are usually not adequate to certain hypothetical model. This hypothetical model is often distorted. Hence the robustness analysis and robust sequential test construction are important problems [1], [6].

2 Robustness in Case of Simple Hypotheses

Let on a probability space $(\Omega, \mathcal{F}, P)$ discrete random variables $x_1, x_2, \ldots$ be defined, $\forall t \in \mathbb{N}$, $x_t \in U = \{u_1, u_2, \ldots, u_M\}$, $M < \infty$, $u_1 < u_2 < \ldots < u_M$. Let these random variables be independent identically distributed, from the discrete probability distribution with a parameter $\theta \in \Theta = \{\theta_0, \theta_1\}$:

$$
P(u; \theta) = P_x\{x_t = u\} = a^{-J(u; \theta)}, \ t \in \mathbb{N}, u \in U,$$

(1)

$a \in \mathbb{N} \setminus \{1\}$; $J(u; \theta): U \times \Theta \rightarrow \mathbb{N}_0$ is a function satisfying $\sum_{u \in U} a^{-J(u; \theta)} = 1$.

Consider two simple hypotheses w.r.t. the parameter $\theta$:

$$
H_0 : \ \theta = \theta_0, \ H_1 : \ \theta = \theta_1.
$$

(2)
Such a problem appears in applications, where one of the two possible regimes can be realized.

Introduce the notation:

\[ \Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} \lambda_i; \]
\[ \lambda_i = \log_2 \left( \frac{P(x_i; \theta_1)}{P(x_i; \theta_0)} \right) = J(x_i; \theta_0) - J(x_i; \theta_1) \in \mathbb{Z}. \]

To test these hypotheses by \( n \) \((n = 1, 2, \ldots)\) observations let us consider the sequential probability ratio test (SPRT) [9]:

\[ d_n = 1_{[C_+, +\infty)}(\Lambda_n) + 2 \cdot 1_{(-\infty, C_-)}(\Lambda_n), \]  
\[ \text{where } 1_D(\cdot) \text{ is the indicator function of the set } D. \]

The decisions \( d_n = 0 \) and \( d_n = 1 \) mean stopping of the observation process and the acceptance of the appropriate hypothesis. The decision \( d_n = 2 \) means that it is necessary to make the \((n + 1)\)-th observation. In (3) the thresholds \( C_- < C_+ \) are the given values (parameters of the test). According to [9], we use

\[ C_+ = [\log_2 ((1 - \beta_0)/\alpha_0)], \quad C_- = [\log_2 (\beta_0/(1 - \alpha_0))], \]

where \( \alpha_0, \beta_0 \) are given maximal possible values of the probabilities of type I and type II errors respectively. In fact, the true values \( \alpha, \beta \) for the probabilities of type I and type II errors differ from \( \alpha_0, \beta_0 \).

For \( n \in \mathbb{N} \), define the random sequence

\[ \xi_n = C_+ 1_{[C_+, +\infty)}(\Lambda_n) + C_- 1_{(-\infty, C_-)}(\Lambda_n) + \Lambda_n 1_{(C_- C_+)}(\Lambda_n) \in \mathbb{Z}. \]  
\[ \text{Introduce the notation: } I_k \text{ is the identity matrix of the } k\text{-th order; } 0_{m \times n} \text{ is the } (m \times n)\text{-matrix all elements of which are equal to 0}; 1(u) \text{ is the unit step function; } 1_k \text{ is the } k\text{-vector-column all elements of which are equal to 1}. \]

Define the one-step transition probabilities matrix and the initial probabilities vector for the transient states

\[ P^{(k)} = \left( p_{ij}^{(k)} \right) = \begin{pmatrix} I_2 & 0_{2 \times (N-2)} \\ R^{(k)} & Q^{(k)} \end{pmatrix}, \quad \pi^{(k)} = \begin{pmatrix} \pi_{C_+ +1}^{(k)} \\ \vdots \\ \pi_{C_- -1}^{(k)} \end{pmatrix}, \]  
\[ \text{where the blocks } R^{(k)}, Q^{(k)}, \text{ and the vector } \pi^{(k)} \text{ are} \]

\[ p_{ij}^{(k)} = \begin{cases} \sum_{u \in U} \delta_j^{(u; \theta_0) - J(u; \theta_1) - i} P(u; \theta_k), & i, j \in (C_-, C_+), \\ \sum_{u \in U} 1(C_+ = i + J(u; \theta_1) - J(u; \theta_0)) P(u; \theta_k), & j = C_- \}
\left\{ \begin{array}{l} \sum_{u \in U} 1(J(u; \theta_0) - J(u; \theta_1) + i - C_+) P(u; \theta_k), & j = C_+, \\ \sum_{u \in U} \delta_j^{(u; \theta_0) - J(u; \theta_1) - i} P(u; \theta_k), & i \in (C_-, C_+). \end{array} \right. \]

For the hypothesis \( H_k \), let \( t^{(k)} \) be the expected stopping time of the decision process (expected number of observations), and \( B^{(k)} \) be the \((N-2) \times 2\)-matrix of absorption probabilities: its \((i, j)\)-th element equals to the probability of absorption at the state \( j \) (acceptance of the hypothesis \( H_j \)) starting from the state \( \xi_i = i \in (C_-, C_+) \). Let us denote the \( i \)-th column of a matrix \( W \) by \( W(i) \).
Theorem 1 If under conditions (1) — (2) the true hypothesis is the hypothesis \( H_k \), and the matrix \( S^{(k)} = I_{N-2} - Q^{(k)} \) is nonsingular, then for the test (3)

\[
i^{(k)} = (\pi^{(k)})' (S^{(k)})^{-1} 1_{N-2} + 1, \quad B^{(k)} = (S^{(k)})^{-1} R^{(k)}.
\]

(8)

Corollary 1 Under the Theorem 1 conditions the error probabilities of type I and type II are \( \alpha = (\pi^{(0)})' B^{(0)}_{(2)}, \beta = (\pi^{(1)})' B^{(1)}_{(1)} \).

Let the hypothetical model (1), (2) be under distortions of Tukey–Huber type [1]: instead of (1) the observations \( x_1, x_2, \ldots \) are taken from the contaminated discrete probability distribution

\[
\tilde{P}(u; \theta) = \tilde{P}_\theta \{ x_t = u \} = (1 - \varepsilon)P(u; \theta) + \varepsilon\tilde{P}(u; \theta), \quad \tilde{P}(u; \theta) = a^{-\tilde{J}(u; \theta)},
\]

(9)

\( \tilde{J}(u; \theta): U \times \Theta \rightarrow N_0 \) is a function different from \( J(\cdot) \), satisfying \( \sum_{u \in U} a^{-\tilde{J}(u; \theta)} = 1 \).

Define the matrix \( \tilde{P}^{(k)} \) analogous to (6), substituting \( P(\cdot) \) with \( \tilde{P}(\cdot) \).

Theorem 2 If the hypothetical model (1), (2) is distorted according to (9), and the matrices \( S^{(k)}, \tilde{S}^{(k)} = I_{N-2} - Q^{(k)} - \varepsilon(\hat{Q}^{(k)} - Q^{(k)}) \) are nonsingular, then the expected number of observations \( \tilde{t}^{(k)} \) and the absorption probabilities matrix \( \tilde{B}^{(k)} \) for the distorted model differ from the same characteristics for the hypothetical model by the values of the order \( O(\varepsilon) \):

\[
\tilde{t}^{(k)} - t^{(k)} = \varepsilon \left((\tilde{\pi}^{(k)} - \pi^{(k)})' + (\pi^{(k)})' (S^{(k)})^{-1} (\hat{Q}^{(k)} - Q^{(k)}) \right) \times \\
\times (S^{(k)})^{-1} (1 \ldots 1)' + O(\varepsilon^2); \quad \tilde{B}^{(k)} - B^{(k)} = \varepsilon (S^{(k)})^{-1} \times \\
\times ((\hat{Q}^{(k)} - Q^{(k)})(S^{(k)})^{-1} R^{(k)} + \tilde{R}^{(k)} - R^{(k)}) + O(\varepsilon^2),
\]

where \( \hat{Q}^{(k)}, \tilde{R}^{(k)} \) are the blocks of the matrix \( \tilde{P}^{(k)} \).

Corollary 2 Under the Theorem 2 conditions the error probabilities \( \bar{\alpha}, \bar{\beta} \) of types I and II for the distorted model differ from the same characteristics for the hypothetical model by the values of the order \( O(\varepsilon) \):

\[
\bar{\alpha} - \alpha = \varepsilon \left( (\pi^{(0)})' (S^{(0)})^{-1} ((\hat{Q}^{(0)} - Q^{(0)}) (S^{(0)})^{-1} R^{(0)}) + \\
+ \hat{R}^{(0)} - R^{(0)})_{(2)} + (\pi^{(0)} - \pi^{(0)}) B^{(0)}_{(2)} \right) + O(\varepsilon^2),
\]

\[
\bar{\beta} - \beta = \varepsilon \left( (\pi^{(1)})' (S^{(1)})^{-1} ((\hat{Q}^{(1)} - Q^{(1)}) (S^{(1)})^{-1} R^{(1)}) + \\
+ \hat{R}^{(1)} - R^{(1)})_{(1)} + (\pi^{(1)} - \pi^{(1)}) B^{(1)}_{(1)} \right) + O(\varepsilon^2).
\]

Using the theory presented above the minimax robust sequential test is constructed in [2]. The generalizations of the results for the case of arbitrary discrete distributions and for dependent observations are discussed in [5].
3 Testing of Composite Hypotheses

Suppose a sequence \( x_1, x_2, \ldots \) of i.i.d. random variables is observed from a continuous distribution with the p.d.f. \( p(x \mid \theta) \), where \( \theta \in \Theta \subseteq \mathbb{R} \) is an unknown value of random parameter. Consider two composite hypotheses

\[
H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1; \tag{10}
\]

\( \Theta_0, \Theta_1 \subseteq \Theta, \Theta_0 \cap \Theta_1 = \emptyset \). Assume that the prior p.d.f. \( p(\theta) \) is known.

One of the possible techniques to test the hypotheses (10) is using of weight functions proposed by Wald [9]. Introduce the notation:

\[
W_i = \int_{\Theta_i} p(\theta) d\theta, \quad w_i(\theta) = \frac{1}{W_i} p(\theta) \cdot 1_{\Theta_i}(\theta), \quad \theta \in \Theta, \quad i = 0, 1; \tag{11}
\]

\[
\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \ln \frac{\int_{\Theta_1} w_1(\theta) \prod_{i=1}^n p(x_i \mid \theta) d\theta}{\int_{\Theta_0} w_0(\theta) \prod_{i=1}^n p(x_i \mid \theta) d\theta}. \tag{12}
\]

For testing the hypotheses (10), under the notation (11), (12) the following parametric family of tests is used:

\[
N = \min\{n \in \mathbb{N} : \Lambda_n \notin (C_-, C_+)\}, \tag{13}
\]

\[
d = 1_{[C_+, +\infty)}(\Lambda_N), \tag{14}
\]

where (13) gives the stopping rule, \( N \) is the random number of the observation, at which the decision \( d \) is made according to (14); \( d = i \) means that the hypothesis \( H_i \), \( i = 0, 1 \), is accepted; \( C_- < 0, C_+ > 0 \) are parameters of the test, which are usually chosen in practice according to (4).

Introduce the discretization parameters

\[
m \in \mathbb{N}, \ C_1, \ldots, C_{m-1} \in \mathbb{R}, \ C_- = C_0 < C_1 < C_2 < \ldots < C_{m-1} < C_m = C_+;
\]

\[
A_i = [C_{i-1}, C_i], \quad i = 1, \ldots, m, \quad A_- = (-\infty, C_-), \quad A_+ = (C_+, +\infty); \quad h = \max_{1 \leq i \leq m} \text{mes} A_i. \tag{15}
\]

**Theorem 3** Let the parameters of discretization (15) and the p.d.f. \( p(x \mid \theta) \) satisfy the condition:

\[
\forall n \in \mathbb{N} : P\{\Lambda_1 \in A^1, \ldots, \Lambda_n \in A^n\} > 0, \tag{16}
\]

and the statistic (12) can be presented in the form: \( \Lambda_n = \Psi_n(\bar{x}(n)), \nu_n(\cdot) : \mathbb{R} \rightarrow \mathbb{R} \) is a (strictly) increasing function. Then \( \forall n \in \mathbb{N}, \forall k \in \{1, \ldots, n-1\}:

\[
P\{\Lambda_{n+1} \in A^{n+1} \mid \Lambda_n \in A^n, \ldots, \Lambda_{n-k} \in A^{n-k}\} = P\{\Lambda_{n+1} \in A^{n+1} \mid \Lambda_n \in A^n\} + \mathcal{O}(h^2), \tag{17}
\]

where \( A^j \subseteq \{A_1, \ldots, A_m\}, j = 1, \ldots, n, \ A^{n+1} \subseteq \{A_1, \ldots, A_m, A_-, A_+\} \).
Theorem 3 states that the Markov property holds for the random sequence $\Lambda_n$ with the accuracy of the order $O(h^2)$.

Using the results of Theorem 3 the special approximating Markov chain is constructed enabling to calculate the performance characteristics of the test (11) — (14). The robustness of the test is analysed under distortions of prior probability density function as well as under distortions of the p.d.f. of observations [3]. The minimax robust sequential test are constructed within the proposed family of generalized sequential tests [4].

4 Sequential Detection of a Given Parametric Set

Accessibility

Let an $N$-dimensional time series be defined on a probability space $(\Omega, \mathcal{F}, P)$: $x_t = (X_{t1}, \ldots, X_{tN})' \in \mathbb{R}^N$, $t \in T \subseteq \mathbb{N}$, that satisfies the following trend model:

$$X_t = X_0 + F(t, \theta) + \Xi_t, \quad t \in T,$$

(18)

where $X_0 = (x_{0i}) \in \mathbb{R}^N$ is a known (not random) initial state of the time series, $F(t, \theta) = (F_i(t; \theta))$: $T \times \mathbb{R}^m \rightarrow \mathbb{R}^N$ is a known vector function of the trend with unknown $m$-vector of parameters $\theta = (\theta_i) \in \mathbb{R}^m$, $\Xi_t = (\Xi_{ti}) \in \mathbb{R}^N$ is a Gaussian random vector:

$$E\{\Xi_t\} = 0_N, \quad E\{\Xi_t \cdot X_{it}'\} = \Sigma = (\sigma_{ij}), \quad i, j \in \{1, \ldots, N\} : \mathcal{L}\{\Xi_t\} = \mathcal{N}(0_N, \Sigma).$$

(19)

As a special case the linear trend may be considered:

$$F(t; \theta) = \theta \cdot t, \quad \theta \in \mathbb{R}^m, \quad t \in T.$$

(20)

Let us use the ellipsoid metric (also known as the Machalanobis distance): $\rho(x, y) = \sqrt{(x - y)'B(x - y)} \geq 0$, $x, y \in \mathbb{R}^N$, where $B = (b_{ij})$ is a symmetric positively defined $N \times N$-matrix. Let $A = (a_i) \in \mathbb{R}^N$ be a fixed point (etalon):

$$A_\varepsilon = \{x \in \mathbb{R}^N : \rho(x, A) \leq \varepsilon\}$$

(21)

be a given $\varepsilon$-neighborhood of point $A$ ($\varepsilon > 0$ is a given real number).

Introduce a set in the parameter space:

$$\Theta_0 = \left\{ \theta \in \mathbb{R}^m : \min_{t \in T} \rho(X_0 + F(t, \theta), A) \leq \varepsilon \right\}.$$  

(22)

It follows from (22) that $\Theta_0$ is a set of the parameter $\theta$ values, for wich the trend trajectories $X_0^t = X_0 + F(t, \theta), \ t \in T$, $X_0^0 = X_0$, starting from point $X_0$ at the time moment $t = 0$, for some $t^* \in T$ gets into the $\varepsilon$-neighborhood (21) of the point $A$: $X_0^{t^*} \in A_\varepsilon$, and has the minimal distance from the point $A$ (within the time moments $t \in T$, for each fixed $\theta$).

Finding of time moment $t^* \in T$ means solving the following etreme problem:

$$\rho_t = (X_0 + F(t; \theta) - A)' \cdot B \cdot (X_0 + F(t; \theta) - A) \rightarrow \min_{t \in T},$$

that is solved numerically in the general case.
Theorem 4 For the model (18) — (20) the likelihood function of the time series $X_1, \ldots, X_n \in \mathbb{R}^N$ being observed during the $n$ sequential time intervals is

$$L(\theta, \Sigma) = (2\pi)^{-nN/2} \cdot |\Sigma|^{-1/2} \cdot \exp \left\{ -\frac{1}{2} \sum_{t=1}^{n} (X_t - X_0 - \theta \cdot t)' \Sigma^{-1} (X_t - X_0 - \theta \cdot t) \right\}.$$ 

In construction of the sequential statistical test for hypothesis $H_0$ against the general alternative, the problem of the conditional optimization appears: $L(\theta; \Sigma) \rightarrow \max_{\theta \in \Theta_0, \Sigma=\Sigma'>0}$. The parametric form of the solution is found analytically, and the correspondent parameter value is obtained numerically. The sequential test and its robust versions are constructed.

The results are illustrated via computer experiments.

The research is partially supported by the ISTC Project B-1910.

References


LONG-MEMORY DISCRETE-VALUED TIME SERIES: MODELS AND METHODS

Yu.S. Kharin
Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University
Minsk, BELARUS
e-mail: Kharin@bsu.by

Abstract
Discrete-valued time series with long memory are considered. To avoid the "curse of dimensionality" of the parameter space we propose to use the parsimonious models of high-order Markov chains that are determined by small number of parameters. Methods for statistical analysis of these models are developed and illustrated on simulated and real statistical data.

1 Introduction

Adequate models of discrete-valued time series are required for computer modeling and analysis of complex systems and processes in many applied fields [4]: in bioinformatics for recognition and analysis of long genetic sequences, in medical diagnostics to predict some events by recorded sequences of digital data, in meteorology to forecast the weather, in sociology for modeling of dynamics in social behavior, in Internet-traffic analysis for optimizing protocols.

Discrete-valued time series is a random process \( x_t \in A \) on some probability space \( (\Omega, F, P) \) with discrete time \( t \in \mathbb{N}_0 = \{0, 1, 2, \ldots\} \) and a discrete state space \( A \) with the cardinality \( N = |A|, 2 \leq N \leq +\infty \). Much attention to discrete-valued time series considered as categorical time series was paid by B. Kedem and K. Fokianos [4].

If \( A \) is any countable set \( (N = +\infty) \), then we have a countable valued time series \( x_t \). One of the well-known model for such a class of time series in the case with \( A = \mathbb{N}_0 \) is the so-called Integer Autoregression Model of order \( p \in \mathbb{N}_0 \) (INAR(\( p \))) determined by the following stochastic difference equation [1]:

\[
x_t = \sum_{i=1}^{p} \sum_{j=1}^{t-i} \xi_{tij} + \eta_t, \quad t \geq p + 1,
\]

where \( \{\xi_{tij} : t \geq p + 1, j \geq 1, i = 1, 2, \ldots, p\} \) are jointly independent Bernoulli random variables with \( P\{\xi_{tij} = 1\} = 1 - P\{\xi_{tij} = 0\} = \alpha_i \in (0, 1), \{\eta_t\} \) are jointly independent nonnegative random variables.

If \( A \) is any finite set \( (N < +\infty) \), then we have a finitely valued time series, also called categorical time series [4]. We consider here this case and without loss of generality assume that \( A = \{0, 1, \ldots, N - 1\} \).

In modern theory of statistical time series analysis because of practice demands much attention is paid to long-memory models. In the well developed theory of “continuous” time series with the state space \( A = (-\infty, +\infty) \) the most studied long-memory model is stationary time series with a “slow decaying” covariance function [10]:

\[
\rho(\tau) = \text{Cov}\{x_t, x_{t+\tau}\} = C \cdot |\tau|^{2d-1}, \quad d \in (-0.5, 0.5), \quad \tau \in \mathbb{Z}.
\]
An example of such a time series is the fractionally integrated ARIMA process $\text{FARIMA}(p, d, q)$ [10].

Theory of long-memory discrete-valued time series is developed weaker. An universal long-memory model for discrete-valued time series $x_t$ is the homogeneous Markov chain $\text{MC}(s)$ of the order $s \in \mathbb{N}_0$, determined by the generalized Markov property ($t > s$):

$$\mathbf{P}\{x_t = i_t | x_{t-1} = i_{t-1}, \ldots, x_1 = i_1\} = \mathbf{P}\{x_t = i_t | x_{t-1} = i_{t-1}, \ldots, x_{t-s} = i_{t-s}\} = p_{i_{t-s}, \ldots, i_{t-1}, i_t},$$

(1)

where $s$ is the memory length; $i_1, i_2, \ldots, i_t \in \mathbf{A}$ are values of the process at the time moments $1, 2, \ldots, t$; $\mathbf{P} = (p_{i_{t-s}, \ldots, i_{t-1}, i_t})$ is an $(s + 1)$-dimensional matrix of one-step transition probabilities. Number of independent parameters for the $\text{MC}(s)$-model increases exponentially w.r.t. the memory length $s$: $D_{\text{MC}(s)} = N^s(N-1)$.

To identify this model we need to have huge data sets and the computation work of size $O(N^s+1)$. To avoid this “curse of dimensionality” we propose to use the parsimonious (“small-parametric”) models of high-order Markov chains that are determined by small number of parameters $d \ll D_{\text{MC}(s)}$.

2 Jacobs-Lewis model

Jacobs-Lewis model is determined by a stochastic difference equation [3] ($t > s$):

$$x_t = \mu_t x_{t-\eta_t} + (1 - \mu_t) \xi_t,$$

(2)

where $\{\xi_t, \eta_t, \mu_t\}$ are independent random variables with probability distributions:

$$\mathbf{P}\{\mu_t = 1\} = 1 - \mathbf{P}\{\mu_t = 0\} = \rho; \quad \mathbf{P}\{\xi_t = k\} = \pi_k, \quad k \in \mathbf{A}, \quad \sum_{k \in \mathbf{A}} \pi_k = 1;$$

$$\mathbf{P}\{\eta_t = i\} = \lambda_i, \quad i \in \{1, 2, \ldots, s\}, \quad \sum_{i=1}^s \lambda_i = 1, \quad \lambda_s \neq 0.$$

(3)

Number of parameters depends linearly on $s$: $D_{\text{JL}} = N + s - 1$.

In [3] only moments and stationary distributions were analyzed. We proved [6, 7] probabilistic and statistical properties of the model (2), (3) by the $\text{MC}(s)$-model.

**Theorem 1.** The discrete time series $x_t$ determined by (2), (3) is a homogeneous Markov chain of the order $s$ with the initial probability distribution $\pi_{i_1, \ldots, i_s} = \pi_{i_1} \cdots \pi_{i_s}$ and the $(s + 1)$-dimensional matrix of transition probabilities $P(\pi, \lambda, \rho) = (p_{i_{t-1}, \ldots, i_{t+1}})$:

$$p_{i_{t-1}, \ldots, i_{s+1}} = (1 - \rho)\pi_{i_{s+1}} + \sum_{j=1}^s \lambda_j \delta_{i_{t-j+1}, i_{s+1}}, \quad i_1, \ldots, i_{s+1} \in \mathbf{A},$$

where $\delta_{j,k}$ is the Kronecker symbol.
Corollary 1. Maximum likelihood estimators (MLE) $(\hat{\pi}, \hat{\lambda}, \hat{\rho})$ by the data $X^n_i = (x_1, \ldots, x_n)'$ are determined by the solution of the maximization problem:

$$l(\pi, \lambda, \rho) = \sum_{t=1}^{s} \ln \pi_{xt} + \sum_{t=s+1}^{n} \ln \left( 1 - \rho \right) \pi_{xt} + \rho \sum_{j=1}^{s} \lambda_j \delta_{x_{t-j}, x_t} \rightarrow \max \pi, \lambda, \rho.$$  

Using these MLE we had constructed [6,7] a consistent generalized probability ratio test for hypotheses on the true values of parameters $(\rho, \lambda, \pi)$ in (3).

3 Raftery model

Mixture Transition Distribution -model (MTD-model) was proposed in 1985 by A. Raftery [11] as a special parsimonious representation of the matrix $P$:

$$p_{i_1, \ldots, i_s, i_{s+1}} = \sum_{j=1}^{s} \lambda_j q_{j, i_{s+1}}, \quad i_1, \ldots, i_{s+1} \in A,$$

where $Q = (q_{i,k})$ is a stochastic $(N \times N)$-matrix, $0 \leq q_{i,k} \leq 1$, $\sum_{k \in A} q_{i,k} = 1$, $i, k \in A$, $\lambda = (\lambda_1, \ldots, \lambda_s)'$ is a discrete probability distribution, $\lambda_1 > 0$.

The MTDg (generalized MTD)-model:

$$p_{i_1, \ldots, i_s, i_{s+1}} = \sum_{j=1}^{s} \lambda_j q^{(j)}_{i_{s-j+1}, i_{s+1}}, \quad i_1, \ldots, i_{s+1} \in A,$$

where $Q^{(j)} = \left(q^{(j)}_{i,k}\right)$ is a stochastic matrix for the $j$-th lag.

Number of parameters for the MTDg: $D_{MTDg} = s(N(N - 1)/2 + 1) - 1$.

We have constructed a simple criterion for the ergodicity of the MTD-model and found a useful property of the stationary probability distribution [6].

Theorem 2. For the MTDg-model (5), if $\exists k \in N : \left( (Q^{(1)})^K \right)_{ij} > 0$, $\forall i, j \in A$, then the $s$-dimensional stationary probability distribution satisfies the equation $(i_1, \ldots, i_s) \in A$:

$$\pi^*_{i_1, \ldots, i_s} = \prod_{l=0}^{s-1} \left( \pi^*_{x_{i_{l+1}} - \lambda_j } q^{(j)}_{i_{j-1}, i_{s+1}} - \sum_{r=0}^{N-1} q_{r, i_{s+1}} \pi^*_{i_{s+1}} \right).$$

Corollary 2. For the ergodic MTD-model (4) the 2-dimensional stationary probability distribution of the random vector $(x_{t-m}, x_t)'$ is $\pi^*_{ki}(m) = \pi^*_{ki} + \pi^*_{k} \lambda_{s-m+1} (q_{ki} - \pi^*_{i})$, $i, k \in A$, $1 \leq m \leq s$.

Based on Corollary 2 we have constructed statistical estimators for $\lambda$, $Q$ by an observed time series $X^n_i = (x_1, \ldots, x_n)'$ of the length $n$:

$$\hat{\pi}_i = \frac{1}{n - 2s + 1} \sum_{t=s+1}^{n} \delta_{x_{t-i}}; \quad \hat{\pi}_{ki} = \frac{1}{n - 2s + 1} \sum_{t=s+1}^{n} \delta_{x_{t-j}} \delta_{x_{t-i}};$$
Theorem 3. For the ergodic MTD-model (4) the estimators $\hat{Q}$, $\hat{\lambda}$ determined by (6) at $n \to \infty$ are consistent and asymptotically unbiased.

MLE $\hat{Q}$, $\hat{\lambda}$ are solutions of the nonlinear maximization problem:

$$l(Q, \lambda) = \sum_{t=s+1}^{n} \ln \sum_{j=1}^{s} \lambda_j q_{x_{t-s+j-1}, x_t} \to \max_{Q, \lambda}.$$  (7)

The estimators $\hat{Q}$, $\hat{\lambda}$ are used as initial values in the iterative computation of the MLEs $\hat{Q}$, $\hat{\lambda}$ in (7). Generalized probability ratio test of the asymptotic size $\epsilon \in (0, 1)$ for $H_0 = \{Q = Q^0, \lambda = \lambda^0\}$, $H_1 = H_0$ is constructed as in the previous section.

4 Markov chain MC$(s, r)$ of the order $s$ with $r$ partial connections

The MC$(s, r)$ proposed by Yu. Kharin in 2004 [5] is determined by the following parsimonious reparametrization of the $(s + 1)$-dimensional transition probability matrix:

$$p_{j_1^{s+1}} = p_{j_1, \ldots, j_s, j_{s+1}} = q_{j_1^{s+1}},$$  (8)

where $J_1^{s+1} = (j_1, \ldots, j_{s+1}) \in A^s$ is the $(s + 1)$-dimensional index vector; $r$ is the number of connections ($1 \leq r \leq s$); $M^0_1 = (m^0_1, \ldots, m^0_1) \in M$ is some integer-valued vector with $r$ ordered components, $1 = m^0_1 < m^0_2 < \ldots < m^0_{r} \leq s$, called the template of connections; $Q = (q_{j_1^{s+1}})_{J_1^{s+1} \in A^{r+1}}$ is an $(r + 1)$-dimensional stochastic matrix. If $r = s$, we have the general MC$(s)$-model.

In [5,8] the probabilistic properties of the MC$(s, r)$-model are found.

Theorem 4. The MC$(s, r)$ defined by (8) is an ergodic Markov chain iff there exists $i \in N$ such that

$$\min_{J_1^{s+1} \in A^{s+1}} \sum_{J_1^{s+1} \in A^{s+1}} \prod_{k=1}^{s+i} q_{j_{k+1}^{s+1}, j_{k+1}^{s+1}} > 0.$$  

Stationary probability distribution $\pi^*_J$ satisfies the equations: $\pi^*_{j_1^{s+1}} = \sum_{j_1 \in A} \pi^*_J q_{j_1^{s+1}, j_1^{s+1}}, J_1^{s+1} \in A^s$.

Corollary 3. For a stationary Markov chain the stationary probability distribution has the multiplicative form: $\pi^*_J = \prod_{i=1}^{s} \pi^*_{j_i}, J_1^{s} \in A^s$, iff $\pi^*_{j_1^{s+1}} = \sum_{j_1 \in A} \pi^*_J q_{j_1^{s+1}}, J_1^{s+1} \in A^r$.

Corollary 4. If $Q$ is doubly stochastic: $\sum_{j_1 \in A} q_{j_1^{s+1}} \equiv 1$, $\sum_{j_1 \in A} q_{j_1^{s+1}} \equiv 1$, then the stationary probability distribution is uniform: $\pi^*_J \equiv N^{-s}$.
Introduce the notation: \( F(J_t^{i+s-1}; M_r) = (j_{i+m_1-1}, \ldots, j_{i+m_r-1}) \) is the selector-function;

\[
\nu_{J_r}^t(X^n_1; M_r) = \sum_{t=1}^{n-s} \delta_F(X_t^{i+s-1}; M_r), J_r^t \delta_{x_t}, J_r^{t+1} \tag{9}
\]

is the frequency statistic for the template \( M_r \in M; \mu_{J_r}^t(M_r) = P \{ F(X_t^{i+s-1}; M_r) = J_r^t, x_t = j_{r+1} \} \) is the probability distribution of the \((r+1)\)-tuple; the dot used instead of any index means summation on all its values: \( \mu_{J_r}^t(M_r) = \sum_{J_r^{t+1} \in A} \mu_{J_r}^t(M_r); \)

\( \mu_{J_r}^t(M_r) = \nu_{J_r}^t(X^n_1; M_r) / (n-s) \) is the frequency estimator for the probability \( \mu_{J_r}^t(M_r), J_r^{t+1} \in A^{r+1}, M_r \in M. \)

**Theorem 5.** If the template of connections \( M_r^0 \) is known, then the MLE for the matrix \( Q \) is \( \hat{Q} = \left( \hat{q}_{J_r}^{t+1} \right)_{J_r^{t+1} \in A^{r+1}} : \hat{q}_{J_r}^{t+1} = \left\{ \hat{\mu}_{J_r}^t(M_r^0)/\hat{\mu}_{J_r}^t(M_r^0), \text{ if } \hat{\mu}_{J_r}^t(M_r^0) > 0; 1/N \text{ else} \right\}. \)

Under the stationarity condition, \( \left\{ \hat{q}_{J_r}^{t+1} : J_r^{t+1} \in A^{r+1} \right\} \) are asymptotically \((n \to \infty)\) unbiased and consistent with covariances \( \text{Cov} \left( \hat{q}_{J_r}^{t+1}, \hat{q}_{K_t}^{t+1} \right) = \hat{\sigma}_{J_r}^{t+1, K_t^{t+1}}/(n-s) + O(1/n^2), \)

\[
\hat{\sigma}_{J_r}^{t+1, K_t^{t+1}} = \hat{\delta}_{J_r^{t+1}, K_t^{t+1}} q_{J_r}^{0} \left( \hat{\delta}_{J_r^{t+1}, K_t^{t+1}} - q_{K_t^{t+1}}^{0} \right) / \mu_{J_r}^{t} \left( M_r^0 \right), \quad J_r^{t+1}, K_t^{t+1} \in A^{r+1}. \]

Moreover, the probability distribution of the \( N \)-dimensional random vector \( \left( \sqrt{n-s} \left( \hat{q}_{J_r}^{t+1} - q_{J_r}^{0} \right) \right)_{J_r^{t+1} \in A^{r+1}} \) at \( n \to \infty \) converges to the normal probability distribution with zero mean and the covariance matrix \( \Sigma = \left( \hat{\sigma}_{J_r}^{t+1, K_t^{t+1}} \right) \).

The consistent statistical test for the hypotheses \( H_0 = \{ Q = Q^0 \} \), where \( Q^0 = \left( q_{J_r}^{0} \right)_{J_r^{t+1} \in A^{r+1}} \) is some given matrix; \( H_1 = \overline{T}_0 \), consists of the following steps.

1. Computation of the statistics \( \nu_{J_r}^{t+1}(X^n_1; M_r^0), J_r^{t+1} \in A^{r+1}, \) by (9).

2. Computation of the statistic \( D_{J_r} = \left\{ j_{r+1} \in A : q_{J_r}^{0}(j_{r+1}) > 0 \right\} \)

\[
\rho = \sum_{J_r^{t+1} \in A^{r}, J_r^{t+1} \in D_{J_r}^{t}} \nu_{J_r}^{t}(X^n_1; M_r^0) \left( \hat{q}_{J_r}^{t+1} - q_{J_r}^{0} \right)^2 / q_{J_r}^{0}. \tag{10}
\]

3. Computation of the P-value: \( P = 1 - G_U(\rho) \), where \( G_U(\cdot) \) is the standard \( \chi^2 \)-distribution function with \( U = \sum_{J_r^{t+1} \in A^{r}} \left( \left| D_{J_r}^{t} \right|-1 \right) \) degrees of freedom.

4. The decision rule with an asymptotic significance level \( \varepsilon \): if \( P \geq \varepsilon \), then to conclude that the hypothesis \( H_0 \) is true; otherwise, the alternative \( H_1 \) is true.

**Corollary 5.** Under stationary Markov family \( MC(s, r) \) and contiguous family of alternatives \( H_{1n} = \{ Q = Q_{1n} \} \), where \( Q_{1n} = \left( q_{J_r}^{1n} \right)_{J_r^{t+1} \in A^{r+1}}, q_{J_r}^{1n} = q_{J_r}^{0} \left( 1 + d_{J_r}^{t+1} / \sqrt{n-s} \right), \quad \sum_{J_r^{t+1} \in A} d_{J_r^{t+1}} q_{J_r}^{0} = 0, \)

68


\[ \sum_{j_{t+1}^r \in \mathcal{A}^{r+1}} |d_{j_{t+1}}| > 0, \]  

if \( H_{1n} \) is true, then at \( n \to \infty \) the power of the developed test \( w \to 1 - G_{U,a}(G_U^{-1}(1-\varepsilon)) \), where \( G_{U,a}(\cdot) \) is the probability distribution function of \( \chi^2 \)-distribution with \( U \) degrees of freedom and the noncentrality parameter \( \alpha = \sum_{j_{t+1}^r \in \mathcal{A}^{r+1}} \mu_{j_{t+1}^r}(M) d_{j_{t+1}}^2 \).

Introduce the notation: \( M \) is the set of all admissible templates \( M_r \);

\[ H(M_r) = - \sum_{j_{t+1}^r \in \mathcal{A}^{r+1}} \mu_{j_{t+1}^r}(M_r) \ln \left( \frac{\mu_{j_{t+1}^r}(M_r)}{\mu_{j_{t+1}^r}(M_r)} \right) \geq 0 \quad (10) \]

is the conditional entropy of the future symbol \( x_{t+s} \in \mathcal{A} \) relative to the past derived by the selector \( F \left( X_{t+s-1}^t; M_r \right) \in \mathcal{A}^r, M_r \in M \); \( \hat{H}(M_r) \) is the “plug-in” estimator of the conditional entropy, which is generated by substitution of the estimators \( \hat{\mu}_{j_{t+1}^r}(M_r) \) instead of true probabilities \( \mu_{j_{t+1}^r}(M_r) \) in (10).

**Theorem 6.** If the order \( s \) and the number of connections \( r \) are known, then the MLE \( \hat{M}_r = \arg \min_{M_r \in M} \hat{H}(M_r) \). Under the stationarity condition of the MC\((s,r)\) the estimator \( \hat{M}_r \) at \( n \to \infty \) is consistent: \( \hat{M}_r \xrightarrow{P} M^0_r \).

Let \( s \in [s_-, s_+] \), \( r \in [r_-, r_+] \), \( 1 \leq s_- < s_+ < \infty \), \( 1 \leq r_- < r_+ < s_+ \). To estimate parameters \( r, s \) we use the Bayesian Information Criterion (BIC) [8]:

\[ BIC(s, r) = 2(n - s)\hat{H}(\hat{M}_r) + U \ln(n - s), \quad (11) \]

where \( U = \sum_{j_{t+1}^r \in \mathcal{A}} \left( |D_{j_{t+1}}| - 1 + \delta_{j_{t+1}^r}(\hat{M}_r, 0) \right) ; D_{j_{t+1}} = \{ j_{r+1} \in \mathcal{A} : \hat{\mu}_{j_{t+1}^r}(\hat{M}_r) > 0 \} \).

Consistent estimators \( \hat{s}, \hat{r} \) are determined by minimization: \( BIC(s, r) \rightarrow \min_{s,r} \).

### 5 Computer experiments on real statistical data

The developed methods were successfully tested on simulated data and also applied to real statistical data [6,7]. We present here only two examples.

**Modeling of wind direction.** The discrete-valued time series of the daily average wind speed at Malin Head (North of Ireland) during the period 1961 – 1978 [11] \( x_t \in \{0, 1, 2\} \) of the length \( n = 6574 \) was fitted by the MC\((s,r)\)-model for \( s = \{1, 2, \ldots, 7\} \), \( r = \{1, 2, \ldots, 7\} \). Table 1 presents the values of the BIC for this real different pairs \((s,r)\).

The best fitted model is the MC\((3,2)\) with \( \hat{M}_r = (1, 3) \) and the matrix

\[
\hat{Q}' = \begin{pmatrix}
0.27 & 0.08 & 0 & 0.22 & 0.04 & 0 & 0.21 & 0.02 & 0 \\
0.73 & 0.86 & 0.63 & 0.78 & 0.82 & 0.52 & 0.79 & 0.72 & 0.43 \\
0 & 0.06 & 0.37 & 0 & 0.14 & 0.48 & 0 & 0.26 & 0.57
\end{pmatrix}.
\]

The fitted model MC\((3,2)\) detects significant dependencies in this data.
Table 1: Modeling of the wind speed data

<table>
<thead>
<tr>
<th>Model</th>
<th>BIC</th>
<th>Model</th>
<th>BIC</th>
<th>Model</th>
<th>BIC</th>
<th>Model</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC(1,1)</td>
<td>8127.52</td>
<td>MC(4,2)</td>
<td>8139.12</td>
<td>MC(5,5)</td>
<td>8621.97</td>
<td>MC(7,1)</td>
<td>9041.43</td>
</tr>
<tr>
<td>MC(2,1)</td>
<td>8777.63</td>
<td>MC(4,3)</td>
<td>8164.79</td>
<td>MC(6,1)</td>
<td>9016.23</td>
<td>MC(7,2)</td>
<td>8163.07</td>
</tr>
<tr>
<td>MC(2,2)</td>
<td>8096.08</td>
<td>MC(4,4)</td>
<td>8332.77</td>
<td>MC(6,2)</td>
<td>8148.48</td>
<td>MC(7,3)</td>
<td>8197.91</td>
</tr>
<tr>
<td>MC(3,1)</td>
<td>8849.90</td>
<td>MC(5,1)</td>
<td>8984.10</td>
<td>MC(6,3)</td>
<td>8190.78</td>
<td>MC(7,4)</td>
<td>8323.19</td>
</tr>
<tr>
<td><strong>MC(3,2)</strong></td>
<td><strong>8079.81</strong></td>
<td><strong>MC(5,2)</strong></td>
<td><strong>8129.83</strong></td>
<td><strong>MC(6,4)</strong></td>
<td><strong>8350.82</strong></td>
<td><strong>MC(7,5)</strong></td>
<td><strong>8599.09</strong></td>
</tr>
<tr>
<td>MC(3,3)</td>
<td>8143.13</td>
<td>MC(5,3)</td>
<td>8177.92</td>
<td>MC(6,5)</td>
<td>8576.92</td>
<td>MC(7,6)</td>
<td>8973.15</td>
</tr>
<tr>
<td>MC(4,1)</td>
<td>8956.11</td>
<td>MC(5,4)</td>
<td>8349.62</td>
<td>MC(6,6)</td>
<td>8969.54</td>
<td>MC(7,7)</td>
<td>9575.64</td>
</tr>
</tbody>
</table>

Genomic sequencing for the drosophila genome sequence (www.fruitfly.org): $N=4$, $n=5 \cdot 10^5$, $s_{-}=1$, $s_{+}=8$, $r_{-}=1$, $r_{+}=8$. The best fitted model is the MC(6, 3) with the template $\hat{M}_r = (1, 5, 6)$ and the matrix $\hat{Q}$ visualized in Figure 1. Here on “x-axes” the values of $\hat{M}_r$-prehistory are indicated, “y-axes” gives the values of one-step transition probabilities to four states indicated by different levels of grey color.

Figure 1: The matrix $\hat{Q}$ for the genomic sequencing
References


ON ACQUISITION
OF NOCICEPTIVE EVOKED POTENTIALS
IN RATS CEREBRAL CORTEX

A.V. Kolchin, E.G. Ionkina
Steklov Institute of Mathematics and
Sechenov First Moscow State Medical University
Moscow, RUSSIA
e-mail: kolchin@mi.ras.ru

Abstract

We succeeded in constructing a portable system for acquisition of electrical activity of a brain, e.g., nociceptive evoked potentials in rats' cerebral cortex. This research was partially supported by the program ‘Modern Problems of Theoretical Mathematics’ of Department of Mathematical Sciences of the Russian Academy of Sciences.

In practical medicine, the pain relief is one of the most important problems. Recently, under active investigation are nociceptive mechanisms related to activation of immune reactions in a body.

In our series of experiments on male Wistar rats, the stimulation of the immune response is carried out by the intraperitoneal injection of a lipopolysaccharide (LPS).

The role of the cerebral cortex in generating the nociceptive reactions is studied with the use of the analysis of evoked potentials registered in certain regions of the cerebral cortex upon electric stimulation of the tail of immobilised rats before the lipopolysaccharide injection and at the first and the seventh day after it.

The nociceptive evoked potentials were analysed with the use of the non-parametric Wilcoxon signed rank test; those tests which use the normal approximation turn out to be hardly applicable to our case.

Due to functional peculiarities of the form of the graphs of the evoked potentials, the question arises how to choose an adequate quantitative characteristic for them. We have chosen, as such a characteristic, the integral under the graph of the potential; we also plan to study in detail their spectral characteristics.

In order to carry out the investigation, we succeeded in constructing a portable system which is able to register the electrical activity of the brain. This system is open-source based and consists of an Intel Pentium IV portable computer (IBM ThinkPad G40) running a real-time Linux version (kernel 2.6.xx), a four-channel amplifier, and an analogue-to-digital converter usbdx-fast (both made according to specifications by courtesy of Incite Technology Ltd., Department of Computing and Maths of University of Stirling, United Kingdom, see [1, 2]; this amplifier was originally developed for teaching ECG at the Medical Faculty of the Ruhr University Bochum). The data acquisition software used is based on the Comedi project (see [3]). Most part of the software was written by us specially for this investigation.
Another problem was to implement a protection against a very intense electromagnetic pollution which exists virtually everywhere in a megalopolis. It was solved by means of many-layer shielding of the analogue part of the system together with the laboratory animal, cables, and the amplifier; in addition, the whole system is powered by a direct, not alternating, current source only.

The system is able to acquire electrical potentials of the brain utilising sixteen analogue input channels, with the rate up to 100 kHz each, thus minimising the analogue brain electrical activity data loss. The results are written to a simple text-type file, which, of course, grows to a very large size.

The experiments were organised under the principles of GLP.

In conclusion, we give an example of dynamics of nociceptive evoked potentials registered in the somatosensory area of the rat’s cerebral cortex before the intraperitoneal injection of the lipopolysaccharide and at the first and the seventh days after it.

Before the LPS injection  Day 1 after the LPS injection  Day 7 after the LPS injection

References

POWERC-LAW GRAPHS ROBUSTNESS AND FOREST FIRES

M.M. LERI$^{1,2}$, YU.L. PAVLOV$^{1,2}$

$^1$Institute of Applied Mathematical Research,
Karelian Research Centre, Russian Academy of Sciences
$^2$Petrozavodsk State University
Petrozavodsk, RUSSIA

e-mail: leri@krc.karelia.ru, pavlov@krc.karelia.ru

Abstract

We study the robustness of power-law random graphs to “random breakdowns” and to “targeted attacks” by computer simulation. We also consider one of the aspects connected with forest fire models.

1 Introduction

We consider power-law random graphs with vertex degrees $\xi_1, \xi_2, \ldots, \xi_N$ drawn independently from the following distribution:

$$P\{\xi \geq k\} = k^{-\tau}, \quad k = 1, 2, \ldots, \quad i = 1, 2, \ldots, N,$$  \hspace{1cm} (1)

where $N$ is the number of vertices numbered from 1 to $N$ and $\tau > 1$ is the distribution parameter.

The graphs are constructed as follows. Each vertex is given a certain degree in accordance with the degree distribution (1). The vertex degree is the number of stubs (or semiedges) that are numbered in an arbitrary order. The graph is constructed by joining all the stubs equiprobably to form edges. If the sum of vertex degrees is odd one stub is added to a random vertex.

Theoretical research in the field of power-law random graphs includes the study of the limit behaviour of the structure characteristics of such graphs [3, 7] as well as analysis of graph robustness to different types of breakdowns [2, 3, 6]. Along with the theoretical approach simulation modeling is also one of the tools for studying these objects.

2 Robustness or vulnerability?

To study the robustness of power-law graphs we considered the so called random graphs of the Internet type [7] with the vertex degree distribution parameter $\tau \in (1, 2)$. These graphs are deemed to be a good implementation of the Internet topology [4]. For these values of parameter $\tau$ the distribution (1) has finite expectation and infinite variance. Internet graphs are known to have one giant component of connected vertices [3, 6, 7]. We showed [5] that the structure of these graphs change dramatically with the change of parameter $\tau$, being less dependent on the graph size $N$. In particular, the structure...
of Internet graphs is one of the reasons why it is interesting to see how this structure changes when some graph vertices are removed.

We considered two types of breakdowns: “random breakdown” when graph vertices are removed equiprobably and “target attack” with a removal of vertices with the highest degrees. We considered graphs of several sizes $N$ from 500 to 5000 with 9 values of parameter $\tau$ from the interval $(1, 2)$ with a step of 0.1 (for each pair $(N, \tau)$ 100 graphs were generated to form statistical data). A graph is deemed destroyed when the size of the second biggest component becomes greater or equal to half the size of the giant component.

Below are the results of the estimation of the regression dependence of the graph destruction probability $p$ on the percentage of removed vertices $r$ and parameter $\tau$ (see Fig. 1,2) with $R^2 = 0.84$ and $R^2 = 0.76$ respectively.

$$p = \begin{cases} 
0, & r < 37/\sqrt{\tau}, \\
-0.2 + 1.5 \cdot 10^{-4} \tau r^2, & 37/\sqrt{\tau} \leq r < 89/\sqrt{\tau}, \\
1, & r \geq 89/\sqrt{\tau},
\end{cases}$$

$$p = \begin{cases} 
0, & \ln r < 1, 85 - \tau, \\
-0.38 + 0.06 \tau r, & 1, 85 - \tau \leq \ln r \leq 3, 13 - \tau, \\
1, & \ln r > 3, 13 - \tau,
\end{cases}$$

The results show that Internet graphs are much more vulnerable to “target attacks” than to “random breakdowns”. In order to “destroy” such a graph by deleting high-degree vertices it is enough to remove $3 - 7\%$ of them. If however graph nodes are broken randomly, it may be not ruined even if more than $60\%$ of its vertices had been removed. Furthermore, Internet graphs’ robustness also depends on the value of parameter $\tau$. In both breakdown cases the graph proved to be more resistant if the value of $\tau$ was closer to 1 and more vulnerable as it moved closer to 2.

## 3 Forest fires

Here we used power-law random graphs to study one of the aspects of forest fires [1]. Let’s consider graph vertices as trees in a certain area of a real forest. Two vertices are connected if on related trees a fire can move from one tree to another. We pose
the question of finding out how many trees should initially be set on a certain area to provide their maximum survival in case of a fire.

The preliminary conditions are as follows. We consider power-law random graphs with vertex degree distribution (1) where parameter \( \tau > 1 \). Let’s view a square area sized 100 × 100. And let’s place on this lattice 10000 trees in an orderly manner. This means that under a full packing every tree has 8 adjacent neighbours. If an average vertex degree \( i \) is less than 8, some graph edges will be missing. With these assumptions and knowing that \( i = \zeta(\tau) \) (where \( \zeta \) is Riemann zeta function), we found that on the interval \( i \in (1, 8] \) graph size \( N \leq 10000 \) is related to the parameter \( \tau \) by the following regression function \( N = 9256 \tau^{-1.05} \) with \( R^2 = 0.97 \) (see the table and Fig. 3):

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.75</td>
<td>3350</td>
</tr>
<tr>
<td>2.96</td>
<td>3000</td>
</tr>
<tr>
<td>2.53</td>
<td>3750</td>
</tr>
<tr>
<td>2.32</td>
<td>3900</td>
</tr>
<tr>
<td>2.19</td>
<td>4000</td>
</tr>
<tr>
<td>2.05</td>
<td>4200</td>
</tr>
<tr>
<td>1.73</td>
<td>5000</td>
</tr>
<tr>
<td>1.49</td>
<td>4780</td>
</tr>
<tr>
<td>1.42</td>
<td>4489</td>
</tr>
<tr>
<td>1.29</td>
<td>5578</td>
</tr>
<tr>
<td>1.23</td>
<td>6700</td>
</tr>
<tr>
<td>1.21</td>
<td>7500</td>
</tr>
<tr>
<td>1.18</td>
<td>8350</td>
</tr>
<tr>
<td>1.16</td>
<td>8911</td>
</tr>
<tr>
<td>1.15</td>
<td>10000</td>
</tr>
</tbody>
</table>

Fig.3 Regression relationship between \( N \) and \( \tau \).

The graph is constructed as it is described in the Introduction. We consider two cases of fire startup: “random fire start” when the first removed vertex is chosen equiprobably and “targeted fire start” with fire starting from a vertex with the biggest degree. As “fire” starts it spreads along the edges to the connected vertices with a given probability \( 0 < p \leq 1 \). “Burned” vertices are removed from the graph. The aim was to find the value of parameter \( \tau = \tau(p) \) for which the number of remaining vertices \( g \) reaches its maximum.

\[
g = 1476.37^2 + 7247.8r - 3313.4 \\
R^2 = 0.86
\]

Fig.4 Relation between the number of remaining vertices \( g \) and parameter \( \tau \) (“random fire start”, \( p = 1 \)).

\[
g = 1296\tau^2 + 7025r - 6431 \\
R^2 = 0.99
\]

Fig.5 Relation between the number of remaining vertices \( g \) and parameter \( \tau \) (“targeted fire start”, \( p = 1 \)).
In Fig. 4 and Fig. 5 are shown the results obtained for $p = 1$ in the cases of “random fire start” and “targeted fire start”, correspondingly. For example, in the “random” case the number of remaining vertices is the maximal in the power-law graphs with parameter $\tau = 2.45$. The initial graph size $N$ then equals 3605 and an average number of vertices remaining after the fire is $g \approx 3580$.

In both cases of “breakdown start” were obtained regression dependencies of the number of vertices remained in a graph $g$ on $\tau$ and the probability of fire spread $p$. Below are given these models in the cases of “random start” (2) and “target start” (3):

$$g = 6008.8 - 1915.3 \, p - 217.4 \, \tau^2, \quad (R^2 = 0.91); \quad (2)$$

$$g = 2938 - 894.2 \, \ln p - 74.5 \, \tau^2, \quad (R^2 = 0.95). \quad (3)$$

Obviously the number of remained vertices is decreasing with the increase of $p$. There were found relations that describe the dependencies of $g$ on $\tau$ for different values of $p$ and dependencies of $g$ on $p$ for different $\tau$. This allowed to find the relation between $\tau_{\max} = \tau_{\max}(p)$ of parameter $\tau$ for with $g$ reaches its maximum $g_{\max}$ on $p$ and, respectively, find the values of $g_{\max} = g_{\max}(p)$. For example, for $p = 1$ and $p = 0.6$ under “random start” $\tau_{\max}(1) = 2.46$, $\tau_{\max}(0.6) = 1.1$, $g_{\max}(1) = 3585$, $g_{\max}(0.6) = 4468$, and under “target start” $\tau_{\max}(1) = 2.74$, $\tau_{\max}(0.6) = 2.23$, $g_{\max}(1) = 3216$, $g_{\max}(0.6) = 3995$.

The study was supported by the Russian Foundation for Basic Research, grant 13-01-00009 and by the Strategic Development Programme of the Petrozavodsk State University for years 2012–2016.

References


TESTING FOR STRUCTURAL HETEROGENEITY IN VECTOR AUTOREGRESSIVE MODEL BY MEANS OF STATISTICAL CLASSIFICATION METHODS

V. Malugin  
Belarusian State University  
Minsk, BELARUS  
e-mail: Malugin@bsu.by

Abstract

The problems of analysis of complex systems, described by VARX model with structural heterogeneity, are studied in this article. It is supposed that structural heterogeneity may be conditioned as a result of two reasons: 1) presence of different system state classes; 2) presence of structural changes. It is supposed that every state of the system is characterized by the stationary model, i.e. the general model is a segmented stationary VARX. The methods of discriminant analysis for VARX models are suggested for the classification of the complex system states. The results of the classification are afterwards used for the estimation of the moments of single or multiple structural changes by means of statistical testing. The simulation results show the acceptable accuracy of the proposed methods.

1 Segmented stationary VARX and problems of their analysis

The problem of structural heterogeneity in multivariate econometric models has attracted considerable attention of researchers during the last decade. A number of well-known results concerning testing and estimating the moments of the structural change have been obtained for the so called segmented stationary models [5]. The theoretical research mainly concern studying of limiting distributions for the test statistics. The proposed algorithms are quite complex and their investigation on the small samples has only been approached with the Monte Carlo experiments.

In this paper, we propose quite simple algorithms for estimation of switching regime (structural changes) moments. These algorithms are based on statistical methods for classification of the multivariate observations [2]. The results of the asymptotic risk analysis for the considered type of decision rules with respect to the multivariate regression observations classification are presented in [4].

Suppose that for a complex system with \( L (L \geq 2) \) classes of states (regime of functioning) the class of state at time \( t \) is described by a random variable \( d_t \in S(L) \). The relationship between endogenous \( x = (x_1, ..., x_N)' \in X \subseteq \mathbb{R}^N \) and exogenous \( z = (z_1, ..., z_M)' \in Z \subseteq \mathbb{R}^M \) variables is described by a vector autoregressive model with a \( p \)-th order \( VARX(p) \) [3] that has the following heterogeneous structure:
\[ x_t = A'_{d(t)} X_{t-1} + B_{d(t)} z_t + \eta_{d(t),t}, \quad t = 1, \ldots, T, \quad \eta_{\alpha,t} \sim \text{NI}(0, \Sigma_{\alpha}), \]  
where \( X_{t-1}' = (x'_{t-1}, \ldots, x'_{t-p}) \in \mathbb{R}^{Np} \) - stacked vector of lagged variables, \( X_0' = (x_0', \ldots, x_1', \ldots, x_{t-p}') \in \mathbb{R}^{Np} \) - given initial value, \( z_t \in \mathcal{Z} \subseteq \mathbb{R}^M \) - stationary vector time series; 
\( A'_{d(t)} = \left( A'_{d(t),1}, \ldots, A'_{d(t),p} \right) \) is a block matrix with dimensions \( N \times pN \), where matrices \( \{ A_{\alpha,l} \} (l = 1, \ldots, p) \) satisfy the stationarity conditions for \( \text{VAR} \forall \alpha \in S(L) \) [3].

It is assumed that the model (1) is valid for the structural heterogeneity assumption:

\[ P\{B_\alpha z = B_\gamma z\} = 0, \quad z \in \mathcal{Z} \subseteq \mathbb{R}^M, \quad B_\alpha \neq B_\gamma, \quad \alpha \neq \gamma, \quad \alpha, \gamma \in S(L). \]

The model (1) is a segmented stationary model within the range \( \mathcal{Z} = [1, \ldots, T] \). The range \( \mathcal{Z} \) includes \( s \geq 1 \) switching of regimes (structural changes) at unknown moments \( \{ \tau_l \}(l = 1, \ldots, s) \) \( 1 < \tau_1 < \ldots < \tau_s < T \).

Both state vector \( d_t \in S(L) \) and parameters \( \{ A_{d(t)}, B_{d(t)}, \Sigma_{d(t)} \} \) are unknown. A learning realization of time series \( (X^L_\alpha, Z^L_\alpha) \) is given, provided that \( X^L_\alpha = (x_{1\tau_1}, \ldots, x_{T\tau_s}) \in X_{T\tau_s}^L, \quad Z^L_\alpha = (z_{1\tau_1}, \ldots, z_{T\tau_s}) \in Z_{T\tau_s}^L, \quad T^L \to \infty (T^L = T^L_\alpha + \ldots + T^L_s). \)

We consider the following problems:

Problem 1. Estimation of state vector \( d = (d_t) \in S^T(L); \)

Problem 2. Estimation of the switching regime (structural changes) moments \( \{ \tau_l \} (l = 1, \ldots, s); \)

Problem 3. Estimation of the parameters \( \{ A_{d(t)}, B_{d(t)}, \Sigma_{d(t)} \} \) for the given moments \( \{ \tau_l \} \) and testing of the statistical significance of the structural changes.

The estimator of state vector \( \hat{d} = (\hat{d}_t) \in S^T(L) \) is obtained by means of statistical classification rules applied to the tested time series \( \{ x_t, z_t \}\) \( (t = 1, \ldots, T) \) (Problem 1). The solution of Problem 2 is based on the elimination of false signals for the structural changes in the sequence \( \hat{d}_1, \ldots, \hat{d}_T \in S(L) \). The proportion of false signals for different classes of states is determined by conditional probabilities of errors \( \{ r^{(l)} \}(l \in S(L)) \) for the used classification decision rules. These probabilities are estimated on the learning sample. To recognize a real signal of structural changes we use a statistical test. Problem 3 is solved by means of traditional estimation and hypothesis testing methods with the use of dummy variables for the given values \( \hat{\tau}_1, \ldots, \hat{\tau}_s \).

Peculiarities of application for the described approach depend on the version of the model (1) and the conditions for solving of problems 1, 2 and 3, which may be as follows:

C.1. Equation (1) describes a model with \( L \) different classes of states, provided that \( L \) is less than the total number of regimes, i.e. \( 2 \leq L < s + 1 \) \( (s \geq 2) \). The learning realizations of time series \( (X^L_\alpha, Z^L_\alpha) \forall \alpha \in S(L) \) are given.

C.2. Equation (1) describes a model with multiple structural changes where \( L = s + 1 > 2 \). In this case \( T^L_l = |J_{l-1,l}| = \tau_l^- - \tau_{l-1}^+ + 1 \) \( (l = 1, \ldots, s + 1) \) where \( J_{l-1,l} = (\tau_{l-1}^+, \ldots, \tau_l^-) \) and \( J_l = [\tau_l^-, \ldots, \tau_l^+] \) is an interval including \( l \)-th structural change.
2 Statistical Classification Rules

For solving of Problems 1 and 2 under conditions C.1, C.2 we propose a "plug-in-rule" Bayesian classifier for the “point-wise classification” of observations described by the model (1) provided that \( P \{d_t = \alpha \} = \pi_\alpha > 0 (\alpha \in S(L)) \). The problem of the “point-wise classification” is represented in the estimation of vector \( d_t \in S(L) \) for the given values \( (x_t, z_t), \ x_{t-p}, ..., x_{t-1}, t = 1, ..., T \).

**Theorem 1.** Let model (1) satisfy the mentioned above assumptions and \( \{\hat{A}_{l,t}\}, \ \{\hat{B}_{l}\} \ (l = 1, ..., p) \) be LS (MLE)-estimates of its parameters on the learning sample \( (X^L_\alpha, Z^L_\alpha) \), then the consistent “point-wise classification” plug-in-rule for the given values \( (x_t, z_t), \ x_{t-p}, ..., x_{t-1} \) is described by the following relations:

\[
\hat{d}_t \equiv \hat{d}(x_t; z_t) = \arg\min_{\alpha\in S(L)} \left\{ \text{tr}(\hat{\Sigma}_\alpha^{-1} S_\alpha) + \ln |\hat{\Sigma}_\alpha| - \frac{2\ln \hat{\pi}_\alpha}{T} \right\}, \\
S_\alpha = \tilde{\eta}_{l,\tau} \tilde{q}_{l,\tau}, \ \tilde{\eta}_{l,\tau} = x_t - \hat{A}_{l,t} X_{t-1} - \hat{B}_{l} z_t, \ t = 1, ..., T. \tag{2}
\]

\[
\text{Step 1. Estimation of the state vector:} \ \text{Plug-in-rule (2)–(3) is used to classify} \ \{x_t, z_t\}(t = 1, ..., T). \ \text{The results of this step is the estimate} \ \hat{d} = (\hat{d}_1, ..., \hat{d}_T) \in S^T(L).
\]

**Step 2. Estimation of the switching regime (structural changes) moments.** The following steps are carried out at this step:

1. Random sequence \( \hat{d}_1, ..., \hat{d}_T \in S(L) \) is divided into \( \gamma = T/m \) fragments with a fixed length \( m \) \((m \ll T)\). For each fragment \( \hat{d}_j = (\hat{d}_{(j-1)m+1}, ..., \hat{d}_{jm}) \in \mathbb{R}^m \ (j = 1, ..., \gamma) \) the following two statistics are calculated:

\[
k_j^{(l)} = \sum_{i=(j-1)m+1}^{jm} \delta_{d_i, 1}, \ k_j^{(l)} = \frac{k_j}{m} \ (j = 1, ..., \gamma), \ \{\eta_l\} \ (l = 1, ..., s); \tag{4}
\]

where \( k_j^{(l)} \in \{0, 1, ..., m\} \) – frequency errors, provided that the complex system is in a state \( \Omega_l \), while the alternative state is \( \Omega_{l+1} \), and the conditional probabilities of errors are equal to \( \hat{r}_0^{(l)} \ (l = 0, 1, ..., L) \). It is known that statistics \( k_j^{(l)} \) has a binomial distribution with parameters \( m, \hat{r}_0^{(l)} \) [1].

2. For each fragment \( j \ (j = 1, ..., \gamma) \) the following hypotheses are tested with respect to the given values \( 0 < \alpha_0, \ \beta_0 \ll 1 \) for the probabilities of the first and the second error type respectively:
where $r_0^{(l)}$ and $r_1^{(l)} = 1 - r_0^{(l+1)}$ are acceptable and unacceptable levels of errors for class $\Omega_l$ correspondingly, provided that there is no strong evidence of the structural change ($H_0$ is not rejected), $r_0^{(l)}$ and $r_1^{(l)}$ are determined by the formulas:

$$r_0^{(l)} = \begin{cases} \hat{r}^{(l)}, & \text{if } \hat{T}_l \rightarrow \infty, \\ \delta^+_{l}(q, \hat{T}_l), & \text{if } \hat{T}_l < \infty, \end{cases}$$

$$r_1^{(l)} = \begin{cases} 1 - \hat{r}_1^{(l+1)}, & \text{if } \hat{T}_{l+1} \rightarrow \infty, \\ 1 - \delta^+_{l+1}(q, \hat{T}_{l+1}), & \text{if } \hat{T}_{l+1} < \infty, \end{cases}$$

(6)

$$\delta^+(q, T^E) = \frac{1}{TE + \phi_q^2} \times \left( k + \frac{\phi_q^2}{2} \mp \phi_q \sqrt{\frac{k(T^E - k_1)}{TE} + \frac{\phi_q^2}{4}} \right),$$

(7)

where $[1]$: $k$ – the number of errors in the point-wise classification of $T^E$ observations; $\phi_q \equiv \Phi^{-1}(q)$ – quantile of the standard normal distribution $N_1(0,1)$ with the confidence level $q$.

The test for hypotheses (5) is used for the approximation of binomial distribution with normal distribution and has the form

$$\text{hypothesis } H_0 \begin{cases} \text{is not rejected if } \kappa_{j}^{(l)} < \kappa^{(l)}, \\ \text{rejected if } \kappa_{j}^{(l)} \geq \kappa^{(l)}, \end{cases}$$

(8)

where

$$\kappa^{(l)} = r_0^{(l)} + \frac{0.5}{m} + \Phi^{-1}(1 - \alpha_0) \frac{\sigma_0}{\sqrt{m}}, \quad \sigma_0 = \sqrt{r_0^{(l)}(1 - r_0^{(l)})},$$

$$m = \left( \frac{\sigma_0 \Phi^{-1}(1 - \alpha_0) + \sigma_i \Phi^{-1}(1 - \beta_0)}{\omega} \right)^2, \quad \sigma_i = \sqrt{r_i^{(l)}(1 - r_i^{(l)})}(i = 0, 1), \quad \omega = r_1^{(l)} - r_0^{(l)}.$$
Table 1: Estimates of the switching regime moments

<table>
<thead>
<tr>
<th>$\Delta(\hat{z})$</th>
<th>$r_0(\hat{z})$</th>
<th>$\hat{\tau}_1$ ($\tau_1 = 251$)</th>
<th>$\hat{\tau}_2$ ($\tau_2 = 501$)</th>
<th>$\hat{\tau}_3$ ($\tau_3 = 751$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.23</td>
<td>0.2693</td>
<td>286</td>
<td>507</td>
<td>767</td>
</tr>
<tr>
<td>2.46</td>
<td>0.1093</td>
<td>250</td>
<td>505</td>
<td>755</td>
</tr>
<tr>
<td>4.92</td>
<td>0.0069</td>
<td>252</td>
<td>500</td>
<td>752</td>
</tr>
</tbody>
</table>

5 Conclusions

Under certain conditions regarding the possibility of forming the learning sample the proposed algorithm for estimation of the switching regimes (or structural changes) moments in VARX has the following benefits: it is less dependent on both the particular specifications of econometric models and the a priori assumptions about the model of structural changes; it is relatively simple in terms of practical implementation and is better applicable for estimating moments of single structural changes, multiple structural changes and structural changes in real time.

References


PARAMETER ESTIMATION IN THE MODELS WITH LONG-RANGE DEPENDENCE

Y.S. Mishura, K.V. Ralchenko, G.M. Shevchenko
Taras Shevchenko National University of Kyiv
Kyiv, UKRAINE

e-mail: myus@univ.kiev.ua, k.ralchenko@gmail.com, zhora@univ.kiev.ua

Abstract
We consider a problem of statistical estimation of an unknown drift parameter for a stochastic differential equation driven by fractional Brownian motion. Two estimators based on discrete observations of solution to the stochastic differential equations are constructed. It is proved that the estimators converge almost surely to the parameter value, as the observation interval expands and the interval between observations vanishes. A bound for the rate of convergence is given. As an auxiliary result of independent interest we establish global estimates for fractional derivative of fractional Brownian motion.

1 Introduction
Fractional Brownian motion (fBm) with Hurst parameter $H \in (0, 1)$ is a centered Gaussian process $\{B^H_t, t \geq 0\}$ with the covariance $E[B^H_t B^H_s] = \frac{1}{2}(s^{2H} + t^{2H} - |t-s|^{2H})$. Stochastic differential equations driven by fBm has been an active research area for the last two decades. Main reason is that such equations seem to be one of the most suitable tools to model long-range dependence in many applied areas, such as physics, finance, biology, network studies etc.

This paper deals with statistical estimation of drift parameter for a stochastic differential equation with fBm by discrete observation of its solution. We propose two new estimators and prove their strong consistency under the so-called “high-frequency data” assumption that the horizon of observations tends to infinity, while the interval between them goes to zero. Moreover, we obtain almost sure upper bounds for the rate of convergence of the estimators. The estimators proposed go far away from being maximum likelihood estimators, and this is their crucial advantage, because they keep strong consistency but they are not complicated technically and are convenient for the simulations.

2 Preliminaries

2.1 Fractional Brownian motion
Fractional Brownian motion (fBm) with Hurst parameter $H \in (0, 1)$ is a centered Gaussian process $\{B^H_t, t \geq 0\}$ on a complete probability space $(\Omega, \mathcal{F}, P)$ with the covariance $E[B^H_t B^H_s] = \frac{1}{2}(s^{2H} + t^{2H} - |t-s|^{2H})$. It is well known that $B^H$ has a modification
with almost surely continuous paths (even Hölder continuous of any order up to \( H \)), and further we will assume that it is continuous itself.

In what follows we assume that the Hurst parameter \( H \in (1/2, 1) \) is fixed. In this case, the integral with respect to the fBm \( B^H \) will be understood in the generalized Lebesgue–Stieltjes sense (see [4]). Its construction uses the fractional derivatives, defined for \( a < b \) and \( \alpha \in (0, 1) \) as

\[
(D_{a+}^\alpha f)(x) = \frac{1}{\Gamma(1-\alpha)} \left( \frac{f(x)}{(x-a)^\alpha} + \alpha \int_a^x f(x) - f(u) \frac{du}{(x-u)^{1+\alpha}} \right),
\]

\[
(D_{b-}^{1-\alpha} g)(x) = \frac{e^{-i\pi\alpha}}{\Gamma(\alpha)} \left( \frac{g(x)}{(b-x)^{1-\alpha}} + (1-\alpha) \int_x^b g(x) - g(u) \frac{du}{(u-x)^{2-\alpha}} \right).
\]

Provided that \( D_{a+}^\alpha f \in L_1[a, b], \ D_{b-}^{1-\alpha} g_{b-} \in L_\infty[a, b] \), where \( g_{b-}(x) = g(x) - g(b) \), the generalized Lebesgue-Stieltjes integral \( \int_a^b f(x) dg(x) \) is defined as

\[
\int_a^b f(x) dg(x) = e^{i\pi\alpha} \int_a^b \left( D_{a+}^\alpha f \right)(x) \left( D_{b-}^{1-\alpha} g_{b-} \right)(x) dx.
\]

It follows from Hölder continuity of \( B^H \) that for \( \alpha \in (1-H, 1) \) \( D_{b-}^{1-\alpha} B^H_{b-} \in L_\infty[a, b] \) a.s. Then for a function \( f \) with \( D_{a+}^\alpha f \in L_1[a, b] \) we can define integral with respect to \( B^H \) through (1):

\[
\int_a^b f(x) dB^H(x) := e^{i\pi\alpha} \int_a^b \left( D_{a+}^\alpha f \right)(x) \left( D_{b-}^{1-\alpha} B^H_{b-} \right)(x) dx.
\]

### 2.2 Estimate of derivative of fractional Brownian motion

In order to estimate integrals with respect to fractional Brownian motion, we need to estimate the fractional derivative of \( B^H \). Let some \( \alpha \in (1-H, 1/2) \) be fixed until the rest of this paper. Denote for \( t_1 < t_2 \)

\[
Z(t_1, t_2) = (D_{t_2-}^{1-\alpha} B^H_{t_2-})(t_1) = \frac{e^{-i\pi\alpha}}{\Gamma(\alpha)} \left( \frac{B^H_{t_1} - B^H_{t_2}}{(t_2-t_1)^{1-\alpha}} + (1-\alpha) \int_{t_1}^{t_2} \frac{B^H_{t_1} - B^H_{t_2}}{(u-t_1)^{2-\alpha}} du \right).
\]

The following proposition is a generalization of [1, Theorem 3].

**Theorem 1.** For any \( \gamma > 1/2 \),

\[
\xi_{H,\alpha,\gamma} := \sup_{0 \leq t_1 < t_2 \leq t_1+1} \frac{|Z(t_1, t_2)|}{(t_2-t_1)^{H+\alpha-1} \left( \log(t_2-t_1)^{1/2} + 1 \right) \log(t_2+3))}\]

is finite almost surely.

Moreover, there exists \( c_{H,\alpha,\gamma} > 0 \) such that \( E \left[ \exp \left\{ x \xi_{H,\alpha,\gamma}^2 \right\} \right] < \infty \) for \( x < c_{H,\alpha,\gamma} \).
2.3 Estimates for solution of SDE driven by fractional Brownian motion

Consider a stochastic differential equation

\[ X_t = X_0 + \int_0^t a(X_s)ds + \int_0^t b(X_s)dB^{H_s}_s, \]

where \( X_0 \) is a non-random coefficient. In [2], it is shown that this equation has a unique solution under the following assumptions: there exist constants \( \delta \in (1/H - 1, 1], K > 0, L > 0 \) and for every \( N > 0 \) there exists \( R_N > 0 \) such that

(A) \( |a(x)| + |b(x)| \leq K \) for all \( x, y \in \mathbb{R} \),

(B) \( |a(x) - a(y)| + |b(x) - b(y)| \leq L |x - y| \) for all \( x, y \in \mathbb{R} \),

(C) \( |b'(x) - b'(y)| \leq R_N |x - y|^\delta \) for all \( x \in [-N, N], y \in [-N, N] \).

Fix some \( \beta \in (1/2, H) \). Denote for \( t_1 < t_2 \)

\[ \Lambda_\beta(t_1, t_2) = 1 \vee \sup_{t_1 \leq u < v \leq t_2} \frac{|Z(u, v)|}{(v - u)^{\beta + \alpha - 1}}. \]

**Theorem 2.** There exists a constant \( M_{\alpha, \beta} \) depending on \( \alpha, \beta, K, \) and \( L \) such that for any \( t_1 \geq 0, t_2 \in (t_1, t_1 + 1] \)

\[ |X_{t_2} - X_{t_1}| \leq M_{\alpha, \beta} \left( \Lambda_\beta(t_1, t_2)(t_2 - t_1)^\beta + \Lambda_\beta(t_1, t_2)^{1/\beta}(t_2 - t_1) \right). \]

**Corollary 1.** For any \( \gamma > 1/2 \), there exist random variables \( \xi \) and \( \zeta \) such that for all \( t_1 \geq 0, t_2 \in (t_1, t_1 + 1] \)

\[ |X_{t_2} - X_{t_1}| \leq \zeta(t_2 - t_1)^{\beta}(\log(t_2 + 3))^{\alpha}, \quad \Lambda_\beta(t_1, t_2) \leq \xi(\log(t_2 + 3))^{\kappa \beta}, \]

where \( \kappa = \gamma/\beta \). Moreover, there exists some \( c > 0 \) such that \( \mathbb{E} \left[ \exp \{ x\xi^2 \} \right] < \infty \) and \( \mathbb{E} \left[ \exp \{ x\zeta^2 \} \right] < \infty \) for \( x < c \). In particular, all moments of \( \xi \) and \( \zeta \) are finite.

3 Drift parameter estimation

Now we turn to problem of drift parameter estimation in equations of type (4). Let \((\Omega, \mathcal{F})\) be a measurable space and \( X : \Omega \to C[0, \infty) \) be a stochastic process. Consider a family of probability measures \( \{\mathbb{P}^\theta, \theta \in \mathbb{R}\} \) on \((\Omega, \mathcal{F})\) such that for each \( \theta \in \mathbb{R} \), \( \mathcal{F} \) is \( \mathbb{P}^\theta \)-complete, and there is an fBm \( \{B_t^{H, \theta}, t \geq 0\} \) on \((\Omega, \mathcal{F}, \mathbb{P}^\theta)\) such that \( X \) solves a parametrised version of (4)

\[ X_t = X_0 + \theta \int_0^t a(X_s)ds + \int_0^t b(X_s)dB^{H, \theta}_s. \]
Our main problem is the following: to construct an estimator for $\theta$ based on discrete observations of $X$. Specifically, we will assume that for some $n \geq 1$ we observe values $X_{t_n^k}$ at the following uniform partition of $[0, 2^n]$: $t_n^k = k2^{-n}$, $k = 0, 1, \ldots, 2^n$.

Fix the parameters $\alpha \in (1 - H, 1/2)$, $\beta \in (1 - \alpha, H)$, $\gamma > 1/2$ and $\kappa = \gamma/\beta$.

In order to proceed, we need another technical assumption, in addition to conditions (A)–(C):

(D) there exist a constant $M > 0$ such that for all $x \in \mathbb{R}$

\[ |a(x)| \geq M, \quad |b(x)| \geq M. \]

Define an estimator

\[
\hat{\theta}_n^{(1)} = \frac{\sum_{k=1}^{2^{2n}} (t_n^k)^{\lambda - 1} (2^n - t_n^k)^{\lambda - 1} b^{-1} (X_{t_n^{k-1}} - X_{t_n^{k-1}}) (X_{t_n^{k-1}} - X_{t_n^{k-1}})}{\sum_{k=1}^{2^{2n}} (t_n^k)^{\lambda - 1} (2^n - t_n^k)^{\lambda - 1} (X_{t_n^{k-1}} - X_{t_n^{k-1}})} \frac{1}{2^n},
\]

where $\lambda = 1/2 - H$.

**Theorem 3.** With probability one, $\hat{\theta}_n^{(1)} \to \theta$, $n \to \infty$. Moreover, there exists a random variable $\eta$ with all finite moments such that $|\hat{\theta}_n^{(1)} - \theta| \leq \eta n^{\kappa + \gamma 2^{-\rho n}}$, where $\rho = (1 - H) \wedge (2\beta - 1)$.

Consider a simpler estimator:

\[
\hat{\theta}_n^{(2)} = \frac{\sum_{k=1}^{2^{2n}} b^{-1} (X_{t_n^{k-1}} - X_{t_n^{k-1}}) (X_{t_n^{k-1}} - X_{t_n^{k-1}})}{\frac{1}{2^n} \sum_{k=1}^{2^{2n}} b^{-1} (X_{t_n^{k-1}} - X_{t_n^{k-1}})}.
\]

This is a discretized maximum likelihood estimator for $\theta$ in equation (4), where $B^H$ is replaced by Wiener process. Nevertheless, this estimator is consistent as well. Namely, we have the following result.

**Theorem 4.** With probability one, $\hat{\theta}_n^{(2)} \to \theta$, $n \to \infty$. Moreover, there exists a random variable $\eta'$ with all finite moments such that $|\hat{\theta}_n^{(2)} - \theta| \leq \eta' n^{\kappa + \gamma 2^{-\rho n}}$.

**References**


SCALABILITY PROPERTIES OF A NEW FORWARD SEARCH ALGORITHM

D. Perrotta, M. Riani, A. Cerioli, F. Torti

EC Joint Research Centre
Ispra, ITALY

and University of Parma
Parma, ITALY

e-mail: domenico.perrotta@ec.europa.eu

Abstract

We propose some computational improvements for a robust method for the identification of atypical observations known as forward search, which is based on the idea of monitoring quantities of interest, such as parameter estimates and test statistics, as the model is fitted to data subsets of increasing size. We provide a recursive implementation of the procedure which exploits the information of the previous step and we demonstrate the computational advantages of the new approach using both synthetic and real data.

1 Overview

The identification of atypical observations and the immunization of data analysis against both outliers and failures of modelling are important aspects of modern statistics. The forward search is a graphics rich approach that leads to the formal detection of outliers and to the detection of model inadequacy combined with suggestions for model enhancement. The key idea is to monitor quantities of interest, such as parameter estimates and test statistics, as the model is fitted to data subsets of increasing size.

Recent applications of the forward search include systematic outlier detection in official Census data [3], and the analysis of international trade markets [1], where important issues such as incorrect declarations, tax evasion and money laundering are at the forefront. In both these instances the number of datasets to be analyzed is of the order of hundreds of thousands, while the sample size of each dataset ranges from less than 10 observations to more than 100000. It is thus crucial to improve the computational features of the methodology and to dramatically reduce its computation time. Otherwise, on line monitoring and outlier detection, which are essential requirements for the successful implementation of statistical methods in these fields, would be unfeasible.

Therefore, the goal of this work is to provide the computational and algorithmic advances that are necessary to apply the forward search to the massive datasets arising in applications like those sketched above. In particular, we provide a recursive implementation of the forward search procedure which exploits the information of the previous step. The output is a set of efficient routines for fast updating of the model parameter estimates, which do not require any data sorting, and fast computation of
likelihood contributions, which do not require any inverse matrix or qr decomposition. It is shown that the new algorithms enable a reduction of the computation time by more than 80%. Furthermore, the running time now increases almost linearly with the sample size.

All the routines discussed are included in the FSDA toolbox for MATLAB which is freely down-loadable from Internet [2].

The scalability properties of the traditional and new forward search algorithms and of comparable robust estimators will be demonstrated on datasets of different sizes extracted from a repository of trade declarations of the customs services of the EU Member States.

References


GOODNESS OF FIT TESTS BASED ON KERNEL DENSITY ESTIMATORS

R. Rudzkis, A. Bakshaev
Vilnius University, Institute of Mathematics and Informatics
Akademijos 4, LT-08663 Vilnius, Lithuania
e-mail: rimantas.rudzkis@mii.vu.lt
e-mail: aleksej.bakshaev@gmail.com

Abstract

The paper is devoted to goodness of fit tests based on kernel estimators of probability density functions. In particular, univariate case is investigated. The test statistic is considered in the form of maximum of the normalized deviation of the estimate from its expected value. Produced comparative Monte Carlo power study show that the proposed test is a powerful competitor to the existing classical criteria testing goodness of fit against a specific type of alternative hypothesis. An analytical way for establishing the asymptotic distribution of the test statistic is proposed, using the theory of high excursions of Gaussian random processes and fields introduced by Rudzkis [17,18]. The extension of the proposed methods to the multivariate case are discussed.

1 Introduction.

The problem of testing goodness of fit is well known and has generated plenty of attention from researchers both in theoretical and applied statistical literature. In its typical purpose these tests are used to determine whether an underlying probability distribution differs from a hypothesized one. Both for one- and multidimensional cases a wide range of solutions of this problem have been provided. However, the choice of the most efficient test, among the available criteria, is regarded as one of the basic problems of statistics. It is well known, that for a variety of problems arising in statistical theory and practice the uniformly most powerful tests are unknown. Therefore creation of new test procedures, especially in multivariate case, sensitive to a particular type of hypotheses remains actual and in our days.

Classical approaches to solve the goodness of fit problem use the empirical process theory. Most of the popular tests such as the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics are based on the empirical distribution function $F_n(x)$.

In this paper, we consider another type of tests based on the kernel density estimator. The idea of using nonparametric kernel density estimators for goodness of fit tests goes back to Bickel and Rosenblatt [13,14]. Since that time a great number of publications has appeared, devote mostly to the $L_p$, $p = 1, 2$ distance between the density estimate $\hat{f}(x) = \hat{f}(x,X^n)$ of the underlying density $f_0(x)$ and its expected value under the null hypothesis. A review of the methods could be found in [1–7,10] and references therein. Thereby much less attention were devoted to a consideration of the deviations in the uniform metric as the loss function for $\hat{f}(x)$, which is an object of this work.
2 Statement of the problem.

Let $X_1, \ldots, X_n$ be a sample of independent observations of a random variable $X$ with an unknown probability density function $f(x)$, $x \in \mathbb{R}$. Using the given sample, it is required to test a simple hypothesis of goodness of fit

$$H_0 : f(x) = f_0(x)$$

against the complex alternative

$$H_1 : f(x) = (1 - \epsilon)f_0(x) + \epsilon g(x),$$

where $f_0(x)$ is a given probability density function, $\epsilon$ is small enough and $g(x)$ is an arbitrary distribution concentrated on a small interval, e.g. $\sigma^2_g \ll \sigma^2_{f_0}$, where $\sigma^2_f$ is a variance of distribution $f$.

The choice of uniform metric for the loss function for $\hat{f}(x)$ is justified by investigation of specific type of alternative hypothesis (1). Such alternatives are of a particular interest in some social and economic studies, e.g. determination of small high income clusters of people, in population income distribution. Meaningful applications could be also achieved in multivariate case, dealing with multimodal distributions for detection of tight clusters.

We consider a test based on the well-known Parzen-Rosenblatt kernel density estimator of $f$, defined for any $x \in \mathbb{R}$ by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $K(\cdot)$ is a probability kernel and $h = h(n)$ is a bandwidth parameter.

The form of the alternative hypothesis motivates us to consider the test statistic in the form of maximum of the normalized deviation in the uniform metric of the estimate $\hat{f}(x)$ from its expected value $E_0\hat{f}(x)$

$$\zeta_h = \max_{x \in I} |\xi_h(x)|,$$

where $\xi_h(x) = \frac{\hat{f}_h(x) - E_0\hat{f}_h(x)}{\sqrt{D_0\hat{f}_h(x)}}$, $E_0$ and $D_0$ denote a mathematical expectation and variance defined in the case of null hypothesis and $I$ is a fixed interval.

Efficient use of kernel estimators requires the choice of an appropriate kernel and a bandwidth parameter. It is well-known that selection of the smoothing parameter rather than the form of the kernel is critical, as under- or over-smoothing can substantially reduce precision. In this work, a certain method to avoid the problem of selection of a bandwidth parameter is proposed. It is suggested to examine the test statistic $\zeta_h$ with different choices of a smoothing parameter $h$ and thereby make the decision of rejecting the null hypothesis, based on the maximum of $\zeta_h$ values with respect to $h$.

This leads us to the following improved form of the test statistic

$$M = \max_{h \in J} \left[ \max_{x \in I} \left| \xi_h(x) \right| - \mu(h) \right] \gamma(h),$$

where $\gamma(h)$ is a function that depends on the choice of bandwidth parameter $h$.
where
\[ \mu(h) = \mathbb{E}_0 \max_{x \in I} |\xi_h(x)| \quad \gamma^2(h) = \mathbb{D}_0 \max_{x \in I} |\xi_h(x)| \] (5)
and maximum with respect to \( h \) is calculated in a certain interval \( J \) defined by a researcher.

We should reject the null hypothesis in the case of large values of our test statistics, that is if \( M > c_\alpha \), where \( c_\alpha \) can be found from the equation
\[ P_0(M > c_\alpha) = \alpha, \] (6)
where \( P_0 \) is a probability distribution corresponding to the null hypothesis and \( \alpha \) is a pre-specified size of the test.

In practice the critical region of the test could be established by means of Monte Carlo simulations. The problem of analytical approximation of the distribution of the test statistics under the null hypothesis is discussed in section 3, using the theory of high excursions of Gaussian (and, in some sense, close to Gaussian) random processes and fields developed by Rudziks [17,18]. Besides some of the already mentioned references, the asymptotic distributions of deviations of kernel density estimators in uniform metric were also considered in [8,9,11,12].

3 Analytical approximation of the null distribution of the test statistic.

This section is devoted to the analytical approximation of the functions \( \mu(\cdot) \) and \( \gamma(\cdot) \) in (5) and the null distributions of statistics (3) and (4) to determine the critical region of the tests. First we will be concerned with the asymptotics of the probability
\[ P_h(u) = P_0 \left\{ \max_{x \in I} |\xi_h(x)| < u \right\} \] (7)
as \( n \to \infty \). Note that \( \hat{f}_h \) is a consistent estimator, its finite dimensional distributions are asymptotically normal, and
\[ \text{cor}(\hat{f}_h(x_1), \hat{f}_h(x_2)) \to 0, \]
if \( x_1 \neq x_2 \) and \( n \to \infty \). The fact that \( \hat{f}_h \) is close to the Gaussian random process in a certain sense suggests us to apply the results from the theory of high excursions of Gaussian processes introduced in [16] to approximate the probability \( P_h(u) \). Let \( \xi(x) \) be a differentiable in the mean square sense Gaussian random process with zero mean, unit variance and continuous trajectories, and \( \mu_i(x), i = 1, 2 \) are smooth enough functions. Rudziks has shown that, under some smoothness and regularity conditions, the probability \( P\{-\mu_1(x) < \xi(x) < \mu_2(x), x \in [a,b]\} \), could be approximated by
\[ P\{-\mu_1(x) < \xi(x) < \mu_2(x), x \in [a,b]\} \cong G(\mu_1, \mu_2), \] (8)
where
\[ G(\mu_1, \mu_2) = [\Phi(\mu_1(a)) + \Phi(\mu_2(a)) - 1] \exp \left\{ -\sum_{i=1}^2 \int_a^b q(\mu_i(t)) dt \right\}, \] (9)
here
\[ q(\mu_i(t)) = \phi(\mu_i(t)) \left[ \beta(t)\phi \left( \frac{\mu_i'(t)}{\beta(t)} \right) - \mu_i'(t)\Phi \left( -\frac{\mu_i'(t)}{\beta(t)} \right) \right], \tag{10} \]

where \( \Phi(\cdot) \) is a probability distribution function of the standard normal distribution, \( \phi(x) = \Phi'(x) \) and \( \beta^2(x) = D\xi'(x) \).

Consider the empirical random process \( \xi_h(\cdot) \). Using approximation (8) we have
\[ P_h(u) \approx [2\Phi(u) - 1] \exp \left\{ -\exp(-u^2/2)/\pi \int \beta(z)dz \right\} =: \widehat{P}_h(u), \tag{11} \]

where
\[ \beta^2(x) = \frac{D\hat{f}_h(x)}{\sigma^2(x)} - \frac{[(\sigma^2(x))^2]}{4\sigma^4(x)}, \quad \sigma^2(x) = D\hat{f}_h(x). \tag{12} \]

For practical usage the exact expression for \( \beta(x) \) could be found in [19].

From (11) it follows, that the functions \( \mu(h) \) and \( \gamma(h) \) could be approximated using the formulas
\[ \mu(h) = \int u\widehat{P}_h(u) \quad \gamma^2(h) = \int u^2d\widehat{P}_h(u) - \mu^2(h). \tag{13} \]

Produced simulation analysis of the accuracy of proposed approximations show that asymptotic distributions (11) provide a really good approximation to the null distribution of the statistic \( \zeta_h \) even for small and moderate sample sizes. However practical experiments showed that the choice of the smoothing parameter could play a crucial role for the goodness of approximation especially for small sample sizes.

Coming back to the null distribution of statistic (4), for the approximations of the probability
\[ P(u) = \mathbb{P}_0 \left\{ \max_{h \in J} \left[ \frac{\max_{x \in I} |\xi_h(x)| - \mu(h)}{\gamma(h)} \right] < u \right\} \tag{14} \]

and determination of the critical region of test, the following obvious inequality could be used
\[ P(u) \leq \max_{h \in J} P_h(u). \tag{15} \]

Produced Monte Carlo simulations show that suggested estimate provide sufficiently good approximation for the null distribution of \( M \) statistic for small sizes of the test (\( \alpha < 0.05 \)).

\section{Multivariate case.}

In this section a brief discussion about the extension of proposed testing procedures to multivariate case is presented. The precise investigations will be provided in our further research.

Since Pearson criteria, goodness of fit tests have been developed mostly for univariate distributions and, except for the case of multivariate normality, few references can be found in the literature about multivariate tests of fit. The main difficulty here is that many tests statistics based on the empirical distribution function of the sample have the limit distribution dependent on the data’s underlying distribution in a nontrivial way.
Therefore analytical establishment of the asymptotic distribution of test statistic may lead to substantial calculational difficulties. To extend the classical univariate goodness of fit tests, e.g. Kolmogorov-Smirnov, Cramer-von Mises, etc., to multivariate case usually initial sample is first transformed to $p$-dimensional cube. After that multivariate hypothesis of uniformity is verified. One of the most popular transformations is the Rosenblatt transformation introduced in [15]. However, this approach also have some disadvantages. One of them is lack of the uniqueness. The method is not invariant with respect to relabelling of the components of $p$-dimensional vector, which lead to a different Rosenblatt transformation and values of statistics. Another disadvantage is connected with the influence of any transformation to the power of the test, as transformation may significantly change the structure of alternative distribution. A generalization of the proposed approach for testing the null hypothesis against the transformation may significantly change the structure of alternative distribution. A replacement of univariate density estimate (2) with its multivariate analog may help to avoid the mentioned problems.

Let now $x \in \mathbb{R}^p$ and $I$ be a $p$-dimensional interval. It is worth noting, that after the replacement of univariate density estimate (2) with its multivariate analog

$$\hat{f}_H(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K(H^{-1}(x - X_i)), \quad (16)$$

where $K(\cdot)$ is the kernel function, $H$ is a smoothing $p \times p$ symmetric and positive definite matrix and $|H|$ its determinant, all the formulas (3) - (5) after corresponding changes remain valid. After that, approximation of the null distribution of $\max_{x \in I} |\xi_H(x)|$ and functions $\mu(H)$, $\gamma(H)$ could be obtained by the direct application of the methods of high excursions of Gaussian fields presented in [18].

It has been shown that if a differentiable (in the mean square sense) Gaussian random field $\{\eta(t), t \in T\}$ with $\mathbb{E}\eta(t) = 0, \mathbb{D}\eta(t) = 1$ and continuous trajectories defined on the $p$-dimensional interval $T \subset \mathbb{R}^p$ satisfies certain smoothness and regularity conditions, then $\mathbb{P}\{-v_1(t) < \eta(t) < v_2(t), t \in T\} \approx e^{-Q}$, as $\forall t \in T v_{1,2}(t) > \chi, \chi \to \infty$, where $v_{1,2}(\cdot)$ are smooth enough functions and $Q$ is a certain constructive functional depending on $v_{1,2}$, $T$ and the matrix function $R(t) = \text{cov}(\eta'(t), \eta'(t))$. Stated result leads to the following approximation of probability $P_H(u)$

$$P_H(u) = \mathbb{P}_0 \left\{ \max_{x \in I} |\xi_H(x)| < u \right\} \approx e^{-2Q(u)} =: \overline{P}_H(u), \quad (17)$$

where $Q$ depends on $u$, $I$ and the matrix function $R(x) = \text{cov}(\xi_H'(x), \xi_H'(x))$. The exact expression of functional $Q$ in the general form could be found in [18]. In bivariate case $Q$ has the form

$$Q(u) = \frac{1}{2\pi} (1 - \Phi(u) + u \phi(u)) \int_I \det(R)^{1/2}dx_1dx_2 +$$

$$+ \frac{\phi(u)}{2\sqrt{2\pi}} \left[ \int_{I_1} (R_{1,1}^{1/2}(y, a_2) + R_{1,1}^{1/2}(x_1, b_2))dx_1 + \int_{I_2} (R_{2,2}^{1/2}(a_1, x_2) + R_{2,2}^{1/2}(b_1, x_2))dx_2 \right],$$

where $R = R(x_1, x_2) = \text{cov}(\xi_H'(x), \xi_H'(x))$ is the covariance matrix of the bivariate random field $\xi_H'(x)$ with elements $R_{i,j} = R_{i,j}(x_1, x_2), i, j = 1, 2$ and $I = I_1 \times I_2 =$
\([a_1, b_1] \times [a_2, b_2]\). Here by \(\phi(\cdot, \cdot|R)\) we denote the probability density functions of the bivariate normal distribution \(N(0, R)\) with covariance function \(R\).

Finally, the approximations for the functions \(\mu(H)\) and \(\gamma(H)\) could be derived from (13) using \(\hat{P}_H(u)\) instead of \(\hat{P}_h(u)\).

Considered method for approximation of the null distribution of test statistic provide a straightforward way for establishment of the critical region of the test without any initial transformations.

5 Simulation study.

Let us switch to a short description of the comparative Monte Carlo power study in detail presented in [19]. The analyzed test \(M\) is compared with the classical criteria: Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov, Shapiro-Wilk, D’Agostino (for the case of normality test) and Bickel-Rosenblatt using the stated type of alternative hypothesis. In the study we restrict to the usage of Epanechnikov kernel \(K\) in (2), as a kernel optimal in the minimum variance sense. The functions \(\mu(\cdot)\) and \(\gamma(\cdot)\) in (4) are calculated using the obtained approximations (13). The critical region of the test is established by means of Monte Carlo simulations. Considered several variants of the tightness of distribution cluster \(g\) and mixing probabilities \(\epsilon\) in (1), give us a wide range of departures from the null hypothesis and allow us to test the sensitivity of criteria to each of them.

The results of simulations show that the proposed test is a powerful competitor to the existing classical ones. For small sample size \((n = 200)\) the proposed test performance is very similar to the Bickel-Rosenblatt criterion being more powerful in comparison with all the other tests. In general, the Bickel-Rosenblatt criterion, as a test also based on the kernel density estimator, is considered to be the main competitor in the study. Detecting a small tight distribution cluster, using kernel estimators in uniform metrics, implies strong and expectable tendencies in increasing of the comparative power of the proposed test, while either sample size is growing and / or mixing distribution \(g\) becomes more concentrated for all mixing probabilities. As a result for large sample size \((n = 1000)\) \(M\) test is the most powerful in our comparative analysis for all considered variants of alternative hypothesis.

References


SOME REMARKS ON ROBUST ESTIMATION OF POWER SPECTRA

G.L. Shevlyakov, N.S. Lyubomishchenko, P.O. Smirnov
St. Petersburg State Polytechnic University
St. Petersburg, RUSSIA
E-mail: Georgy.Shevlyakov@phmf.spbstu.com

Abstract

Various robust modifications of the classical methods of power spectra estimation, both nonparametric and parametric, are considered. Their performance evaluation is studied in autoregressive models with contamination. It is found out that prospective robust estimates of power spectra are based on robust highly efficient estimates of autocovariances and on robust filtering algorithms. Several open problems for future research are formulated.

1 Introduction

Robust methods ensure high stability of statistical inference under uncontrolled deviations from the assumed distribution model. Much less attention is devoted in the literature to robust estimation of data spectra as compared to robust estimation of location, scale, regression and covariance [3, 4, 8]. However, it is necessary to study these problems due to their both theoretical and practical importance (estimation of time series power spectra in various applications, such as communication, geophysics, medicine, etc.), and also because of the instability of classical methods of power spectra estimation in the presence of outliers in the data [5].

There are several classical approaches to estimation of the power spectra of time series, e.g., via the nonparametric periodogram and the Blackman-Tukey formula methods, as well as via the parametric Yule-Walker and filter-based methods [2]. Thereafter, we may consider their various robust versions: to the best of our knowledge, a first systematic study of them is made in the dissertation of Bernhard Spangl [14].

In what follows, we partially use the aforementioned study as a baseline, mostly follow the classification of robust methods of power spectra estimation given in [14], specify them and propose some new approaches with their comparative performance evaluation. Basically, to obtain good robust estimates of power spectra, we use highly efficient robust estimates of scale and correlation.

Our main goals are both to outline the existing approaches to robust estimation of power spectra and to indicate open problems, so our paper is partially a review and partially a program for future research.

The remainder of the paper is as follows. In Section 2, classical methods of power spectra estimation are briefly enlisted. In Section 3, robust modifications of classical approaches are formulated. In Section 4, a few preliminary results on the comparative study of the performance evaluation of various robust methods are represented. In Section 5, some conclusions and open problems for future research are drawn.


2 Classical Estimation of Power Spectra

2.1 Nonparametric Estimation of Power Spectra

The nonparametric approach to estimation of power spectra is based on smoothed periodograms [2].

Let \( x_t, t = 1, \ldots, n \) be a second-order stationary time-series with zero mean. Assume that the time intervals between two consecutive observations are equally spaced with duration \( \Delta t \). Then the periodogram is defined as follows:

\[
\hat{S}_P(f) = \frac{\Delta t}{n} \left| \sum_{t=1}^{n} x_t \exp\{-i 2\pi f t \Delta t\} \right|^2
\]  

over the interval \([-f(n), f(n)]\), where \( f(n) \) is the Nyquist frequency: \( f(n) = 1/(2\Delta t) \).

The Blackman-Tukey formula gives the representation of formula (1) via the sample autocovariances \( \hat{c}_{xx} \) of the time series \( x_t \) [1]:

\[
\hat{S}_P(f) = \hat{S}_{BT}(f) = \frac{\Delta t}{n} \sum_{h=-(n-1)}^{n-1} \hat{c}_{xx}(h) \exp\{-i 2\pi f h \Delta t\}.
\]  

It can be seen that the periodogram \( \hat{S}_P(f) \) (1) at the frequency \( f = f_k = k/(n \Delta t) \), where \( k \) is an integer such that \( k \leq \lfloor n/2 \rfloor \), is equal to the squared absolute value of the discrete Fourier transform \( X(f_k) \) of the sequence \( x_1, \ldots, x_n \) given by the following formula

\[
X(f_k) = \Delta t \sum_{t=1}^{n} x_t \exp\{-i 2\pi f_k t \Delta t\}.
\]  

To reduce the bias and variance of the periodogram \( \hat{S}_P(f) \), the conventional techniques based on tapering and averaging of periodograms is used [2].

2.2 Parametric Estimation of Power Spectra

The widely used form of a parametric power spectra estimation procedure exploits an autoregressive model of order \( p \) for the underlying power spectrum \( S(f) \). A stationary \( AR(p) \) process \( x_t \) with zero mean is described by the following equation

\[
x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \epsilon_t,
\]  

where \( \epsilon_t \) are i.i.d. Gaussian white noises with zero mean and variance \( \sigma^2 \). The power spectrum estimate \( \hat{S}_{AR}(f) \) has the form [2]

\[
\hat{S}_{AR}(f) = \frac{\Delta t \hat{\sigma}^2}{\left| 1 - \sum_{j=1}^{p} \hat{\phi}_j \exp\{-i 2\pi f j \Delta t\} \right|^2}, \quad |f| \leq f(n),
\]  

where \( \hat{\phi}_1, \ldots, \hat{\phi}_p \) and \( \hat{\sigma}^2 \) are the maximum likelihood estimates of the model parameters.
3 Robust Estimation of Power Spectra

3.1 Preliminaries

A natural way to provide robustness of the classical estimates of power spectra is based on using highly robust and efficient estimates of location, scale and correlation in the classical estimates. Here we enlist several highly robust and efficient estimates of scale and correlation.

Robust Scale: The median absolute deviation $\text{MAD}_n(x) = \text{med} |x - \text{med} x|$ is a highly robust estimate of scale with the maximal value of the breakdown point 0.5, but its efficiency is only 0.37 at the normal distribution [3]. In [7], a highly efficient robust estimate of scale $Q_n$ has been proposed: it is close to the lower quartile of the absolute pairwise differences $|x_i - x_j|$, and it has the maximal breakdown point 0.5 as for $\text{MAD}_n$ but much higher efficiency 0.82. The drawback of this estimate is its low computation speed; the computation of $Q_n$ requires an order of greater time than of $\text{MAD}_n$.

In [12], an $M$-estimate of scale denoted by $FQ_n$ whose influence function is approximately equal to the influence function of the estimate $Q_n$ is proposed

$$FQ_n(x) = 1.483 \text{MAD}_n(x) \left(1 - (Z_0 - n/\sqrt{2})/Z_2\right),$$

$$Z_k = \sum_{i=1}^{n} u_i^k e^{-u_i^2/2}, \quad u_i = (x_i - \text{med} x)/(1.483 \text{MAD}_n), \quad k = 0, 2; \quad i = 1, \ldots, n.$$ 

The efficiency and breakdown point of $FQ_n$ are equal to 0.81 and to 0.5, respectively.

Robust Correlation: A remarkable robust minimax bias and variance $\text{MAD}$ correlation coefficient with the breakdown point 0.5 and efficiency 0.37 is given by

$$r_{\text{MAD}}(x, y) = (\text{MAD}^2(u) - \text{MAD}^2(v))/\text{MAD}^2(u) + \text{MAD}^2(v),$$

where $u$ and $v$ are the robust principal variables [10]

$$u = \frac{x - \text{med} x}{\sqrt{2} \text{MAD} x} + \frac{y - \text{med} y}{\sqrt{2} \text{MAD} y}, \quad v = \frac{x - \text{med} x}{\sqrt{2} \text{MAD} x} - \frac{y - \text{med} y}{\sqrt{2} \text{MAD} y}.$$

Much higher efficiency 0.81 with the same breakdown point 0.5 can be provided by using the $FQ$ correlation coefficient [11]

$$r_{FQ}(x, y) = (FQ^2(u) - FQ^2(v))/(FQ^2(u) + FQ^2(v)).$$

3.2 Robust Analogs of the Discrete Fourier Transform

Since computation of the discrete Fourier transform (DFT) (3) is the first step in periodogram estimation of power spectra, consider several robust analogs of the DFT.

Robust $L_p$-Norm DFT Analogs: As the classical DFT (3) $X(f)$ can be obtained via the $L_2$-norm approximation to the data $y_t(f) = x_t \exp\{-i2\pi f t\Delta t\}$, $t = 1, \ldots, n$:

$$X(f) \propto \arg \min_Z \sum_{t=1}^{n} \left|y_t(f) - Z\right|^2,$$
the $L_p$-norm analog of $X(f)$ (up to the scale factor) is defined as follows:

$$X_{L_p}(f) \propto \arg \min_Z \left\{ \sum_{t=1}^n |y_t(f) - Z|^p \right\}^{1/p}, \quad 1 \leq p < \infty. \quad (9)$$

The case of $1 \leq p < 2$, and especially the $L_1$-norm or the median Fourier transform, are of our particular interest [9, 13, 14]. The other possibilities such as the component-wise, spatial medians, and trimmed mean analogs of the DFT are also considered in [9, 14].

**Robust Cross-Product DFT Analog**: Here we exploit the well-known relation connecting the cross-product, covariance and means:

$$\sum x_t z_t = ncov(x,z) + nxz. \quad (10)$$

Since the DFT is decomposed into the real (cosine) and imaginary (sine) parts as $X(f) = X^c(f) + iX^s(f)$, we apply formula (10) to them using robust estimates of covariances and means. Denoting the results of application of formula (10) to cosine and sine parts as $X_{CP}^c(f)$ and $X_{CP}^s(f)$, respectively, we define robust cross-product analog of the classical DFT as follows:

$$X_{CP}(f) = X_{CP}^c(f) + iX_{CP}^s(f). \quad (11)$$

In the case of the conventional estimate of the covariance in (10), namely, the sample covariance $n^{-1} \sum (x_t - \overline{x})(z_t - \overline{z})$, we get the classical definition of the DFT.

### 3.3 Robust Nonparametric Estimation via Periodograms and the Blackman-Tukey Formula

Now we apply the aforementioned robust analogs of the DFT as well as highly robust and efficient estimates of scale and correlation to the classical nonparametric estimation of power spectra.

**Robust Nonparametric Estimation via Periodograms**: Here we apply the robust $L_p$-norm and cross-product analogs of the DFT to the classical periodogram $\hat{S}_P(f)$ (1):

$$\hat{S}_{L_p}(f) \propto |X_{L_p}(f)|^2, \quad \hat{S}_{CP}(f) \propto |X_{CP}(f)|^2. \quad (12)$$

In what follows, the $L_1$- or the median periodogram is of our particular interest.

**Robust Nonparametric Estimation via the Blackman-Tukey Formula**: In order to construct robust modifications of the Blackman-Tukey formula, we have to consider robust estimates of autocovariances $\tilde{c}_{xx}(h)$ instead of the conventional ones used in (2). These robust estimates are based on the highly robust MAD and FQ estimates of scale and correlation (6) - (8):

$$\tilde{c}_{MAD}(h) = r_{MAD}(x_t, x_{t-h})MAD(x_t)MAD(x_{t-h}) = r_{MAD}(h)MAD^2(x), \quad (13)$$

$$\tilde{c}_{FQ}(h) = r_{FQ}(x_t, x_{t-h})FQ(x_t)FQ(x_{t-h}) = r_{FQ}(h)FQ^2(x).$$

To provide the required Teplitz property (symmetry, semipositive definiteness, equal elements on sub-diagonals) of the autocovariance matrix $\tilde{C}_{xx}$ built of the element-wise robust autocovariances (13), a new effective transform is used [6]. Thus, the Teplitz transformed estimates are substituted into formula (2), and the corresponding robust estimates of power spectra are denoted as $\tilde{S}_{MAD}(f)$ and $\tilde{S}_{FQ}(f)$, respectively.
3.4 Robust Parametric Estimation of Power Spectra via the Yule-Walker Equations

A classical approach to estimation of autoregressive parameters $\phi_1, \ldots, \phi_p$ in (4) is based on the solution of the linear system of the Yule-Walker equations [2]:

\[
\begin{align*}
\hat{c}(1) &= \hat{c}(0)\hat{\phi}_1 + \hat{c}(1)\hat{\phi}_2 + \cdots + \hat{c}(p-1)\hat{\phi}_p \\
\hat{c}(2) &= \hat{c}(1)\hat{\phi}_1 + \hat{c}(2)\hat{\phi}_2 + \cdots + \hat{c}(p-2)\hat{\phi}_p \\
&\vdots \\
\hat{c}(p) &= \hat{c}(p-1)\hat{\phi}_1 + \hat{c}(p-2)\hat{\phi}_2 + \cdots + \hat{c}(0)\hat{\phi}_p \\
\end{align*}
\]

The estimate of the innovation noise variance is defined by the following equation

\[
\hat{c}(0) = \hat{c}(1)\hat{\phi}_1 + \hat{c}(2)\hat{\phi}_2 + \cdots + \hat{c}(p)\hat{\phi}_p + \sigma^2.
\]

Substituting robust estimates of autocovariances (13) into (14) and (15), we get the robust analogs of the Yule-Walker equations. Solving these equations, we arrive at the robust estimate of power spectra in the form (5).

3.5 Robust Parametric Estimation via Filtering

A wide collection of robust methods of power spectra estimation is given by various robust filters (Kalman, Masreliez, ACM-type, robust least squares, filter-cleaners, etc.) providing preliminary cleaning the data with the subsequent power spectra estimation. An extended comparative experimental study of robust filters is made in [14]; below we compare some of those results with ours.

4 Performance Evaluation

In Monte Carlo experiment, the autoregressive model AR(2): $x_t = x_{t-1} - 0.9 x_{t-2} + \epsilon_t$ and AR(4): $x_t = x_{t-1} - 0.9 x_{t-2} + 0.5 x_{t-3} - 0.1 x_{t-4} + \epsilon_t$ are used, where the innovation noise $\epsilon \sim (1 - \alpha) N(x; 0, 1) + \alpha N(x; 0, 10)$ with different levels of contamination $\alpha$. Nonparametric and parametric estimates are tested on different sample sizes $n$ and numbers of trials $M$. Some results are exhibited in Figs. 1-4.

5 Concluding Remarks

- From Fig. 1 it follows that the classical periodogram is catastrophically bad under contamination. From Figs. 1-2 it can be also seen that the robust Yule-Walker method considerably outperforms robust filter methods.

- From Fig. 3-4 it follows that the bias of estimation by the robust Yule-Walker method increases with growing dimension and contamination. Under heavy contamination, the nonparametric median periodogram and the robust Blackman-Tukey method outperform the Yule-Walker method in estimating the frequency location, although they have a considerable bias in amplitude.
Figure 1: Power spectra estimation with robust filter-cleaners: \( n = 100, M = 400 \).

Figure 2: Power spectra estimation by the Yule-Walker method: \( n = 128, M = 2000 \).

- The obtained results are preliminary indicating many open problems, most theoretical: analysis of the statistical asymptotic properties of the proposed methods, reducing their bias and variance on finite samples, study of the properties of the
Figure 3: Power spectra estimation by the Yule-Walker method: \( n = 128, M = 2000 \).

Figure 4: Power spectra estimation by nonparametric methods: \( n = 1024, M = 50 \).
direct and inverse $L_p$-norm analogs of the Fourier transform ($1 < p < \infty$) as well as of the direct $L_1$-norm analog, testing cross-product analogs of the DFT.

References


RANK TESTS FOR MULTI-SAMPLE LOCATION PROBLEM

EUGENIA STOIMENOVA
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad.
G.Bontchev str., block 8, 1113 Sofia, Bulgaria
e-mail: jeni@math.bas.bg

Abstract

This paper deals with testing nonparametric hypotheses against ordered alternatives. The test statistics proposed are based on the number of observations from one sample that precede or exceed a threshold specified by the other samples. The approach consists of defining two sets of rankings: the first is induced by the alternative and the other by the data itself. The test statistic measures the distance between the two sets.

1 Multi-sample location problem

Let \( X_{i,1}, \ldots, X_{i,n_i} \), \( i = 1, \ldots, r \), be \( r \) independent random samples with \( X_{i,l} \), \( l = 1, \ldots, n_i \) having an absolutely continuous distribution function \( F_i(x) \). In the parametric case, we may have \( F_i(x) = F(x - \theta_i) \). We shall be concerned with testing the hypothesis of no treatment effect against the alternative that there is a monotone treatment effect.

\[
H_0 : \quad F_1(x) \equiv \ldots \equiv F_r(x) \\
H_A : \quad F_1(x) \geq \ldots \geq F_r(x),
\]

where each inequality is strict for some \( x \). Equivalently, in the parametric case, the hypotheses become ordered location alternative

\[
H_0 : \quad \theta_1 = \ldots = \theta_r \\
H_A : \quad \theta_1 \leq \ldots \leq \theta_r,
\]

where at least one inequality is strict.

A number of distribution free tests are available in the literature. Jonckheere \[6\] and Terpstra \[11\] were among the first to develop a nonparametric test for the nondecreasing ordered alternative (referred to as the JT test) based on the linear combination of the Mann-Whitney \( U \)-statistics associated with the \( r(r - 1)/2 \) possible pairs between the \( r \) ordinal groups. Distances between two sets are naturally involve in nonparametric hypothesis testing. A general approach to hypothesis testing based on ranks was proposed by Critchlow \[2\] and shown to lead to a number of new test statistics.

2 Algebraic structure of rank tests

In this section, we focus on the test procedure using distances between permutations of natural integers modelled along Critchlow’s \[2\] basic construction. His general approach
consists of defining two sets of rankings: the first is induced by the alternative and the other by the data itself.

To test any nonparametric hypothesis $H$ versus an alternative $A$, the first step is to collect all data relevant to the problem, and rank order all of these observations. This produces a single permutation $\alpha \in S_n$. Second step is to identify two suitable sets of permutations. The equivalence class $[\alpha]$ consists of all permutations in $S_n$ which are equivalent (for the particular testing problem) to the observed permutation $\alpha$. The set $E$ of extremal permutations consists of all permutations in $S_n$ which are least in agreement with $H$ and most in agreement with $A$. Then the proposed test statistic is the minimum interpoint distance between the sets $[\alpha]$ and $E$:

$$d([\alpha], E) = \min_{\pi \in [\alpha]} \min_{\sigma \in E} d(\pi, \sigma),$$

where $d$ is an arbitrary metric on $S_n$. The test rejects the null hypothesis for small values of $d([\alpha], E)$.

### 2.1 Multi-sample rank tests

We apply this construction for testing problem (1)

1. Denote the rank of $X_{i,j}$ in the pooled sample by $R_{i,j}$ ($i = 1, \ldots, r; j = 1, \ldots, n_j$). Then $\alpha = \langle R_{1,1}, \ldots, R_{1,n_1}, \ldots, R_{r,1}, \ldots, R_{r,n_r} \rangle \in S_n$.

2. Denote by $S_{n_1}, \ldots, S_{n_r}$ the subgroups of $S_n$ given by:

$$S_{n_1} = \{ \pi \in S_n : \pi(i) = i, \ \forall \ i \notin N_1 \}$$
$$S_{n_2} = \{ \pi \in S_n : \pi(i) = i, \ \forall \ i \notin N_2 \}$$
$$\ldots$$
$$S_{n_r} = \{ \pi \in S_n : \pi(i) = i, \ \forall \ i \notin N_r \},$$

where $N_1, \ldots, N_r \subset \{1, \ldots, n\}$ are:

$$N_1 = \{1, \ldots, n_1\}$$
$$N_2 = \{n_1 + 1, \ldots, n_1 + n_2\}$$
$$\ldots$$
$$N_r = \{n_1 + \cdots + n_{r-1} + 1, \ldots, n\}.$$

Let $S = S_{n_1} \times \cdots \times S_{n_r} \subset S_n$.

The equivalence class $[\alpha]$ is the left coset $\alpha S$. It assigns the same set of ranks to each of the populations as $\alpha$.

3. The extremal set $E$ is the subgroup $S = S_{n_1} \times \cdots \times S_{n_r}$.

4. Let $d$ be a metric on $S \subset S_n$. Then appropriate defined distance between classes $[\alpha]$ and $E$ is used as test statistics for $H_0$ vs. $H_A$.

The test rejects $H_0$ for small values of $d([\alpha], E)$. 

106
2.2 Metrics on permutations

Some common statistical measures on permutations are obtained by applying an affine transformation to the functions:

\[ F(\alpha, \beta) = \sum_i | \alpha(i) - \beta(i) | \quad \text{Spearman’s footrule} \]

\[ R^2(\alpha, \beta) = \sum_i (\alpha(i) - \beta(i))^2 \quad \text{Spearman’s rho} \]

\[ T(\alpha, \beta) = \text{number of pairs } (i, j) \text{ such that } \alpha(i) < \alpha(j) \text{ and } \beta(i) > \beta(j) \quad \text{Kendall’s tau} \]

\[ M(\alpha, \beta) = \max_{1 \leq i \leq n} | \alpha(i) - \beta(i) | \quad \text{Chebyshev’s metric} \]

These and other metrics on permutations and their use in statistics are discussed by Diaconis [4], Critchlow [1, 2], Marden [7], Deza ([3], etc.

2.3 Multi-sample tests induced by metrics

The test statistics for \( H_0 : F_1(x) \equiv \ldots \equiv F_r(x) \) vs. \( H_A : F_1(x) \geq \ldots \geq F_r(x) \) induced by the Kendall’s tau metric, Spearman’s footrule, Spearman’s rho and Chebyshev’s metric are:

\[ T([\alpha], E) = \sum_{j<k} \#\{(i, m) \in N_j \times N_k : \alpha(i) > \alpha(m)\} \]

This is equivalent to the Jonckheere-Terpstra test statistic.

\[ F([\alpha], E) = \sum_{j=1}^r \sum_{i=1}^{n_j} |R_{j,i}^* - k_j + i| \]

\[ R^2([\alpha], E) = \sum_{j=1}^r \sum_{i=1}^{n_j} (R_{j,i}^* - k_j + i)^2 \]

\[ M([\alpha], E) = \max_{1 \leq j \leq r-1} \left[ \max \left\{ R_{j,n_j}^* - k_j, k_j + 1 - R_{j+1,1}^* \right\} \right] \]

where \( R_{1,1}^* \leq \cdots \leq R_{1,n_1}^* \) are the ordered ranks of the first sample, \( \ldots \), \( R_{r,1}^* \leq \cdots \leq R_{r,n_r}^* \) are the ordered ranks of the last sample. Here \( k_j = \sum_{i=1}^{n_j} j \), \( j = 1, \ldots, r - 1 \).

The first three tests are derive by Critchlow [2].

3 Multi-sample test induced by Chebyshev metric

Applying Critchlow’s procedure to Chebyshev metric

\[ M(\alpha, \beta) = \max_{1 \leq i \leq n} | \alpha(i) - \beta(i) | , \]

we have derived (Stoimenova [8])

\[ M([\alpha], E) = \max_{1 \leq j \leq r-1} \left[ \max \left\{ R_{j,n_j}^* - k_j, k_j + 1 - R_{j+1,1}^* \right\} \right] \quad (3) \]
Figure 1: Deviation from the alternative hypothesis.

where $R_{1,1}^* \leq \cdots \leq R_{1,n_1}^*$ are the ordered ranks of the first sample, $\ldots$, $R_{r,1}^* \leq \cdots \leq R_{r,n_r}^*$ are the ordered ranks of the last sample and $k_j = \sum_{i=1}^{j} n_i$, $j = 1, \ldots, r - 1$.

Despite of complicated formula, $M$-statistic is easy for calculation. The test statistic is based on two kind of counts: number of observations from each sample that exceed the largest observation from the other samples and number of observations from each sample that precede the smallest observation from the other samples. Figure 2 illustrate two set corresponding to ranks of the the observed sample and to the extremal ranking for three samples. The extremal set is represented by the identity permutation (points on the diagonal) and observed data is represented by the ordered ranks in each sample placed in series.

3.1 Two-sample test

We illustrate the relation of the proposed tests to the tests based on exceeding observations for the the two-sample problem.

Given two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ with distributions $F_1$ and $F_2$, respectively, we test the hypothesis $H_0$ that $F_1$ and $F_2$ are identical against the alternative $H_A$ that $F_1(x) \geq F_2(x)$, with strict inequality for some $x$. For this problem $M$-test (Stoimenova [9]) defined by (3) has the form,

$$M = \max\{R_{1,m}^* - m + n, n + 1 - R_{2,1}^*\},$$

with $R_{1,m}^*$ the maximum rank among $X_1, \ldots, X_m$ and $R_{2,1}^*$ the minimum rank among $Y_1, \ldots, Y_n$.

Rank tests based on exceedances (Hájek and Šidák [5]) involve the following statistics. Denote by $A$ and $B'$ the number of observations among $X_1, \ldots, X_m$ larger than $\max_{1 \leq j \leq n} Y_j$ or smaller than $\min_{1 \leq j \leq n} Y_j$, and by $A'$ and $B$ denote the number of observations among $Y_1, \ldots, Y_n$ larger than $\max_{1 \leq i \leq m} X_i$ or smaller than $\min_{1 \leq i \leq m} X_i$. 

108
According to this notation the $M$-test is equivalent to
\[ M = \max\{n - A', m - B'\}. \]

The $M$-test allows a simple comparison of two distribution functions. It is sensitive for both deviations from the null hypothesis: shift of $Y$'s to the right with respect to $X$'s or shift of $X$'s to the left with respect to $Y$'s. However, the extreme sample values may get inflated by possible outliers, which may adversely affect these test statistics. For this reason, we may want to reduce their influence by defining thresholds above the smallest and below the largest observed values in the samples. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent random samples from continuous distributions $F$ and $G$, respectively. Thresholds based on $(r+1)$-th order statistic from the $Y$-sample and $(m-s)$-th order statistic from the $X$-sample define the exceedance and precedence statistics of the form
\[ A_s = \text{the number of } Y\text{-observations larger than } X_{(m-s)}, \]
\[ B_r = \text{the number of } X\text{-observations smaller than } Y_{(1+r)}, \]
where $0 \leq s < m$ and $0 \leq r < n$.

To make the $M$-statistic stable to unusual extreme values, Stoimenova and Balakrishnan [10] modify it (in terms of Precedence- and Exceedance-statistics) as
\[ M_r = \max\{n - A_s, m - B_r\}. \]

Evidently, small values of $M_r$ lead to the rejection of $H_0$ in favor of the stochastically ordered alternative $H_A$. Its equivalent formulation in terms of ranks is
\[ M_\rho = \max\{R^*_1, m-s - m + s, m + r + 1 - R^*_2\}, \]
where the threshold statistics $X_{(m-s)}$ and $Y_{(1+r)}$ are determined as $s = [\rho m]$ and $r = [\rho n]$ for some $0 \leq \rho < 1$, with $[\cdot]$ denoting the integer part.

Various values of $\rho$ yield a family of test statistics. It is reasonable for $\rho$ to be small since we want to reduce the possible influence of a small number of potential outliers.

We may use the exact distributions where they are available and also carry out Monte Carlo simulation to estimate the critical values approximately. Some R-functions in this study are provided by Georgi Boshnakov.

4 Šidák-type class of tests for multi-sample problem

Another family of rank statistics for the multi-sample problem is the test statistic defined as a sum of the largest deviations from the extremal set. It has the form
\[ V_\rho = \sum_{j=1}^{r-1} (R^*_{j,nj-sj} - k_j + s_j) + k_j + 1 + s_{j+1} - R^*_{j+1,1+s_{j+1}} \]
Various values of $\rho$ yield a family of test statistics which we refer to as Šidák-type tests. For $\rho = 0$, the $V_{\rho}$-statistic is equivalent to the Šidák test statistic (see Šidák and Vondráček [12] or Hájek and Šidák [5]).

For small samples the null distribution and critical values for rejecting $H_0$ could be estimated using Monte Carlo simulations. For large samples, the gamma approximation turns out to be reasonable. Figure 2 shows the distributional shapes for $n_1 = n_2 = n_3 = 40$ and different choices of the proportion coefficient $\rho$. The first to the last figure correspond to the proportion coefficient $\rho = 0, 0.05, 0.1, 0.15, 0.2, 0.25$, respectively.

Further, the powers of $M_{\rho}$- and $V_{\rho}$-tests are compared to the power of Jonckheere-Terpstra test.

References


Providing data with high utility and no disclosure risk for the public and researchers: an evaluation by advanced statistical disclosure risk methods

M. Templ
Vienna University of Technology & data-analysis OG
Austria
e-mail: matthias@data-analysis.at

Abstract

The demand of data from surveys, registers or other data sets containing sensible information on people or enterprises have been increased significantly over the last years. However, before providing data to the public or to researchers, confidentiality has to be respected for any data set containing sensible individual information. Confidentiality can be achieved by applying statistical disclosure control (SDC) methods to the data. The research on SDC methods becomes more and more important in the last years because of an increase of the awareness on data privacy and because of the fact that more and more data are provided to the public or to researchers. However, for legal reasons this is only visible when the released data has (very) low disclosure risk.

In this contribution existing disclosure risk methods are review and summarized. These methods are finally applied on a popular real-world data set - the Structural Earnings Survey (SES) of Austria. It is shown that the application of few selected anonymisation methods leads to well-protected anonymised data with high data utility and low information loss.

1 Introduction

A microdata file is defined as a data set on individual level. For each observation a set of variables is typically available. Concerning SDC, these variables can be split into three categories.

- **Direct Identifiers:** Variables that definitely identify a statistical unit. For example, the social insurance number, name of companies or people or addresses are considered as direct identifiers.

- **Key variables:** A set of variables that - when considered together - may be used to identify an individual unit. For example with the combination of gender, age, region and occupation some individuals may be identified. Other examples for (confidential) key variables could be income, health information, nationality or political preferences. For the description of the methods, it is advantageous to distinguish between categorical and continuous scaled key variables.

- **Non-confidential variables:** All variables that are not classified in any of the former two groups.
The goal of anonymizing a microdata set is to prevent that confidential information can be linked to a specific respondent. The ultimative aim is to release a safe microdata set that has both, low risk of linking confidential information to individual respondents and high data utility.

2 Measuring disclosure risk

Measuring risk in an microdata set is of course of great concern when having to decide on whether a microdata set is safe to be released. To be able to assess the disclosure risk it is required to make realistic assumptions on the information data users might have at hand to match against the microdata set. These assumptions are called ’disclosure risk scenarios’. Based on a specific disclosure risk scenario one must define a set of identifying variables (key variables) that can be used as input for the risk evaluation procedure.

Typically risk evaluation is based on the concept of “rareness/uniqueness” in the sample and/or in the population. The interest is on units/individuals/observations that possess rare combinations of key variables. Those can be assumed to be identified easier and thus have higher risk. It is possible to cross tabulate all identifying variables and have a look at its cast. Patterns\(^1\) with only very few individuals are in this sense considered risky if they have also low sampling weights, i.e. if the expected individuals with the same pattern is expected to be low in the population.

2.1 Frequencies Counts

Consider a random sample of size \(n\) drawn from a finite population of size \(N\). Let \(\pi_j, j = 1, \ldots, N\) be the (first order) inclusion probabilities, i.e. the probability that the element \(u_j\) of a population of the size \(N\) is chosen in a sample of the size \(n\).

All possible combinations of categories in the key variables \(X_1, \ldots, X_m\) can be calculated by cross tabulation of these categorical variables. Let \(f_i, i = 1, \ldots, n\) be the frequency counts obtained by cross tabulation and let \(F_i\) be the frequency counts of the population which belong to the same category. If \(f_i = 1\) applies the corresponding observation is unique in the sample. If \(F_i = 1\) applies then the observation is unique in the population. Note that \(F_i\) is usually unknown since usually information on samples is collected and only few information about the population is known from registers and/or external sources.

2.2 The \(k\)-Anonymity Concept

Based on a set of key variables a desired characteristic of a protected microdata set might be to achieve \(k\)-anonymity [Samarati and Sweeney, 1998, Sweeney, 2002]. This means that each possible combination of the values of the key variables features at least \(k\) units in the microdata, meaning that all \(f_i \geq 3, i = 1, \ldots, n\). A typical value is \(k = 3\).

\(^1\) a pattern is defined as a specific combination of values of all key variables
$k$-anonymity is typically provided by recoding categorical key variables and by additionally suppressing specific values in the key variables of individual units.

An extension of $k$-anonymity is $l$-diversity [Machanavajjhala et al., 2007]. Consider for one group of observations with the same pattern in the key variables and let the group fulfill $k$-anonymity. A possible data intruder can therefore not identify an individual in this group. However, if all observations have the same entries in a sensitive variable (such as cancer in the variable medical diagnosis) then the attack is successful anyway.

2.3 Considering Sample Frequencies on Subsets: SUDA2

SUDA (Special Uniques Detection Algorithm) estimates a disclosure risk for each individual. SUDA2 [see, e.g., Manning et al., 2008] is a recursive algorithm for finding Minimal Sample Uniques. The algorithm generates all possible variable subsets of defined categorical key variables and scans them for unique patterns in the subsets of variables. The risk of an observation is then dependent on two aspects.

(a) The lower the amount of variables needed to receive uniqueness, the higher the risk (and the higher the *suda score*) of the corresponding observation.

(b) The larger the number of minimal sample uniqueness contained within an observation, the higher the risk of the observation.

(a) is calculated for each observation $i$ by $l_i = \prod_{k=MSU_{\text{min}_i}}^{m-1}(m - k)$, $i = 1, \ldots, n$, for $m$ the depth (the maximum size of variable subsets of the key variables), $MSU_{\text{min}_i}$ the number of minimal uniques of observation $i$ and $n$ the number of observations of the data set. Since each observation is treated independently, the $l_i$ that belongs to one pattern are summed up to result in a common suda score for each of the observation belonging to this pattern (this summation is the contribution of (b)).

To result in the final SUDA score, the suda score are normalized due division by $p!$, with $p$ being the number of key variables. The so called DIS suda score is then calculated from the suda and the so called DIS scores [we refer to Elliot, 2000, for details]. SUDA2 does not consider sampling weights and biased estimates may therefore result.

2.4 Considering Population Frequencies - The Individual Risk

To define if an individual unit is at risk, typically a threshold approach is used. If the individual risk of re-identification for an individual is above a certain threshold value, the unit is said to be at risk. To calculate the individual risks it is necessary to estimate the frequency of a given key in the population. In the previous section, Section 2.1, the population frequencies have already been estimated. However, one can show that these estimates almost always overestimate small population frequency counts [details can be found in Templ and Meindl, 2010] and should not be used to estimate the disclosure risk.
A better approach is to use so-called super-population models in which population frequency counts are modeled given a certain distribution. The whole estimation procedure of sample counts given the population counts can be modeled, for example, by using a Negative Binomial distribution [see, e.g., Rinott and Shlomo, 2006]. It is out of scope of the paper to explain the final measurement of individual risk in this contribution but it can be found in Franconi and Polettini [2004] and Templ and Meindl [2010].

2.5 Measuring the Global Risk

Although the individual risk have to be respected since a data intruder should not be able to identify individuals, often also a measure of the global risk is estimated to express the risk of the whole data set with one number.

2.5.1 Measuring the Global Risk Based on the Individual Risks

The first approach is to determine a threshold for the individual risk and to calculate the percentage of individuals that have larger individual risk than this threshold.

2.5.2 Measuring the Risk Using Log-Linear Models

The sample frequencies, considered for each of $M$ patterns $m$, $f_m, m = 1, ..., M$ can be modeled by a Poisson distribution, and the global risk may be defined as [see Skinner and Holmes, 1998]

$$
\tau_1 = \sum_{m=1}^{M} \exp \left( -\frac{\mu_m (1 - \pi_m)}{\pi_m} \right), \quad \text{with} \quad \mu_m = \pi_m \lambda_m .
$$

For simplicity, the inclusion probabilities are assumed to be equal, $\pi_m = \pi, m = 1, ..., M$. $\tau_1$ can be estimated by log-linear models including the main effects and possible interactions. The model is

$$
\log(\pi_m \lambda_m) = \log(\mu_m) = x_m \beta .
$$

To estimate the $\mu_m$'s, the regression coefficients $\beta$ have to be estimated, for example, by using the iterative proportional fitting. Global risk measure 1 is then given by $\hat{\tau}_1 = \sum_{i=1}^{n} I(f_i = 1) e^{-(1-\pi)\hat{\mu}}$ (corresponding to the risk $P(F_i = 1|f_i = 1)$) and the second one by $\hat{\tau}_2 = \sum_{i=1}^{n} I(f_i = 1) e^{1-(1-\pi)\hat{\mu}} / ((1-\pi)\hat{\mu})$ (corresponding to the risk $E(1/F_i|f_i = 1)$).

2.6 Measuring Risk for Continuous Key Variables

Applying the concept of uniqueness and $k$-anonymity on quantitative variables results that every observation in the data set is unique. Hence, this approach will fail for continuous key variables.

If detailed information about a value of a continuous scaled variable is available, one may be able to identify (by linking information) and eventually gain further information about an individual. For continuous key variables it is assumed that an intruder has information about a statistical unit.
2.6.1 Distance-Based Record Linkage

By using distance based record linkage methods the aim is to find the nearest neighbors between observations from two data sets. Domingo-Ferrer and Torra [2001] has shown that these methods outperform probabilistic methods. Generally, it is evaluated if the original value falls within an interval centered on the masked value. Such an interval might be based on the standard deviation of the variable [see also Mateo-Sanz et al., 2004].

Almost all data sets from Official Statistics consists of statistical units whose values in at least one variable are quite different from the main part of the observations. This leads to the fact that these variables are very asymmetric distributed. Such outliers might be enterprises with a very large value for turnover, for example, or persons with extremely high income or even multivariate outliers. Other disclosure risk methods that are not used in this contribution take the “outlyingness” of an observation into account [for details, see, Templ and Meindl, 2008].

3 Application to the Sturctural Earnings Survey

The Structural Earnings Survey (SES) is conducted in almost all European countries and it includes variables on earnings of employees and other variables on employees and employment level (e.g. region, size of the enterprise, economic activities of the enterprise, gender and age of the employees, …).

Generally such linked employer-employee data are used to identify the determinants/differentials of earnings but also some indicators are directly derived from the hourly earnings like the gender pay gap or the Gini coefficient [Gini, 1912]. The most classical example is the income inequality between genders as discussed in Groshen [1991], for example.

A correct identification of factors influencing the earnings could lead to relevant evidence-based policy decisions. The research studies are usually focused on examining the determinants of disparities in earnings.

The Austrian SES 2006 survey data consists of 199.909 observations obtained from a two-stage design - in the first stage of the design, the enterprises are chosen with certain inclusion probabilities depending on the enterprise size and location, in the second stage employee’s in the selected enterprises are chosen with different inclusion probabilities [for more information have a look at Geissberger, 2009].

3.1 Disclosure Risk and Information Loss for SES

The following variables are chosen as key variables:

**Categorical key variables:** size of enterprise (5 ordered categories), age (66 ordered categories), location (3 categories), economic activity (53 categories)

**Continuous key variables:** hourly earnings, earnings
Table 1: Disclosure risk and information loss on SES

Table 1 shows the resulting disclosure risk and information loss of the SES data. The the columns in Table 1 corresponds to the following data

**orig**: original data (key variables)

**rec1**: (recoding) the variable *economic activity* is recoded to 14 reasonable categories.

**rec2**: (recoding) additionally, the variable *size of employment* is recoded into three reasonable categories (10-49, 50-249, 250-...).

**rec3**: (recoding) additionally, age is discretised into six reasonable categories.

**supp**: (suppression) additionally, local suppression is applied so that no observation violates 3-anonymity.

**mdav**: microaggregation (method mdav, see e.g. Domingo-Ferrer and Mateo-Sanz [2002]) with aggregation level 3.

**add**: additive noise (noise parameter equals 10, see Templ et al. [2012])

**corr**: correlated noise (defaults of Templ et al. [2012])

**sh**: shuffling [Muralidhar and Sarathy, 2006]

The rows of Table 1 corresponds to disclosure risk and information loss measures - **R:2-a** (**R:3-a**): percent of observations violating 2(3)-anonymity, **R:ind**: percent of observations with individual risk below 0.01, **R:suda**: percent of observations having suda dis score lower than 0.1, **R:glob1, R:glob2**: global risks from log-linear models, **R:num**: distance-based disclosure risk, **IL1**: information loss IL1, **IL:eig**: information loss based on differences in the eigenvalues and **IL:lm**: model-based estimation information loss. The mentioned measures of information loss are briefly explained in the following.

<table>
<thead>
<tr>
<th></th>
<th>orig</th>
<th>+rec1</th>
<th>+rec2</th>
<th>+rec3</th>
<th>+supp</th>
<th>mdav</th>
<th>add</th>
<th>corr</th>
<th>sh</th>
</tr>
</thead>
<tbody>
<tr>
<td>R:2-a</td>
<td>2.49</td>
<td>0.47</td>
<td>0.24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R:3-a</td>
<td>5.65</td>
<td>1.12</td>
<td>0.56</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R:ind</td>
<td>2.48</td>
<td>0.67</td>
<td>0.52</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td>R:suda</td>
<td>0.87</td>
<td>0.15</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R:glob</td>
<td>0.83</td>
<td>0.14</td>
<td>0.08</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R:glob</td>
<td>1.35</td>
<td>0.23</td>
<td>0.13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R:num</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.73</td>
<td>7.86</td>
<td>61.86</td>
<td>12.26</td>
</tr>
<tr>
<td>IL1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>11.29</td>
<td>0.11</td>
<td>1.02</td>
</tr>
<tr>
<td>IL:eig</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>5.68</td>
<td>0.06</td>
<td>1.77</td>
</tr>
<tr>
<td>IL:lm</td>
<td>0</td>
<td>0.24</td>
<td>0.24</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>240.29</td>
<td>0.2</td>
<td>8.53</td>
</tr>
</tbody>
</table>

The table presents the disclosure risk and information loss measures for the SES data after applying various data transformation techniques. The measures are calculated based on different methods and parameters, including microaggregation, additive noise, correlated noise, local suppression, and shuffling. The table includes measures such as percent of observations violating 2(3)-anonymity, percent of observations with individual risk below 0.01, and global risks from log-linear models. The information loss is measured using different metrics, including differences in eigenvalues and model-based estimation.
IL1: $IL1 = \frac{1}{p} \sum_{i=1}^{p} \frac{|x_{ij} - x'_{ij}|}{\sqrt{2S_j}}$, scaled distances between original and perturbed values for all $p$ continuous key variables.

IL:eig: The relative absolute differences between the eigenvalues of the covariance standardized continuous key variables of the original and the perturbed variables.

IL:lm: $|\bar{\hat{y}}_w^o - \bar{\hat{y}}_w^p|/\bar{\hat{y}}_w^o$, with $\bar{\hat{y}}_w$ the (Horwitz-Thompson) weighted mean of exponentials of the fitted values from the model $\log(\text{earnings Hour}) \sim \text{age} + \text{Location} + \text{Sex} + \text{education} + \text{Occupation} + \text{economic Activity} + \text{Length} + \text{Size}$ (using weighted least squares estimation considering the sampling weights) obtained from the original (index $o$) and the perturbed data (index $p$).

Table 1 let us to the following interpretation. The original unmodified SES data contains about 5.35 % of observations that violate 3-anonymity and about 2.48 % of risky observations (using the individual risk approach). For the original data, the global model-based risk is 0.83 (and 1.35) which is quite similar to the percentage of observations having high disclosure score (0.87). Of course, the risk on continuous key variables is 100 % and the information loss on that variables is zero. When recoding economic activity into less categories, the risk reduces by almost the factor of 5. When additionally recoding the variable age the risk reduces dramatically. After applying local suppression additionally, the risk for all risk methods zero, expect the individual risk.

The risk on continuous variables is evaluated for any method independently. It is very low for adding additive noise to the data but in the same time the information loss is inacceptable large. The information loss is very small for adding correlated noise, but the risk is still high. For microaggregation, the information loss is (almost) zero, but the risk is high. However, always three observations are aggregated and therefore anonymisation might be fine but the disclosure risk method is not suitable for microaggregation. The performance of shuffling is good, but the model based estimates differ more than 8 % after shuffling the data.

Probably the most interesting information loss measure - the measure which accounts for fitting a linear model on the data (IL:lm) reports that the information loss very low expect for the adding additive noise method and shuffling.

4 Conclusion

In this contribution, popular disclosure risk methods have been summarized. We stressed to measure the disclosure risk after the application of any SDC method to the data. Because of the limit of pages we only briefly focused on measuring the data utility and information loss, but it should be clear that the aim is both, to provide a data set with low disclosure risk and high data utility.

In the practical example, a very popular data set was used and the disclosure risk and data utility/information loss is evaluated. Hereby, the whole range of disclosure
risk methods has been applied to the data, which is done the first time to our knowledge. The results show that by application of few selected anonymisation methods, the disclosure risk dramatically decreases and in the same time, the information loss is considerable small.

All estimations/calculations have been made with the R-package sdcMicro Templ et al. [2012]. The SES data were provided by Statistics Austria.

References


R TOOLS FOR ROBUST STATISTICAL ANALYSIS OF HIGH-DIMENSIONAL DATA

V. Todorov, P. Filzmoser
United Nations Industrial Development Organization (UNIDO)
Vienna, AUSTRIA
e-mail: v.todorov@unido.org

Abstract

The present work discusses robust multivariate methods specifically designed for high dimensions. Their implementation in R is presented and their application is illustrated on examples. The first group of classes are algorithms for outlier detection, already introduced elsewhere and implemented in other packages. The value added of the new package is that all methods follow the same pattern and thus can use the same graphical and diagnostic tools. The next topic covered is sparse principal components including an object oriented interface to the standard method proposed by Zou et al [14] and the robust one proposed by Croux et al [2]. Robust partial least squares (Hubert and Vanden Branden [6]) as well as partial least squares for discriminant analysis conclude the scope of the new package.

1 Introduction

High-dimensional data are typical in many contemporary applications in scientific areas like genetics, spectral analysis, data mining, image processing, etc. and introduce new challenges to the traditional analytical methods. First of all, the computational effort for the anyway computationally intensive robust algorithms increases with increasing number of observations $n$ and number of variables $p$ towards the limits of feasibility. Some of the robust multivariate methods available in R (see Todorov and Filzmoser [11]) are known to deteriorate rapidly when the dimensionality of data increases and others are not applicable at all when $p$ is larger than $n$.

The present work discusses robust multivariate methods specifically designed for high dimensions. Their implementation in R is presented and their application is illustrated on examples. A key feature of this extension of the framework is the object model which follows the one already introduced by rrcov and based on statistical design patterns. The first group of classes are algorithms for outlier detection, already introduced elsewhere and implemented in other packages. The value added of the new package is that all methods follow the same pattern and thus can use the same graphical and diagnostic tools. The next topic covered is sparse principal components including an object oriented interface to the standard method proposed by Zou et al [14] and the robust one proposed by Croux et al [2]. These are presented and illustrated in Section 2. Robust partial least squares ([6], [10]) as well as partial least squares for discriminant analysis are presented in Section 3. Section 4 concludes.
2 Robust sparse Principal Component Analysis

Principal component analysis (PCA) is a widely used technique for dimension reduction achieved by finding a smaller number $q$ of linear combinations of the originally observed $p$ variables and retaining most of the variability of the data. It is important to be able to interpret these new variables, referred to as principal components, especially when the original variables have physical meaning. The link between the original variables and the principal components is given by the so called loadings matrix used for transforming the data and thus it should serve as a means for interpreting the PCs. Dimension reduction by PCA is mainly used for: (i) visualization of multivariate data by scatter plots (in a lower dimensional space); (ii) transformation of highly correlated variables into a smaller set of uncorrelated variables which can be used by other methods (e.g. multiple or multivariate regression); (iii) combination of several variables characterizing a given process into a single or a few characteristic variables or indicators.

The classical approach to PCA measures the variability through the empirical variance and is essentially based on computation of eigenvalues and eigenvectors of the sample covariance or correlation matrix. Therefore the results may be extremely sensitive to the presence of even a few atypical observations in the data. The outliers could artificially increase the variance in an otherwise uninformative direction and this direction will be determined as a PC direction. These discrepancies will carry over to any subsequent analysis and to any graphical display related to the principal components such as the biplot.

PCA was probably the first multivariate technique subjected to robustification, either by simply computing the eigenvalues and eigenvectors of a robust estimate of the covariance matrix or directly by estimating each principal component in a robust manner. Different approaches to robust PCA are discussed in many review papers, see for example [11] and [5], and examples are given how these robust analyses can be carried out in R. Details about the methods and algorithms can be found in the corresponding references. However, PCA usually tends to provide PCs which are linear combinations of all the original variables (by giving them non-zero loadings). Regarding the interpretability of the results it would be very helpful to reduce not only the dimensionality but also the number of used variables (ideally to relate each PC to only a few variables). It is not surprising that vast research effort was devoted to this issue and various proposals have been introduced in the literature. A straightforward informal method is to set to zeros those PC loadings which have absolute values below a given threshold (simple thresholding). Jolliffe et al [9] proposed SCoTLASS which applies a lasso penalty on the loadings in a PCA optimization problem, and recently Zou et al [14] reformulated PCA as a regression problem and used the elastic net to obtain a sparse version - SPCA.

Despite of being more or less successful in achieving sparsity, all these methods suffer a common drawback - all are based on the classical approach to PCA which measures the variability through the empirical variance, and is essentially based on computation of eigenvalues and eigenvectors of the sample covariance or correlation matrix. Therefore the results may be very sensitive to the presence of even a few
atypical observations in the data. The outliers could artificially increase the variance in an otherwise uninformative direction and this direction will be determined as a PC direction. To cope with the possible presence of outliers in the data, recently Croux et al [2] proposed a method which is sparse and robust at the same time. It utilizes the projection pursuit approach where the PCs are extracted from the data by searching the directions that maximize a robust measure of variance of data projected on it. An efficient computational algorithm was proposed by Croux et al [1].

Example  We will use a real data example to illustrate the standard and robust sparse methods—the low-dimensional cars data set, which is available in the package rrcovHD as the data frame cars. For $n = 111$ cars, $p = 11$ characteristics were measured, including the length, width, and height of the car. After looking at pairwise scatterplots of the variables, and computing pairwise Spearman rank correlations $\rho(X_i, X_j)$ we see that there are high correlations among the variables, for example, $\rho(X_1, X_2) = .83$ and $\rho(X_3, X_9) = .87$. We conclude that PCA could be an appropriate method for finding the most important sources of variation in this data set (see also Hubert et al [8]). The first four classical PCs explain more than 96% of the total variance and the first four robust PCs explain more than 95%, therefore we decide to retain four components in both cases. Next we need to choose the degree of sparseness which is controlled by the regularization parameter $\lambda$. Since the sparse PCs have to provide a good trade-off between sparseness and achieved percentage of explained variance we can proceed similarly as in the selection of the number of principal components with the scree plot - we compute the sparse PCA for many different values of $\lambda$ and plot the percent of explained variance against $\lambda$. We choose $\lambda = 0.78$ for classical PCA and $\lambda = 2.27$ for robust PCA, thus attaining 83 and 82 percent of explained variance, respectively, which is only an acceptable reduction compared to the non-sparse PCA. Retaining $k = 4$ principal components as above and using the selected parameters $\lambda$, we can construct the so called diagnostic plots which are especially useful for identifying outlying observations. The diagnostic plot is based on the score distances and orthogonal distances computed for each observation.

The diagnostic plot shows the orthogonal distances versus the score distance, and indicates with a horizontal and vertical line the cut-off values that allow to distinguish regular observations (those with small score and small orthogonal distance) from the different types of outliers: bad leverage points with large score and large orthogonal distance, good leverage points with large score and small orthogonal distance and orthogonal outliers with small score and large orthogonal distance (for detailed description see [8]). In Figure 1 the classical and robust diagnostic plot as well as their sparse alternatives are presented. The diagnostic plots for the standard PCA reveals only several orthogonal outliers and identifies two observations as bad leverage points. Three more observations are identified as bad leverage points by the sparse standard PCA which is already an improvement, but only the robust methods identify a large cluster of outliers. These outliers are masked by the non-robust score and orthogonal distances and cannot be identified by the classical methods. It is important to note that the sparsity feature added to the robust PCA did not influence its ability to detect
properly the outliers.

Figure 1: Distance-distance plots for standard and sparse PCA and their robust versions for the cars data.

3 Robust linear regression and classification in high dimensions

Regression problems become challenging when the number of explanatory variables $p$ exceeds the number of observations $n$. The standard tool to use in these situations is partial least squares regression (PLS). PLS was developed by [13] in the 1960s in the
context of econometric path modeling but some twenty years later it was successfully adopted for regression problems in chemometrics and spectroscopy. It performs a dimensionality reduction to the original regressor variables $X$ by searching for directions $w$. The objective is to maximize the covariance between the scores $Xw$ and a linear projection of the responses $Y$. This ensures that the newly derived regressor variables contain relevant information for the prediction of the responses. There are different models and estimators for the PLS regression problem. The idea of PLS regression is a decomposition of the predictor matrix $X$ and the response matrix $Y$:

$$X = TP^T + Ex$$  \hspace{1cm} (1)$$

$$Y = TQ^T + Ey$$  \hspace{1cm} (2)$$

where $T = XW \in \mathbb{R}^{n \times K}$ is a score matrix, and $W = (w_1, \ldots, w_K) \in \mathbb{R}^{p \times K}$ is a matrix of direction (loading) vectors. The equations (1) and (2) can be regarded as ordinary least squares problems, so $P \in \mathbb{R}^{p \times K}$ and $Q \in \mathbb{R}^{q \times K}$ are matrices of coefficients, whereas $Ex \in \mathbb{R}^{n \times p}$ and $Ey \in \mathbb{R}^{n \times q}$ are matrices of random errors. Again, $K$ denotes the number of components, with $K \leq \min\{n, p, q\}$.

If we rewrite equation (2),

$$Y = TQ^T + Ey = XWQ^T + Ey,$$  \hspace{1cm} (3)$$

we see that $WQ^T \in \mathbb{R}^{p \times q}$ is a matrix of coefficients that relates $Y$ to the original data $X$ according to the original model:

$$Y = XB + E,$$  \hspace{1cm} (4)$$

where the response variable is an $n \times q$ matrix $Y$ of univariate responses. Accordingly, the matrix of regression coefficients $B$ is of dimension $p \times q$, and the error matrix $E$ has the same dimension as $Y$.

In order to successively find direction vectors $w$ that maximize the covariance between the explanatory variables and the responses, the SIMPLS criterion of [3] is used. The first normalized weight vectors $r_1$ and $q_1$ are obtained as linear combinations of $X$ and $Y$ that maximize

$$\text{cov}(Xr_1; Yq_1).$$  \hspace{1cm} (5)$$

The solution of this maximization problem is found by taking $r_1$ and $q_1$ as the first left and right singular eigenvectors of $S_{xy} = X^TY/(n - 1)$, the cross-covariance matrix of the explanatory variables and response variables. For each observation the first coordinate of the score $t_{i1}$ is computed as $t_{i1} = x_i r_1$. The other weight vectors $r_a$ and $q_a$ for $a = 2, \ldots, K$ are obtained by imposing an orthogonality constraint to the elements of the scores.

In general, the weight vectors $r_a$ and $q_a$ are obtained as the left and right singular vector of $S_{xy}^a$ where $S_{xy}^a$ is the deflated covariance matrix:

$$S_{xy}^a = S_{xy}^{a-1} - P_{[a-1]}(P_{[a-1]}^T P_{[a-1]})^{-1} P_{[a-1]}^T S_{xy}^{a-1}$$  \hspace{1cm} (6)$$

125
Finally, when the scores are $K$-dimensional, multivariate linear regression is performed of the responses $y_i$ on these scores $t_i$:

$$B_{p \times q} = R_{p \times K} (T_{K \times n}^T T_{n \times K})^{-1} T_{K \times n}^T Y_{n \times q}$$

$$B_0 = \bar{y} - B_{q \times p}^T \bar{x},$$

where $R_{p \times K} = [r_1, \ldots, r_K]$. A robust alternative to PLS was proposed by Hubert and Vanden Branden [6]. It modifies the SIMPLS algorithm in only two steps. First, robust estimates of $S_{xy}$ and $S_x$ will be obtained by using robust PCA in order to compute a robust covariance matrix in high dimensions. Then a robust multivariate regression method is performed in the second stage. Another approach to robust PLS is the method called partial robust M (PRM) regression [10]. The main idea is to use an M estimator for regression not on the complete but only for a partial information of the explanatory variables. This partial information is obtained via latent variables that need to be extracted in a robust manner (see also [4]).

**Classification in high dimensions**  The prediction of group membership and/or describing group separation on the basis of a data set with known group labels (training data set) is a common task in many applications and linear discriminant analysis (LDA) has often been shown to perform best in such classification problems. However, very often the data are characterized by far more variables than objects and/or the variables are highly correlated which renders LDA (and the other similar standard methods) unusable due to their numerical limitations. Let us assume that $Y$ is univariate and categorical, i.e. $\forall 1 \leq i \leq n : y_i \in \{1, \ldots, G\}$ where $G$ is the number of groups. For high dimensional data sets, classical linear discriminant analysis cannot be performed due to the singularity of the estimated covariance matrix $\hat{\Sigma}$, as it requires the inverse of $\hat{\Sigma}$. To overcome the high dimensionality problem in classification context one can reduce the dimensionality by either selecting a subset of "interesting" variables (variable selection) or construct $K$ new components, $K \ll p$ which represent the original data with minimal loss of information (feature extraction, dimension reduction). Many methods for dimension reduction were considered in the literature but the two most popular are principal component analysis (PCA) and partial least squares (PLS). It is intuitively clear that a supervised method (which uses the group information while constructing the new components) like PLS should be preferred to unsupervised methods like PCA.

PLS was not originally designed to be used in the context of statistical discrimination but nevertheless was routinely applied with evident success by practitioners for this purpose. Taking into account the grouping variable(s) when decomposing the data one would intuitively expect an improved performance for group separation. Since the response variable in case of a classification problem is a categorical variable, none of the robust PLS methods proposed above can be used. Therefore, in order to obtain a robust PLS-DA we proposed to apply any of the outlier detection methods described in Filzmoser and Todorov [5], which are implemented in package rrcovHD, and then
use classical PLS on the already cleaned data set. Hubert and Van Driessen [7] used a

Figure 2: Prediction histograms for class D for the fruit data using classical and robust PLS-DA.

data set containing the spectra of three different cultivars of the same fruit. The three cultivars (groups) are named D, M and HA, and their sample sizes are 490, 106 and 500 observations, respectively. The spectra are measured at 256 wavelengths. The fruit data is thus a high-dimensional data set which was used to illustrate a new approach for robust linear discriminant analysis, and it was studied again by Vanden Branden and Hubert [12]. From these studies it is known that the first two cultivars D and M are relatively homogenous and do not contain atypical observations, but the third group HA contains a subgroup of 180 observations which were obtained with a different illumination system. In Figure 2 are shown the prediction histograms for class D for the fruit data using classical and robust PLS-DA.

4 Summary and conclusions

An object oriented framework for robust multivariate analysis developed in the S4 class system of the programming environment R already exists implemented in the package rrcov and is described in [11]. The main goal of this framework is to support the usage, experimentation, development and testing of robust multivariate methods as well as simplifying comparisons with related methods. In this article we investigated several robust multivariate methods specifically designed for high dimensions. All considered methods and data sets are available in the R package rrcovHD. A key feature of this extension of the framework is that the object model follows the one already introduced by rrcov which is based on statistical design patterns.

Acknowledgements

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization.
References


GENERALIZED METHOD OF WAVELET MOMENTS FOR THE
ESTIMATION OF COMPOSITE STOCHASTIC MODELS

MARIA-PIA VICTORIA-FESER
Research Center for Statistics
HEC-Genève - University of Geneva
Switzerland
e-mail: maria-pia.victoriafeser@unige.ch

Joint work with Stéphane Guerrier, Yannick Stebler, Jan Skalud, Roberto Milinari

Abstract

We present a new estimation method for the parameters of a times series model called the Generalised Method of Wavelet Moments (GMWM) estimator; see Guerrier et al. (2013). We consider here composite Gaussian processes that are the sum of independent Gaussian processes which in turn explain an important aspect of the time series, as is the case in engineering and natural sciences. The proposed estimation method offers an alternative to classical estimation based on the likelihood, that is straightforward to implement and often the only feasible estimation method with complex models. The estimator results as the optimization of a criterion based on a standardized distance between the sample wavelet variances (WV) estimates and the model based WV. Indeed, the WV provides a decomposition of the variance process through different scales, so that they contain the information about different features of the stochastic model. We derive the asymptotic properties of the proposed estimator for inference and perform a simulation study to compare our estimator to the MLE and the LSE with different models. We also propose a robust alternative, based on a robust WV estimator. We use the new estimator to estimate the stochastic error’s parameters of the sum of three first order Gauss-Markov processes by means of a sample of over 800,000 issued from gyroscopes that compose inertial navigation systems.

References

ON THE EXACT TOLERANCE INTERVALS FOR UNIVARIATE NORMAL DISTRIBUTION

VIKTOR WITKOVSKÝ
Institute of Measurement Science, Slovak Academy of Sciences
Bratislava, SLOVAKIA
e-mail: witkovsky@savba.sk

Abstract

Statistical tolerance interval is another type of interval estimator used for making statistical inference on an unknown population. Simply stated, it is an interval estimator, based on a sample from preliminary experiment, which can be asserted with confidence level $1 - \alpha$ (for example 0.95), to contain at least a specified proportion, say $1 - \gamma$ (for example 0.99), of the items in the population under consideration. The limits of a statistical tolerance interval are called statistical tolerance limits. The confidence level $1 - \alpha$ is the probability that a statistical tolerance interval constructed in the prescribed manner (i.e. based on result of an experiment conducted under identical conditions) will contain at least a proportion $1 - \gamma$ of possibly infinite sequence of items coming from the considered (unknown) population (i.e. realizations of independent random variables from the given distribution). In contrast with other statistical intervals commonly used for statistical inference, like e.g. the confidence intervals for the parameters and/or the prediction intervals for future observation(s), the tolerance intervals are used relatively rarely. One reason is that the theoretical concept and computational complexity is significantly more difficult, if compared with the commonly used confidence and prediction intervals. In this paper we briefly describe the theoretical background and computational approaches for computing the tolerance factors and limits for statistical tolerance intervals based on samples from univariate normal (Gaussian) populations.

1 Introduction

Statistical tolerance intervals are interval estimators used for making statistical inference on population(s), which can be fully described by a probability distribution from given family of distributions (as e.g. the family of normal distributions). For more details on different types of statistical intervals consult the books by Hahn and Meeker (1991), Krishnamoorthy and Mathew (2009), and Liu (2011).

Although the concept of statistical tolerance intervals has been well recognized for long time, surprisingly, it seems that their applications remain still limited. The reliable algorithms for computing the exact tolerance factors are missing in the commonly used statistical packages (even for inferences on normal populations), however, more or less accurate approximations are available. Implementations of such algorithms (mainly based on approximate and/or Monte Carlo methods) are currently under fast development, as e.g. in the package tolerance for R, see Young (2010).

Thus, possible applications should rely either on implemented approximate methods or on published collections of tables for tolerance factors (if available), see e.g. the
book by Odeh and Owen (1980) which gives many of the most required factors in the context of the normal distribution, however with limited precision. Due to the recognized importance of statistical tolerance intervals in technical applications, ISO (the International Organization for Standardization) has currently prepared a revised version of the standard ISO 16269-6 (Statistical interpretation of data — Part 6: Determination of statistical tolerance intervals), which also provides detailed tables of tolerance factors for selected tolerance intervals.

The theory of statistical tolerance intervals (tolerance regions), as well as the computational methods and algorithms, have been significantly developed during the last three decades. This, together with the fast growing computational power of the personal computers, allows development of fast, efficient and reliable implementations of the algorithms for highly precise computing of the exact tolerance factors and limits required for the statistical tolerance intervals. For a comprehensive overview of the recent advances and developments in this area see the book by Krishnamoorthy and Mathew (2009).

In this paper we shall briefly describe the theoretical background and computational approaches for computing the exact two-sided tolerance factors and limits for statistical tolerance intervals based on sample from normal (Gaussian) population(s). The described approaches are based mainly on the results presented in Krishnamoorthy and Mathew (2009).

Based on those methods and algorithms we have developed MATLAB algorithm for efficient and highly precise computation of the exact tolerance factors for population(s) with assumed univariate normal distribution.

2 Two-sided tolerance intervals for univariate normal distribution

Let us consider a simple calibration experiment, say \( \mathcal{E} \), which is represented by a random sample of size \( n \) from a population whose distribution is characterized by a univariate normal distribution \( N(\mu, \sigma^2) \), i.e. \( \mathbf{X} = (X_1, \ldots, X_n) \) where \( X_i \) are independent random variables normally distributed as \( X_i \sim N(\mu, \sigma^2) \), where \( \mu \) and \( \sigma^2 \) are unknown parameters (mean and variance) of the population distribution.

Notice that the available information on distribution of the unknown population, based on the result of the calibration experiment \( \mathcal{E} \), is fully characterized by the random sample \( \mathbf{X} \) or equivalently by the sufficient statistics \( \bar{X} \) (the sample mean) and \( S^2 \) (the sample variance), i.e. \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \) and \( S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \). Under given assumptions, it is well known that the sufficient statistics are independent random variables and their distribution is given by \( \bar{X} \sim N(\mu, \delta^2 \sigma^2) \), where \( \delta^2 = \frac{1}{n} \), and \( S^2 \sim \sigma^2 \frac{1}{\nu} \chi^2_\nu \), where \( \nu = n - 1 \) denotes the degrees of freedom (DFs) and \( \chi^2_\nu \) represents a chi-square distribution with \( \nu \) DFs.

Given the result of the calibration experiment \( \mathcal{E} \), we wish to construct the two-sided tolerance interval (i.e. a random interval with its limits \( (L_\mathcal{E}, U_\mathcal{E}) \) depending on the result of the experiment \( \mathcal{E} \)) which can be asserted with confidence level \( 1 - \alpha \) (for
example 0.95) to contain at least a specified proportion 1 − γ (for example 0.99) of the items in the population under consideration.

That is, we wish to construct the two-sided (1 − γ, 1 − α)-tolerance interval which will cover pre-specified proportion of possibly infinite sequence of independent future realizations of the response variable $X = \mu + \sigma \varepsilon$ (with $\varepsilon \sim N(0,1)$ assumed to be independent of the calibration experiment $\mathcal{E}$) such that

$$\Pr(\varepsilon) \left( \Pr(X) \left( L_\mathcal{E} \leq X \leq U_\mathcal{E} | \mathcal{E} \right) \geq 1 - \gamma \right) = 1 - \alpha.$$  \hfill (1)

Notice that the confidence level 1 − α is related to the random nature of the outcome (result) of the calibration experiment $\mathcal{E}$. That is, the required two-sided tolerance interval will cover more than $(1 - \gamma) \times 100\%$ proportion of the items of the unknown (normal) population, and this will be true in $(1 - \alpha) \times 100\%$ cases of the hypothetical calibration experiments.

In general, there are potentially many possible approaches how to find a solution to the problem as specified by (1). There is no unique solution until the form of the tolerance limits of the two-sided tolerance interval $(L_\mathcal{E}, U_\mathcal{E})$ is reasonably restricted. Commonly, the tolerance limits are considered in the form

$$L_\mathcal{E} = \bar{X} - \kappa \sqrt{S^2}, \quad U_\mathcal{E} = \bar{X} + \kappa \sqrt{S^2},$$  \hfill (2)

where $\kappa$ denotes the tolerance factor (the parameter which is a subject of the required solution) which depend on the stated coverage and confidence probabilities, 1 − γ and 1 − α, and further on the parameters characterizing the design of the experiment, $\delta^2$ and $\nu$. So, if necessary, we can emphasize the dependence of the tolerance factor $\kappa$ on other parameters by writing either $\kappa(1 - \gamma, 1 - \alpha, \delta^2, \nu)$, or $\kappa(\delta^2, \nu)$, etc.

Consequently, the following conditional probability statement should be fulfilled for $(1 - \alpha) \times 100\%$ of the possible results of the calibration experiment (i.e. $\bar{X}$ and $S^2$)

$$1 - \gamma \leq \Pr(X) \left( L_\mathcal{E} \leq X \leq U_\mathcal{E} \right)$$

$$= \Pr(\varepsilon) \left( \bar{X} - \kappa \sqrt{S^2} \leq \mu + \sigma \varepsilon \leq \bar{X} + \kappa \sqrt{S^2} \right)$$

$$= \Pr(\varepsilon) \left( (\bar{X} - \mu)/\sigma - \kappa \sqrt{S^2}/\sigma^2 \leq \varepsilon \leq (\bar{X} - \mu)/\sigma + \kappa \sqrt{S^2}/\sigma^2 \right)$$

$$= \Phi(\delta Z + \kappa \sqrt{Q/\nu}) - \Phi(\delta Z - \kappa \sqrt{Q/\nu}) \equiv C(\kappa | Z, Q),$$  \hfill (3)

where $Z = (\bar{X} - \mu)/\delta \sigma$, $Q = \nu S^2/\sigma^2$, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. So, $C(\kappa | Z, Q)$ represents the proportion of the population covered by the tolerance interval for given tolerance factor $\kappa$ and for the given result of the calibration experiment $\mathcal{E}$. $C(\kappa | Z, Q)$ is commonly known as a content function.

The content function $C(\kappa | Z, Q)$ cannot be evaluated directly for given $\kappa$ and the observed result of the experiment $\mathcal{E}$ ($\bar{X}$, and $S^2$), as it depends on the unknown parameters $\mu$ and $\sigma^2$. However, if we are interested in the stochastic properties of the tolerance intervals based on a large number of results of the hypothetical calibration
experiments (i.e. the variability of the results of independent calibration experiments is to be considered), then \( Z \) and \( Q \) are independent pivotal random variables with known probability distributions independent of the unknown parameters \( \mu \) and \( \sigma^2 \), i.e. \( Z \sim N(0, 1) \) and \( Q \sim \chi^2_\nu \).

So, \( C(\kappa | Z, Q) \) can be used directly for checking the properties (i.e. the true confidence) of any proposed value of tolerance factor \( \kappa \). Or alternatively, by applying a suitable iterative optimization procedure, \( C(\kappa | Z, Q) \) can be used also for computing the exact value of the tolerance factor \( \kappa \), such that it fulfills the required properties given by (1) and (2). This may be realized by using repeated Monte Carlo simulations, or repeated two-dimensional numerical integration.

The below presented formula for computing the tolerance factors \( \kappa \) of the two-sided \((1 - \gamma, 1 - \alpha)\)-tolerance intervals for a univariate normal distribution requires (repeated) evaluation of one-dimensional integral, only. As we shall discuss in more details in the next Sections, the approach can be generalized also for computing the tolerance factor for other models based on normal distribution (as e.g. the non-simultaneous (pointwise) tolerance intervals as well as the simultaneous tolerance intervals for normal linear regression models), however with possibly needed evaluation of multivariate integrals.

Derivation is based on the results presented in Mathew and Krishnamoorthy (2009) (for more details see the equations (1.2.3), (1.2.4), also (2.5.7) and (2.5.8)).

Note that, for a fixed \( \delta \) and \( Z \), the function \( \Phi(\delta Z + r) - \Phi(\delta Z - r) \) is an increasing function of \( r \). Therefore, \( C(\kappa | Z, Q) \geq 1 - \gamma \) if and only if \( \kappa \sqrt{Q/\nu} > r \) (or equivalently \( Q \geq \nu r^2/\kappa^2 \)), where \( r \) is the solution to the equation

\[
\Phi(\delta Z + r) - \Phi(\delta Z - r) = 1 - \gamma
\]

or equivalently,

\[
\Pr(\varepsilon) \left( (\varepsilon - \delta Z)^2 \geq r^2 \mid Z \right) = 1 - \gamma
\]

where \( \varepsilon \sim N(0, 1) \). Thus, for fixed \( Z \), the random variable \( (\varepsilon - \delta Z)^2 \sim \chi^2_1(\delta^2 Z^2) \), i.e. it has a non-central chi-square distribution with one degree of freedom and the noncentrality parameter \( \delta Z \). Consequently, \( r = \sqrt{\chi^2_{1,1-\gamma}(\delta^2 Z^2)} \), where \( \chi^2_{1,1-\gamma}(\delta^2 Z^2) \) denotes the \((1 - \gamma)\)-quantile of the distribution \( \chi^2_1(\delta^2 Z^2) \).

Based on that, the tolerance factor \( \kappa \) defined by (1) and (2) is given implicitly as a solution to the equation

\[
E_{\{Z\}} \left( \Pr_{\{Q\}} \left( Q \geq \frac{\nu}{\kappa^2} \chi^2_{1,1-\gamma}(\delta^2 Z^2) \right) \right) = E_{\{Z\}} \left( 1 - F_{\chi^2_{\nu}} \left( \frac{\nu}{\kappa^2} \chi^2_{1,1-\gamma}(\delta^2 Z^2) \right) \right) = 1 - \alpha
\]

where \( E_{\{Z\}}(f(Z)) \) denotes the expectation of the function \( f(Z) \), with \( Z \sim N(0, 1) \) and \( F_{\chi^2_{\nu}}(\cdot) \) denotes the CDF of a chi-square distribution with \( \nu \) degrees of freedom.

The formula (6) can be further generalized for computing the common tolerance factor \( \kappa \) of the \( m \) simultaneous tolerance intervals for the independent normal populations with possibly different means \( \mu_i \) and common variance \( \sigma^2 \).

However, it is true only under rather restrictive assumption that the result of calibration experiment \( \mathcal{E} \) is based on \( m + 1 \) sufficient statistics, i.e. \( \bar{X}_1, \ldots, \bar{X}_m \) with
\[ \bar{X}_i = \frac{1}{n} \sum_{j=1}^{m} X_{ij} \] (the sample means) and \[ S^2 = \frac{1}{\nu} \sum_{i=1}^{m} (n - 1) S^2_i \] (the pooled sample variance), with \[ S^2_i = \frac{1}{n-1} \sum_{j=1}^{m} (X_{ij} - \bar{X}_i)^2 \] and \( \nu = m(n - 1) \), where \( n \) is the common sample size for each population. Computation of the tolerance factors for simultaneous tolerance intervals based on an unbalanced sampling should be treated differently (similarly as for the simultaneous tolerance intervals for linear regression).

The generalized formula is derived from (6) by considering the distribution of the random variable \( Z_{\text{max}} = \max(|Z_1|, \ldots, |Z_m|) \) (where \( Z_i \sim N(0, 1) \), \( i = 1, \ldots, m \), are independent random variables) instead of \( Z \sim N(0, 1) \).

In summary, the required tolerance factor \( \kappa \) can be computed as a solution to the equation

\[
2m \int_{0}^{\infty} \left( 1 - F_{X^2} \left( \frac{\nu}{\kappa^2} \chi^2_{1:1-\gamma} \left( \delta^2 z^2 \right) \right) \right) \left( 2\Phi(z) - 1 \right)^{m-1} \phi(z) \, dz = 1 - \alpha, \tag{7}
\]

where \( \delta^2 \) is the scale parameter of the variance of the mean estimator (e.g. \( \delta^2 = \frac{1}{n} \) for the sample mean), \( \nu \) represents the degrees of freedom associated with the variance estimator (e.g. \( \nu = n - 1 \) for the sample variance and \( \nu = m(n - 1) \) for the pooled sample variance), \( m \geq 1 \) represents the parameter of simultaneity (the number of independent, equally sampled normal populations with possibly different means \( \mu_i \), common variance \( \sigma^2 \), and with common sample size \( n \) for all \( m \) populations, used in the calibration experiment).

Further, \( \Phi(\cdot) \) denotes the CDF and \( \phi(\cdot) \) the PDF of the standard normal distribution, \( F_{X^2}(\cdot) \) denotes the CDF of a chi-square distribution with \( \nu \) degrees of freedom. The probability \( 1 - \alpha \) is the required nominal confidence level and \( \chi^2_{1:1-\gamma} \) denotes the \( (1 - \gamma) \)-quantile of the non-central chi-squared distribution with 1 degree of freedom and the non-centrality parameter \( \sqrt{\delta^2 z^2} (1 - \gamma) \) is the required coverage/content of the tolerance interval). From computational point of view, the value \( r^2 = \chi^2_{1:1-\gamma} \) can be computed more efficiently by directly solving the equation (4), i.e. \( \Phi(\delta z + r) - \Phi(\delta z - r) = 1 - \gamma \).

Notice that for \( m = 1 \), the tolerance factors given by the solution to the equation (7) are equivalent to the factors given by the solution to the equation (6), i.e. to the (non-simultaneous) tolerance intervals as defined by (1) and (2).

### 3 The MATLAB algorithm ToleranceFactorGK

Based on (7), the MATLAB algorithm ToleranceFactorGK computes the tolerance factors \( \kappa \) for the two-sided tolerance intervals by using an adaptive Gauss-Kronod quadrature. Usage of the complementary incomplete Gamma function (for computing the CDF of chi-square distribution) and the complementary error function (for computing the CDF of standard normal distribution) allows precise evaluation of the tolerance factors even for extremely small values of the probabilities \( \gamma \) and/or \( \alpha \) (i.e. for extremely high coverage and confidence). The complementary error function is also used to find a solution (root) \( r \), of the equation \( 1 - (\Phi(x + r) - \Phi(x - r)) = 1 - \gamma \), by using the Halley’s method (root-finding algorithm based on two function deriva-
The current version of the algorithm is available at the web page http://www.mathworks.com/matlabcentral/fileexchange/24135-tolerancefactor.

For illustration and possible comparisons with other algorithms, here we present several values of the tolerance factor $\kappa$ (presented with up to 15 decimal places) computed by the algorithm ToleranceFactorGK for the two-sided $(1-\gamma, 1-\alpha)$-tolerance interval for univariate normal population, based on a calibration experiment characterized by the parameters $n$, $\delta^2$, $\nu$, and $m$.

**Example 1.** Let us consider the following parameters: $\gamma = 0.01$, $\alpha = 0.05$, $n = 10$, $m = 1$, $\nu = n-1$, and $\delta^2 = \frac{1}{n}$. The tolerance factor, defined as a solution to the equation (7), is calculated in MATLAB by using the algorithm ToleranceFactorGK:

```matlab
gamma = 0.01; alpha = 0.05;
n = 10; m = 1; nu = n-1; delta2 = 1/n;
kappa= ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta2)
kappa = 4.436908728948544
```

**Example 2.** As was explained in Section 2, by solving the equation (7), it is possible to compute the common tolerance factor also for the simultaneous tolerance intervals of $m$ populations, assuming that the common sample size for all $m$ populations is $n$. Let us consider the following parameters: $\gamma = 0.01$, $\alpha = 0.05$, $n = 10$, $m = 4$, $\nu = m(n-1)$, and $\delta^2 = \frac{1}{n}$. The (simultaneous) tolerance factor is calculated by

```matlab
gamma = 0.01; alpha = 0.05;
n = 10; m = 4; nu = m*(n-1); delta2 = 1/n;
options.Simultaneous = true;
kappa= ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta2,options)
kappa = 3.574857233534562
```

**Example 3.** The information from $m$ independent sources can be effectively used also if we are interested to calculate a non-simultaneous tolerance interval for one particular population. However, we wish to use the pooled sample variance estimator (i.e. with more degrees of freedom than could be achieved from one sample). So, let us consider the following parameters: $\gamma = 0.01$, $\alpha = 0.05$, $n = 10$, $m = 4$, $\nu = m(n-1) = 36$ $\delta^2 = \frac{1}{n}$. Now, the (non-simultaneous) tolerance factor is $\kappa = 3.38557968498129$. Notice that the tolerance factor can be calculated also if we directly set $m = 1$ and $\nu = 36$ (if $m = 1$ the calculated tolerance factor is non-simultaneous).

```matlab
gamma = 0.01; alpha = 0.05;
n = 10; m = 4; nu = m*(n-1); delta2 = 1/n;
options.Simultaneous = false;
kappa= ToleranceFactorGK(n,1-gamma,1-alpha,m,nu,delta2,options)
kappa = 3.38557968498129
```
Example 4. In order to illustrate the ability to compute the tolerance factors even for extremely large values of the coverage and confidence probabilities, let us consider the following parameters: $\gamma = 10^{-5}$, $\alpha = 10^{-18}$, $n = 250$, $m = 1$, $\nu = n - 1$, $\delta^2 = \frac{1}{n}$. The calculated value of the tolerance factor is given by

```matlab
gamma = 1e-5; alpha = 1e-18;
n = 250; m = 1; nu = n-1; delta2 = 1/n;
options.TailProbability = true;
kappa= ToleranceFactorGK(n,gamma,alpha,m,nu,delta2,options)
kappa = 6.967664575030617
```

Example 5. Finally we note, the the algorithm can be directly used for computing the tolerance factors of the non-simultaneous (point-wise) tolerance intervals for normal (homoscedastic) linear regression models. The tolerance interval should be calculated for each value (vector) of explanatory variables $x$ (which is of interest) separately, by setting $\delta^2_x = x'(X'X)^{-1}x$ and $\nu = n - k$, where $X$ is a design matrix of the linear regression model used for calibration experiment and $k$ represents the number of parameters.

4 Discussion on simultaneous tolerance intervals for normal linear regression model

Although the motivation for computing the exact tolerance factors for simultaneous tolerance intervals in (normal) linear regression models is rather strong, the required methods and algorithms for computation are more complicated than for the non-simultaneous tolerance intervals. The efficient algorithms are still missing in the commonly used statistical packages.

A good motivation for computing such simultaneous tolerance intervals is the multiple-use calibration and the associated calibration confidence intervals. In many experimental sciences, acquisition of the measurement results frequently requires measurement procedures involving instrument calibration which can be modeled as linear (polynomial) regression problem. Then, the required measurement result, say $x$, is obtained through measuring the observable response variable, say $y$, and by inverting the fitted regression (calibration) function. A problem of constructing and computing the appropriate (exact) confidence intervals for the unobservable values of the explanatory variable $x$, based on given fitted calibration function (a result of the calibration experiment), for possibly unlimited sequence of future observations of the response variable $y$ is known as the multiple-use calibration problem.

As proposed by Scheff (1973), such calibration intervals for $x$ values can be obtained from simultaneous tolerance intervals for the considered regression (calibration) function, with warranted minimum $(1 - \gamma)$-content for all such intervals simultaneously, and with confidence at least $(1 - \alpha)$. However, the known simultaneous tolerance intervals in regression are typically conservative in that the actual confidence level exceeds the
required nominal level \((1 - \alpha)\), and as such are generally broader than they necessarily should be. For an overview of the problem and methods for simultaneous tolerance intervals in linear regression and suggested improvements see e.g. Mee et al. (1991), Mee and Eberhardt (1996), Mathew and Zha (1997), Mathew and Krishnamoorthy (2009) and Chvostekov (2013).

The implementation of the generalization of the algorithm suitable also for efficient computation of the simultaneous tolerance factors for linear regression model is currently under development by the author.

**Acknowledgements**

The work was supported by the Slovak Research and Development Agency, projects APVV-0096-10, SK-AT-0025-12, and by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and the Slovak Academy of Sciences, projects VEGA 2/0038/12 and VEGA 2/0043/13.

**References**


DIFFERENCES BETWEEN THE PEARSON STATISTICS DISTRIBUTION FUNCTION AND NON–CENTRAL CHI–SQUARE DISTRIBUTION FUNCTION

M.V. Filina, A.M. Zubkov
Steklov Mathematical Institute
Moscow, Russia
e-mail: mfilina@mi.ras.ru, zubkov@mi.ras.ru

Abstract

We investigate the distribution of the Pearson statistics in the goodness-of-fit test of discrete distribution uniformity in the case of alternative hypothesis. Exact formulas for first two moments are obtained. Results of exact numerical comparisons of Pearson statistics distribution and non-central chi-square distribution will be discussed.

1 Introduction

Let $\nu_1, \ldots, \nu_N$ be frequences of all $N$ outcomes of a multinomial scheme in a sample of size $T$. A standard goodness-of-fit test for the hypothesis

$H$: probabilities of outcomes are equal to $p_1, \ldots, p_N$

is based on the Pearson statistics

$$X^{2}_{N,T} = \sum_{k=1}^{N} \frac{(\nu_k - Tp_k)^2}{Tp_k}.$$ 

It is well-known that if this hypothesis is valid, then the distribution of $X^{2}_{N,T}$ converges to the chi-square distribution with $N - 1$ degrees of freedom as $T \to \infty$.

If the hypothesis $H$ is not valid and the true probabilities of outcomes are equal to $x_1, \ldots, x_N$, then the Pearson statistics $X^{2}_{N,T}$ converges weakly to the random variable $X^{2}_{N,T,\lambda}$ having the noncentral chi-square distribution with noncentrality parameter

$\lambda = N(T-1) \sum_{k=1}^{N} a_k^2$ if $p_k - x_k = a_k T^{-1/2}, k = 1, \ldots, N$, as $T \to \infty$ and $p_1, \ldots, p_N$ are fixed.

We consider the case $p_1 = \ldots = p_N = \frac{1}{N}$, which is of particular interest for testing the uniformity of discrete distributions.

The first two moments of the Pearson statistics in this case are equal to

$$\mathbb{E} \sum_{k=1}^{N} \frac{(\nu_k - T/N)^2}{T/N} = N(T-1) \sum_{k=1}^{N} \left(x_k - \frac{1}{N}\right)^2 + N - 1,$$ (1)
\[
D \sum_{k=1}^{N} \left( \frac{\nu_k - T/N}{T/N} \right)^2 = \frac{2N^2(T-1)}{T} \left[ \left( 1 - \sum_{k=1}^{N} x_k^2 \right) \sum_{k=1}^{N} x_k^2 + 2(T-2) \left[ \sum_{k=1}^{N} x_k \left( x_k - \frac{1}{N} \right)^2 - \left( \sum_{k=1}^{N} x_k \left( x_k - \frac{1}{N} \right)^2 \right)^2 \right] \right].
\]

Note that (2) differs slightly from the formula

\[
D \sum_{k=1}^{N} \left( \frac{\nu_k - T/N}{T/N} \right)^2 = 2N^2 \left[ \sum_{k=1}^{N} x_k^2 + 2(T-1) \sum_{k=1}^{N} x_k^3 - (2T-1) \left( \sum_{k=1}^{N} x_k^2 \right)^2 \right] \quad (3)
\]

proved in [2] and cited in [1], 30.24, because (3) was obtained in [2] as the main term of the asymptotic expansion of variance. We see that exact formula is not very complicated compared to the asymptotic one (and is somewhat more transparent). We hope to derive analogous exact formulas without the assumption \( p_1 = \ldots = p_N = \frac{1}{N} \).

By means of the same methods as in [3], [4] we investigated the differences between \( P\{X_{N,T,\lambda}^2 \leq y\} \) and the distribution function \( F(y; N-1, \lambda) \) of noncentral chi-square distribution with \( N-1 \) degrees of freedom, noncentrality parameter \( \lambda \) for \( p_1 = \ldots = p_N = \frac{1}{N} \) and different sets of “true” probabilities \( x_1, \ldots, x_N \).

We discuss some interesting features of the differences between exact Pearson statistics distributions and corresponding non-central chi-square distributions obtained in our computational experiments.

References


ROBUST AND MULTIVARIATE DATA ANALYSIS
DETECTION OF OUTLIERS WITH BOXPLOTS
K. Andrea, G.L. Shevlyakov, P.O. Smirnov
St. Petersburg State Polytechnic University
St. Petersburg, RUSSIA
E-mail: Kliton.Andrea@gmail.com

Abstract
Low-complexity robust modifications to the Tukey boxplot based on fast highly efficient robust estimates of scale are proposed. The performance of the Tukey boxplot and its modified robust versions is measured relative to identification of outliers in Monte Carlo experiments at contaminated normal distributions. The obtained results show that the proposed methods outperform the conventional Tukey boxplot and the classical Grubbs test.

1 Introduction
Robust statistics provides stability of statistical inferences under departures from the accepted distribution models. Although robust statistical procedures involve highly refined asymptotic tools, they exhibit satisfactory behavior within small samples and therefore are quite useful in real-world applications.

In parallel with robust statistics, practical methods for analyzing data evolved known as Exploratory Data Analysis (EDA). A significant feature of EDA is that it does not assume an underlying probability distribution for the data which is typical in classical statistical methods and therefore is flexible in practical settings.

Our work represents new results in robust data analysis technologies, providing alternatives to the boxplot technique. The univariate Tukey method summarizes the characteristics of a data distribution allowing for a quick visual inspection of streams of data over windows. Despite being a simple data analysis tool, it concisely summarizes information about the location, scale, asymmetry, tails, and outliers in the data distribution. In our study, we concentrate on visualization of distribution tails and on detection of outliers in the data.

The remainder of the paper is organized as follows. In Section 2, two new robust versions of the Tukey boxplot based on the highly efficient robust estimates of scale are proposed. In Section 3, two new rules for detection of outliers based on the proposed robust boxplots are introduced and examined on the contaminated Gaussian data. In Section 4, some conclusions are drawn.

2 Robust Modifications of the Tukey Boxplot
The Tukey univariate boxplot [5] is specified by five parameters: the two extremes, the upper $UQ$ (75th percentile) and lower $LQ$ (25th percentile) quartiles and the median (50th percentile). The lower and upper extremes of a boxplot are defined as

$$x_L = \max\left\{x_{(1)}, LQ - \frac{3}{2} IQR\right\}, \quad x_U = \min\left\{x_{(n)}, UQ + \frac{3}{2} IQR\right\}. \quad (1)$$
Different streams of data are compared via their respective boxplots in a quick and convenient way. It is a common practice to identify outliers by those points which are located beyond the extremes (maximum and minimum) and mark them within the corresponding boxplots.

Although the Tukey boxplot is a widely used tool for anomaly detection, it can be modified for better performance. For estimating the width of the central part of a data distribution, (the box part of the boxplot), the sample interquartile range (IQR) can hardly be improved, since it is a natural choice for representation of the half of the data distribution mass.

The remaining possibilities of improving most refer to the choice of robust estimates of scale used for visualization of tail areas and anomalies in the data (the boxplot lower and upper extremes). In this case, the sample interquartile range \( \text{IQR} \) as a robust estimate of scale is not the best choice as its efficiency and robustness can be considerably improved [2].

Since the interquartile range is less resistant to outliers than the median absolute deviation \( \text{MAD}_n \), a more robust rule for constructing the boxplot extremes can be given by

\[
x_L = \max \{x(1), LQ - k_{\text{MAD}} \text{MAD}_n\}, \quad x_U = \min \{x(n), UQ + k_{\text{MAD}} \text{MAD}_n\},
\]

where \( k_{\text{MAD}} \) is a threshold coefficient chosen from additional considerations.

Although the median absolute deviation \( \text{MAD}_n \) is a highly robust estimate of scale with the maximal value of the breakdown point \( \varepsilon^* = 0.5 \), its efficiency is only 0.37 at the normal distribution. In [3], a highly efficient robust estimate of scale \( Q_n \) has been proposed: it is close to the lower quartile of the absolute pairwise differences \( |x_i - x_j| \), and it has the maximal breakdown point 0.5 as for \( \text{MAD}_n \) but much higher efficiency 0.82. The drawback of this estimate is its low computation speed; the time complexity of \( Q_n \) is of a greater order than of \( \text{MAD}_n \).

In [4], an \( M \)-estimate of scale denoted by \( FQ_n \) whose influence function is approximately equal to the influence function of the estimate \( Q_n \) is proposed

\[
FQ_n = 1.483 \text{MAD}_n \left( 1 - \frac{Z_0 - n/\sqrt{2}}{Z_0} \right),
\]

where

\[
Z_k = \sum_{i=1}^n u_i^k e^{-u_i^2/2}, \quad u_i = \frac{x_i - \text{med} x}{1.483 \text{MAD}_n}, \quad k = 0, 2; \quad i = 1, \ldots, n.
\]

The efficiency and breakdown point of \( FQ_n \) are equal to 0.81 and to 0.5, respectively.

Based on the highly efficient robust estimate \( FQ_n \) of scale, we propose a new rule for the boxplot extremes defined as

\[
x_L = \max \{x(1), LQ - k_{FQ} FQ_n\}, \quad x_U = \min \{x(n), UQ + k_{FQ} FQ_n\}.
\]

### 3 Performance Evaluation

The proposed robust boxplots as alternatives to the Tukey boxplot, differ in estimating tail areas and consequently in detecting outliers. Therefore, we undertake a comparison
study involving the robust and Tukey versions relative to detection of outliers.

In statistics, an outlier is an observation that is numerically distant from the rest of the data. A frequent cause of outliers is a mixture of two distributions, namely, a combination of "good data" and "bad data".

Within the classical approach to detection of outliers, an observation $x$ is taken as an outlier if $|x - \overline{x}|/S > k_\alpha$, where $\overline{x}$ is the sample mean, $S$ is the standard deviation, and the threshold $k_\alpha$ is determined from the given false alarm rate at the normal distribution. This rule is the classical Grubbs test [1].

In this paper, we most consider the boxplot (BP) detection tests of the form: an observation $x$ is regarded as an outlier if $x < x_L$ or $x > x_U$, where $x_L$ and $x_U$ are the lower and upper extremes, respectively. In this setting, these thresholds also depend on a free parameter $k$, which is chosen from the false alarm rate $\alpha = 0.1$.

The Monte Carlo experiments are conducted by generating 300 samples of observations from the mixture of normal distributions (Tukey’s model of gross errors)

$$f(x) = (1 - \varepsilon)N(x; 0, 1) + \varepsilon N(x; \mu, s),$$

where $0 \leq \varepsilon < 1$ is the probability of outliers in the data and $s > 1$ is their scale.

For evaluating the performance of different tests, the sensitivity (SE) and specificity (SP) measures are used in the comparative study. Note that the sensitivity is nothing but the test power, and the specificity is just unit minus the false alarm probability. These two metrics are combined into a single measure, namely, the harmonic mean between SE and SP: $H$-mean $= 2 \text{SE} \text{SP}/(\text{SE} + \text{SP})$. The introduced $H$-mean is an analog to the widely used in IR studies $F$-measure, which is the harmonic mean between the recall (R) and the precision (P): $F = 2RP/(R+P)$. The $H$-mean can be naturally used for performance evaluation in detection of outliers, since tests with different values of the false alarm probability can be effectively compared. In our study, the false alarm rates for the Tukey and modified boxplots are $\alpha = 0.06$ and $\alpha = 0.1$, respectively.

The results of Monte Carlo experiment are given in Tables 1-2 with the best performing statistics represented in boldface.

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>1000</th>
<th>10000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tukey BP</td>
<td>0.64</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>$MAD$-BP</td>
<td>0.67</td>
<td>0.72</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>$FQ$-BP</td>
<td>0.66</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>Grubbs test</td>
<td>0.17</td>
<td>0.29</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
</tr>
</tbody>
</table>

From Table 1 it follows that under scale contamination, the performances of boxplot tests, generally, are close to each other, and all of them outperform the classical Grubbs test, which is catastrophically bad. This effect can be explained by non-robustness of the Grubbs test forming statistics, the sample mean and standard deviation, under contamination.
Further, the robust MAD and FQ versions are slightly but systematically better than the Tukey boxplot test. Similar results are also obtained for the gross error models with shift contamination.

In Table 2, it is observed that with small and moderate levels of shift contamination, the FQ-boxplot is marginally better than its competitors. For larger fractions of contamination ($\varepsilon \geq 0.3$), the MAD-boxplot outperforms its competitors. It can be explained by the fact that the MAD is a minimax bias estimate of scale under the Tukey gross error model [2].

4 Conclusions

The two robust versions of the Tukey boxplot are proposed. Both versions aim at the symmetric distribution as their classical counterpart, the first MAD-BP being preferable under heavy contamination, while the second FQ-BP – under moderate contamination. The thresholds $k$ can be adjusted to the adopted level of the false alarm probability $\alpha$: we recommend the values $k_{MAD} = 1.44$ and $k_{FQ} = 0.97$ corresponding to the rate $\alpha = 0.1$ under normality. All the boxplot tests considerably outperform the classical Grubbs test, which is catastrophically bad under contamination.

References


ERROR PROBABILITIES FOR SEQUENTIAL TESTING OF SIMPLE HYPOTHESES UNDER FUNCTIONAL DISTORTIONS IN THE $L_2$-METRIC

Belarusian State University
Minsk, BELARUS
e-mail: chernovsky@tut.by

Abstract

The problem of error probabilities evaluation for the sequential probability ratio test under functional distortions in the $L_2$-metric is analyzed. The asymptotic expansions for the error probabilities are constructed.

Keywords: Sequential Test, Simple Hypotheses, Outlier, $L_2$-metric.

1 Introduction

The sequential method of hypotheses testing is widely used for data analysis in medicine, statistical quality control, biology and finance. A profit of sequential procedures is that the average number of observations is less than for the equivalent tests procedures based on a fixed number of observations [1]. The sequential probability ratio test (SPRT) proposed by A. Wald [1] is considered in the paper because it is used frequently for practical purposes [3].

The SPRT can be applied effectively if the observed data satisfy the theoretical model. One of the essential sequential approach disadvantages is that the probabilities of types I and II are influenced greatly by the distortions in observations. In [6] we develop the approach proposed in [5] for robustness analysis when observed data is contaminated by outliers of Tukey and Huber type [2]. In [7] functional distortions in the $L_2$-metric were already considered, however in this paper the family of observation distributions is extended and new asymptotic expansions and more precise robustness analysis for error probabilities under this type of distortions are provided.

2 Mathematical Model

Let random observations $x_1, x_2, \ldots, x_k \in \mathbb{D} \subseteq \mathbb{R}$, be independent and identically distributed on a measurable space $(\Omega, \mathcal{F})$ according to a probability density function $f(x, \theta)$ with a parameter $\theta \in \Theta = \{\theta_0, \theta_1\}$. The true value of $\theta$ is unknown. Let the cumulative probability distribution function $F(x, \theta)$ corresponds to $f(x, \theta)$.

There are two simple hypotheses on the value of the parameter: $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$.

Denote the accumulated likelihood ratio statistic:

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{k=0}^{n} \lambda_k, \quad \lambda_k = \lambda(x_k) = \ln \frac{f(x_k, \theta_1)}{f(x_k, \theta_0)},$$

(1)
where \( \lambda_k \) is the likelihood ratio statistic calculated with the observation \( x_k \). Let \( \lambda_0 = 0 \).

To test hypotheses \( H_0 \) and \( H_1 \) by \( n = 1, 2, \ldots \) observations sequentially, the SPRT is used:

\[
N = \min\{n \in \mathbb{N} : \Lambda_n \notin (C_-, C_+)\},
\]

\[
d = 1_{[C_+, +\infty]}(\Lambda_N),
\]

where \( N \) is the stopping time, at which the decision \( d \) is made according to (3). In (2) the thresholds \( C_- \) and \( C_+ \) are the parameters of the test defined according to [1]:

\[ C_- = \ln \frac{\beta_0}{1-\alpha_0}, \quad C_+ = \ln \frac{1-\beta_0}{\alpha_0}, \]

where \( \alpha_0 \) and \( \beta_0 \) are given maximal possible values of probabilities of types I and II errors respectively. It is known [1] that \( \alpha_0 \) and \( \beta_0 \) are the approximate values of the actual error probabilities of types I and II only.

Let us make the following assumption:

A1) the function \( \lambda(x) \), \( \lambda : \mathbb{D} \to \mathbb{R} \), defined by (1), is measurable w.r.t. the Borel \( \sigma \)-algebra \( \mathcal{B}(\mathbb{R}) \), \( \mathbb{D} \in \mathcal{B}(\mathbb{R}) \).

Let \( \lambda^{-1}(A) = \{x \in \mathbb{D} : \lambda(x) \in A\} \) be the pre-image of set \( A \), \( |A| \) be the Lebesgue measure of set \( A \), \( A \in \mathcal{B}(\mathbb{R}) \). Denote \( \lambda^{-1}(a, b) = \lambda^{-1}((a, b)) \) and \( \lambda^{-1}(a) = \lambda^{-1}(\{a\}) \), \( a, b \in \mathbb{D}, a < b; h_k = |\lambda^{-1}(kh - h, kh)|, k = -m + 1, m \).

Without loss of generality, we suppose that the hypothesis \( H_0 \) is true, so the value of the probability of type I error is considered. Denote \( F(x) = F(x, \theta_0), f(x) = f(x, \theta_0) \).

To approximate the error probability \( \alpha \) of the sequential probability ratio test (2), (3), the approach mentioned in [4] and developed in [8] is applied.

Denote \( h = (C_+ - C_-)/m \), where \( m \in \mathbb{N} \) is the parameter of a fragmentation (approximation). Introduce the random sequences

\[
\Lambda_n^- = \sum_{k=0}^n \lambda_k^-, \quad \Lambda_n^+ = \sum_{k=0}^n \lambda_k^+, \quad \lambda_0^- = C_- + \left[\frac{\lambda_0 - C_-}{h}\right], \quad \lambda_k^- = \left[\frac{\lambda_k}{h}\right] h, \quad \lambda_k^+ = \lambda_k^- + h, \quad k \in \mathbb{N}.
\]

Construct absorbing Markov chains \( L_n^- \) and \( L_n^+ \) with the states \( \{0, 1, \ldots, m, m + 1\} \) where 0 and \( m + 1 \) are absorbing ones:

\[
L_n^- = \begin{cases} 0, & \text{if } \Lambda_n^- \leq C_-, h, \\ i, & \text{if } \Lambda_n^- = C_- + (i - 1)h, \\ m + 1, & \text{if } \Lambda_n^+ \geq C_+,
\end{cases} \quad L_n^+ = \begin{cases} 0, & \text{if } \Lambda_n^+ \leq C_-, \\ i, & \text{if } \Lambda_n^+ = C_- + i h, \\ m + 1, & \text{if } \Lambda_n^+ \geq C_+ + h.
\end{cases}
\]

**Theorem 1.** [8] If assumption A1 is fulfilled for the considered model, then the initial probabilities and the one-step transition probabilities of the Markov chains \( L_n^- \) and \( L_n^+ \) are

\[
\pi_i = \delta_{[-C_-/h] + 1, i}, \quad i, j = 1, m, \\
p_{i,j}^- = \int_{x^{-1}((j - i)h, (j - i + 1)h)} f(y)dy, \quad p_{i,j}^+ = \int_{x^{-1}(j - i - 1)h, (j - i)h)} f(y)dy, \\
p_{i,0}^- = \int_{x^{-1}(-\infty, (1 - i)h)} f(y)dy, \quad p_{i,0}^+ = \int_{x^{-1}(-\infty, -1)h} f(y)dy, \\
p_{i,m+1}^- = 1 - \int_{x^{-1}(-\infty, (m - i + 1)h)} f(y)dy, \quad p_{i,m+1}^+ = 1 - \int_{x^{-1}(-\infty, (m - i)h)} f(y)dy.
\]

Let the probabilities \( p_{i,j}^\pm \) be the elements of matrices \( P^\pm \in \mathbb{R}^{m \times m} \), \( R^\pm \in \mathbb{R}^{m \times 2} \), and \( S^\pm = I - P^\pm \), where \( I \) is the identity matrix of correspondent size. According to
the introduced notation, the probabilities of absorption of Markov chains $L^{-}_{m}$ and $L^{+}_{m}$ in state $m + 1$ are equal to \( \alpha^{\pm}_{m} = \pi(I - P^{\pm})^{-1}R^{\pm}_{2} = \pi(S^{\pm})^{-1}R^{\pm}_{2} \). It is known that \( \alpha^{-}_{m} \leq \alpha \leq \alpha^{+}_{m} \) and \( \alpha^{+}_{m} - \alpha^{-}_{m} \rightarrow 0 \), if \( m \rightarrow +\infty \), so instead of an unknown value of \( \alpha \) one can use \( \hat{\alpha}_{m} = (\alpha^{+}_{m} + \alpha^{-}_{m})/2 \).

### 3 Mathematical Model Distorted by Outliers

Let the observations \( x_{n}, n \in \mathbb{N} \), in the SPRT (2), (3) have a probability density function \( s(x) \) that is unknown and may differ from \( f(x) \) – the theoretical one, but the distance between \( s(x) \) and \( f(x) \) in the \( L_{2} \)-metric is supposed to be not more than \( \varepsilon \):

\[
\left( \int_{\mathbb{R}} |s(x) - f(x)|^{2}dx \right)^{1/2} \leq \varepsilon, \tag{7}
\]

where \( 0 \leq \varepsilon \leq \varepsilon_{0} \) and \( \varepsilon_{0} \) is fixed in advance.

Denote by \( L_{2}(f, \varepsilon) \) the family of probability distributions \( s(x) \), satisfying (7) for the fixed value of \( \varepsilon \). Let the cumulative distribution function \( S(x) \) corresponds to \( s(x) \).

Denote by \( L^{-}_{m}(s) \) and \( L^{+}_{m}(s) \) Markov chains \( L^{-}_{m} \) and \( L^{+}_{m} \), if observations have the probability density function \( s(\cdot), s(\cdot) \in L_{2}(f, \varepsilon) \). Let \( \pi(s), P^{\pm}(s), R^{\pm}(s) \) be the vector of initial probabilities and matrices of one-step transition probabilities of Markov chains \( L^{-}_{m}(s) \) and \( L^{+}_{m}(s) \), \( \alpha^{-}(s, \varepsilon) \) and \( \alpha^{+}(s, \varepsilon) \) – their absorption probabilities in state \( m + 1 \).

Let us investigate how much the value of type I error probability \( \alpha \) is influenced by the introduced distortion in the \( L_{2} \)-metric. Construct the least favorable distribution that maximizes \( \alpha \) and belongs to \( L_{2}(f, \varepsilon) \). The following optimization problem should be solved. Since value of \( \alpha \) is approximated by \( \alpha^{-}_{m} \) and \( \alpha^{+}_{m} \) (see the previous section), \( \alpha^{+}_{m} \) is considered instead unknown value of \( \alpha \):

\[
\begin{cases}
\alpha^{+}(s, \varepsilon) \rightarrow \max, \\
\int_{\mathbb{D}} |s(x) - f(x)|^{2}dx \leq \varepsilon^{2}, \\
\int_{\mathbb{D}} s(x)dx = 1. \tag{8}
\end{cases}
\]

The probability of absorption \( \alpha^{+}(s, \varepsilon) \) is defined by the distribution of random variable \( \lambda^{\pm}_{k} \), i.e. by the set \( \{p^{\pm}_{k}(s)\} \), \( k = -m + 1, m \), the set of elements of matrices \( \pi(s), P^{\pm}(s), R^{\pm}(s) \). Since for the fixed \( m \) (and \( h \) \( p^{\pm}_{k}(s) = P\{\lambda^{\pm}_{k} = kh\} = \int_{\lambda^{-1}(kh-h, kh)} s(y)dy \), i.e. the probability \( p^{\pm}_{k}(s) \) depends only on the probabilistic weight, concentrated on the \( \lambda^{-1}(kh-h, kh) \), the difference \( s(x) - f(x) \) can be represented as a constant \( a_{k} \) on every set \( \lambda^{-1}(kh-h, kh) \):

\[
s(x) = f(x) + \varepsilon \cdot \sum_{k=-m+1}^{m} a_{k} 1_{\lambda^{-1}(kh-h, kh)}(x),
\]

therefore \( P^{+}(s) = P^{+} + A\varepsilon, R^{+}(s) = R^{+} + B\varepsilon \), where \( A \) and \( B \) are matrices of the appropriate size consisting of elements \( a_{k}h_{k} \). As \( p^{+}_{k}(s) = \int_{\lambda^{-1}(kh-h, kh)} s(y)dy = \int_{\lambda^{-1}(kh-h, kh)} s(y)dy \)
\[ \int_{\lambda^{-1}(kh_{-1},kh_{+1})} f(y)dy + \varepsilon \cdot \int_{\lambda^{-1}(kh_{-1},kh_{+1})} a_k dy = p_k^+ (f) + \varepsilon a_k h_k, \] 
then the probability density function \( s(x) \) is defined by the vector \( a = (a_{-m+1}, a_{-m+2}, \ldots, a_m) \). Let \( \alpha^+(s, \varepsilon) = \alpha^+(a, \varepsilon) \), therefore the problem (8) is replaced by the problem

\[
\begin{cases}
\alpha^+(a, \varepsilon) \rightarrow \max, \\
a_{-m+1}^2 h_{-m+1} + \ldots + a_m^2 h_m \leq 1,
\end{cases}
\]

(9)

Since it is rather difficult to obtain the analytic solution of problem (9), \( \alpha(s, \varepsilon) \) is expanded into the series in terms of powers of \( \varepsilon \):

\[ \alpha^+(a, \varepsilon) = \alpha_m^+ + c_1^+(a)\varepsilon + O(\varepsilon^2), \]

where \( c_1(a) = \pi(S^+)^{-1}A(S^+)^{-1}R^+_2 + \pi(S^+)^{-1}B \). Then the value \( \alpha^+(a, \varepsilon) \) in the optimization problem (9) is replaced by the dominant term \( c_1^+(a) \) of this expansion:

\[
\begin{cases}
c_1^+(a) \rightarrow \max, \\
a_{-m+1}^2 h_{-m+1} + \ldots + a_m^2 h_m \leq 1,
\end{cases}
\]

(10)

**Theorem 2.** The optimization problem (10) has the following solution:

\[
\max\{c_1^+(a)\} = \sqrt{|D| \cdot D\{\xi\}},
\]

\[ a_k = (c_k - E\{\xi\}) / \sqrt{|D| \cdot D\{\xi\}}, \]

where \( \xi \) is a discrete random variable, \( P\{\xi = c_k\} = h_k / |D| \), where \( c_k \) can be calculated using matrices \( \pi, P^+ \) and \( R^+ \).

**References**


SEMIPARAMETRIC ESTIMATION IN MIXTURE MODELS WITH VARYING MIXING PROBABILITIES

A.V. DORONIN
Kyiv National Taras Shevchenko University
Kyiv, Ukraine
e-mail: al_doronin@ukr.net

Abstract
We consider a model of mixture with concentrations varying from observation to observation. Semiparametric estimation problems are considered for this model. We present three types of estimators, namely the moment, quantile and adaptive ones. Their performance is compared both analytically and by simulations.

1 Introduction
We consider the series of samples from $N$ observations $\{\xi_{1:N}, \ldots, \xi_{N:N}\}_{N \geq 1}$. Each observation $\xi_{j:N}$ can have a CDF $F_m(\cdot)$, $m = 1, M$ with some predefined probability (concentration) $p_{j:N}$. Thus, the CDF of $\xi_{j:N}$ is $P[\xi_{j:N} \in A] = \sum_{m=1}^{M} p_{j:N}^m F_m(A)$, $A \subset \mathbb{R}$.

In what follows we assume that the CDF of the first component is parametrized with some Euclidean parameter $t \in \Theta \subset \mathbb{R}^d$ (i.e. $F_1(A) = F_1(A, t)$). The true value of parameter we denote as $\vartheta \in \Theta$. The CDFs of the rest of the components are assumed to be fully unknown.

Moment, quantile and adaptive estimators for $\vartheta$ are discussed in Sections 2-4. Performance of these estimators is assessed via simulations in Section 5.

2 Moment estimators
Weighted empirical CDF $\hat{F}_{m:N}(x)$ with minimax weight coefficients $a_{j:N}^m$ is defined in [4] as $\hat{F}_{m:N}(x) := \frac{1}{N} \sum_{j=1}^{N} a_{j:N}^m 1_{\{\xi_{j:N} \leq x\}}$.

Improved weighted empirical CDF is introduced in [3] as $\hat{F}_{m:N}^+(x) := \min \{1, \sup_{y \leq x} \hat{F}_{m:N}(y)\}$.

Consider some measurable function $h : \mathbb{R} \to \mathbb{R}^d$.

We define moment estimator as a solution of moment equation

$$\hat{\vartheta}_N^{simple} := \arg_{t \in \Theta} \left\{ \int h(x) \hat{F}_{1:N}(dx) = \int h(x) F_1(dx, t) \right\}.$$  \hspace{1cm} (1)

Alternatively, we define improved moment estimator

$$\hat{\vartheta}_N^{impr} := \arg_{t \in \Theta} \left\{ \int h(x) \hat{F}_{1:N}^+(dx) = \int h(x) F_1(dx, t) \right\}.$$  \hspace{1cm} (2)

Consistency and asymptotic normality are demonstrated for both $\hat{\vartheta}_N^{simple}$ and $\hat{\vartheta}_N^{impr}$ in [4] and [3].
3 Quantile estimators

Estimators for quantiles

It is proposed in [4] to define an estimator for a quantile \( \hat{Q}_{m,N}(\alpha) \) of level \( \alpha \) as a value of a function, inverted to piece-wise linear interpolation of improved CDF \( \hat{F}^+_m_N \) defined in section 2. Consistency and asymptotic normality of this estimate are demonstrated in [4].

Asymptotic normality of a vector \((\hat{Q}^m_N(\alpha_i))_{i=1}^q\) is shown in [2], and it is found its dispersion matrix.

Theorem 1. (Theorem 1 from [2])

Assume that

1. \( \sup_{j:N} |a^m_{j,N}| < \infty \).
2. The limits \( \langle p^k p^l a^m \rangle, \langle p^k a^m \rangle \) exist for all \( k, l = 1, M \).
3. \( F_k(\cdot) \) are continuous on \( \mathbb{R} \), \( k = 1, M \).
4. The unbiasedness condition \( \langle a^m p^k \rangle = \mathbb{I}_{k=m} \) holds.
5. Functions \( F_k(\cdot) \), \( k = 1, M \) are monotone increasing in some neighborhoods \( I_1, \ldots, I_q \) of points \( Q^F_m(\alpha_i), \ldots, Q^F_m(\alpha_q) \) respectively.
6. Function \( F_m(\cdot) \) has a continuous derivative \( f_m(\cdot) \) on \( I_1, \ldots, I_q \), and \( f_m(Q^F_m(\alpha_i)) \neq 0 \), \( i = 1, q \).

Then the vector \( (\sqrt{N}(\hat{Q}^m_N(\alpha_i) - Q^F_m(\alpha_i)))_{i=1}^q \) weakly converges as \( N \to \infty \) to a Gaussian random vector with zero mean and covariance matrix \( S = (S_{r,s})_{r,s=1}^q \) with elements

\[
S_{r,s} = \frac{1}{f_m(Q^F_m(\alpha_r))f_m(Q^F_m(\alpha_s))} \times \\
\times \left( \sum_{k=1}^M \langle p^k a^m \rangle^2 F_k(\min\{Q^F_m(\alpha_r), Q^F_m(\alpha_s)\}) - \\
- \sum_{k,l=1}^M \langle p^k p^l a^m \rangle F_k(Q^F_m(\alpha_r)) F_l(Q^F_m(\alpha_s)) \right). \tag{3}
\]

Quantile estimator (for Gaussian distribution)

Let \( F_{\vartheta}(x; \theta) \) be the CDF of a Gaussian distribution with the true value of a parameter \( \vartheta = (\mu, \sigma)^T \in \mathbb{R}^2 \).

We denote the interquartile range of standard Gaussian distribution as \( \gamma := Q^{\mathcal{N}(0,1)}(3/4) - Q^{\mathcal{N}(0,1)}(1/4) \) (approximately 1.34898), and \( \vartheta := \begin{pmatrix} 0 & 1 & 0 \\ -1/\gamma & 0 & 1/\gamma \end{pmatrix} \). Quantile estimator for Gaussian component is defined in [2] as

\[
\hat{\vartheta}^{\text{quant}}_N := (\hat{\mu}^{\text{quant}}_N, \hat{\sigma}^{\text{quant}}_N)^T := \vartheta \cdot (\hat{Q}^1_N(1/4), \hat{Q}^1_N(1/2), \hat{Q}^1_N(3/4))^T. \tag{4}
\]

Corollary 1. (from theorem 1)

Let \((\alpha_i)_{i=1}^q = (1/4, 1/2, 3/4)^T\). Assume that the assumptions 1-6 from theorem 1 hold.

Then the vector \( \sqrt{N}(\hat{\vartheta}^{\text{quant}}_N - \vartheta) \) weakly converges to Gaussian distribution with zero mean and covariance matrix \( E_\vartheta \cdot S \cdot E_\vartheta^\top \) with \( S \) defined by (3).
4 Adaptive estimator (from GEE method)

Adaptive estimators are constructed in [1] as GEE estimators with the estimating function adapted by data to derive optimal dispersion matrices. For practical needs, it is recommended in [1] to consider a vector of some predefined parametrized functions $u(x; t) \in \mathbb{R}^d$, and choose the estimating function as a linear combination of $u(x; t)$: $B(t) \cdot u(x; t)$, where $B(t)$ is some $d$-by-$R$ matrix. Approximate adaptive estimators are obtained from pilot estimators as one-step Newton type approximate solutions of adapted estimating equations. Any $\sqrt{N}$-consistent estimator such as a moment or a quantile one can be used as the pilot estimator $\tilde{\vartheta}_N$. Thus, adaptive estimator takes form

$$\hat{\vartheta}_N^{\text{adapt}} := \tilde{\vartheta}_N - \hat{B}_N(\tilde{\vartheta}_N) \cdot \hat{u}_{1:N}(\tilde{\vartheta}_N)$$

(5)

where $\hat{B}_N(\tilde{\vartheta}_N)$ and $\hat{u}_{1:N}(\tilde{\vartheta}_N)$ are estimations for $B(\vartheta)$ and $\int u(x; \vartheta)F_1(dx; \vartheta)$ respectively.

Consistency and asymptotic normality of the adaptive estimator defined by (5) are demonstrated in [1].

5 Numerical examples

We assessed performance of the following estimators by simulations.

1. Simple estimate $\hat{\vartheta}_N^{\text{simple}}$ defined by (1) with $h(x) := (x, x^2)^T$.

2. Improved estimate $\hat{\vartheta}_N^{\text{impr}}$ defined by (2) with $h(x) := (x, x^2)^T$.

3. Quantile estimate $\hat{\vartheta}_N^{\text{quant}}$ defined by (4).

4. Adaptive estimate $\hat{\vartheta}_N^{\text{adapt}}$ defined by (5) with $\hat{\vartheta}_N^{\text{impr}}$ as a pilot.

5. Adaptive estimate $\hat{\vartheta}_N^{\text{adapt}}$ defined by (5) with $\hat{\vartheta}_N^{\text{quant}}$ as a pilot.

Experiments were conducted on two types of two-component mixture from Gaussian distributions with the following parameters:

**Experiment 1.** Component 1: $\mu = -3$, $\sigma = 1$; component 2: $\mu = 3$, $\sigma = 2$.

**Experiment 2.** Component 1: $\mu = 0$, $\sigma = 2$; component 2: $\mu = 1$, $\sigma = 2$.

The estimates were calculated for different sizes of a sample (value $N$): 50, 100, 250, 500, 750, 1000, 2000, 5000. The dispersion of constructed estimates was calculated from 1000 samples (for each value of $N$). The set of concentration was uniform: $p_{1:N}^1 = \{\frac{j}{N}\}_{j=1}^N$, $p_{1:N}^2 = 1 - p_{1:N}^1$.

For adaptive estimate as a vector $u(x; t)$ is taken a vector from 8 functions. First 5 of them are cubic B-splines with support $(t_1 - 4t_2, t_1 + 4t_2)$ and uniform subdivision of this support into 8 intervals. The last 3 functions: $1$, $(x - t_1)/t_2$, $(x - t_1)^2/t_2^2$.

The results of simulation are presented on figure 1.

So, in our experiments the adaptive estimators outperformed the other ones in almost all cases for sample sizes larger then 100.
Figure 1: Dispersion of estimates: □ – simple estimates, ■ – improved estimates, △ – quantile estimates, • and ◦ – adaptive estimates with improved and quantile as pilot ones respectively. Asymptotic values are presented by dotted lines.

References


APPLICATION OF RESIDUAL EMPIRICAL PROCESSES TO ROBUST LINEAR HYPOTHESES TESTING IN AUTOREGRESSION

D.M. ESAULOV
Lomonosov Moscow State University
Moscow, Russia
e-mail: daniel.yesaulov@gmail.com

Abstract
The paper deals with the property of asymptotic uniform linearity of residual empirical processes for AR(p) when observations contain outliers. We apply the result to construct robust GM–tests for linear hypotheses. The scheme of data contamination by additive single outliers with the intensity $O(n^{-1/2})$, $n$ is data level, is considered.

1 Introduction. Formulation of the problem

In this paper we construct nonparametric robust generalized M-test (GM-test) for linear hypotheses in autoregressive AR(p) model. Consider the model

$$u_t = \beta_1 u_{t-1} + \cdots + \beta_p u_{t-p} + \varepsilon_t, \quad t \in \mathbb{Z}. \quad (1.1)$$

Here $\{\varepsilon_t\}$ — i.i.d. random variables with unknown distribution function $G$ and density $g$, $E\varepsilon_1 = 0$, $E\varepsilon_1^2 < \infty$; $\beta = (\beta_1, \ldots, \beta_p)^T$ — vector of unknown parameters. We suppose the coefficients $\beta_1, \ldots, \beta_p$ are such that the roots of the characteristic equation

$$x^p = \beta_1 x^{p-1} + \cdots + \beta_p$$

are less than one in absolute value. This condition guarantees that (1.1) has an a.s. unique stationary solution. Let the observations contain outliers and be of the following form

$$y_t = u_t + z_t^n \xi_t, \quad t = 0, 1, \ldots, n. \quad (1.2)$$

In (1.2) $\{u_t\}$ is the sample from (1.1); $\{z_t^n\}$ — i.i.d. random variables with Bernoulli distribution $Br(\gamma_n)$, $\gamma_n = \min(1, n^{-1/2})$, parameter $\gamma \geq 0$ is unknown; $\{\xi_t\}$ — i.i.d. random variables with an unknown distribution $\mu$; the sequences $\{u_t\}$, $\{z_t^n\}$, $\{\xi_t\}$ are independent to each other. Sequence $\{\xi_t\}$ is interpreted as a sequence of outliers (contamination), $\gamma_n$ is a contamination level. Scheme (1.2) is the local variant of contamination scheme from [6]. Let’s represent $\beta$ as $\beta^T = (\beta^{(1)}^T, \beta^{(2)}^T)$, where the vectors $\beta^{(i)}$, $i = 1, 2$, have the dimensions $m$ and $p - m$ respectively, $1 \leq m < p$. The linear hypothesis has the form

$$H_0: \beta^{(2)} = \beta^{(2)}_0.$$  

Here $\beta^{(2)}_0$ is a known vector and $\beta^{(1)}$ is an interfering parameter. Introduce the local alternatives $H_{1n}(\tau)$: $\beta = \beta_n := \beta_0 + n^{-1/2} \tau$, where $\beta_0^T = (\beta^{(1)}_0, \beta^{(2)}_0)^T$, $\tau^T = (\tau^{(1)}^T, \tau^{(2)}^T) \in \mathbb{R}^p$ is a constant vector with subvectors of dimensions $m$ and $p - m$ respectively. Thus, the unknown parameter $\beta^{(1)}$ is admitted to variate with a magnitude of order $O(n^{-1/2})$ in the alternatives $H_{1n}(\tau)$.  

153
In order to verify $H_0$ in case of absence of outliers least squares tests, rank, sign, and M-tests are often used. One of the general aims of this paper is to construct GM-test which would be robust against outliers $\{\xi_t\}$ using $\{y_t\}$.

Usually GM-tests are constructed from GM-estimators. Let’s describe this method. Consider model (1.1) without outliers. For some functions $\varphi$, $\psi$ and arbitrary $\alpha \in \mathbb{R}^p$ introduce vector

$$L_n(\alpha) := (L_{n1}(\alpha), L_{n2}(\alpha), \ldots, L_{np}(\alpha))^T,$$

$$L_{nj}(\alpha) := n^{-1/2} \sum_{t=1}^n \varphi(u_{t-j})\psi(u_t - \alpha_1 u_{t-1} - \cdots - \alpha_p u_{t-p}).$$

GM-estimator is defined as $\sqrt{n}$-consistent solution of the equation $L_n(\alpha) = 0$. Denote the solution as $\hat{\beta}_{n,GM}$. For $\varphi$ and $\psi$ satisfying definite conditions estimator $\hat{\beta}_{n,GM}$ is asymptotically Gaussian (see details in [3]). Using this property last $p - m$ coordinates of $\hat{\beta}_{n,GM}$ can be used for verifying $H_0$ in standard way.

Consider statistic $L_n(\alpha)$ which is constructed from $\{y_t\}$ in the same way as $L_n(\alpha)$ is constructed from $\{u_t\}$. In this paper tests are based not on GM-estimators but on statistic $L_n(\hat{\beta}_n)$ transformed in a special way. Here $\hat{\beta}_n$ is an arbitrary $\sqrt{n}$-consistent estimator of $\beta$. Note that for tests of this kind functions $\varphi$ and $\psi$ should satisfy weaker regularity conditions than for tests based on GM-estimators. In case of $\varphi(x) = x$ our test is asymptotically equivalent to M-test constructed in [5].

Asymptotic properties of constructed test are considered in the alternatives $H_{1n}(\tau)$ using representation of $L_n(\alpha)$ as an integral functional of residual weighted empirical process. We obtain asymptotic uniform linearity (AUL) of these processes and hence $L_n(\alpha)$. In general the property of AUL of special-type residual processes has been widely used for parameter estimation and GM–estimation particularly in different models without contamination. For example nonlinear models with additive noises were considered in [4], ARCH models — in [2].

Denote the power of constructed test in alternatives $H_{1n}(\tau)$ as $W_n(\tau, \gamma, \mu)$. In this work we obtain its limit power $W(\tau, \gamma, \mu) := \lim_{n \to \infty} W_n(\tau, \gamma, \mu)$. Let $W(\tau) := W(\tau, 0, \mu)$ be the limit power of the test in scheme (1.1) without outliers. It turns out that for small values of $\gamma$ limit powers $W(\tau, \gamma, \mu)$ and $W(\tau)$ are closed uniformly with respect to $\mu \in \mathcal{M}_2$ (see statement (2.5)). Here $\mathcal{M}_2$ is a class of outliers with finite second moment. Expression (2.5) means that the test has qualitatively robust limit power. This definition of robustness against outliers had been used before. For example in [1] a sign test for linear hypotheses in AR(p) was constructed and corresponding relation of the type of (2.5) for this test was proved.

2 Main results

Consider the scheme of the type (1.2). Using observations $\{y_t\}$ construct the vector $Y_{t-1} := (y_{t-1}, y_{t-2}, \ldots, y_{t-p})^T$, $t = 1, \ldots, p$. For the sake of brevity denote vector function $\varphi(x_1, x_2, \ldots, x_p) := (\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_p))^T$. Define weighted residual empirical
process as $v_n(\alpha, x, \gamma, \mu) := n^{-1/2} \sum_{t=1}^{n} \varphi(Y_{t-1}) I(y_t - \alpha^T Y_{t-1} \leq x)$, $I(\cdot)$ is an indicator function. Then the integral functional is constructed as

$$L_n^Y(\alpha) := \int_{-\infty}^{\infty} \psi(x) dv_n(\alpha, x, \gamma, \mu). \tag{2.1}$$

Now define $\sigma$-algebras $\mathcal{F}_{t-1} := \sigma(\alpha, i \leq t - 1; (\xi_j, z_j^n), 0 \leq j \leq t)$, $t = 1, \ldots, n$. Denote $\eta(\beta, \gamma) := z^n t \xi_t - \beta_1 z^n t - 1 \xi_t - \beta_2 z^n t - 2 \xi_t - \cdots - \beta_p z^n t - p \xi_t - p$ and introduce empirical process which is conditionally centered with respect to $\mathcal{F}_{t-1}$

$$u_n(\theta, x, \gamma, \mu) := n^{-1/2} \sum_{t=1}^{n} \varphi(Y_{t-1}) \left[ I(\varepsilon_t \leq x + n^{-1/2} \theta^T Y_{t-1} - \eta(\beta, \gamma)) - G(x + n^{-1/2} \theta^T Y_{t-1} - \eta(\beta, \gamma)) \right]. \tag{2.2}$$

To formulate the property of AUL when $\gamma \neq 0$ we need following conditions to hold:

**Condition (i)**. $\sup_x |\varphi(x)| < \infty$.

**Condition (ii)**. $G$ is twice differentiable, $G' = g$, $\sup_x |g'(x)| < \infty$.

**Condition (iii)**. $E|\\delta_1|^2 < \infty$.

**Theorem 1.** Let the conditions (i)-(iii) hold. Let $H_{1m}(\tau)$ be valid. Then

$$\sup_{|\theta| \leq \Theta, x \in \mathbb{R}^1} |u_n(\theta, x, \gamma, \mu) - u_n(0, x, \gamma, \mu)| \overset{P}{\to} 0, \quad n \to \infty, \quad 0 \leq \Theta < \infty$$

Let $u_1^n, \ldots, u_n^n$ be a sample from the strictly stationary solution to equation (1.1) for $\beta = \beta_0$. Denote

$$\Delta(\mu) := (\Delta_1(\mu), \Delta_2(\mu), \ldots, \Delta_p(\mu))^T,$$

$$\Delta_j(\mu) := E \varphi(u_0^0 + \xi_2 - j) E \psi(\varepsilon_1 + \xi_1)$$

$$+ \sum_{i=1, i \neq j}^p E \varphi(u_0^0 + \xi_2 - j) E \psi(\varepsilon_2 - \beta_0 \xi_2 - i)$$

$\Delta(\mu)$ characterizes the outliers influence on $L_n^Y(\alpha)$ and test statistic.

Define the following condition for the function $\psi$.

**Condition (iv)**. Variation $\text{Var} |\infty_0 \varphi | < \infty, E \psi(\varepsilon_1) = 0, \int_{-\infty}^{\infty} g(x) d\psi(x) \neq 0$.

Using $\{u_0^0\}$ construct vectors $U_{t-1}^0 := (u_0^0, \ldots, u_{t-1}^0)^T$, $t = 1, \ldots, n$. Let $\tilde{L}_n(\beta_0) := n^{-1/2} \sum_{t=1}^{n} \varphi(U_{t-1}^0) \psi(\varepsilon_t)$. Define matrix

$$C := \int_{-\infty}^{\infty} g(x) d\psi(x) E \varphi(U_0^0)(U_0^0)^T.$$

Under theorem 1 and relation (2.1) we obtain

**Theorem 2.** Let the conditions (i)-(iv) hold. Let $\varphi(x)$ be continuous a.s. If $H_{1m}(\tau)$ is valid then for any $0 \leq \Theta < \infty$ the following uniform convergence holds

$$\sup_{|\theta| \leq \Theta} |L_n^Y(\beta + n^{-1/2} \theta) - \tilde{L}_n(\beta_0) + C \theta - \gamma \Delta(\mu)| \overset{P}{\to} 0, \quad n \to \infty. \tag{2.3}$$
Now proceed to robust GM-test construction. Denote \( \hat{\beta}_{n0}^T = (\hat{\beta}_{n0}^{(1)T}, \beta_0^{(2)T}) \), where \( \hat{\beta}_{n0}^{(1)} \) is an arbitrary \( \sqrt{n} \)-consistent estimator of interfering parameter \( \beta^{(1)} \). Denote consistent estimator (in \( H_{l_n}(\tau) \)) of matrix \( C \) as \( \hat{C}_n \). Let \( \pi \) be the orthogonal projection onto last \( (p - m) \) components. Let \( \det C \neq 0 \). Take
\[
\Lambda_{n,Y}^\pi := \left[ \pi \circ \hat{C}_n^{-1}L_n^Y(\hat{\beta}_{n0}) \right]^T J_n^{-1} \left[ \pi \circ \hat{C}_n^{-1}L_n^Y(\hat{\beta}_{n0}) \right]
\] (2.4)
as a statistic of the test. Here \( \hat{J}_n \) is an arbitrary consistent estimator of \( J \) which is a covariance matrix of \( \pi \circ \hat{C}_n^{-1}L_n^Y(\hat{\beta}_{n0}) \). Denote \( a(\gamma, \mu) := \gamma C^{-1} \Delta(\mu) \), \( a^T(\gamma, \mu) = (a^{(1)T}, a^{(2)T}) \). With the help of expansion (2.3) one can prove the following

**Theorem 3.** Let the conditions of the Theorem 2 hold. Then for \( n \rightarrow \infty \)
\[
\Lambda_{n,Y}^\pi \overset{d}{\rightarrow} \chi^2(p - m, \lambda^2), \quad \lambda^2 = |J^{-1/2}(\tau^{(2)} + a^{(2)})|^2
\]

In the alternatives \( H_{l_n}(\tau) \) the power of the test based on \( \Lambda_{n,Y}^\pi \) is \( W_n(\tau, \gamma, \mu) = P_{\beta_n}(\Lambda_{n,Y}^\pi > \chi_1^{p-m}) \), \( \chi_1^{p-m} \) is \((1 - \alpha)\)–quantile of the distribution of \( \chi^2(p - m) \). Due to the theorem 3 \( \lim_{n \rightarrow \infty} W_n(\tau, \gamma, \mu) = W(\tau, \gamma, \mu) = 1 - F_{\gamma}^{p-m}(\chi_1^{p-m}, \lambda^2) \), and the test has the asymptotic confidence level \( \alpha \). Let \( W_n(\tau) \) be the power of test in scheme (1.1) without outliers. Then due to the theorem 3 there exists \( \lim_{n \rightarrow \infty} W_n(\tau) = W(\tau) := W(\tau, 0, \mu) \).

The following theorem characterizes the robustness of GM–test (2.4) against outliers or, if to be more precise, the qualitative robustness of its limit power (see [1]).

**Theorem 4.** Let the conditions of the Theorem 2 hold, then
\[
\sup_{\mu \in \mathfrak{M}_2} |W(\tau, \gamma, \mu) - W(\tau)| \rightarrow 0, \quad \gamma \rightarrow 0.
\] (2.5)

Thus the family of limit powers \( \{W(\tau, \gamma, \mu)\} \) is equicontinuous in \( \gamma \) at \( \gamma = 0 \).

**References**


156
ON D- AND A-OPTIMAL DESIGNS OF EXPERIMENTS

V.P. KIRLITSA
Belarusian State University
Minsk, BELARUS
e-mail: Kirlitsa@bsu.by

Abstract

It is shown that for linear multiple model of heteroscedastic observations for which variances changing linearly $D$-optimal designs take place when all points of spectrum of these designs lay in vertices of unite cube and not only. The structure such $D$- and $A$-optimal designs for heteroscedastic observations is investigated.

1 Introduction

Consider the linear multiple model of heteroscedastic observations:

$$y_i = \theta_1 x_{1i} + \ldots + \theta_m x_{mi} + \varepsilon(x^{(i)}), i = 1, 2, \ldots, n, n \geq m,$$  \hspace{1cm} (1)

where $y_i$ are observed variables; $\theta_1, \ldots, \theta_m$ are unknown parameters; $x^{(i)} = (x_{1i}, \ldots, x_{mi})$ are $m$-vectors of controllable variables and components of these vectors taken from the interval $[-1, 1]$; $\varepsilon(x^{(i)})$ are uncorrelated random errors of observations with mean zero and limited variance

$$D(\varepsilon(x^{(i)})) = a_0 + a_1 x_{1i} + \ldots + a_m x_{mi} > 0,$$ \hspace{1cm} (2)

for each realization $x^{(i)}$ from $m$-dimensional unit cube and $a_0 > 0, |a_1| + \ldots + |a_m| < a_0$. In the article [1] it is proved that for homoscedastic observations (1) the $D$-optimal design is obtained by setting $x_{ij} = \pm 1$. In this paper we obtain the same result for heteroscedastic observations (1), (2). In the paper [2] it is investigated the structure of saturated $D$-optimal designs $\varepsilon_0^0$ for homoscedastic model of observations (1) when $n = 3, 4, 5, 6$. In our article it is shown that the set of designs $\varepsilon_3^0$ for homoscedastic observations (1) is much wider than the set which is described in [2]. The set of such designs $\varepsilon_3^0$ infinitely and has power of a continuum. The structure of $A$-optimal designs for a line of regress with square-law change of a variance of observations is investigated. These results generalize results received in [3] for linear change of variance of observations.

2 The model of heteroscedastic observations

We have the following theorem for heteroscedastic observations (1), (2).

Theorem 1. Exist exact $D$-optimal design $\varepsilon_n^0$ of experiments (1), (2) all spectrum points of it lay in the vertices of $m$-dimensional unit cube.
Consider that $\varepsilon_n^0$ is $D$-optimal design of experiments (1), (2) and that $|x_{ij}| < 1$. Make this point "floating", i.e. $x_{11} = x \in [-1, 1]$. Denote the design of experiment being obtain from $\varepsilon_n^0$ by changing $x_{11}$ as $\varepsilon_x$. Information matrix of design $\varepsilon_x$ can be written as $M(\varepsilon_x) = A(x) + B$, where

$$A(x) = \frac{1}{d_1(x)} \left( \begin{array}{c} x \\ x' \end{array} \right), \quad B = \sum_{i=2}^{n} \frac{x^{(i)}(x^{(i)})'}{d_i},$$

$d_1(x) = a_1x + c_1, c_1 = a_0 + a_21 + \ldots + a_mx_{m1}, d_i = a_0 + a_1x_{11} + \ldots + a_mx_{mi}, t' = (x_{21}, \ldots, x_{m1})'$. Matrix $B$ does not depend on $x$ and $\text{rank} A(x) = 1$. Matrixes $B$ and $A(x)$ can be written as

$$B = (b_1, \ldots, b_m), \quad A(x) = (a_1(x), \ldots, a_m(x)), $$

where $b_i$ and $a_i(x)$ are $m$-dimensional columns, $i = 1, 2, \ldots, m$. Determinant $\left| A(x) + B \right|$ may be presented as a sum of determinants of matrixes containing different combinations of columns matrixes $A(x)$ and $B$. Determinants of these combinations containing more than one column of $A(x)$ is equal to zero. Hence

$$\left| A(x) + B \right| = \left| a_1(x), b_2, b_3, \ldots, b_m \right| + \left| b_1, a_2(x), b_3, \ldots, b_m \right| + \ldots + \left| b_1, b_2, b_3, \ldots, a_m(x) \right|.$$  \hspace{1cm} (3)

Calculating determinants in the right part of (3) with respect elements of columns $a_1(x), a_2(x), \ldots, a_m(x)$ we obtain

$$f(x) = | A(x) + b | = \frac{\alpha x^2 + \beta x + \gamma}{a_1x + c_1} + m, \hspace{1cm} \text{(4)}$$

where

$$m = | B |, \quad \alpha = \sum_{i=2}^{n} \frac{x^{(i)}(x^{(i)})'}{d_i} \geq 0, \quad z^{(i)} = (x_{21}, \ldots, x_{m1})'$$

and $\beta, \gamma$ are constants which do not depend on $x$. If the function $f(x)$ do not depend on $x$ then we can set $x = \pm 1$ in the $D$-optimal design $\varepsilon_n^0$. Hence we obtain the result of theorem. If the function $f(x)$ depend on $x$ then $f(-1)$ or $f(1)$ take value more than $f(x_{11})$. It is impossible. We have contradiction. Indeed, the derivative of function $f(x)$ is

$$\frac{df(x)}{dx} = \frac{\alpha a_1x^2 + 2\alpha c_1x + c_1\beta - a_1\gamma}{(a_1x + c_1)^2}. \hspace{1cm} \text{(5)}$$

Let is designate though $D = 4\alpha(\alpha c_1^2 - a_1(c_1\beta - a_1\gamma))$ numerator discriminant in (5). If $D \leq 0$, the derivative (5) does not change his signum in the interval $[-1, 1]$, i.e. the function $f(x)$ is increasing or decreasing in $[-1, 1]$. If $D > 0$, than the function $f(x)$ is convex downwards. Indeed,

$$\frac{d^2f(x)}{dx^2} = \frac{D}{2\alpha(a_1x + c_1)^2} > 0.$$ 

Hence, the function $f(x)$ takes maximum value at $x = \pm 1$ and it value more than $f(x_{11})$. Theorem 1 is proved.

Theorem 1 says that it is possible that among $D$-optimal designs exist design in which not all points lay in the vertices of unite cube. It will be shown in following section that such possibility is realized for saturated design $\varepsilon_n^1$.
3 Saturated $D$-optimal designs

Information matrix for saturated design $\varepsilon_n$ (when $n = m$) for model heteroscedastic observations (1), (2) is

$$M = X_1'X_1, X_1 = \begin{pmatrix}
\frac{x_{11}}{\sqrt{d_1}} & \frac{x_{21}}{\sqrt{d_1}} & \ldots & \frac{x_{n1}}{\sqrt{d_1}} \\
\ldots & \ldots & \ldots & \ldots \\
\frac{x_{1n}}{\sqrt{d_n}} & \frac{x_{2n}}{\sqrt{d_n}} & \ldots & \frac{x_{nn}}{\sqrt{d_n}}
\end{pmatrix}$$

Determinant of matrix $M$ is equal to $|M| = (|X_1|)^2$. So, $D$-optimal saturated design $\varepsilon_0^n$ is such design which takes maximum of $|X_1|$ in absolute value. Using theorem 1 we can prove the following theorem.

**Theorem 2.** Saturated $D$-optimal designs $\varepsilon_0^3$ for homoscedastic observations (1) form infinite set of designs having following structure. Three spectrum points of this design should lay on one of six sides of unite cube, $-1 \leq x_i \leq 1, i = 1, 2, 3$. Two points are the ends of same edge and the third point is any point laying on opposite edge of the same side.

It has been shown in article [2], that the absolute value of a determinant matrix $X_1$ corresponding to design $\varepsilon_0^3$ at which all points of spectrum lays in vertices of unite cube is equal 4. It is easy to check up, that such situation takes place for the matrixes $X_1$ which structure is described in theorem 2. Hence, such matrixes correspond to optimal designs and such designs there are an infinite set. Theorem 2 is proved.

Following theorem is correct for heteroscedastic observations.

**Theorem 3.** Saturated $D$-optimal design $\varepsilon_0^3$ for heteroscedastic observations coincides with design $\varepsilon_0^3$ for homoscedastic observations at which spectrum points lay in vertices of unite cube $|x_i| \leq 1, i = 1, 2, 3$ and at which product of variances of observations in these points is minimal.

Construction of $\varepsilon_0^3$ is reduced to maximization of absolute value of

$$|X_1| = \frac{|X|}{\sqrt{d_id_jd_k}}, \tag{6}$$

where $X$ is the matrix of design $\varepsilon_3$,

$$X = \begin{pmatrix}
x_{1i} & x_{2i} & x_{3i} \\
x_{1j} & x_{2j} & x_{3j} \\
x_{1k} & x_{2k} & x_{3k}
\end{pmatrix}. \tag{7}$$

In matrix (7) $i, j, k$ are various numbers of vertices of unite cube. Determinants of matrixes $X$ with such structure are equal on absolute value to zero or 4. The matrixes (7) which determinants are equal $\pm 4$ and product of observations minimum define the optimum designs $\varepsilon_0^3$. Theorem 3 is proved.

For example, for variance $d(x) = 10 - 2x_1 + x_2 - x_3$ design with spectrum points $x^{(1)} = (1, 1, 1), x^{(2)} = (1, -1, 1), x^{(3)} = (1, -1, -1)$ is saturated $D$-optimal and it is unique.
4 Structure of \( A \)-optimal designs for regress line with heteroscedastic observations

Consider regress line

\[
y_i = \theta_0 + \theta_1 x_i + \varepsilon(x_i), \quad |x_i| \leq 1, i = 1, 2, \ldots, n,
\]

with variance of observations

\[
D(\varepsilon(x_i)) = a_2 x_i^2 + a_1 x_i + a_0 = a_0 (k_2 x_i^2 + k_1 x_i + 1) > 0, a_0 > 0,
\]

where \( k_2 = \frac{a_2}{a_0}, k_1 = \frac{a_1}{a_0} \) and constants are such that for which one of following conditions is carried out

i) \( |k_1| < 2\sqrt{k_2}, k_2 > 0, \)

ii) \( |k_1| < 1 + k_2, -1 < k_2 \leq 0, \)

iii) \( 2\sqrt{k_2} < |k_1| < 1 + k_2, 0 < k_2 < 1. \)

These conditions guarantee that inequality (9) will be carried out. In [3] it is investigated structure of \( A \)-optimal designs for model (8), (9) when \( a_2 = 0 \). It is possible to generalize the result obtained in [3] and receive the following theorem.

**Theorem 4.** Let one of conditions i), ii), iii) is carried out. Then for \( p_{s-1} \leq k_1 \leq p_s \) \( A \)-optimal designs for model of observations (8), (9) are following:

\[
\varepsilon_n^0 = \begin{pmatrix} -1; & 1 \\ n_1 - s; & n - n_1 + s \end{pmatrix}, s = 0, \pm 1, \pm 2, \ldots,
\]

where \( n_1 = \left\lfloor \frac{n}{2} \right\rfloor \) is the whole part from \( \frac{n}{2} \), \( n_1 - s = 1, 2, \ldots \), and

\[
p_s = \frac{n(1 + k_2)(n - 2n_1 + 2s + 1)}{2n_1(n_1 - 2s - n - 1) + n(2s + n + 1) + 2s(s + 1)}.
\]

**References**


ON ROBUST KALMAN FILTERING WITH USING WAVELET ANALYSIS

V.I. LOBACH  
Belarussian State University  
Minsk, Belarus  
e-mail: lobach@bsu.by

Abstract

One presents a nonlinear filtering algorithm that propagates the entire conditional probability density functions. These functions are recursively computed in efficient manner using the discrete wavelet transform. With the multiresolution analysis we can speed up the computation by ignoring the high-frequency details of the probability density function up to a certain level. The level of the wavelet decomposition can be determined at each time step adaptively.

Keywords: Kalman filter, wavelet transform, nonlinear filtering, probability density function.

1 Introduction

The classical Kalman filter has been applied in various fields such as control systems, reliability engineering and wireless communications. The Kalman filter is a maximum likelihood estimate assuming a linear model, quadratic performance criterion and Gaussian probability distribution for the system process and observation noises. In practical systems the assumed model is only approximate one and the two types of noises may follow a thick-tailed, non-Gaussian probability distribution inducing innovation and observation outliers.

Innovation and observation outliers arise naturally in many areas of engineering [1]. Examples of these occurrences include hardware discontinuities in digital control systems, faults in the sensors of a control system including target estimation and tracking in aerospace applications. In the presence of all these outliers the Kalman filter may provide a strongly biased solution or even diverge.

The theory of nonlinear filtering concerns the estimation of a signal corrupted by the white noise and has diverse applications in signal processing, automatic control, finance and so on. The basic setting of the theory involves a Markov signal process observed in independent corrupting noise. The calculation of the resulting filters is a classical topic in stochastic analysis [2]. In order for the theory to be practically useful it is important to establish that the filtered estimates are not too sensitive to the choice of underlying model.

In this paper we consider an approach of representing and recursively generating the approximation of the conditional probability density function using the wavelet transform [3]. The center idea is to recognize that the conditional probability density waveform may be an approximation to original conditional probability density function.
2 The statement of the problem

Consider the discrete-time nonlinear system model

\[ x(k + 1) = f(k, x(k)) + u(k), \]
\[ y(k) = h(k, x(k)) + v(k), \]

(1)

where \( k = 0, 1, 2, \ldots \), \( x(k) \in R^n \) is the state, \( u(k) \in R^n \) is the input noise, \( y(k) \in R^m \) is the measured output, \( v(k) \in R^m \) is the measurement noise; with \( u \) and \( v \) being independent white noise random processes. A fundamental problem associated with such system is that of state estimation, i.e., the optimal estimation of the state \( x(k) \) from the noisy measurements \( \{ y(i), i = 0, 1, \ldots, k \} \); the corresponding state estimate is denoted \( \hat{x}(k|k) \). Here \( f(x, t) \) and \( h(x, t) \) are nonlinear vector-valued vector functions. The probability density function \( p_{x(0)}(x) \) of the uncertain initial state \( x(0) \) is assumed to be known. The process noise \( u(t) \) and the measurement noise \( v(t) \) are modelled as the independent random vectors with known probability density functions \( p_{u(k)}(x) \) and \( p_{v(k)}(x) \) respectively; \( p_{x(0)}(x) \) and \( p_{u(k)}(x) \) and \( p_{v(k)}(x) \) are not necessary Gaussian.

A fundamental problem associated with such systems is that of state estimation, i.e., the optimal estimation of the state \( x(k) \) from the noisy measurements \( y_k = \{ y(i), i = 0, 1, \ldots, k \} \); the corresponding state estimate is denoted \( \hat{x}(k|k) \). The model (1) is called the state space model of the time serie.

It is known that the optimal estimate of the state \( x(k) \) is conditional mathematical expectation

\[ \hat{x}(k|k) = E\{x(k)|y_k\} = \int_{-\infty}^{\infty} x \cdot p_{x(k)|y_k}(x) dx. \]

(2)

3 Approximation of conditional probability density function using the wavelet transform

It is not difficult to show that the predicted conditional probability density function \( p_{x(k+1)|y_k}(x) \) and filtered conditional probability function \( p_{x(k)|y_k}(x) \) can be obtained recursively staring from \( p_{x(0)|y_0}(x) = p_{x(0)}(x) \) using formulas:

\[ p_{x(k+1)|y_k}(x) = \int_{-\infty}^{\infty} p_{x(k)}(x - f(y, k))p_{x(k)|y_k}(y) dy, \]

(3)

\[ p_{x(k+1)|y_{k+1}}(x) = c \cdot p_{v(k)}(y(k + 1) - h(x, k)) \cdot p_{x(k+1)|y_k}(x), \]

(4)

where \( c \) is the normalizing constant.

The main problems arise from the calculation (3), because with every step the integrand function is complicated. Even in many, relatively simple cases, expressions obtained directly from (3) and (4) are too cumbersome, which greatly hinders their practical application. However, using the wavelet transform, it is possible to significantly reduce the number of the calculation. We will use the discrete wavelet transform.
With variable resolution of the wavelet transform we accelerate the calculations, missing high density elements up to primary level. If we use the approximation of the wavelet of the level \( j \), the number of operations is reduced by \( 2^j \) times [4].

Present an algorithm of calculation (2).

1. Select the mesh nodes \( \{-a \leq x_0 < x_1 < \ldots < x_{N+1} < x_N \leq a\} \), \( a > 0 \).

2. For each of the conditional probability density function (time) to construct vectors
\[
g^{(k)} = \left( g_0^{(k)}, g_1^{(k)}, \ldots, g_{n-1}^{(k)} \right)^T \quad \text{and} \quad r^{(k)} = \left( r_0^{(k)}, r_1^{(k)}, \ldots, r_{N-1}^{(k)} \right)^T,
\]
elements of which are given by
\[
\begin{align*}
q_i^{(k)} &= p_u(x_i), & i &= 0, N-1, \\
q_0^{(k)} &= p_x(0), & i &= 0, N-1, \\
q_i^{(k)} &= \sum_{m=0}^{N-1} r_i^{(k)} g_{m}^{(k-m)}, & i &= 0, N-1.
\end{align*}
\]

3. Apply direct wavelet transform for each of the vectors \( r^{(k)}, q^{(k)} \) and construct the wavelet transforms \( a_j^{(k)} \) and wavelet coefficients \( d_j^{(k)} \):
\[
a_{j,n}^{(k)} = r_n^{(k)}, \quad i = 0, N-1, \\
a_{j,n}^{(k)} = \sum_i h_i a_{j-1,2n+i}, \\
d_{j,n}^{(k)} = \sum_i g_i d_{j-1,2n+i},
\]
where \((h_i, g_i)\) – the base of wavelet transform.

4. Apple indirect wavelet transform
\[
a_{j-1,n}^{(k)} = \sum_i h_{n-2i} a_{j,i} + \sum_i g_{n-2i} d_{j,i}.
\]

5. Construct the vectors
\[
R^{(k)} = \left( r^{(k)}, Z r^{(k)}, \ldots, Z^{N-1} r^{(k)} \right)^T,
\]
where the matrix \( Z \) \((N \times N)\):
\[
Z = \begin{pmatrix}
0 & 0 & \ldots & 0 & 0 \\
1 & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & 0 \\
0 & 0 & \ldots & 1 & 0
\end{pmatrix}.
\]
6. Conditional probability density functional is obtained in the form of the grid function

\[ p_{x|y}(x) = R \cdot g. \]

7. Optimal estimate is defined as

\[ \hat{x}(k|k) = \int_{-\infty}^{\infty} xp_{x|y}(x)dx \approx \sum_{l=0}^{N-1} x_l \cdot p_{x|y}(x_l) \Delta x_l. \]

The results of numerical simulation show the effectiveness of the proposed method of estimation of the state of the unobserved process.

References


ROBUST INTERPOLATION PROBLEM FOR
STOCHASTIC SEQUENCES WITH
STATIONARY INCREMENTS

M.M. Luz, M.P. Moklyachuk
Taras Shevchenko National University of Kyiv
Kyiv, UKRAINE

e-mail: maksim_luz@ukr.net, mmp@univ.kiev.ua

Abstract

The problem of optimal estimation of the functional \( A_N \xi = \sum_{k=0}^{N} a(k) \xi(k) \) depending on the unknown values of stochastic sequence \( \xi(k) \) with stationary \( n \)th increments from observations of the sequence \( \xi(k) + \eta(k) \) at points of time \( k = N + 1, N + 2, \ldots \) and observations of the sequence \( \xi(k) \) at points of time \( k = -1, -2, \ldots \) is considered. Formulas for calculating the mean square error and the spectral characteristic of the optimal linear estimate of the functional are proposed under condition of spectral certainty, where spectral densities of the sequences \( \xi(k) \) and \( \eta(k) \) are exactly known. Minimax (robust) method of estimation is used in the case where spectral densities are not known exactly, but sets of admissible spectral densities are given. Formulas that determine the least favorable spectral densities and the minimax spectral characteristics are proposed for some definite sets of admissible densities.

1 Introduction

Wide sense stationary and related stochastic processes have been studied by many scientists. The results are applied for solving practical problems in different areas. Particularly, seasonal time series, ARIMA sequences, which are the examples of stochastic processes with stationary increments studied by A.M. Yaglom [6], play an important role in description and prediction of economic data.

The classical methods of estimation of unobserved values of stationary processes and sequences were developed by A.N. Kolmogorov, N. Wiener, A.M. Yaglom. A.N. Kolmogorov [3] proposed a method of projections in Hilbert space for solving the interpolation problem which consists in finding an estimate of missed values of a stationary sequence from available observations. This method may be applied in the case where the spectral density of the process is known. However, in practice often the spectral density is not exactly known. In this case one can use the minimax-robust approach to estimation proposed by Ulf Grenander [1]. J. Franke and H. V. Poor [2] investigated the minimax extrapolation and filtering problems for stationary sequences with the help of convex optimization methods. M. P. Moklyachuk [5] applied the minimax approach to extrapolation, interpolation and filtering of functionals that depend on the unknown values of stationary processes and sequences.

In the article by M.M. Luz and M.P. Moklyachuk [4] the problem of optimal estimation of linear functional \( A_N \xi = \sum_{k=0}^{N} a(k) \xi(k) \) depending on unknown values of a stochastic sequence \( \xi(k) \) with stationary \( n \)-th increments from observations of the
sequence at points of time \(Z \setminus \{0, 1, \ldots, N\}\) is investigated. Formulas for calculation the least favorable spectral densities and the minimax spectral characteristics are given for some special sets of admissible spectral densities. In the presented paper we consider the interpolation problem for linear functional \(A_N\xi\) under condition that we have observations of the sequence \(\xi(k)\) at points of time \(k \leq -1\) and observations of the sequence \(\xi(k) + \eta(k)\) at points of time \(k \geq N + 1\), where \(\eta(k)\) is an uncorrelated with \(\xi(k)\) stochastic sequence with \(n\)-th stationary increments.

### 2 Interpolation problem

**Definition 1.** For a given stochastic sequence \(\{\xi(m), m \in Z\}\) the function \(\xi^{(n)}(m, \mu) = (1 - B_\mu)^n\xi(m) = \sum_{l=0}^n (-1)^l C_n^l \xi(m - l\mu)\), where \(B_\mu\) is a backward operator with step \(\mu \in Z\), \(B_\mu\xi(m) = \xi(m - \mu)\), is called stochastic \(n\)-th increment with step \(\mu \in Z\).

**Definition 2.** Stochastic increment \(\xi^{(n)}(m, \mu)\) of stochastic sequence \(\{\xi(m), m \in Z\}\) is called wide sense stationary increment if mathematical expectations \(E\xi^{(n)}(m_0, \mu) = e^{(n)}(\mu)\) and \(E\xi^{(n)}(m_0 + m, \mu_1)\xi^{(n)}(m_0, \mu_2) = D^{(n)}(m, \mu_1, \mu_2)\) exist for each \(m_0, \mu, m, \mu_1, \mu_2 \in Z\) and do not depend on \(m_0\). Stochastic sequence \(\{\xi(m), m \in Z\}\) is called sequence with stationary \(n\)-th increments.

**Theorem 1.** Stationary \(n\)-th increment \(\xi^{(n)}(m, \mu)\) admits the representation

\[
\xi^{(n)}(m, \mu) = \int_{-\pi}^{\pi} e^{im\lambda}(1 - e^{-im\lambda})^n(i\lambda)^{-n} dZ(\lambda),
\]

where \(Z(\lambda)\) is an orthogonal random measure on \([-\pi, \pi]\).

Let uncorrelated stochastic sequences \(\xi(m)\) and \(\eta(m)\) have stationary increments \(\xi^{(n)}(m, \mu)\) and \(\eta^{(n)}(m, \mu)\), \(\mu > 0\), with absolutely continuous spectral functions \(F(\lambda)\) and \(G(\lambda)\) which have spectral densities \(f(\lambda)\) and \(g(\lambda)\) respectively. Consider the problem of finding the mean square sense optimal linear estimate of the functional \(A_N\xi = \sum_{k=0}^{N} a(k)\xi(k)\) depending on unknown values \(\xi(k), k = 0, 1, \ldots, N\), from observations of the sequence \(\xi(k)\) at points of time \(k = -1, -2, \ldots\), and of the sequence \(\xi(k) + \eta(k)\) at points of time \(k = N + 1, N + 2, \ldots\).

Let \(\{d_\mu(k) : k \geq 0\}\) be coefficients from the relation

\[
\sum_{k=0}^{\infty} d_\mu(k) x^k = \left(\sum_{j=0}^{\infty} x^{j\mu}\right)^n
\]

and let \(q(k, \mu), k < 0\), denote the least integer number among numbers which are greater than or equal to \(-\frac{k}{\mu}\). Then the functional \(A_N\xi\) can be presented as \(A_N\xi = B_N\xi - V_N\xi\), where \(B_N\xi = \sum_{k=0}^{N} b_\mu(k)\xi^{(n)}(k, \mu), V_N\xi = \sum_{k=-\mu}^{-1} v_\mu(k)\xi^{(n)}(k, \mu), v_\mu(k) = \sum_{l=q(k, \mu)}^{n} (-1)^l C_n^l b_\mu(l\mu + k), b_\mu(k) = \sum_{m=k}^{N} a(m) d_\mu(m - k) = (D_N^\mu a)_k, k = 0, 1, \ldots, N\).
$D_N^h$ is a linear operator defined by elements $(D_N^h)_{k,j} = d_{\mu}(j-k)$ if $0 \leq k \leq j \leq N$, and $(D_N^h)_{k,j} = 0$ if $j < k$, $k,j = 0, 1, \ldots, N$; vector $a = (a(0), a(1), \ldots, a(N))$; $b_{\mu}(k) = 0$ if $k > N$.

The mean square optimal linear estimate of the functional $A_NX$ is presented as

$$
\hat{A}_N \xi = - \sum_{k=-\mu n}^{-1} \nu_{\mu}(k) \xi(k) + \int_{-\pi}^{\pi} h_{\mu(a,1)}^a(\lambda) dZ_{\xi^{(n)}}^{(a)}(\lambda) + \int_{-\pi}^{\pi} h_{\mu(a,2)}^a(\lambda) dZ_{\xi^{(n)}}^{(a)}(\lambda).
$$

(3)

Using Kolmogorov’s method of projection in Hilbert space we found the following formulas for calculation the optimal spectral characteristic $h_{\mu(a)}^a(\lambda) = (h_{\mu(a,1)}^a(\lambda), h_{\mu(a,2)}^a(\lambda))$ and the value of the mean square error $\Delta(h_{\mu(a)}^a; f, g) = E|A_NX - \hat{A}_N \xi|^2$:

$$
h_{\mu(a,1)}^a(\lambda) = \sum_{k=0}^{N}(D_N^h(a)_{k}) e^{i\lambda k} + \sum_{k=0}^{N+\mu}(G_{\mu})^{-1}[D_N^h(a)_{\infty}]_{N+\mu-k} e^{i\lambda k} - \sum_{k=0}^{N+\mu}(G_{\mu})^{-1}[D_N^h(a)_{\infty}]_{N+\mu-k} e^{i\lambda k} - \sum_{k=-\infty}^{N+\mu}(G_{\mu})^{-1}[D_N^h(a)_{\infty}]_{N+\mu-k} e^{i\lambda k}.
$$

(4)

$$
h_{\mu(a,2)}^a(\lambda) = \sum_{k=0}^{N}(D_N^h(a)_{\infty})_{k} e^{i\lambda k} + \sum_{k=-\infty}^{N+\mu}(G_{\mu})^{-1}[D_N^h(a)_{\infty}]_{N+\mu-k} e^{i\lambda k} - \sum_{k=-\infty}^{N+\mu}(G_{\mu})^{-1}[D_N^h(a)_{\infty}]_{N+\mu-k} e^{i\lambda k}.
$$

(5)

$$
\Delta(h_{\mu(a)}^a; f, g) = \int_{-\pi}^{\pi} \left| \sum_{k=0}^{N}(D_N^h(a)_{\infty})_{k} e^{i\lambda k} \right|^2 d\lambda + \int_{-\pi}^{\pi} \left| \sum_{k=-\infty}^{N+\mu}(G_{\mu})^{-1}[D_N^h(a)_{\infty}]_{N+\mu-k} e^{i\lambda k} \right|^2 d\lambda,
$$

(6)

where $[D_N^h(a)_{\infty}]$ is an element in $\ell_2$ space received from the $(N+1)$-dimension vector $D_N^h(a)$ by adding zeros, $F_{\mu} = G_{\mu}^{-1}G_{\mu}^{\epsilon}(G_{\mu}^{\epsilon})^{-1}G_{\mu}^{\epsilon} - G_{\mu}^{\epsilon}$, $G_{\mu} = (G_{\mu}^{\epsilon})^{-1}G_{\mu}^{\epsilon}$, $G_{\mu}$ and $F_{\mu}$ are the linear operators in $\ell_2$ defined by coefficients $(G_{\mu}^{\epsilon})_{l,k} = g_{\mu}(N + \mu - l - k)$, $(G_{\mu})_{l,k} = g_{\mu}(k-l)$, and $(F_{\mu})_{l,k} = f_{\mu}(N + \mu - l - k)$, $l,k \geq 0$; $\{f_{\mu}(k) : k \geq 0\}$ are the Fourier coefficients of the functions $\frac{1- e^{i\lambda \mu}}{1- e^{i\lambda \mu}}$ under the condition of integrability of these functions on $[-\pi, \pi]$.

### 3 Minimax-robust method of interpolation

In the previous section we proposed solution of the considered interpolation problem in the case when the spectral densities $f(\lambda)$ and $g(\lambda)$ are exactly known. In the case where the spectral densities are not known, but a set $D_f \times D_g$ of possible spectral densities is given, the minimax (robust) approach to estimation of functionals of the unknown values of random sequence with stationary increments is reasonable. In other words we found an estimate that minimizes the mean square error for all spectral densities from the given classes $D_f \times D_g$ simultaneously.
Denote by $f_0$, $g_0$ the least favorable densities in the class $D_f \times D_g$ and by $h^0 = h^{(e)}_\mu(f_0, g_0)$ the minimax-robust spectral characteristic (see [5]). Then $f_0$ and $g_0$ are solutions to the conditional extremum problem

$$
\tilde{\Delta}(f, g) = -\Delta(h^{(e)}_\mu(f_0, g_0); f, g) \rightarrow \inf, \quad (f, g) \in D_f \times D_g,
$$

(7)

Let us consider the following set of admissible densities:

$$
D = \left\{ f(\lambda)\left| \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\lambda^{2n}d\lambda}{1 - e^{i\lambda\mu}2n f(\lambda)} \geq P_1 \right. \right\} \times \left\{ g(\lambda)|\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\lambda^{2n}d\lambda}{1 - e^{i\lambda\mu}2n g(\lambda)} \geq P_2 \right\}.
$$

In this case the least favorable densities

$$
f^0_\mu(\lambda) = \frac{\lambda^{2n}}{1 - e^{i\lambda\mu}2n} \left( \sum_{k=-\infty}^{\infty} f_\mu^0(|k|)e^{i\lambda k} \right)^{-1},
$$

(8)

$$
g^0_\mu(\lambda) = \frac{\lambda^{2n}}{1 - e^{i\lambda\mu}2n} \left( \sum_{k=-\infty}^{\infty} g_\mu^0(|k|)e^{i\lambda k} \right)^{-1},
$$

are defined by coefficients $f^0_\mu(k)$, $g^0_\mu(k)$, $k \geq 0$, that satisfy the following equations:

$$
\left( (G^c_\mu)^0 + (F^c_\mu)^0 \right) \left( (G^c_\mu)^0 \right)^{-1} p^\mu_1 = (G^c_\mu)^0 p^\mu_1 + [D^\mu_N a]_{\infty},
$$

(9)

$$
\left( (G^c_\mu)^0 + (F^c_\mu)^0 \right) - (G^c_\mu)^0 \left( (G^c_\mu)^0 \right)^{-1} (G^c_\mu)^0 [D^\mu_N a]_{\infty} = p^\mu_2,
$$

(10)

where $p^\mu_1$ is a vector from $\ell_2$ with coordinates $p^\mu_1(0) = p_1$ and $p^\mu_1(k) = 0$, $k > 0$, and $p^\mu_2$ is a vector from $\ell_2$ with coordinates $p^\mu_2(N + \mu n) = p_1 - p_2$ and $p^\mu_2(k) = 0$, $k \neq N + \mu n$; $p_1$ and $p_2$ are the Lagrange multipliers.

Equations (9), (10) can be simplified and solved in some special cases under additional conditions. Other sets $D$ of admissible spectral densities are considered.

References


ON STATISTICAL ESTIMATION FOR MARKOV CHAIN OF CONDITIONAL ORDER

M.V. MALTSEW
Belarusian State University
Minsk, Republic of Belarus
e-mail: maltsew@mail.ru

Abstract

The paper deals with Markov chain of conditional order, which is a special case of a high-order Markov chain with a small number of parameters. Estimator for the order is constructed and its consistency is proved. Numerical results of comparison of different Markov models via Bayesian information criterion are presented.

1 Introduction

Markov chain of the order $s$ ($s \geq 1$) [3] is a well-known mathematical model widely used in various spheres [7, 1]. But this model has exponential complexity, since the number of independent parameters $D$ of the $N$-state Markov chain of the order $s$ increases exponentially when $s$ grows: $D = (N - 1)N^s$. And in applications one often needs the sample of vast size $n \geq D$ unobtainable in practice. Therefore small-parametric or parsimonious models was developed to overcome this difficulty. These models are special cases of fully-connected $s$-order Markov chain, but the number of parameters required to determine transition matrix is less than $D$. Let us give some examples of such models: the Markov chain of the order $s$ with $r$ partial connections [5], Raftery model [6]. This paper is devoted to another parsimonious model — the Markov chain of conditional order proposed in [4].

2 Mathematical model

At first let us introduce the notation we will use: $N$ is the set of natural numbers, $2 \leq N < \infty$, $A = \{0, 1, \ldots, N - 1\}$ is the finite state space with $N$ elements; $J^m_n = (j_n, \ldots, j_m) \in A^{m-n+1}$, $m \geq n$, is the multiindex; $\{x_t \in A : t \in N\}$ is a homogeneous Markov chain of the order $s$, $(2 \leq s < \infty)$ with $(s + 1)$-dimensional matrix of transition probabilities $P = (p_{j_{t+1}^s}): p_{j_{t+1}^s} = P\{x_{t+s} = j_{s+1} | x_{t+s-1} = j_s, \ldots, x_t = j_1\}$, $j_{t+1}^s \in A^{s+1}$, $t \in N$. $L \in \{1, 2, \ldots, s - 1\}$, $K = N^L - 1$ are natural numbers; $Q^{(1)}, \ldots, Q^{(K)}$ are $M$ ($1 \leq M \leq K + 1$) different square stochastic matrices of the order $N$: $Q^{(m)} = (q^{(m)}_{i,j})$, $0 \leq q^{(m)}_{i,j} \leq 1$, $\sum_{j \in A} q^{(m)}_{i,j} \equiv 1$, $i, j \in A$, $1 \leq m \leq M$; $< J^m_n > = \sum_{k=n}^{m} N^{k-n} j_k \in \{0, 1, \ldots, N^{m-n+1} - 1\}$ is the numeric representation of the multiindex $J^m_n \in A^{m-n+1}$; $I\{C\}$ is the indicator of the event $C$.
The Markov chain \( \{x_t \in A : t \in \mathbb{N}\} \) is called the Markov chain of conditional order [4] if its one-step transition probabilities have the following form:

\[
p_{j_{i}^{t+1}} = \sum_{k=0}^{K} I\{< J_{s-t-L+1}^k > = k \} \cdot q_{s-t-L+1}^{(m_k)}, \tag{1}
\]

where \( 1 \leq m_k \leq M, 1 \leq b_k \leq s - L, 0 \leq k \leq K, \min_{0 \leq k \leq K} b_k = 1; \) and all elements of the set \( \{1, 2, \ldots, M\} \) occur in the sequence \( m_0, \ldots, m_K \). The sequence of elements \( J_{s-t-L+1}^k \) is called the base memory fragment (BMF) of the random sequence, \( L \) is the length of BMF; the value \( s_k = s - b_k + 1 \) is called the conditional order. Thus the state \( x_t \) at time \( t \) doesn’t depend on all previous states, but depends only on \( L + 1 \) states \( (j_{b_k}, J_{s-t-L+1}^k) \). The transition matrix \( P \) of the Markov chain of conditional order is defined by \( d = 2(NL+1) + MN(N-1) \) independent parameters. Note that if \( L = s-1, s_0 = s_1 = \cdots = s_K = s \), we have fully-connected Markov chain of the order \( s \). In the rest of the paper we will consider \( M = K+1 \), i.e. \( K+1 \) independent stochastic matrices correspond to \( K+1 \) different values of BMFs and \( m_k = k+1, k = 0, 1, \ldots, K \).

3 Estimators for \( s \) and \( L \)

We used Bayesian information criterion (BIC) [2] to estimate the order \( s \) and the BMF length \( L \) of the Markov chain of conditional order (1) from the sample \( X^n_t \in A^n \) of length \( n \):

\[
(\hat{s}, \hat{L}) = \arg \min_{2 \leq s \leq s_+, 1 \leq L \leq L_+} BIC(s', L'), \tag{2}
\]

\[
BIC(s', L') = -2 \sum_{j_0^L+1 \in A} \sum_{k=0}^{K} I\{< J_1^L > = k \} \nu_{L+2,g(s_k,L)}(J_0^L+1) \ln \hat{q}_{j_0^L,J_{s+1}^L+1}^{(k+1)} + d \ln(n-s'),
\]

where \( g(x,y) = x - y - 1; S_+ \geq 2, 1 \leq L_+ \leq S_+ - 1 \) are maximum admissible values of \( s \) and \( L \) respectively, \( \nu_{L+2,g(s_k,L)}(J_1^L) = \sum_{t=1}^{n-s} I\{x_{t+s-l-y+1} = j_1; X_{t+s-l-y+2} = J_2^L, l \geq 2, y \geq 0, l+y \leq s+1 \) are frequency of the state \( J_1^L \in A^l \) with a gap of length \( y \) between \( j_1 \) and \( J_2^L \). If \( s \) and \( L \) are known, maximum likelihood estimators for transition probabilities \( q_{j_0,J_{s+1}^L+1}^{(k+1)} \) and conditional orders \( s_k \) are constructed in [4]:

\[
\hat{q}_{j_0,J_{s+1}^L+1}^{(k+1)} = \begin{cases} 
\sum_{J_1^L \in A} I\{< J_1^L > = k \} \nu_{L+2,g(s_k,L)}(J_0^L+1), & \text{if } \nu_{L+1,g(s_k,L)}(J_0^L) > 0, \\
1/N, & \text{if } \nu_{L+1,g(s_k,L)}(J_0^L) = 0,
\end{cases} \tag{3}
\]

\[
\hat{s}_k = \arg \max_{L+1 \leq s \leq s} \sum_{J_1^L \in A} I\{< J_1^L > = k \} \sum_{j_0,J_{s+1}^L \in A} \nu_{L+1,g(s_k,L)}(J_0^L) \ln(\hat{q}_{j_0,J_{s+1}^L+1}^{(k+1)}). \tag{4}
\]

Theorem 1. If the Markov chain of conditional order (1) is stationary, then (2) gives consistent estimators at \( n \to \infty \).
Proof. Let us give only a scheme of the proof. Let 

\[ \pi_l,g(J_l^I) = P\{x_t = j_1, X_{l+y+1}^t = J_l^I\}, \]

for all \(l \geq 2\), \(y \geq 0\). Then 

\[ q_{j_0,J_{l+1}}^{(k+1)} = \frac{\pi_{l+2,g(s_0,l)}(J_{l+1}^l)}{\pi_{l+1,g(s_0,l)}(J_{l}^l)} \]

where \( J_l^I \geq k \). Note that if \( X_t^I = J_t^I \) is fixed, then 

\[ -\frac{1}{n} \sum_{j_0}^{K} \sum_{L^t+2,y}^K I\{< J_t^I \geq k\} \nu_{L^t+2,y}(J_0^I+1) \ln \frac{\pi_{L^t+2,y}(J_0^I+1)}{\nu_{L^t+1,y}(J_0^I+1)} P_{J_t^I} \]

is a conditional entropy \( H_{J_t^I} \{x_{L^t+1}|x_0\} \) of \( x_{L^t+1} \) given \( x_0 \). Using asymptotic properties of the estimators (3) and (4) [4] it is easy to show that at \( n \to \infty \)

\[ \sum_{L^t+1}^K \sum_{y_k}^K I\{< J_t^I \geq k\} H_{J_t^I} \{x_{L^t+1}|x_0\}, \]

\( L^t+1 \leq y_k \leq s' \). And using properties of entropy and methods described in [2] we can prove that \( P\{(\hat{s}, \hat{L}) \in ([2, s+] \times [1, L+]) \setminus (s, L)\} \to 0 \) at \( n \to \infty \).

4 Numerical results

In this section we present the results of comparison of different Markov models via BIC. The data we examined was from human DNA. Partition of genes into coding segments (exons) and noncoding segments (introns) is an important task in bioinformatics, and fitting a stochastic model of a gene is a fruitful approach to this problem. For instance, it is used in well-known program GenScan. The sequence of introns from human gene HSHMG17G [8] was tested. The length of the sequence \( n = 6922\), \( S_+ \leq 6\), the size of state space \( A \) is 4 (0 corresponds to nucleotide A, 1 to C, 2 to G, 3 to T). We used the following models: fully-connected \( s \)-order Markov chain (MC(\( s \)), the Markov chain of order \( s \) with \( r \) partial connectons (MS(\( s, r \))) and the Markov chain of conditional order with BMF length \( L \) (MCCO(\( s, L \))). For each model the value of BIC was calculated. Results are presented in table (1). Minimum value of BIC is marked by bold type.

As we can see, the most adequate model is the Markov chain of conditional order with parameters: \( s = 6\), \( L = 1\). Estimators for conditional orders are: \( \hat{s}_0 = 4\), \( \hat{s}_1 = 3\), \( \hat{s}_2 = 3\), \( \hat{s}_3 = 6\). Estimators for transition matrices are:

\[ \hat{Q}^{(1)} = \begin{pmatrix} 0.484 & 0.376 & 0.083 & 0.005 & 0.005 \\ 0.463 & 0.405 & 0.085 & 0.047 \\ 0.251 & 0.181 & 0.373 & 0.195 \\ 0.312 & 0.201 & 0.294 & 0.193 \end{pmatrix}, \]

\[ \hat{Q}^{(2)} = \begin{pmatrix} 0.372 & 0.485 & 0.040 & 0.103 \\ 0.309 & 0.509 & 0.081 & 0.101 \\ 0.220 & 0.265 & 0.240 & 0.275 \\ 0.216 & 0.329 & 0.108 & 0.347 \end{pmatrix}, \]

\[ \hat{Q}^{(3)} = \begin{pmatrix} 0.254 & 0.210 & 0.270 & 0.266 \\ 0.170 & 0.370 & 0.285 & 0.175 \\ 0.205 & 0.320 & 0.320 & 0.155 \\ 0.196 & 0.253 & 0.306 & 0.245 \end{pmatrix}, \]

\[ \hat{Q}^{(4)} = \begin{pmatrix} 0.201 & 0.181 & 0.331 & 0.287 \\ 0.099 & 0.326 & 0.276 & 0.299 \\ 0.125 & 0.230 & 0.342 & 0.303 \\ 0.193 & 0.206 & 0.215 & 0.386 \end{pmatrix}. \]
Table 1: Values of BIC

<table>
<thead>
<tr>
<th>model</th>
<th>BIC</th>
<th>model</th>
<th>BIC</th>
<th>model</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC(1)</td>
<td>17792.7</td>
<td>MC(4, 3)</td>
<td>18162.9</td>
<td>MCCO(3, 1)</td>
<td>17557.5</td>
</tr>
<tr>
<td>MC(2)</td>
<td>17595.7</td>
<td>MC(5, 1)</td>
<td>18108.2</td>
<td>MCCO(4, 1)</td>
<td>17472.6</td>
</tr>
<tr>
<td>MC(3)</td>
<td>18293.1</td>
<td>MC(5, 2)</td>
<td>17553.8</td>
<td>MCCO(4, 2)</td>
<td>18205.2</td>
</tr>
<tr>
<td>MC(4)</td>
<td>22252.5</td>
<td>MC(5, 3)</td>
<td>18219.8</td>
<td>MCCO(5, 1)</td>
<td>17482.5</td>
</tr>
<tr>
<td>MC(5)</td>
<td>39894.1</td>
<td>MC(5, 4)</td>
<td>21896.6</td>
<td>MCCO(5, 2)</td>
<td>18170.6</td>
</tr>
<tr>
<td>MC(6)</td>
<td>116798.2</td>
<td>MC(6, 1)</td>
<td>18119.8</td>
<td>MCCO(5, 3)</td>
<td>22616.9</td>
</tr>
<tr>
<td>MC(2, 1)</td>
<td>18112.9</td>
<td>MC(6, 2)</td>
<td>17568.9</td>
<td>MCCO(6, 1)</td>
<td>17448.8</td>
</tr>
<tr>
<td>MC(3, 1)</td>
<td>18116.7</td>
<td>MC(6, 3)</td>
<td>18150.0</td>
<td>MCCO(6, 2)</td>
<td>18139.9</td>
</tr>
<tr>
<td>MC(3, 2)</td>
<td>17535.8</td>
<td>MC(6, 4)</td>
<td>21849.5</td>
<td>MCCO(6, 3)</td>
<td>22520.2</td>
</tr>
<tr>
<td>MC(4, 1)</td>
<td>18123.6</td>
<td>MC(6, 5)</td>
<td>26457.0</td>
<td>MCCO(6, 4)</td>
<td>41618.7</td>
</tr>
<tr>
<td>MC(4, 2)</td>
<td>17532.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References


SPARSE PRINCIPAL BALANCES FOR HIGH-DIMENSIONAL COMPOSITIONAL DATA

Can Mert¹, Peter Filzmoser¹, Karel Hron²

¹Department of Statistics and Probability Theory
Vienna University of Technology Vienna, AUSTRIA
²Department of Mathematical Analysis and Applications
Palacký University, Olomouc, CZECH REPUBLIC
e-mail: mehmet.mert@tuwien.ac.at

Abstract

Extracting the most essential information out of compositional data can be done by a method called principal balances [5]. This method is, however, computationally only feasible for low-dimensional data. For high-dimensional compositional data we introduce the concept of sparse principal balances, a method that relies on sparse principal component analysis to construct principal directions with many zero loadings. Sparse principal balances are fast to compute even for very high-dimensional data, and their interpretation is easier than principal balances.

1 Introduction

Principal component analysis (PCA) is an efficient method for dimension reduction with the aim of providing maximum explained variance. Here we are interested in a dimension reduction of high-dimensional compositional data. Compositional data are not represented in the usual Euclidean geometry, which implies that PCA cannot be applied directly. Compositional data are represented in the D-part simplex, they consist of relative information, and all its strictly positive components sum up to a constant value, which can be chosen e.g. as 1. According to [1], the sample space of compositional data is defined as

\[ S^D = \left\{ x = (x_1, \ldots, x_D), x_i > 0, \sum_{i=1}^{D} x_i = 1 \right\}. \] (1)

[2] introduced the centered logratio (clr) transformation to transform the data from the simplex \( S^D \) to the real Euclidean space \( \mathbb{R}^D \) by taking the logarithm of the ratio parts over the geometric mean. The collinearity shortcoming of the clr-transformation led [3] to introduce a further transformation, namely the ilr (isometric logratio) transformation. The ilr-transformation preserves all the advantageous properties of clr transformation like isometry and symmetry while solving the collinearity problem.

The PCA scores of the clr-transformed data which also corresponds to the ilr-coordinates are difficult to interpret. The reason is that due to the geometric mean as denominator, all compositional parts are involved in the PCA scores. To overcome the interpretation problem, [4] introduced the concept of balances where only the relative
information of specific groups of parts is involved. Balances are defined by two non-
overlapping groups of parts $x_-$ and $x_+$. The number of positive parts is given by $r$, and
the number of negative parts by $s$ where $r + s \leq D$. The coordinates of the
composition, also called balance, is defined as

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{g_m(x_+)}{g_m(x_-)},$$

(2)

where $g_m$ stands for the geometric mean. The basis vector of a balance is defined in
the clr-space as $clr(e) = (a_1, \ldots, a_D)$ and its components are defined as:

$$a_i = \begin{cases} 
0 & \text{if } x_i \notin x_+ \text{ and } x_i \notin x_- \\
\sqrt{\frac{s}{r(r+s)}} & \text{if } x_i \in x_+ \\
-\sqrt{\frac{r}{s(r+s)}} & \text{if } x_i \in x_- 
\end{cases}$$

(3)

Balances provide a simplicity for the interpretation of the groups but do not guarantee
the maximum explained variance unconditionally. After finding the first balance, the
parts $x_+$ and $x_-$ are split into non-overlapping parts further until a complete set of
$D-1$ basis vectors is reached. In order to get the maximum explained variance, each
possibility of the combinations has to be considered. Increasing dimension of the data
causes an exponential growth of the number of the binary partitions which makes the
algorithm cost-intensive especially for high-dimensional data [5].

[5] defines principal balances (PBs) that construct an orthonormal basis on the sim-
plex with a compromise of maximum explained variance and simplified interpretability.
Three algorithms are suggested by [5] to construct PBs: Angular-Proximity (AP), Hi-
erarchical Clustering Components (HC), Maximum Explained Variance (MV). AP and
MV are based on all possible combinations of the parts of balances which is compu-
tationally feasible only for low-dimensional data. The algorithm HC is based on the
result of hierarchical cluster analysis of the variation matrix, containing the variances
of all pairwise log-ratios. This algorithm also works with high-dimensional data. For
a detailed description of these algorithms we refer to [5].

2 Constructing sparse principal balances

In this paper we propose an algorithm based on sparse PCA for computing interpretable
orthonormal directions with high explained variance. For sparse PCA we use the
algorithm introduced by [6], which is implemented as function SPC in the R package
PMA. An overview for constructing sparse principal balances (SPBs) can be defined as
follows: The compositional data in $S^D$ are transformed with the clr-transformation to
$R^D$. Then sparse PCA is applied on the clr-transformed data. The first $k$ components
are computed, and the resulting loadings matrix is denoted by $D \times k$ matrix $V = [v_{ij}]$.
Subsequently, the loadings matrix is simplified in order to obtain balances with a better
interpretation. The procedure is described in the following.

For every row $i \in \{1, \ldots, D\}$ in $V$, the smallest $j$ for which $v_{ij} \neq 0$, is searched
and the entries of $v_{il}$, where $l > j$ and $v_{ij} \neq 0$, are set to zero. If row $i$ has zero
entries in all loadings $j$, store $i$ in the index set $I_0$ which contains all zero-lines in the loadings matrix $V$. Secondly, the indices of the columns containing only positive or only negative entries are detected and stored in the index sets $J_p$ and $J_n$, respectively. The size of these sets are given by $|J_p|$ and $|J_n|$. Then the entries $\min(i, j) \neq 0$ of $V$ are identified and set to zero, as long as $|I_0| < |J_p| + |J_n|$, and $I_0$, $J_p$ and $J_n$ are updated. For every $j_p \in J_p$ we set $v_{i_0j_p}$ to any positive value, where $i_0 \in I_0$ and $j_p \in J_p$, and $I_0 := I_0 \{i_0\}$ is updated. Analogously, $v_{i_0j_n}$ is set to any negative value, where $j_n \in J_n$. Consequently, we guarantee positive and negative signs in each column of $V$, and at most one non-zero entry in the rows of $V$. The signs of the resulting matrix $V$ are used for constructing balances.

3 Example

In this section we generate a simulated dataset and apply the methods HC and SPB. We compare those methods also with the results of CoDa-PCA, i.e. a PCA on the clr-transformed data. We generate four matrices of size $\frac{D}{2} \times (D - 1)$ where the observations follow multivariate normal distribution with mean vectors $(0, \ldots, 0, 1/\sqrt{2}, 1/\sqrt{2})$, $(0, \ldots, 0, -1/\sqrt{2}, 1/\sqrt{2})$, $(0, \ldots, 0, -1/\sqrt{2}, -1/\sqrt{2})$, respectively, and the same diagonal covariance matrix $C_Z = \text{diag}(0.5, 0.51/2, \ldots, 0.51/k, 0.01, \ldots, 0.01)$. The matrix $Z_{ilr}$ of size $n \times (D - 1)$ results from joining the four matrices together, and it represents the PCA scores in the ilr-space.

Then a loadings matrix in the ilr-space, $L_{ilr}$, of size $(D - 1) \times (D - 1)$, with uniformly distributed values in $[-1, 1]$ is generated, where the columns are standardized to unit length vectors. $X_{ilr}$ defines the reconstructed matrix $X_{ilr}$ in the ilr-space by $X_{ilr} = Z_{ilr}L_{ilr}^t$, and it is transformed to the simplex by the inverse ilr-transformation; the resulting matrix is denoted by $X_s$. Clr-transformation is applied to $X_s$, SPB and CoDa-PCA are applied to $X_{clr}$, and HC to $X_s$.

We take the dimension of the simulated dataset as $D = 500$ with $n = 100$ observations and consider $k = 5$ PCs. Table 1 shows the cumulative explained variances of the three methods. It turns out that HC yields the smallest explained variances, followed by SPB and CoDa-PCA. Although the explained variance of CoDa-PCA is clearly higher than that of SPB, the results are in general not interpretable since all compositional parts are involved to some extent.

Table 1: Cumulative explained variance (in %) of the first 5 components for CoDa-PCA, SPB, and HC for simulated data with dimension 500.

<table>
<thead>
<tr>
<th>Comp.</th>
<th>CoDa-PCA</th>
<th>Comp.2</th>
<th>Comp.3</th>
<th>Comp.4</th>
<th>Comp.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9.4</td>
<td>16.4</td>
<td>22.2</td>
<td>27.4</td>
<td>30.3</td>
</tr>
<tr>
<td>SPB</td>
<td>7.4</td>
<td>11.8</td>
<td>14.1</td>
<td>15.1</td>
<td>15.6</td>
</tr>
<tr>
<td>HC</td>
<td>4.6</td>
<td>7.7</td>
<td>9.9</td>
<td>11.6</td>
<td>13.1</td>
</tr>
</tbody>
</table>

Figure 1 shows the projection of the data in the space of the first two PCs (balances)
of CoDa-PCA, HC, and SPB. PCA and SPB clearly reflect the pattern of the for simulated groups in contrast to HC, which means that the main structure of the data is preserved in the first two components. For SPB we obtain 191 non-zero loadings in the first balance, and 129 non-zeros in the second balance, which simplifies the interpretation. In contract, HC has no zeros in the first balance, and 356 non-zeros in the second balance.

Figure 1: First two components of CoDa-PCA, SPB, and HC for simulated data with a dimension 500.

References


THE JOINT DISTRIBUTION OF THE
STANDARDIZED MAXIMUM AND
STANDARDIZED MINIMUM FOR A NORMAL
SAMPLE

L.K. SHIRYAEVA
Samara State Economic University
Samara, RUSSIA
e-mail: Shiryeva_LK@mail.ru

Abstract
We find the joint distribution of Grubbs test statistics for a normal sample. Those statistics are standardized maximum and standardized minimum. We note some properties of the joint distribution function which can be applied for possible application of this function.

1 Introduction
Let \( X_1, X_2, \ldots, X_{n-1}, X_n \) be a random sample from a normal \( N(\mu, \sigma^2) \) distribution with mean \( \mu \) and variance \( \sigma^2 \). Grubbs proposed the standardized maximum and standardized minimum [1]:

\[
T_n^{(1)} = \frac{\max_{1 \leq i \leq n} \{X_i\} - \overline{X}}{S}; \quad T_n^{(1)} = \frac{\overline{X} - \min_{1 \leq i \leq n} \{X_i\}}{S},
\]

where \( \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \) is the sample mean and \( S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \) is the sample variance.

It is known that distributions of statistics \( T_n^{(1)} \) and \( T_{n(1)} \) coincide, i.e.

\[
P(T_n^{(1)} < t) = P(T_{n(1)} < t). \tag{1}
\]

Let \( F_n^{(1)}(t) \) be the distribution function of \( T_n^{(1)} \), then [2]:

\[
F_n^{(1)}(t) = P(T_n^{(1)} < t) = \begin{cases} 
0, & t \leq \frac{1}{\sqrt{n}}, \ n \geq 2; \\
\int_{\frac{1}{\sqrt{n}}}^{t} F_{n-1}^{(1)}(g_n(x)) f_{T_n}(x) dx, & \frac{1}{\sqrt{n}} < t \leq \frac{n-1}{\sqrt{n}}, \ n \geq 3; \\
1, & t > \frac{n-1}{\sqrt{n}}, \ n \geq 2;
\end{cases} \tag{2}
\]

where

\[
f_{T_n}(x) = \frac{1}{n - 1} \left[ \frac{n}{\pi} \Gamma \left( \frac{n - 1}{2} \right) / \Gamma \left( \frac{n - 2}{2} \right) \left( 1 - \frac{n}{(n - 1)^2} x^2 \right)^{\frac{n-1}{2}} \right] |x| < \frac{n-1}{\sqrt{n}}; \tag{3}
\]

\[
g_n(x) = \frac{n}{n - 1} x / \sqrt{\frac{n - 1}{n - 2} \left( 1 - \frac{n}{(n - 1)^2} x^2 \right)}, \ |x| < \frac{n-1}{\sqrt{n}}, \ n \geq 3. \tag{4}
\]

We find the joint distribution of Grubbs test statistics \( T_{n(1)} \) and \( T_n^{(1)} \) for a normal sample.
The joint distribution of Grubbs test statistics for a normal sample

Define
\[ T_i = \frac{X_i - \bar{X}}{S}, \quad (i = 1, 2, \ldots, n). \]

Random variables \( T_1, T_2, \ldots, T_n \) are identically distributed. Let \( f_{i:n}(x) \) be density function of studentized deviation \( T_i, \quad i = \bar{1}, n \), then [3]:

\[ f_{i:n}(x) = f_{\Delta_n}(x), \quad (i = 1, 2, \ldots, n), \]

where \( f_{\Delta_n}(x) \) can be calculated with using (3).

We have the following equalities for statistics \( T_n^{(1)} \) and \( T_{i,(1)} \):

\[ T_{i,(1)} = - \min_{1 \leq i \leq n} \{ T_i \}, \quad (5) \]
\[ T_n^{(1)} = \max_{1 \leq i \leq n} \{ T_i \}. \quad (6) \]

It follows from relationships (1) and (2) that

\[ \frac{1}{\sqrt{n}} \leq T_{i,(1)} \leq \frac{n-1}{\sqrt{n}}, \quad \frac{1}{\sqrt{n}} \leq T_n^{(1)} \leq \frac{n-1}{\sqrt{n}}. \quad (7) \]

Let \( \Lambda_n(t_1, t_2) = P(T_{n,(1)} < t_1, T_n^{(1)} < t_2) \) be the joint distribution function of Grubbs test statistics \( T_n^{(1)} \) and \( T_{i,(1)} \). The following theorem describes our main result.

**Theorem 1.** If \( X_1, X_2, \ldots, X_n \) is a random sample from a normal \( N(a, \sigma^2) \) distribution, then the joint distribution function of Grubbs test statistics \( T_n^{(1)} \) and \( T_{i,(1)} \) for the case \( n = 2 \) is given by

\[ \Lambda_2(t_1, t_2) = \begin{cases} 1, & (t_1, t_2) \in \Delta_2, \quad \Delta_2 = \left[ \frac{\sqrt{2}}{\sqrt{n}} < t_1 < \infty; \frac{\sqrt{2}}{\sqrt{n}} < t_2 < \infty \right]; \\ 0, & (t_1, t_2) \notin \Delta_2, \end{cases} \]

and for the case \( n > 2 \)

\[ \Lambda_n(t_1, t_2) = \begin{cases} F_n^{(1)}(t_2), & t_1 \geq \frac{n-1}{\sqrt{n}}, \\ F_n^{(1)}(t_1), & t_2 \geq \frac{n-1}{\sqrt{n}}, \\ \frac{n}{\sqrt{n}} \int_0^{t_2} \Lambda_{n-1}(\rho_n(t_1, -x), g_n(x)) f_{\Delta_n}(x) dx, & (t_1, t_2) \in \Delta_n; \\ 0, & (t_1, t_2) \notin \Delta_n, \quad t_1 < \frac{n-1}{\sqrt{n}}, t_2 < \frac{n-1}{\sqrt{n}}. \end{cases} \]

where distribution function \( F_n^{(1)}(t) \) can be calculated with using (2);

\[ \rho_n(u, v) = \left( u + \frac{v}{n-1} \right) / \sqrt{n-1} \left( 1 - \frac{n}{(n-1)^2} v^2 \right), \quad |v| < \frac{n-1}{\sqrt{n}}; \]

functions \( g_n(x) \) and \( f_{\Delta_n}(x) \) can be calculated with using (4) and (3) correspondingly; \( \Delta_n = [1/\sqrt{n} < t_1 < (n-1)/\sqrt{n}; 1/\sqrt{n} < t_2 < (n-1)/\sqrt{n}], \) if \( n > 2. \)
Proof. Suppose that $n = 2$. In that case $T_{2(1)} = T_2^{(1)} = \frac{1}{\sqrt{2}}$. We have $\Lambda_2(t_1, t_2) = P\{(1/\sqrt{2} < t_1) \cap (1/\sqrt{2} < t_2)\}$, that proves the formula (8).

Suppose that $n > 2$. Then for $t_1 \geq \frac{n-1}{\sqrt{n}}$ by (7) we have

$$\Lambda_n(t_1, t_2) = P\{(T_{n,1} < (n - 1)/\sqrt{n}) \cap (T_n^{(1)} < t_2)\} = P\left(T_n^{(1)} < t_2\right).$$

Therefore,

$$\Lambda_n(t_1, t_2) = F_n^{(1)}(t_2), \quad t_1 \geq \frac{n-1}{\sqrt{n}}, \quad (11)$$

where $F_n^{(1)}(t_2)$ is defined by (2).

Similarly, in the case $t_2 \geq (n - 1)/\sqrt{n}$ we have $\Lambda_n(t_1, t_2) = P\left(T_{n,1} < t_1\right)$. It follows from (1) that

$$\Lambda_n(t_1, t_2) = F_n^{(1)}(t_1), \quad t_2 \geq (n - 1)/\sqrt{n}. \quad (12)$$

Next, it follows from conditions (7), that

$$\Lambda_n(t_1, t_2) = 0, \quad \text{if } (t_1 \leq 1/\sqrt{n}) \text{ or } (t_2 \leq 1/\sqrt{n}). \quad (13)$$

Let be $\Delta_n = [1/\sqrt{n} < t_1 < (n - 1)/\sqrt{n}; 1/\sqrt{n} < t_2 < (n - 1)/\sqrt{n}]$. We have with use (5) and (6) that $\Lambda_n(t_1, t_2) = P\{\min_{1 \leq j \leq n} \{T_j\} > -t_1) \cap (\max_{1 \leq j \leq n} \{T_j\} < t_2)\}$. Hence, the joint distribution function of statistics $T_{n,1}$ and $T_n^{(1)}$ is defined by

$$\Lambda_n(t_1, t_2) = P\{\cap_{i=1}^n (-t_i < T_i < t_2)\}. \quad (14)$$

Next, using formulas (14), (6) and (7), we have for $(t_1, t_2) \in \Delta_n$

$$\Lambda_n(t_1, t_2) = P\left\{\cap_{i=1}^n (-t_i < T_i \leq \max_{1 \leq j \leq n} \{T_j\}) \cap \left( \frac{1}{\sqrt{n}} < \max_{1 \leq j \leq n} \{T_j\} < t_2\right)\right\}. \quad (15)$$

Then

$$\Lambda_n(t_1, t_2) = nP\left\{\cap_{i=1}^{n-1} (-t_i < T_i < T_n) \cap \left( \frac{1}{\sqrt{n}} < T_n < t_2\right)\right\}. \quad (15)$$

Define

$$T_i^* = \frac{X_i - \overline{X}^*}{S^*}, \quad (i = 1, 2, \ldots, n - 1),$$

where $\overline{X}^* = \frac{1}{n-1} \sum_{k \neq n} X_k$, $S^* = \frac{1}{n-2} \sum_{k \neq n} (X_k - \overline{X}^*)^2$.

It is valid that [3]

$$T_i = T_i^* \sqrt{\frac{n-2}{n-1} \left(1 - \frac{n}{(n-1)^2} T_n^2\right)} - \frac{1}{n-1} T_n, \quad i = 1, n-1, \quad |T_n| < \frac{n-1}{\sqrt{n}}.$$ 

Therefore the relationship (15) can be written as:

$$\Lambda_n(t_1, t_2) = nP\left\{\cap_{i=1}^{n-1} (\rho_n(-t_i, T_n) < T_i^* < g_n(T_n)) \cap (1/\sqrt{n} < T_n < t_2)\right\}, \quad (16)$$

179
where the function \( g_n(x) \) is defined by (4) and \( \rho_n(u,v) \) — by (10).

It is well known that random variable \( T_n \) is the function of \( \bar{X}^*, S^* \) and \( X_n \) [3]. Besides, random variables \( \{T^*_1, T^*_2, \ldots, T^*_{n-1}\} \) is independent of \( \{\bar{X}^*, S^*, X_n\} \). Hence, the random variable \( T_n \) is independent of \( \{T^*_1, T^*_2, \ldots, T^*_{n-1}\} \) [3]. Therefore the relationship (16) can be written as

\[
\Lambda_n(t_1, t_2) = n \int_{1/\sqrt{n}}^{t_2} \frac{1}{n-1} \{\rho_n(-t_1, x)\} f_{T_n}(x)dx.
\]

It follows from (14) that

\[
\Lambda_n(t_1, t_2) = n \int_{1/\sqrt{n}}^{t_2} \Lambda_{n-1}\{\rho_n(t_1, -x)\}, g_n(x)\} f_{T_n}(x)dx, \quad (t_1, t_2) \in \Delta_n.
\]

If we combine relationships (11), (12), (13) and (17), then we receive (9), that proves the theorem.

Note some properties of the joint distribution function \( \Lambda_n(t_1, t_2) \) which can be derived from (9).

1. Function \( \Lambda_n(t_1, t_2) \) is symmetrical, i.e. \( \Lambda_n(t_1, t_2) = \Lambda_n(t_2, t_1) \).

2. It is valid for \((t_1, t_2) \in D_n \) and \( n > 2 \):

\[
\Lambda_n(t_1, t_2) = F_n^{(1)}(t_1) - n \int_{t_2}^{t_1} F_n^{(1)}(\rho_n(t_1, -x)) f_{T_n}(x)dx,
\]

where \( D_n = [\frac{1}{\sqrt{n}} < t_1 < \frac{n-1}{\sqrt{n}}; \tau^*_n \leq t_2 < \frac{n-1}{\sqrt{n}}; \tau^*_n = \sqrt{\frac{(n-1)(n-2)}{2n}}] \).

3. It is valid for \((t_1, t_2) \in \Xi_n \) and \( n > 2 \):

\[
\Lambda_n(t_1, t_2) = F_n^{(1)}(t_1) + F_n^{(1)}(t_2) - 1,
\]

where \( \Xi_n = [t^*_n \leq t_1 < \frac{n-1}{\sqrt{n}}; t^*_n \leq t_2 < \frac{n-1}{\sqrt{n}}; t^*_n = \sqrt{\frac{n-1}{2}}] \).

\[
F_n^{(1)}(t) = 1 - n \int_{t}^{t_n} f_{T_n}(x)dx, \quad t \geq t^*_n.
\]

References


Abstract

The results of an experimental comparison of the accuracy of the classic Kalman filter and a simple non-causal smoother are presented, and a new version of the Kalman smoother, which does not need of a time lag, is described. The offered filter is based on a local approximation of a noisy trajectory by parametric curves. The state parameters of the filter are coefficients of the local regression that is used to build a priori and a posteriori estimates. Results of experimental comparison of the new filter with the linear Kalman one are given.

1 Introduction

The Kalman filter is now one of the most popular statistic tools to estimate noisy trajectories of nonstationary dynamic systems. It is applicable to virtually all areas of engineering and science. This includes electrical, mechanical, chemical and aerospace engineering, robotics, economics, ecology, biology, medicine etc. Applications here are limited only by the engineer’s imagination, which is why theory of Kalman filters has become such a widely researched and applied discipline in the past few decades [1, 2, 3].

From the one hand, the Kalman filter gives an optimal in the mean square sense estimate of the state vector of a dynamic system. From the other, it is one of the best causal filters. We tried to estimate experimentally how much it is better than simple non-causal filters and found for some elementary models its accuracy is only slightly better. After, a new version of the Kalman smoother has been described, which is visibly more precise than the classic Kalman linear filter in the case when trajectories of the system are smooth. Besides, it provides smoother estimates. The description of the filter is placed in Section 3.

Two statistical criteria were used to compare experimentally properties of the estimates. One of them evaluates the least square error of the estimators, the other measures their smoothness. Results of experiments are reflected in Section 2. As seen from Table 1 the offered modification of the filter provides smoother and more accurate recovered trajectories.
2 Naive Experimental Analysis of Accuracy of the Kalman Filter

To perform direct comparison and avoid the problem of tuning many parameters we use the classic linear Kalman filter without control

\[ \begin{align*}
    x_{k+1} &= Fx_k + w_k, \\
    y_{k+1} &= Hx_{k+1} + v_k,
\end{align*} \tag{1} \]

where, as usual, \( F \) is the state transition matrix, the matrix \( H \) specifies the observation model, random vectors \( w_k \) and \( v_k \) are independent \( (0, Q_k) \) and \( (0, R_k) \) Gaussian. This choice of the filter can be justified not only by its simplicity but also by its theoretical and experimental optimality for actually linear models [3, p.480]. Smooth functions \( f(x) \) of one variable distorted by the white noise, so that \( y_k = f_k + \xi_k \), were taken as measurements. They were recovered by the linear Kalman filter and by non-causal and causal moving average methods. The Kalman filter was either of the form

\[ \begin{align*}
    x_{k+1} &= x_k + \xi_k, \\
    y_{k+1} &= x_{k+1} + \eta_k, \quad k = 1 \div N,
\end{align*} \tag{2} \]

where the random variables \( \xi_k \) and \( \eta_k \) are independent \( (0, \sigma^2_\xi) \) and \( (0, \sigma^2_\eta) \) Gaussian or of the vector form (1) with vector-column \( x_k = (x_k, \dot{x}_k, \ddot{x}_k) \) and matrices

\[ F = \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \quad \text{cov}(w_k) = \sigma^2_w I, \quad \text{cov}(v_k) = \sigma^2_v I \tag{3} \]

with \( a, b, c > 0 \) and \( 3 \times 3 \) identity matrix \( I \).

Two criteria were taken to compare accuracy of the filters. They are the estimator least square error and the average \( l_1 \)-distance between derivatives of the estimator and of the original function

\[ \Delta = \sqrt{\frac{1}{N} \sum (f_k - \hat{f}_k)^2}, \quad \delta = \frac{1}{N} \sum |\hat{f}_k - \hat{f}_{k-1} - f_k + f_{k-1}|, \]

where common notation \( \hat{f}_k \) means either the Kalman \( f_{Kalman,k} \) or the moving average \( f_{ME,k} \) estimator. The second criterion rates the smoothness of estimators. Note that \( \Delta = \Delta(\sigma) \), \( \delta = \delta(\sigma) \) for the 1-st Kalman model and \( \Delta = \Delta(\sigma, a, b, c) \), \( \delta = \delta(\sigma, a, b, c) \) for the 2-nd one.

In Table 1 results of experiments with \( \sigma^2_\eta = \sigma^2_v = 4 \) are shown. The Kalman \( 1 \times 1 \) and \( 3 \times 3 \) filters are specified by equations (2) and (1),(3). The variances \( \sigma^2_\eta \) and \( \sigma^2_w \) have been found to minimize the criterion \( \Delta \).

Values of the smoothness criterion \( \delta \) for the optimal \( \sigma^2_\eta \) and \( \sigma^2_w \) were computed and placed in second cell rows of Table 1. The last column contains characteristics of the offered Kalman smoother. The size of kernels of the moving average has been restricted by 401 counts but for the causal version of this filter the optimal size of kernels has never exceeded 11. It can be seen that for smooth slowly varying trajectories the linear Kalman filter is only a bit more accurate than the moving average and therefore in order to get more accurate results other its versions ought to be applied.
Table 1: Results of comparison of Kalman and moving average filters

<table>
<thead>
<tr>
<th>Function</th>
<th>Criteria</th>
<th>Kalman 1 x 1 filter</th>
<th>Kalman 3 x 3 filter</th>
<th>MovAv non-causal</th>
<th>MovAv causal</th>
<th>Kalman Smoother</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>[\Delta_{\text{min}}]</td>
<td>1.230</td>
<td>0.704</td>
<td>0.137</td>
<td>1.549</td>
<td>0.191</td>
</tr>
<tr>
<td></td>
<td>[\delta]</td>
<td>0.766</td>
<td>0.161</td>
<td>0.8713</td>
<td>0.8713</td>
<td>0.010</td>
</tr>
<tr>
<td>[\sqrt{x}]</td>
<td>[\Delta_{\text{min}}]</td>
<td>0.516</td>
<td>0.395</td>
<td>0.130</td>
<td>0.454</td>
<td>0.319</td>
</tr>
<tr>
<td></td>
<td>[\delta]</td>
<td>0.168</td>
<td>0.057</td>
<td>1.988</td>
<td>1.988</td>
<td>0.033</td>
</tr>
<tr>
<td>sin([\pi x/50])</td>
<td>[\Delta_{\text{min}}]</td>
<td>0.757</td>
<td>0.889</td>
<td>0.405</td>
<td>0.739</td>
<td>0.709</td>
</tr>
<tr>
<td></td>
<td>[\delta]</td>
<td>0.273</td>
<td>0.309</td>
<td>0.452</td>
<td>0.452</td>
<td>0.175</td>
</tr>
</tbody>
</table>

3 Kalman Smoother

Evidently smoother trajectories are preferable in many applications. In order to construct them special Kalman filters have been built, which is called Kalman smoothers [3, p.263]. They provide not only smoother estimates, but also possess greater accuracy. However their drawback is some time lag of resulting trajectories.

We describe a new modification of the Kalman smoother that operates without time lag. The state parameter of the smoother is the vector of coefficients of a local regression curve that approximates a few latest observations of the noisy trajectory. This vector is present in the measurement equation of the smoother and specifies a current value of the estimator.

In more detail. Let \(d + 1\) be number of latest observations to build a current local regression curve, and for the current discrete time \(k\) the vector-column of observations of the noisy trajectory is of the form \(y_k = (y_k, y_{k-1}, ..., y_{k-d})\), where \(y_j, j = k \div (k - d)\) are values of the noisy trajectory. Let also \(a_k = (a_{0,k}, a_{1,k}, ..., a_{p,k})\) be coefficients of the regression curve to approximate the vector \(y_k\). Then the Kalman filter with the polynomial regression smoothing is written as

\[
\begin{align*}
    a_{k+1} &= F a_k + w_k, \\
    y_{k+1} &= H a_{k+1} + v_k
\end{align*}
\]

with matrices

\[
F = I_{(p+1) \times (p+1)}, \quad \text{and } H_{(d+1) \times (p+1)} = \begin{pmatrix}
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
(-j)^p & (-j)^{p-1} & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
(-d)^p & (-d)^{p-1} & \cdots & 1
\end{pmatrix}
\]

in the simplest case.

A similar construction has been described in [1, p.155] to build the Fix Memory Polynomial Filter, however, it is only a particular case of the filter (4) without the first equation.
Our reasons to construct the filter were, first, to exploit a local regression for building the a priori estimate in order to improve the smoothness of recovered trajectories but avoid the time lag, and, second, to use geometric characteristics, such as the slope angle and the form of the regression curve, for tuning the filter. These reasons can be implemented by means of an appropriate choice of the matrix $H$ and the covariance matrix $Q_k$. Unlike the classical linear filter in many cases it converges to the model function $f(x)$, when $Q_k \equiv 0$. Results of experiments can be seen in the last columns of Table 1 and in Figure 1.

![Figure 1: The original function $f(x) = x$ sampled with the (0, 4)-Gaussian noise and the trajectory recovered by offered smoother.](image)

In the future we are planning to study properties of the Kalman smoother with non-polynomial regression curves and other types of state and covariance matrices.

References


ASSIGNMENT OF MULTIVARIATE SAMPLES TO THE FIXED CLASSES BY THE MAXIMUM LIKELIHOOD METHOD AND ITS RISK

E.E. Zhuk
Belarusian State University
Minsk, BELARUS
e-mail: zhukee@mail.ru

Abstract
The problem of statistical assignment of samples of identically distributed multivariate observations to the classes determined by their probability distributions is considered. The decision rule by the maximum likelihood method is proposed and its risk is analytically evaluated. The obtained results are illustrated for the well known Fisher model.

1 Introduction: mathematical model and problem of assignment of multivariate samples
Let \( L \geq 2 \) classes \( \{\Omega_1, \ldots, \Omega_L\} \) be determined in the observation space \( R^N (N \geq 1) \). According to the traditional mathematical model [2, 3] the class \( \Omega_i \) contains identically distributed random observations \( x \in R^N \) described by the probability density function:

\[
p_i(x) \geq 0, \quad x \in R^N : \quad \int_{R^N} p_i(x)dx = 1, \tag{1}
\]

which determines \( \Omega_i \) and is called a conditional density [2] of the class \( \Omega_i \) \((i \in S, S = \{1, \ldots, L\} \) is the set of class indices).

Usually the statistical classification problem consists in the construction of the decision rule (DR) \([2, 3]\): \( d = d(x) : R^N \rightarrow S \), which is a statistical estimator for unknown class index \( d^0 \in S \) of observation \( x \in R^N \). It is supposed that classified observation belongs to one of the classes \( \{\Omega_i\}_{i \in S} \) and is described by one of the densities (1).

But, often in practice \([2, 4]\) we need to assign to the classes \( \{\Omega_i\}_{i \in S} \) observations with a density

\[
p(x) \geq 0, \quad x \in R^N : \quad \int_{R^N} p(x)dx = 1, \tag{2}
\]

which is distinguished from the densities (1).

In this paper the following more common problem is investigated. Let a random sample \( X = \{x_1, \ldots, x_n\} \) of size \( n \) consists of independent in total random observations \( x_t \in R^N, t = 1, n \), described by some density (2). The problem consists in assignment of this sample to one of the classes \( \{\Omega_i\}_{i \in S} \) determined by the densities (1).
2 Maximum likelihood method and its risk

To solve the assignment problem the maximum likelihood method can be used [1, 2, 3]:

\[ D(X) = \arg \max_{i \in S} P_i(X); \quad (3) \]

\[ P_i(X) = \prod_{t=1}^{n} p_i(x_t), \quad i \in S, \]

where the maximum likelihood DR (MLDR) (3): \( D = D(X) : R^{nN} \to S \), assigns the sample \( X \) to the class with index \( D(X) \in S \). The functions \( \{ P_i(X) \}_{i \in S} \) from (3) are the conditional likelihood functions [2, 3], which mean conditional probability densities of sample \( X \) evaluated for the classes \( \{ \Omega_i \}_{i \in S} \).

Note, the unconditional (“real”) likelihood function of the sample \( X \) is

\[ P(X) = \prod_{t=1}^{n} p(x_t), \]

where \( p(\cdot) \) is the density (2).

Now let us define the generalization of the traditional risk [2, 3] as an efficiency measure of the MLDR (3). Introduce the following notations: \( \Omega_0 \) is the additional \((L + 1)\)-th class which contains observations from the sample \( X \) and is determined by the density (2) \( (S_0 = \{0\} \cup S = \{0, 1, \ldots, L\} \) is the extended set of class indices);

\[ R_{ij} = R(\Omega_i, \Omega_j) \geq 0, \quad i, j \in S_0, \quad (4) \]

are any interclass distances [2] \( (R_{ij} \) is the nearness measure for the classes \( \Omega_i \) and \( \Omega_j \)); \( U(z) = \{1, z \geq 0; 0, z < 0\} \) is the unit function.

The risk of the MLDR (3) means the maximum probability to do not assign the sample \( X \) to one of the classes from \( \{ \Omega_i \}_{i \in S} \) which are the most close to the class \( \Omega_0 \):

\[ r = 1 - \min_{i \in I} P^{(i)}, \quad I = \{k : R_{0k} = \min_{j \in S} R_{0j}\}, \quad (5) \]

where

\[ P^{(i)} = P\{D(X) = i\} = \int_{R^{nN}} \prod_{k \in S} U(P_i(X) - P_k(X)) P(X) dX, \quad i \in S, \quad (6) \]

mean the probabilities to assign by the MLDR (3) the sample to the classes \( \{ \Omega_i \}_{i \in S} \).

If all distances \( \{ R_{0k} \}_{i \in S} \) between the class \( \Omega_0 \) and the classes \( \{ \Omega_i \}_{i \in S} \) are distinguished then the risk (5) is simplified:

\[ r = 1 - P^{(k)}, \quad k = \arg \min_{j \in S} R_{0j}. \quad (7) \]

The smaller is the risk \( r \) from (5) or (7) \( (0 \leq r \leq 1) \), the more efficiency is the MLDR (3) (the greater is the probability to assign by the MLDR (3) the sample \( X \) to the most closed class).
Note, if the sample \( X \) contains observations from one of the classes \( \{ \Omega_i \}_{i \in S} \) then
the assignment problem is the so-called problem of group classification [1] and the risk
(5), (6) means the conditional error probability:
\[
r = 1 - P(i) = P\{D(X) \neq i | d^0 = i\}, \quad i \in S,
\]
where \( d^0 = i \) is the class index of observations from the sample \( X \) \((p(\cdot) \equiv p_i(\cdot), P(\cdot) \equiv P_i(\cdot))\).

3 The case of the Fisher model

Now let us consider the often meeting in applications Fisher model [2, 3], when the
conditional densities of the classes \( \{ \Omega_i \}_{i \in S} \) from (1) are supposed multivariate normal
(Gaussian):
\[
p_i(x) = n_N(x | \mu_i, \Sigma) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left( -\frac{1}{2} (x - \mu_i)^t \Sigma^{-1} (x - \mu_i) \right), \quad x \in \mathbb{R}^N, \quad i \in S,
\]
where
\[
\mu_i = \mathbb{E}\{x | d^0 = i\} = \int_{\mathbb{R}^N} x p_i(x) dx, \quad i \in S,
\]
are the conditional mathematical mean \( N \)-vectors (the “centers” [2, 3] of the classes
\( \{ \Omega_i \}_{i \in S} \)) and
\[
\Sigma = \mathbb{E}\{(x - \mu_i)(x - \mu_i)^t | d^0 = i\}, \quad i \in S,
\]
is the common for all classes \( \{ \Omega_i \}_{i \in S} \) non-singular covariance \((N \times N)\)-matrix (“\( \rho \) is
the transposition symbol).

The observations from the sample \( X = \{x_t\}_{t=1}^n \) are supposed to be normal too:
\[
p(x) = n_N(x | \mu, \Sigma), \quad x \in \mathbb{R}^N,
\]
where
\[
\mu = \mu_0 = \int_{\mathbb{R}^N} x p(x) dx \in \mathbb{R}^N
\]
is their mathematical mean vector (the “center” of the \((L + 1)\)-th class \( \Omega_0 \)) and \( \Sigma \) is
their covariance matrix: \( \int_{\mathbb{R}^N} (x - \mu)(x - \mu)^t p(x) dx = \Sigma \).

Under the Fisher model (9) as the interclass distances (4) let us define the Mahalanobis
distances between the class “centers” [2, 3]:
\[
R_{ij} = R(\Omega_i, \Omega_j) = \rho(\mu_i, \mu_j), \quad i, j \in S_0,
\]
where
\[
\rho(y, z) = \sqrt{(y - z)^t \Sigma^{-1} (y - z)}, \quad y, z \in \mathbb{R}^N,
\]
is the Mahalanobis metric in \( \mathbb{R}^N \).
Theorem. Under the Fisher model (9) the MLDR (3) takes the form:

\[ D(X) = \arg \min_{i \in S} \rho(\bar{x}, \mu_i), \quad X \in \mathbb{R}^{nN}, \tag{13} \]

where \( \bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t \) is the sample average.

If observations from the sample \( X = \{x_t\}_{t=1}^{n} \) are independent in total with the density (10) and the Mahalanobis interclass distances (11), (12) are used then the risk (5), (6)

\[ r = 1 - \min_{i \in I} P^{(i)} \quad I = \{k : \rho(\mu, \mu_k) = \min_{j \in S} \rho(\mu, \mu_j)\}; \]

\[ P^{(i)} = \int_{\mathbb{R}^N} \prod_{k \in S, k \neq i} U(\rho(x, \mu_k) - \rho(x, \mu_i)) n_N(x|\mu, \frac{1}{n} \Sigma) \, dx, \quad i \in S. \]

Note that the MLDR (13) obtained under the Fisher model is also known as the L-means DR [2, 3].

Corollary. Under the conditions of the theorem in the case of two classes (\( L = 2 \)):

\[ r = \Phi \left( -\sqrt{n} \frac{\rho^2(\mu, \mu_1) - \rho^2(\mu, \mu_2)}{2\rho(\mu_1, \mu_2)} \right), \tag{14} \]

where \( \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp \left( -\frac{w^2}{2} \right) dw \), \( z \in \mathbb{R} \) is the standard Gaussian distribution function.

From the relation (14) it is seen that the risk \( r \) has the maximum (“worst”) value when \( \rho(\mu, \mu_1) = \rho(\mu, \mu_2) \): \( r = \Phi(0) = 1/2 \). If observations from the sample \( X \) belong to one of the classes \( \Omega_1 \) or \( \Omega_2 \) (\( \mu = \mu_1 \) or \( \mu = \mu_2 \)) then the risk (14) is the conditional error probability (8):

\[ r = \Phi \left( -\sqrt{n} \frac{\rho(\mu_1, \mu_2)}{2} \right), \]

where \( \rho(\mu_1, \mu_2) = \sqrt{(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)} \) is the Mahalanobis interclass distance between the classes \( \Omega_1 \) and \( \Omega_2 \).

References


STATISTICAL ANALYSIS OF TIME SERIES AND STOCHASTIC PROCESSES
ON PARAMETERS ESTIMATION OF AUTOREGRESSIVE TIME SERIES UNDER RIGHT CENSORING

I.A. Badziahin
Belarussian State University
Minsk, BELARUS

Abstract

Autoregressive time series observed under right censoring are considered. Statistical estimators of autoregression model parameters are constructed by using the method of moments for special auxiliary time series. Consistency of constructed estimators are proved under some additional general conditions.

1 Introduction

In practise time series are often observed under different distortions of hypothetical model such as outliers, missing values, heteroscedasticity, censoring [2–5]. Because of these distortions classical estimators of the hypothetical model parameters are usually biased and inconsistent [2]. That is why it is necessary to construct new methods for estimation of model parameters robust to certain types of distortions.

In this paper we propose a new method of parameters estimation for autoregressive time series observed under right censoring [4, 5]. This method is based on using the method of moments for special auxiliary time series. Right censoring means that exact values of all observations that are above the certain censoring level are unknown, and it is only known that true values of the censored observations are greater than the censoring level [3, 4].

2 Mathematical model

Consider the autoregressive time series model (AR(p)) defined on probability space $(\Omega, \mathcal{F}, P)$ [1]:

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + u_t, t \in \mathbb{Z},$$

where $\theta_1, \ldots, \theta_p$ are unknown coefficients of the autoregression and all roots of the characteristic polynomial $z^p - \sum_{i=1}^{p} \theta_i z^{p-i}$ are inside the unit circle; $\{u_t\}_{t \in \mathbb{Z}}$ are i.i.d. normal random variables, $E\{u_t\} = 0, \ D\{u_t\} = \sigma^2 < +\infty$.

The autoregressive time series $x_t$ is observed under right censoring. It means that all values exceeded the censoring level $c \in \mathbb{R}$ are not observed, they are censored. Introduce the notation: $T$ — the number of observation time moments; $T_C = \{t : x_t > c\} \subseteq$
\{1, \ldots, T\} — the subset of censoring time moments; \(M = |T_C|\) — the number of censored observations. In the given notation right censoring means that instead of true values \(x_t, \ t \in T_C\) we observe only the following random events:

\[
A^*_t = \{x_t \in A\} \in \mathcal{F}, \ A = [0, +\infty), \ t \in T_C.
\]

The problem considered in this paper is to estimate the autoregressive time series model parameters \((\theta_1, \ldots, \theta_p, \sigma^2)\) when some observations are censored (2).

### 3 Moments of auxiliary time series for right censored AR\((p)\)

Define the auxiliary time series for the right censored AR\((p)\) [4]:

\[
y_t = f_c(x_t) = \begin{cases} x_t, & t \in \{1, \ldots, T\} \setminus T_C, \\ c, & t \in T_C \end{cases} = \min\{x_t, c\} = x_t + (c - x_t)\mathbf{I}_A(x_t),
\]

where \(\mathbf{I}_A(\cdot)\) is the indicator function.

Introduce the notation to simplify further calculations:

\[
m = E\{y_t\}, \ m_\tau = E\{y_t y_{t+\tau}\},
\]

\[
\sigma_\tau = E\{x_t x_{t+\tau}\}, \ \rho_\tau = \frac{\sigma_\tau}{\sigma_0}, \ D_\tau = \sigma_0 - \frac{\sigma_\tau^2}{\sigma_0},
\]

\[
I_{k,\tau} = E\{x_k^\tau \mathbf{I}_A(x_t)\mathbf{I}_A(x_{t+\tau})\}, \ k = 0, 1, 2, \ t, \tau \in \mathbb{Z},
\]

\[
\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \ \Phi(x) = \int_{-\infty}^{x} \varphi(t)dt.
\]

**Theorem 1.** Let \(y_t\) be the auxiliary time series (3) for the autoregressive time series (1), then its moments \(m, m_\tau\) defined by (4) are:

\[
m = c\Phi\left(\frac{-c}{\sqrt{\sigma_0}}\right) - \sqrt{\sigma_0} \varphi\left(\frac{c}{\sqrt{\sigma_0}}\right),
\]

\[
m_\tau = \sigma_\tau - 2\rho_\tau \sigma_0 \Phi\left(\frac{-c}{\sqrt{\sigma_0}}\right) + D_\tau \frac{\sqrt{D_\tau}}{\sqrt{\sigma_0}} \varphi\left(\frac{c}{\sqrt{\sigma_0}}\right) \varphi\left(\frac{c(1 - \rho_\tau)}{\sqrt{D_\tau}}\right) +
\]

\[
+ D_\tau \frac{\rho_\tau c}{\sqrt{\sigma_0}} \varphi\left(\frac{c}{\sqrt{\sigma_0}}\right) \Phi\left(\frac{c(\rho_\tau - 1)}{\sqrt{D_\tau}}\right) + c^2 I_{1,\tau} - 2c I_{1,\tau} + \rho_\tau I_{2,\tau},
\]

where \(I_{1,\tau} = \frac{D_\tau}{(1 - \rho_\tau)\sqrt{\sigma_0}} \varphi\left(\frac{c}{\sqrt{\sigma_0}}\right) \Phi\left(\frac{c(\rho_\tau - 1)}{\sqrt{D_\tau}}\right), \ I_{2,\tau} = \frac{D_\tau}{1 - \rho_\tau} \left(\rho_\tau \frac{\sqrt{D_\tau}}{\sqrt{\sigma_0}} \varphi\left(\frac{c}{\sqrt{\sigma_0}}\right) \varphi\left(\frac{c(1 - \rho_\tau)}{\sqrt{D_\tau}}\right) +
\]

\[
+ (1 + \rho_\tau^2) \frac{c}{\sqrt{\sigma_0}} \varphi\left(\frac{c}{\sqrt{\sigma_0}}\right) \Phi\left(\frac{c(\rho_\tau - 1)}{\sqrt{D_\tau}}\right) + I_{0,\tau}\right).
\]

Note that the mathematical expectation \(I_{0,\tau}\) defined by (5) can be computed numerically.
4 Algorithm of the parameters estimation

Algorithm of the proposed method is the following.

Step 1. To compute sample moments for \( \{ y_1, \ldots, y_T \} \):

\[
\hat{m} = \frac{1}{T} \sum_{t=1}^{T} y_t, \quad \hat{m}_\tau = \frac{1}{T - \tau} \sum_{t=1}^{T-\tau} y_t y_{t+\tau}, \quad \tau = 1, \ldots, p.
\]

Step 2. To find MM estimates \((\hat{\sigma}_0, \hat{\sigma}_1, \ldots, \hat{\sigma}_p)\) as the solution of the system of nonlinear equations by using (6), (7):

\[
\begin{align*}
    m(\sigma_0) &= \hat{m}, \\
    m_\tau(\sigma_0, \sigma_\tau) &= \hat{m}_\tau, \quad \tau = 1, \ldots, p.
\end{align*}
\]

Step 3. To compute estimates \((\hat{\theta}_1, \ldots, \hat{\theta}_p, \hat{\sigma}^2)\) as the solution of the Yule-Walker system of linear algebraic equations [1]:

\[
\begin{align*}
    \hat{\sigma}_0 &= \sum_{i=1}^{p} \theta_i \hat{\sigma}_i + \sigma^2, \\
    \hat{\sigma}_\tau &= \sum_{i=1}^{p} \theta_i \hat{\sigma}_{|\tau-i|}, \quad \tau = 1, \ldots, p.
\end{align*}
\]

**Theorem 2.** Let the parameters of the underlying AR(p) and the level of censoring \(c\) satisfy the condition:

\[
\sum_{u=-\infty}^{+\infty} \left| \int_{\mathbb{R}^4} f_c(x_1) f_c(x_2) f_c(x_3) f_c(x_4) n_4(x \mid 0_4, \Sigma_4(\tau, u)) \, dx \right| < +\infty,
\]

where the matrix \(\Sigma_4(\tau, u)\) is

\[
\begin{pmatrix}
    \sigma_0 & \sigma_\tau & \sigma_u & \sigma_{\tau+u} \\
    \sigma_\tau & \sigma_0 & \sigma_{\tau-u} & \sigma_\tau \\
    \sigma_u & \sigma_{\tau-u} & \sigma_0 & \sigma_u \\
    \sigma_{\tau+u} & \sigma_\tau & \sigma_u & \sigma_0
\end{pmatrix}.
\]

Then the MM estimators \((\hat{\theta}_1, \ldots, \hat{\theta}_p, \hat{\sigma}^2)\) are consistent:

\[
\left(\hat{\theta}_1, \ldots, \hat{\theta}_p, \hat{\sigma}^2\right) \xrightarrow{P} (\theta_1, \ldots, \theta_p, \sigma^2).
\]

5 Computer modeling

To illustrate the proposed method computer experiments are performed. The experiment for given \(T \in \mathbb{N}\) consists in the following: 1) the autoregressive time series of the length \(T\) is generated with parameters: \(p = 1, \theta_1 = 0.6, \sigma^2 = 1\); and all observations \(x_i \geq c\) (\(c = 0\)) are considered as censored; 2) the model parameters estimators \((\hat{\theta}_1, \hat{\sigma}^2)\) are constructed by using the proposed method; 3) the empirical mean and variance of
constructed estimators are found by using the Monte-Carlo method with number of replications $N = 10^3$. The length of the time series $T \in \{100, 200, \ldots, 5000\}$.

Results of the experiments for $\hat{\theta}_1$ are illustrated on the figure 1. As we can see empirical mean of the constructed estimator $\hat{\theta}_1$ tends to the exact value of this parameter ($\theta_1 = 0.6$). So the constructed estimator looks like asymptotically unbiased. Also empirical variance of the constructed estimator $\hat{\theta}_1$ tends to zero and this result is agreed with consistency of the constructed estimator shown in Theorem 2.

![Figure 1: Dependence of empirical mean and variance on time series length, $c = 0$](image)

References


ON QUALITATIVELY ROBUST SIGN TEST IN RANDOM WALK MODEL

M.V. BOLDIN
Moscow State University, Faculty of Mechanics and Mathematics
Moscow, Russia
e-mail: boldin_m@hotmail.com

Abstract

Robustness of the nonparametric sign test in AR(1) with a unit root against outliers is studied. We consider the local scheme of data contamination by independent additive outliers with intensity $O(n^{-1/2})$. The qualitative robustness of the test in term of power equicontinuity is established.

1 Introduction

We consider the nonstationary $AR(1)$ model

$$u_t = \beta u_{t-1} + \varepsilon_t, \ t = 1, \ldots, n; \ u_0 = 0. \ (1)$$

In (1), $\beta$ is an unknown parameter, $\beta \in R^1$, $\{\varepsilon_t\}$ are i.i.d. random variables with an unknown distribution function (d.f.) $G(x)$, satisfying following condition:

Condition (i).

$P(\varepsilon_1 < 0) = P(\varepsilon_1 > 0) = 1/2$.

We will consider the situation when the autoregression observations contain additive outliers, so that the observed variables $y_0, y_1, \ldots, y_n$ have the form

$$y_t = u_t + \gamma^* \xi_t, \ t = 0, 1, \ldots, n, \ n \geq 2. \ (2)$$

In (2) $\{u_t\}$ is a sample from (1); $\{\gamma^* \xi_t\}$ are i.i.d. Bernoulli random variables with parameter

$$\gamma_n = \min(n^{-1/2} \gamma, 1), \ \gamma \geq 0, \ \gamma \text{ is unknown};$$

$\{\xi_t\}$ are i.i.d. random variables with unknown and arbitrary distribution $\mu$; the sequences $\{u_t\}, \{\gamma^* \xi_t\}, \{\xi_t\}$ are assumed to be mutually independent.

Random variables $\{\xi_t\}$ play the role of outliers or contaminations, $\gamma_n$ is the contamination level. The contamination schemes similar to (2) but with contamination level independent of $n$ are often used in time series analysis, see, for example, [4]. We consider the local version of such schemes.

Using $\{y_t\}$, we will test the hypothesis $H_0 : \beta = 1$ against left-sided alternative $\beta \in (-1, 1)$.

Consequently, $H_0$ is the random walk hypothesis for $\{u_t\}$ and the alternative means the stationarity. This problem is very famous and has a long history. It arises, for example, in econometrics, financial statistics, engineering. It is assumed usually that autoregression observations have not outliers and test statistics are based on least
square (LS) estimators. See, for example, famous paper [3] where \( \{\varepsilon_t\} \) form a square-integrable martingale difference and the asymptotic distribution of LS statistic is found under \( H_0 \) and close alternatives

\[
H_{1n}(\tau) : \beta = \beta_n := 1 + n^{-1}\tau, \tau \in R^1.
\]

Further we will consider \( \tau \leq 0 \), then \( H_{1n}(0) \) is \( H_0 \).

Unfortunately, LS procedures are very sensitive to gross outliers in scheme (2). Pathological situations when significance level for some \( \mu \) is close to one for any small \( \gamma > 0 \) are possible for them.

Our goal is to propose the sign test for \( H_0 \) in scheme (2) which has not such defects. We will use the results of [1]. In this monograph the locally optimal nonparametric sign test was proposed in scheme (1) without contaminations for testing general hypothesis \( \beta = \beta_0, \beta_0 \in R^1 \). Now we can form this test in scheme (2).

Namely, let

\[
S_k(\beta_0) = \text{sgn}(y_k - \beta_0 y_{k-1}), \ k = 1, \ldots, n;
\]

\[
\Gamma_{tn}(\beta_0) = (n - t)^{-1} \sum_{k=t+1}^{n} S_{k-t}(\beta_0)S_k(\beta_0), \ t = 1, \ldots, n - 1;
\]

\[
T_n(\beta_0) = \sum_{t=1}^{n-1} \beta_0^{t-1}(n - t)\Gamma_{tn}(\beta_0).
\]

If \( \gamma = 0 \), then the test for hypothesis \( \beta = \beta_0 \) with statistic \( T_n(\beta_0) \) is the locally most powerful among tests based on signs \( \{S_t(\beta_0)\} \) against one-sided alternatives. See §4.3 in [1].

We are interested in the statistic \( T_n(1) \). This statistic is continuous and increasing function from \( |Z_n| \), where

\[
Z_n := n^{-1/2} \sum_{t=1}^{n} S_t(1)
\]

(Lemma 1 in Section 2). Therefore we will study the test with statistic \( |Z_n| \).

Choose the critical region such that the asymptotic level is \( \alpha \) if \( \gamma = 0 \), see Section 2. Denote the local power of the test under \( H_{1n}(\tau) \) in scheme (2) for finite \( n \) by \( W_n(\tau, \gamma, \mu) \). First we find the stochastic expansion of the test statistic \( Z_n \) as \( n \to \infty \), which is uniform over \( 0 \leq \gamma \leq \Gamma < \infty \) and arbitrary \( \mu \). This implies uniform convergence in distribution of the test statistic \( Z_n \) (Theorem 1 in Section 2). Theorem 1 implies uniform convergence of \( W_n(\tau, \gamma, \mu) \) to the limiting power \( W(\tau, \gamma, \mu) \) (Corollary 1). This enables us to establish equicontinuity in \( \gamma \) at \( \gamma = 0 \) of the family \( \{W_n(\tau, \gamma, \mu)\} \) (Theorem 2). We interpret this fact as the qualitative robustness of the sign test in the local scheme (2) under the hypothesis and close alternatives. We call it local qualitative robustness.

Theorem 2 is a new fact. Previously the local qualitative robustness was established only for sign test in stationary \( AR(1) \) for hypothesis \( \beta = \beta_0, |\beta_0| < 1 \), see [2].

It is worth mentioning that commonly used LS tests do not possess local qualitative robustness property.
Main results

Lemma 1  Let $G(x)$ is continuous. Let $Z_n := n^{-1/2} \sum_{t=1}^{n} S_t(1)$. Then a.s.
$$T_n(1) = (n/2)(Z_n^2 - 1).$$

By Lemma 1 tests with statistics $T_n(1)$ and $|Z_n|$ are equivalent. We will base the test on $|Z_n|$. We need to know the limit distribution of $Z_n$ under $H_{1n}(\tau)$.

Condition (ii).
$$E_{\varepsilon_1} = 0, \ E_{\varepsilon_1^2} < \infty.$$

Condition (iii).
D.f. $G(x)$ is twice differentiable with $g(x) := G'(x), g(0) > 0, \sup_{x} |g'(x)| < \infty$.
Let for $\tau < 0$
$$\sigma^2(\tau) := 1 - 4g(0)E[\varepsilon_1][1 + \frac{1 - e^\tau}{\tau}] + 4g^2(0)E[\varepsilon_1^2][1 + 2\frac{1 - e^\tau}{\tau} - \frac{1 - e^{2\tau}}{2\tau}].$$

For $\tau = 0$ define $\sigma^2(0) := 1$.

Let $\nu$ is a standard normal random variable. Let
$$\delta(\mu) := E[G(\xi_1) + G(-\xi_1)] - 1.$$

For symmetric and continuous $G(x)$, of course, $\delta(\mu) = 0$.

Theorem 1 Let Conditions (i)–(iii) be satisfied. Let $H_{1n}(\tau), \tau \leq 0$, be true. Let $0 \leq \Gamma < \infty$. Then:

1. In probability
$$Z_n - n^{-1/2} \sum_{t=1}^{n} \text{sgn} \varepsilon_t + 2g(0)n^{-1/2} \sum_{t=1}^{n} [1 - (1 + \frac{\tau}{n})^{n-t}] \varepsilon_t +$$
$$2\gamma n^{-1/2} \sum_{t=1}^{n} [G(z^\gamma_{t-1} \xi_{t-1} - z^\gamma_t \xi_t) - G(0)] \to 0, \text{ as } n \to \infty;$$

2. In distribution
$$Z_n \to \sigma(\tau)\nu - 2\gamma \delta(\mu), \text{ as } n \to \infty.$$ (4)

The convergence in probability in (3) and the convergence in distribution in (4) are uniform over $0 \leq \gamma \leq \Gamma$ and arbitrary $\mu$.

The proof of Theorem 1 is similar to the proof of Theorem 3.1 in [2].

The critical region for $H_0$ against left-sided alternative has the form
$$|Z_n| < \text{const.}$$ (5)

Choose the constant in (5) is equal to $(1+\alpha)/2$--quantile of the standard normal d.f. $\Phi(x)$. Then the test (5) has asymptotical size $\alpha$ in scheme (1) without contaminations because by Theorem 1 under $H_0$ and $\gamma = 0$ in distribution
$$Z_n \to \nu, \text{ as } n \to \infty.$$
Denote the power of such test under $H_{1n}(\tau)$

$$W_n(\tau, \gamma, \mu) := P_{\beta_n}(|Z_n| < t_{(1+\alpha)/2}).$$

Let

$$W(\tau, \gamma, \mu) := P(|\sigma(\tau)\nu - 2\gamma\delta(\mu)| < t_{(1+\alpha)/2}) =$$

$$\Phi(\sigma^{-1}(\tau)(t_{(1+\alpha)/2} + 2\gamma\delta(\mu))) - \Phi(\sigma^{-1}(\tau)(-t_{(1+\alpha)/2} + 2\gamma\delta(\mu))).$$

Theorem 1 implies

**Corollary 1** Let conditions of Theorem 1 be satisfied. Then uniformly over $0 \leq \gamma \leq \Gamma$ and arbitrary $\mu$

$$W_n(\tau, \gamma, \mu) \rightarrow W(\tau, \gamma, \mu), \text{ as } n \rightarrow \infty.$$  

Hence, $W(\tau, \gamma, \mu)$ is the asymptotic power of sign test in scheme (2).

Let

$$W_n(\tau) := W_n(\tau, 0, \mu)$$

be the power of sign test in scheme (1).

The proof of next Theorem 2 is similar to the proof of Theorem 3.2 in [2] and is based, in particular, on Corollary 1.

**Theorem 2** Let Conditions (i)–(iii) be satisfied. Let $H_{1n}(\tau), \tau \leq 0$, be true. Then

$$\sup_{n \geq 2, \mu} |W_n(\tau, \gamma, \mu) - W_n(\tau)| \rightarrow 0, \text{ as } \gamma \rightarrow 0. \quad (6)$$

Theorem 2 means equicontinuity in $\gamma$ at $\gamma = 0$ of the family $\{W_n(\tau, \gamma, \mu)\}$. This fact can be interpreted as the qualitative robustness of the sign test in the local scheme (2) under the hypothesis and close alternatives. We call it local qualitative robustness.

Note, that definition (6) of local qualitative robustness is similar to the qualitative robustness introduced in [5]. In [5] the case of i.i.d. observations was considered and the qualitative robustness was defined in nonlocal framework.

**References**


MINIMAX EXTRAPOLATION PROBLEM FOR PERIODICALLY CORRELATED PROCESSES

I.I. Dubovetska, M.P. Moklyachuk
Taras Shevchenko National University of Kyiv
Kyiv, UKRAINE
e-mail: idubovetska@gmail.com, mmp@univ.kiev.ua

Abstract

The problem of mean square optimal estimation of linear functional from periodically correlated stochastic process is considered. This problem is investigated under conditions of spectral certainty and spectral uncertainty.

1 Introduction

Methods of solution of the problem of estimation of the unknown values of stationary stochastic processes (extrapolation, interpolation and filtering problems) with known spectral densities of processes are developed by A.N. Kolmogorov [3], N. Wiener, A.M. Yaglom. In the case where complete information on the spectral densities is impossible, but a set of admissible spectral densities is given, the minimax-robust approach to estimation problem is reasonable. That is we find estimate that minimizes the mean square error for all spectral densities from a given class of densities simultaneously. M.P. Moklyachuk, O.Yu. Masyutka [4] studied the extrapolation, interpolation and filtering problems for stationary processes and sequences.

In the paper by E.G. Gladyshev [2] investigation of periodically correlated processes was started. Minimax estimation problems for linear functionals from periodically correlated sequences were studied in articles by I.I. Dubovetska, O.Yu. Masyutka, M.P. Moklyachuk in [1].

2 The problem statement

We consider the problem of mean square optimal linear estimation of the functional

$$A\zeta = \int_{0}^{\infty} a(t)\zeta(t)dt$$

that depends on the unknown values of the mean square continuous periodically correlated stochastic process $\zeta(t)$ based on observations of the process $\zeta(t)$ for $t < 0$. The function $a(t), t \in R_+$, satisfies the condition $\int_{0}^{\infty} |a(t)|dt < \infty$.

3 Periodically correlated processes and generated stationary sequences

Definition 1 [2] Mean square continuous stochastic process $\zeta: R \rightarrow H = L_2(\Omega, \mathcal{F}, P)$, $E\zeta(t) = 0$, is called periodically correlated (PC) with period $T$, if its correlation function...
\( K(t + u, u) = E\zeta(t + u)\bar{\zeta}(u) \) for all \( t, u \in R \) and some fixed \( T > 0 \) is such that

\[
K(t + u, u) = E\zeta(t + u + T)\bar{\zeta}(u + T) = K(t + u + T, u + T).
\]

Let \( \{\zeta(t), t \in R\} \) be a PC process. We construct the sequence of stochastic functions

\[
\{\zeta_j(u) = \zeta(u + jT), u \in [0, T), j \in Z\}. \tag{1}
\]

The sequence (1) forms the \( L_2([0, T); H) \)-valued stationary stochastic sequence \( \{\zeta_j, j \in Z\} \). If we define in \( L_2([0, T); R) \) the orthonormal basis \( \{\vec{e}_k = \frac{1}{\sqrt{T}} e^{2\pi i(-1)^k [\frac{t}{T}]} \}, \) \( k = 1, 2, \ldots, \) \( \langle \vec{e}_j, \vec{e}_k \rangle = \delta_{kj}, \) the generated vector stationary sequence \( \{\zeta_j, j \in Z\} \) can be represented in the form

\[
\zeta_j = \sum_{k=1}^{\infty} \zeta_{kj} \vec{e}_k, \quad \zeta_{kj} = \langle \zeta_j, \vec{e}_k \rangle = \frac{1}{\sqrt{T}} \int_0^T \zeta_j(v) e^{-2\pi i(-1)^k [\frac{v}{T}]} dv. \tag{2}
\]

### 4 The classical method of extrapolation

Taking into account the decomposition (2) of generated stationary sequence \( \{\zeta_j, j \in Z\}, \) the functional \( A\zeta \) can be represented in the form

\[
A\zeta = \int_0^\infty a(t)\zeta(t)dt = \sum_{j=0}^{\infty} \int_0^T a_j(u)\zeta_j(u)du = \sum_{j=0}^{\infty} \sum_{k=1}^{\infty} a_{kj}\zeta_{kj} = \sum_{j=0}^{\infty} \tilde{a}_j \tilde{\zeta}_j,
\]

\[
\tilde{\zeta} = (\zeta_{kj}, k \geq 1)\top, \quad \tilde{a}_j = (a_{kj}, k \geq 1)\top = (a_{1j}, a_{3j}, a_{2j}, \ldots, a_{2k+1,j}, a_{2k,j}, \ldots)\top.
\]

The classical Kolmogorov projection method [3] allows us to find main characteristics of the optimal linear estimate of the functional \( A\zeta. \)

**Theorem 1** Let \( \{\zeta(t), t \in R\} \) be a PC process such that the vector stationary sequence \( \{\zeta_j, j \in Z\} \) generated by relation (1) has the spectral density \( f(\lambda) \) that satisfies the minimality condition \( \int_{-\pi}^{\pi} Tr[(f(\lambda))^{-1}]d\lambda < \infty. \)

The spectral characteristic \( h(f) \) and the mean square error \( \Delta(f) \) of the optimal linear estimate of the functional \( A\zeta \) from observations of the process \( \zeta(t) \) for \( t < 0 \) are given by formulas

\[
h(\lambda) = A\top(e^{i\lambda}) - C\top(e^{i\lambda})[f(\lambda)]^{-1}, \tag{3}
\]

\[
\Delta(f) = \langle c, a \rangle, \tag{4}
\]

where \( A(e^{i\lambda}) = \sum_{j=0}^{\infty} \tilde{a}_j e^{ij\lambda}, \ C(e^{i\lambda}) = \sum_{j=0}^{\infty} \zeta_{kj} e^{ij\lambda}, \ a = \{\tilde{a}_j\}_{j=0}^{\infty}, \ c = \{\zeta_{kj}\}_{j=0}^{\infty} = B^{-1}a, \)

\( B = \{B(l, j)\}_{l,j=0}^{\infty} \) is matrix with elements \( B(l, j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [(f(\lambda))^{-1}]^\top e^{i(l-j)\lambda}d\lambda. \)
5 Least favorable spectral densities in the class $D_0$

Consider the minimax-robust approach to the problem of estimation of the functional $A\zeta$ under the condition that the spectral density $f(\lambda)$ of generated vector stationary sequence $\{\zeta_j, j \in \mathbb{Z}\}$ constructed by relation (1) belongs to the class

\[ D_0 = \left\{ f(\lambda) \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} Tr f(\lambda)d\lambda = P_\zeta \right. \right\}. \]

The spectral density $f(\lambda)$ of the regular stationary sequence $\{\zeta_j, j \in \mathbb{Z}\}$ admits the canonical factorization

\[ f(\lambda) = P(\lambda)P^*(\lambda), \quad P(\lambda) = \sum_{u=0}^{\infty} d(u)e^{-iu\lambda}, \quad d(u) = \{d_{km}(u)\}_{k=1,\infty}^{m=1,\infty}. \]

Applying the Lagrange multipliers method to the conditional extremum problem

\[ \Delta(f) = \|Ad\|^2 \rightarrow \text{max}, \quad f(\lambda) = P(\lambda)P^*(\lambda) \in D_0, \]

(5)

where $\|Ad\|^2 = \sum_{l=0}^{\infty} \| (Ad)l \|^2$, $(Ad)l = \sum_{j=l}^{\infty} a_j^*d(j-l), l \geq 0$, we can derive relations

\[ \sum_{p=0}^{\infty} \sum_{s=0}^{\infty} a_{r+p}^*a_{s+p}^*d(s) = \alpha^2d(r), \quad r = 0, 1, \ldots, \]

(6)

\[ \|d\|^2 = \sum_{u=0}^{\infty} \|d(u)\|^2 = P_\zeta, \]

(7)

that determine solution $f^0(\lambda)$ of (5). The following theorem holds true.

**Theorem 2** The spectral density

\[ f^0(\lambda) = \left( \sum_{u=0}^{\infty} d^0(u)e^{-iu\lambda} \right) \left( \sum_{u=0}^{\infty} d^0(u)e^{-iu\lambda} \right)^* \]

of the moving average sequence $\{\zeta_j, j \in \mathbb{Z}\}$ with components

\[ \zeta_{kj} = \sum_{u=-\infty}^{j} \sum_{m=1}^{M} d^0_{km}(j-u)\bar{\varepsilon}_m(u) \]

is the least favorable in the class $D_0$ for the optimal linear estimation of the functional $A\zeta$. The sequence of matrices $d^0 = \{d^0(u), u = 0, 1, \ldots\}$ is determined by relations (6), (7). The minimax spectral characteristic $h(f^0)$ is given by the formula

\[ h^T(f^0) = A^T(e^{i\lambda}) - S^0(e^{i\lambda})Q^0(\lambda), \]

where $S^0(e^{i\lambda}) = \sum_{l=0}^{\infty} (Ad^0)e^{il\lambda}$, $Q^0(\lambda) = \{q^0_{mk}(\lambda)\}_{k=1,\infty}^{m=1,\infty}$ is the matrix function determined by equation $Q^0(\lambda)P^0(\lambda) = I_M$.

\[ 200 \]
6 Least favorable spectral densities in the class $D_0^-$

Consider the minimax-robust approach to the problem of estimation of the functional $A\zeta$ under the condition that the spectral density $f(\lambda)$ belongs to the class $D_0^- = \{f(\lambda) | 1/2\pi \int_{-\pi}^{\pi} f^{-1}(\lambda) d\lambda = V\}$, where $V = \{v_{kn}\}_{k,n=1}^{\infty}$ is a given nonnegative matrix with $\det V \neq 0$. The least favorable in the class $D_0^-$ spectral density gives solution to the problem

$$\Delta(h(f^0); f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} (C^0(e^{i\lambda}))^\top (f^0(\lambda))^{-1} f(\lambda)(f^0(\lambda))^{-1}C^0(e^{i\lambda}) d\lambda \to \sup, \ f \in D_0$$

and is determined by relation

$$[(f^0(\lambda))^{-1}]^\top C^0(e^{i\lambda}) = [(f^0(\lambda))^{-1}]^\top \vec{\alpha}, \ (8)$$

where $\vec{\alpha}$ is the Lagrange multiplier. Let denote by $V^\top (l-j) = B^0(l,j)$, $(\vec{a}_0)^{-1}$ define from the equality $((\vec{a}_0)^{-1})^\top \vec{a}_0 = 1$. Then the following theorem holds true.

**Theorem 3** Suppose the coefficients $\{\vec{a}_j, j = 0, 1, \ldots\}$, that determine the functional $A\zeta$, are such that the matrix function $\sum_{u=-\infty}^{\infty} V(u) e^{iu\lambda}$, where

$$V(u) = V^*(u) = V(\vec{a}_0)^{-1}\vec{a}_u^\top, \ u = 0, 1, \ldots,$$

is positive definite and has nonzero determinant. Then the spectral density

$$f^0(\lambda) = \left(\sum_{u=-\infty}^{\infty} V(u) e^{iu\lambda}\right)^{-1}$$

is the least favorable in the class $D_0^-$ for the optimal linear estimation of the functional $A\zeta$. The minimax spectral characteristic $h(f^0)$ is given by the formula

$$h(f^0) = -\sum_{u=1}^{\infty} \overline{V(u)(V^\top)^{-1}\vec{a}_0} e^{-iu\lambda}.$$

References


MISCLASSIFICATION PROBABILITY BASED ON LINEAR DISCRIMINANT FUNCTION FOR SAR ERROR MODELS

K. Dučinskas, M. Karaliutė, I. Šimkienė
Department of Statistics
Klaipėda University, Lithuania
e-mail: k.ducinskas@gmail.com

Abstract
Given training sample, the problem of classifying Gaussian spatial data into one of two populations specified by spatial autoregressive model (SAR) with different mean functions is considered. This paper concerns classification procedures associated with Linear Discriminant Function (LDF) under deterministic spatial sampling design. In the case of complete parametric certainty, the overall misclassification probability associated with LDF is derived. Spatial weights based on inverse of Euclidean distance and the third order neighbourhood schemes on regular 2-dimensional lattice are used for illustrative examples. The effect of the prior distribution of class labels on the performance of proposed classification procedure is numerically evaluated.

Keywords: discriminant function, Covariance function, Gaussian random field, misclassification probability, Training labels configuration.

1 Introduction
Many authors (see e.g. Kharin (1996)) have investigated the performance of the linear discriminant function(LDF) in classification of dependent observations (Markov dependence, autoregressive models). Error rates in classification of spatially correlated Gaussian observations associated with LDF is considered in Ducinskas (2009). Here only the parametric models for spatial covariance belonging to Mattern class are considered. In this paper the extension of the latter investigation to the case of spatial Gaussian data specified by widely used SAR model is presented.

In this work an explicit expression for overall misclassification probability(OMP) is derived. By using the derived OMP, the performance of the BDF is numerically analyzed in the case of stationary Gaussian random field on 2-dimensional regular lattice. The dependence of the values of obtained OMP on the Mahalanobis distance for different spatial sampling designs and prior distributions for class labels is investigated. By applying the proposed criterion, the numerical comparison of some training labels configurations (TLC) is carried out.

2 The main concepts and definitions
The main objective of this paper is to classify the Gaussian random field (GRF) observation $Z(s_0)$ in population $\Omega_j$ ($j = 1, 2$), when means satisfy the condition

$$\mu_1(s_0) \neq \mu_2(s_0), s_0 \in D.$$
Denote by \( S_n = \{ s_i \in D; i = 1, ..., n \} \) the set of locations where training sample \( T' = (Z(s_1), ..., Z(s_n)) \) is taken, and call it the set of training locations (STL). It specifies the spatial sampling design for training sample. Assume that \( S_n \) is partitioned into union of two disjoint subsets, i.e. \( S_n = S^{(1)} \cup S^{(2)} \), where \( S^{(j)} \) is the subset of \( S_n \) that contains \( n_j \) locations of feature observations from \( \Omega_j \), \( j = 1, 2 \). So each partition \( \xi(S_n) = \{ S^{(1)}, S^{(2)} \} \) with marked labels determines TLC.

For TLC \( \xi(S_n) \), define the variable \( d = |D^{(1)} - D^{(2)}| \), where \( D^{(j)} \) is the sum of distances between the location \( s_0 \) and locations in \( S^{(j)}, j = 1, 2 \).

As it follows, we assume that STL \( S_n \) and TLC \( \xi \) are fixed. Denote by \( w_{ij} \) a spatial weight specifying the interconnection between locations \( s_i \) and \( s_j \) (\( w_{ii} = 0 \) and \( w_{ij} \neq 0 \) if \( i \approx j \), and 0 otherwise) for \( i, j = 0, ..., n \). Here \( i \approx j \) denotes that location \( s_j \) is a neighbour of location \( s_i \), and let \( W \) denote the \( n \times n \) spatial weights matrix for \( S_n \).

So the training sample \( T \) follows SAR error model

\[
T = M + E,
\]

where \( M \) is the vector of the training sample mean and \( E \) is the \( n \times 1 \) vector of random errors with multivariate Gaussian distribution \( N_n(0, V) \), with \( V = \sigma^2[(I - \lambda W)'(I - \lambda W)]^{-1} \), and \( \sigma^2, \lambda \) are respectively the variance and autoregressive parameters. Let \( t \) denote the realization of \( T \).

Assume that \( Z_0 \) follows SAR error model. Then the conditional distribution of \( Z_0 \) given \( T = t, \Omega_j \) is Gaussian with mean \( \mu^0_{it} = E(Z_0|T = t; \Omega_j) \) and variance \( \sigma^2_t = \text{var}(Z_0|T = t; \Omega_j) \).

Under the assumption of complete parametric certainty of populations the LDF minimizing the OMP is formed by log ratio of conditional likelihoods.

Then LDF is specified by (Anderson, 2003)

\[
W_i(Z_0) = (Z_0 - \frac{1}{2}(\mu^0_{1t} + \mu^0_{2t}))(\mu^0_{1t} - \mu^0_{2t})/\sigma^2_t + \gamma,
\]

where \( \gamma = \ln(\pi_1/\pi_2) \).

Here \( \pi_1, \pi_2 (\pi_1 + \pi_2 = 1) \) are prior probabilities of the populations \( \Omega_1 \) and \( \Omega_2 \) for observation at location \( s_0 \). They specified the prior distribution for class membership for observation at location \( s_0 \).

**Definition 1** The OMP for the LDF \( W_i(Z_0, \Psi) \) is defined as

\[
P_B = \sum_{i=1}^{2} \sum_{j=1,j \neq i}^{2} \pi_i P_{ij},
\]

where for \( i, j = 1, 2 \), \( P_{ij} = P_{it}((-1)^i W_i(Z_0) < 0) \).

Here, for \( i, j = 1, 2 \), the probability measure \( P_{it} \) is based on conditional distribution of \( Z_0 \) given \( T = t, \Omega_i \). Squared Mahalanobis distance between conditional distributions of \( Z_0 \) given \( T = t \) are denote by and \( \Delta^2 = (\mu^0_{1t} - \mu^0_{2t})^2/\sigma^2_t \).

Set \( \Delta = |\mu_1(s_0) - \mu_2(s_0)| \) and denote by \( w_0 \) the \( n \) vector of spatial weights between \( s_0 \) and \( S_n \), i.e. \( w_0 = (w_{01}, w_{02}, ..., w_{0n}) \).
The OMP is useful in providing a guide to the performance of classification rule before it is actually formed from training sample. The OMP is the performance measure to the BDF similar as the mean squared prediction error (MSPE) is the performance measure to the kriging predictor (see Diggle et al 2002).

Make the following assumptions:
(A1) The set of locations \(S_n^0\) form a clique of size \(n + 1\),
(A2) Spatial weights for \(S_n\) and \(S_n^0\) are based on the Euclidean distance between different locations.

Lemma 1 Suppose that observation \(Z_0\) to be classified by BDF and let covariance matrix of \(T_0\). Then under the assumptions (A1) and (A2), OMP takes the form

\[
P_B = \sum_{j=1}^{2} (\pi_j \Phi(-\Delta_0/2 + (-1)^j \gamma/\Delta_0)),
\]

where \(\Phi(\cdot)\) is the standard normal distribution function and \(\Delta_0^2 = \Delta^2(1 + \lambda^2 w_0^\prime w_0)/\sigma^2\).

Proof 1 The proof of lemma is easily done by using the properties of normal distribution.

3 Example and discussions

Numerical example is considered to investigate the influence of the statistical parameters of populations to the proposed LDF in the finite (even small) training sample case. With an insignificant loss of generality the cases with \(n_1 = n_2\) are considered. We also suppose that assumptions (A1) and (A2) hold.

In this example, observations are assumed to arise from stationary Gaussian random field with constant mean. The spatial weights are specified by \(w_{ij} = 1/d_{ij}\), where \(d_{ij}\) is the Euclidean distance between different locations.

Assume \(D\) is regular 2-dimentional lattice with unit spacing. We consider NN(3) spatial structure scheme for \(s_0 = (2, 2)\). STL for this scheme consists of 12 third-order neighbours of \(s_0\) and is denoted by \(S_{12}\). Set \(M1 = \{i : i = 1, ..., n_1, i \approx 0\}\).

Two cases of prior probabilities are considered:

- **CN** \(\pi_1 = \left( \sum_{i \in M1} 1/n \right)/n\),
- **CD** \(\pi_1 = \left( \sum_{i \in M1} 1/d_{0i} \right)/\left( \sum_{i=1,...,n} 1/d_{0i} \right)\).

Case CN is based only on the number of neighbours, while case CD incorporated spatial adjacency (distances) also. So OMP is denoted PBN for the case CN and PBD for the case CD.

Consider two TLC \(\xi_1, \xi_2\) for \(S_{12}\) specified by \(\xi_1 = \{S^{(1)} = \{(1, 3), (2, 4), (2, 3), (2, 1), (3, 2), (4, 2)\}, S^{(2)} = \{(0, 2), (1, 2), (0, 2), (3, 3), (3, 1), (1, 1)\}\}, \xi_2 = \{S^{(1)} = \{(0, 2), (1, 2), (2, 3), (2, 1), (3, 2), (4, 2)\}, S^{(2)} = \{(2, 4), (1, 3), (1, 1), (2, 0), (3, 3), (3, 1)\}\}.

The comparison of two cases of prior distribution for each TLC is done by the values of index \(\eta = PBD/PBN\). The values PBD, PBN and index \(\eta\) for various values of
Table 1: Values of PBD, PBN and $\eta$ for NN(3) neighbourhood scheme.

<table>
<thead>
<tr>
<th>TLC</th>
<th>$\xi_1$</th>
<th>$\xi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta$</td>
<td>PBD</td>
<td>PBN</td>
</tr>
<tr>
<td>0,1</td>
<td>0,4662</td>
<td>0,4847</td>
</tr>
<tr>
<td>0,2</td>
<td>0,4586</td>
<td>0,4694</td>
</tr>
<tr>
<td>0,3</td>
<td>0,4468</td>
<td>0,4542</td>
</tr>
<tr>
<td>0,4</td>
<td>0,4334</td>
<td>0,4390</td>
</tr>
<tr>
<td>0,5</td>
<td>0,4194</td>
<td>0,4239</td>
</tr>
<tr>
<td>0,6</td>
<td>0,4052</td>
<td>0,4089</td>
</tr>
<tr>
<td>0,7</td>
<td>0,3909</td>
<td>0,3941</td>
</tr>
<tr>
<td>0,8</td>
<td>0,3767</td>
<td>0,3794</td>
</tr>
<tr>
<td>0,9</td>
<td>0,3625</td>
<td>0,3649</td>
</tr>
<tr>
<td>1,0</td>
<td>0,3484</td>
<td>0,3505</td>
</tr>
</tbody>
</table>

Parameter $\Delta$ are given in Table 1. The results of calculations with $\lambda = 0,3$, $\sigma^2 = 1$ for $\xi_1$ and $\xi_2$ are presented in Table 1. By the definition variable $d$ represents the asymmetry population labels distribution in training sample. It is easy to obtain that $d = 2(\sqrt{2} - 1)$ for $\xi_1$ and $d = 4(\sqrt{2} - 1)$ for $\xi_2$. So $\xi_1$ is less asymmetric TLC than $\xi_2$.

Analyzing the contents of table 1. We can conclude that prior distribution based on distances to neighbours outperforms the one based only on numbers of neighbours, because PBD values are smaller than corresponding PBN values. Table also shows that for both neighbourhood schemes OMP decreases with the increasing of $\Delta$. Values of index $\eta$ in tables numerically illustrate the comparison of two cases of prior distributions. These values enable us to conclude that effect of the incorporation distances into prior distributions is stronger for more asymmetric TLC i.e. for $\xi_2$. Hence the results of numerical analysis give us strong arguments to expect that proposed derived formula of the OMP could be effectively used for performance evaluation of classification procedures and for the optimal designing of spatial training samples.

References


Approximate formulas for expectation of functionals from solution to stochastic equation with random measure

A.D. Egorov
National Academy of Sciences, Institute of Mathematics
Minsk, BELARUS
e-mail: egorov@im.bas-net.by

Abstract
Approximate formulas for evaluation of mathematical expectation of nonlinear functional from solution to stochastic differential equation driven by Wiener process and Poisson random measure is given.

1 Introduction
Evaluation of functional defined on trajectories of stochastic processes is motivated by probabilistic representations of solutions to partial differential equations and by applications of stochastic differential equations to finance [1,5]. There are a number of papers devoted to modeling markets using jump-diffusion processes considered in the report [6]. As a rule for evaluation of functionals from processes Monte Carlo methods averaging over the trajectories is used. But there is large variance in most interesting cases, so methods allowing the reduce variance obtain significance. Our report is relative to deterministic approach based on generalizing of the quadrature formulas method for ordinary multidimensional integrals and constructing integration formulas which are exact for functional polynomials of a given degree. Calculation of functionals on trajectories of Gaussian processes and processes with independent increments had been considered in frame of numerical methods of functional integration [1]. In [2] this approach was applied to the case of functionals from empirical processes. At the same time the approach had been used for evaluation of expectation of functional from solutions of stochastic differential equations [1]. But in general case the approach based on formulas exact for functional polynomials work only asymptotically. So the problem of constructing the formulas in special cases when the solution of equation be given explicitly is current problem [3]. In this report we represent the results in this direction for new class of stochastic differential equations.

2 The results
We consider Ito’s stochastic differential equation

$$X_t = 1 + \int_0^t a(s)X_{s-} \, ds + \int_0^t b(s)X_{s-} \, dW_s + \int_0^t \int \theta(s, x)X_{s-} \tilde{N}(ds, dx),$$  (1)
where \( W = \{ W_t, t \in [0, 1] \} \) is the Wiener process; \( \tilde{N}(ds, dx) = \tilde{N}(ds, dx) - \nu(dx)ds \) is centered Poisson measure, \( \nu \) is a characteristic of \( \tilde{N}; \theta(s, x) \geq -1 \), \( \int_0^t \theta^2(s, x) \nu(dx)ds < \infty \). The solution of the equation is known [6]:

\[
X_t = \exp \left\{ \int_0^t (a(s) - \frac{1}{2} b^2(s)) ds + \int_0^t b(s) dW_s \right\} \times \\
\exp \left\{ \int_0^t \int_R \ln(1 + \theta(s, x)) \tilde{N}(ds, dx) + \right. \\
\int_0^t \int_R (\ln(1 + \theta(s, x)) - \theta(s, x)) \nu(dx)ds \right\} = Y_1(t)Y_2(t).
\]

Using independence of Wiener process and centered Poisson measure, and the next formulas

\[
E\left[ \exp \left\{ \int_0^t b(s)dW_s \right\} \right] = \exp \left\{ - (1/2) \int_0^t b^2(s)ds \right\}; \\
E\left[ \exp \left\{ \int_0^t \int_R \theta(s, x) \tilde{N}(ds, dx) \right\} \right] = \exp \left\{ \int_0^t \int_R (e^{\theta(s,x)} - 1 - \theta(s, x)) \nu(dx)ds \right\}
\]

we can evaluate all moments of the process \( X_t \): \( E[X_t] = \exp \left\{ \int_0^t a(s)ds \equiv c_0(t) \right\}, \)

\[
E\left[ \prod_{j=1}^n X_{t_j} \right] = \prod_{j=1}^n c_0(t_j) \exp \left\{ \sum_{j=2}^n \sum_{l=1}^{t_j} \int_0^{t_j} b^2(s)ds \right\} \times \\
\prod_{j=1}^n \exp \left\{ - \int_0^{t_j} \int_R \theta(s, x) \nu(dx)ds \right\} \times \\
\prod_{j=1}^n \exp \left\{ \int_0^{t_j} \int_R (1 + \theta(s, x))^{n-j} \theta(s, x) \nu(dx)ds \right\}, \quad n \geq 2, \tag{2}
\]

where \( \bar{t}_1, \ldots, \bar{t}_n \) are the numbers \( t_1, \ldots, t_n \) arranged in ascending order. Let us define approximate formulas for evaluation of \( E[F(X_{\bar{t}})] \):

\[
E[F(X_{\bar{t}})] \approx J(F) \equiv J_1(F(Y_1(\cdot))Y_2(\cdot)), \tag{3}
\]

where the operators \( J_1, J_2 \) operate on \( F(Y_1Y_2) \) independently as functional of variables \( Y_1, Y_2 \) in the following way:

\[
J_1(G(Y_1)) = F(0) + \sum_{k=1}^2 A_k G(c_kc_0(\cdot)) + \\
\frac{1}{2} \sum_{l=1}^2 (-1)^l \int_{-1}^1 G(c_0(\cdot) \left[ \frac{l}{2} b(|u|) \times \\
\exp \left\{ \frac{1}{2} \int_0^{|u|} b^2(\tau)d\tau \right\} 1_{[0, l)](|u|) \text{sign}(u) + \exp \left\{ 2 \int_0^{|u|} b^2(\tau)d\tau \right\} )du, \tag{4}
\]

\[\]

207
where \([t] - \) the integer part of \(\frac{t}{2}\), \(\Lambda F(f) = (1/2)(F(f) - F(-f))\);

\[
J_2(G(Y_2)) = \Delta G(1) - G(0) + \sum_{k=1}^{2} A_k \{ \Lambda G(c_k q(\cdot) p_0) - \}
\]

\[
\int_0^1 \int_0^1 \Lambda G(c_k q(\cdot) \varphi_1(u) \varphi_2(v) 1_{[0,1]}(u) 1_{[1,1]}(v)) du dv + \int_0^1 \Lambda G(c_k q(\cdot) \varphi_1(u) 1_{[0,1]}(u)) du - \int_0^1 \Lambda G(c_k q(\cdot) \varphi_2(u) 1_{[1,1]}(u)) du \}
\]

\[
\left\{ \int_0^1 \Delta G(\varphi(u) 1_{[0,1]}(u)) du, \right. \]

\[
\varphi_1(u) = \left( \int \theta^2(u, x) (1 + \theta(u, x)) \nu(dx) \right)^{1/3} \]

\[
\exp \left\{ (1/3) \int_0^u \int_R \theta^2(s, x) (1 + \theta(s, x)) \nu(dx) ds \right\}, \]

\[
\varphi_2(u) = -\left( \int_R \theta^2(u, x) \nu(dx) \right)^{1/3} (q(u))^{-1/3}, \quad \varphi_3(u) = \left( \int_R \theta^2(u, x) \nu(dx) \right)^{1/2} q(u)^{1/2}, \]

\[
q(t) = \exp \left\{ \int_0^t \int_R \theta^2(s, x) \nu(dx) ds \right\}, \quad p_0 = q(1)^{-1/3}, \]

\[
\Delta F(f) = (1/2)(F(f) + F(-f)), \quad A_1 c_1 + A_2 c_2 = 0, \quad A_1 c_3^2 + A_2 c_2^3 = 1. \]

Recall that functional polynomials of \(X_t, t \in [0, 1]\), in general can be presented by multiple Stieltjes integrals in the form:

\[
P_n(F) = F_0 + \sum_{k=1}^{n} \int_0^1 \cdots \int_0^1 \prod_{j=1}^{k} X_{t_j} dt_1, \ldots, t_k F_k(t_1, \ldots, t_k),
\]

where \(F_k(t_1, \ldots, t_k) - \) a deterministic function of bounded variation.

**Theorem 1.** Approximate formulas (3)-(5) is exact for functional polynomials of third degree.

Proof follows from the comparison of the results of computation by formula from the right side of approximate equality (3) with their exact values for functionals \(F(X) = \prod_{j=1}^{k} X_{t_j}, \quad k = 1, 2, 3\), and for \(F(X) = \text{const}\).
where \(0 = t_0 < t_1 < \ldots < t_N = 1\) is a digitization of \([0,1]\). This approximate formulas is exact for functional polynomial of third degree and converges to exact value of \(E[F(X_t)]\) for some class of smooth functionals satisfying dominated convergence conditions. At last an analogue of the formulas (3)-(5) can be used for approximate evaluation of mathematical expectation of functionals \(F(X_t) = \int_0^1 G(X_s, s) ds\) from solution to equation

\[
X_t = 1 + \int_0^t a(s, X_{s-}) ds + \int_0^t b(s, X_{s-}) dW_s + \int_0^t \int_0^R \theta(s, x, X_{s-}) \tilde{N}(ds, dx).
\]

In this case we use the time digitization and consider linear approximation of coefficients of the equation on sub-intervals [4]. Then we apply our approximate formulas to functionals from solution to linear equation on the sub-interval with suitable initial condition received on preceding step.

### References


PREDICTION OF TIME SERIES BASED ON STATE SPACE MODEL

S.V. Lobach  
Belarussian State University  
Minsk, Belarus  
e-mail: sergey.lobach@mail.ru

Abstract

A method of the prediction of time series based on the state space model is suggested. It provides recursive formulae for the estimation of the state but in some special case recursive formulas can be derived. The AR(2) model is investigated in more details including a simulation study.

Keywords: state space model, Kalman filter, time series, prediction.

1 Introduction

State space models were developed originally by control engineers but are receiving increasing attention in the economics literature [1]. There is a number advantages in representing models in state space form. The likelihood function can be written in terms of the one-step-ahead prediction error $v_t$ and their variance $\sigma_t$. The Kalman filter when applied to a model in state space form provides an algorithm for prediction $v_t$ and its variance. Since many models (for example all ARMA models) can be represented in state space form, the Kalman filter provides a convenient general method of representing the likelihood functions for what may be very complex models and an algorithm for prediction of time series.

2 The linear Gaussian state space model and Kalman filter

Let us consider a linear system of the form

$$
\beta_{t+1} = F_t \beta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, Q_t),
$$

$$
z_t = H_t \beta_t + \eta_t, \quad \eta_t \sim N(0, R_t),
$$

$$
\beta_1 \sim N(a_1, P_1), \quad t = 1, 2, \ldots, n.
$$

Matrices $F_t$, $H_t$, $R_t$ and $Q_t$ are assumed known. Initially, $a_1$ and $P_1$ are assumed known, we may consider what to do when some elements of them are unknown. The $p \times 1$ vector $z_t$ is the observation. The unobserved $m \times 1$ vector $\beta_t$ is called the state. The disturbances $\varepsilon_t$ and $\eta_t$ are independent sequences of independent normal vectors. The equation (1) is called the state equation and the equation (2) is called the observation equation.
The structure of model (1) is a natural one for representing the behavior of many time series as a first approximation. The equation (2) is a standard multivariate linear regression model whose coefficient vector $\beta_t$ varies over time; the development over time of $\beta_t$ is determined by the first-order vector autoregression given in the equation (1). The Markovian nature of the model accounts for many of its properties. In spite of the conceptual simplicity of this model it is highly flexible and has a wide range of applications to problem in practical time series analysis.

Let $Z_t = \{z_1, \ldots, z_t\}, t = 1, 2, \ldots, n$. We may focus on the following items:

1. Kalman filter. This recursively computes $a_{t|t} = E(\beta_t|Z_t)$ and $P_{t|t} = \text{Var}(\beta_t|Z_t)$ for $t = 1, 2, \ldots, n$. Since distributions are normal, these quantities specify the distribution of $\beta_t$ given data up to time $t$ [2].

2. State smoother. This estimates $\hat{\beta}_{t-1|t}=E(\beta_{t-1}|Z_t)$ and $V_{t-1|t}=\text{Var}(\beta_t|Z_t)$ and hence the conditional distribution of $\beta_t$ given all the observations for $t=1,\ldots,n$ [2].

3. Simulation smoother. An algorithm for generating draws from $P(\beta_1,\ldots,\beta_n|Z_n)$. This is an essenon-Gaussian and nonlinear models [1].

4. Missing observation. We show that the treatment of missing observations is particulary simple in the state space approach [3].

5. Forecasting. Prediction is simply treated as a special case of missing observation [4].

6. Initialization. This deals with the case where some elements of $a_1 = E(\beta_1)$ and $V_1 = \text{Var}(\beta_1)$ are unknown.

7. Univariate treatment of multivariate series. This puts a multivariate model into univariate form, which can simplify the treatment of large complex models.

8. Parameter estimation. It may be shown that the likelihood function is easy constructed using the Kalman filter.

The diffuse Kalman filter [2] computes $\hat{\beta}_{t|t} = E(\beta_t|Z_t)$ and $P_{t|t} = \text{Var}(\beta_t|Z_t)$ by the recursion formulae.

We introduce the following notation. Let

$$\hat{\beta}_{t|t-1} = E(\beta_t|Z_{t-1}),$$
$$\hat{z}_{t|t-1} = E(z_t|Z_{t-1}),$$
$$V_{t|t} = E\left(\left(\beta_t - \hat{\beta}_{t|t}\right)\left(\beta_t - \hat{\beta}_{t|t}\right)^T\right),$$
$$V_{t|t-1} = E\left(\left(\beta_t - \hat{\beta}_{t|t-1}\right)\left(\beta_t - \hat{\beta}_{t|t-1}\right)^T\right),$$
$$M_{t|t-1} = E\left(\left(z_t - \hat{z}_{t|t-1}\right)\left(z_t - \hat{z}_{t|t-1}\right)^T\right).$$
Then $\hat{\beta}_{t|t-1}$ and $\hat{z}_{t|t-1}$ are the predictions of $\beta_t$ and $z_t$ formed at time $t-1$. The matrices $V_{t|t}$, $V_{t|t-1}$ and $M_{t|t-1}$ are the mean squared prediction errors and $v_t$ is the residual in the regression of $z_t$ on its past values.

We have the following relations for the diffuse Kalman filter [2]:

$$\hat{\beta}_{t|t} = \hat{\beta}_{t|t-1} + K_t v_t,$$

(3)

$$K_t = V_{t|t-1} H_t^T (H_t V_{t|t-1} H_t^T + R_t)^{-1},$$

(4)

$$V_{t|t} = (I - K_t H_t) V_{t|t-1},$$

(5)

$$\hat{\beta}_{t+1|t} = F_t \hat{\beta}_{t|t},$$

(6)

$$V_{t+1|t} = F_t \hat{\beta}_{t+1|t} + Q_t,$$

(7)

$$\hat{z}_{t+1|t} = H_{t+1} \hat{\beta}_{t+1|t},$$

(8)

$$M_{t+1|t} = H_{t+1} V_{t+1|t} H_{t+1}^T + R_t,$$

(9)

$$v_t = z_t - H_t \hat{\beta}_{t|t-1}.$$ 

To apply the Kalman filter some initial values are required. More precisely, to apply formulas (3) and (4) at time $t = 1$, we need $\hat{\beta}_{1|0}$ and $V_{1|0}$. Intuitively, $\hat{\beta}_{1|0}$ is the optimal prediction of $\beta_1$ based on no information. Thus $\hat{\beta}_{1|0}$ is equal to the mean $\alpha_1$. Then we have $V_{1|0} = P_1$.

Remark. The formulas giving the covariance matrices (4), (5), (7), (8), (9) do not involve the observations $z_t$. Thus these formulas can be used independently of any observations provided the matrices $F_t$, $H_t$ are known.

### 3 Missing observations and forecasting

The missing observations are easy to handle in state space analysis. If observation $Z_t$ is missing for any $t$ from 2 to $n-1$, all we have to do is put $v_t = 0$ and $K_t = 0$ [4].

The forecasting is also easy in state space analysis. Suppose we want to forecast $z_{n+1}, \ldots, z_{n+K}$ given $z_1, z_2, \ldots, z_n$ and calculate mean square forecast errors. We treat $z_{n+1}, \ldots, z_{n+K}$ as missing and proceed using (3) as in the case of the missing observations. We use $H_{n+1} \hat{\beta}_{n+1|n}$, $H_{n+K} \hat{\beta}_{n+K|n+K-1}$ as the forecast and use $V_{n+1}, \ldots, V_{n+K}$ to provide mean square errors [4].

### 4 Prediction of AR(2) process

Consider the second-order autoregressive model

$$z_t - \varphi_1 z_{t-1} - \varphi_2 z_{t-2} = u_t,$$

(10)

where $u_t$ is Gaussian white noise process with variance $\sigma^2$. 


Let \( \beta = \begin{pmatrix} z_t \\ z_{t-1} \end{pmatrix} \).

We have

\[
\begin{align*}
\beta_{t+1} &= \begin{pmatrix} \varphi_1 & \varphi_2 \\ 1 & 0 \end{pmatrix} \beta_t + \begin{pmatrix} u_{t+1} \\ 0 \end{pmatrix}, \\
z_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \beta_t.
\end{align*}
\]

Hence the model can be written in a state space form. Note that there the state vector is observed and that the error term \( \eta_t \) is zero. This form is sometimes called the companion form.

We observe \( z_1, \ldots, z_t \) and we want to predict \( \beta_{t+h} \) and \( z_{t+h} \), \( h \geq 1 \). That is, we want to determine

\[
\hat{\beta}_{t+h|t} = E (\beta_{t+h}|Z_t), \\
\hat{z}_{t+h|t} = E (z_{t+h}|Z_t),
\]

as well as the variance matrices of the prediction errors. When \( h = 1 \), the optimal predictions of \( \beta_{t+1} \) and \( z_{t+1} \) as well as the variance matrices of the corresponding prediction errors are given by formulae (3), (8) and (5), (7) of the Kalman filter. To obtain the prediction values of \( \hat{\beta}_{t+h|t} \) and \( \hat{z}_{t+h|t} \) it suffices to follow the sequence of formulae (6) and (8), when \( \hat{\beta}_{t|t} \) and \( V_{t|t} \) are replaced by \( \hat{\beta}_{t+1|t} \) and \( V_{t+1|t} \) respectively and to iterate this operation \( h - 1 \) times.

Remark. When \( H_t \) and \( F_t \) are functions of \( z_1, \ldots, z_t \) we saw that Kalman filter equations (3)–(9) remain valid. In contrast, when the prediction horizon \( h \) is larger than two, i.e. \( h \geq 2 \), this procedure can no longer be used. This is because the matrices \( F_{t+h} \) and \( H_{t+h} \) are now random conditional on \( z_1, \ldots, z_t \).

The formulas (3)–(9) are used to determine prediction values of \( \hat{z}_{t+h|t} \) in model (10).

Results of the simulation illustrates the usefulness of the proposed recursive method.

References


MODELING OF SPATIAL QUEUES SYSTEM

V.S. Mukha, M.V. Prishchepchik
Belarusian State University of Informatics and Radioelektronics
Minsk, Republic of Belarus
e-mail: mukha@bsuir.by

Abstract
A mathematical model of a scalar random field generated by Poisson random process is presented. Computer modeling for implementation of a proposed random field with a graphic illustration of the results is performed.

1 Introduction
An important problem in the development of spatial queues systems is to construct a mathematical model of the input requirements flow for such a system. At present, there is no suitable mathematical model for those purposes. There is only a definition of a Poisson random point field [4]: this is a random field, in which many points in any set of non-overlapping measurable sets \( B_i \subseteq Z \) represent mutually independent random variables with a Poisson distribution:

\[
P(\xi(B_i) = k) = \frac{(\Lambda(B_i))^k}{k!} \exp(-\Lambda(B_i)),
\]

where \( \Lambda(B_i) = E(\xi(B_i)) \), \( E \) is the symbol of mathematical expectation. Implementation of a random field on the set \( B \subseteq Z \) can be obtained in two stages. Initially, value of the random variable \( \zeta \) is modeled with the Poisson distribution (1) with average \( \Lambda(B) \). If \( \zeta = k \), then points \( z_j, j = 1, k \), in which \( \xi(z_j) > 0 \), are the results of the \( k \)-fold independent “throwing” of points on \( B \) with the probability of contact with any \( C \subseteq B \) equal to \( \Lambda(C)/\Lambda(B) \). In accordance with the above algorithm, the implementation of Poisson random point field can be represented as the presence in some area \( B \) of \( k \) points, selected from a uniform distribution in this area, and the number of \( k \) is selected from the Poisson distribution. Such implementation can describe, for example, the positions of fixed visible stars in the starry sky [5], but not the requirement flow of queues system. A mathematical model, more suitable for spatial queues systems, should reflect the dynamics of a random field on spatiotemporal argument. Such mathematical model can be obtained as a generalization to spatial case of a mathematical model of the Poisson random process [1, 2]. This paper presents a mathematical model of Poisson random field and modeling of its implementations.

2 Poisson random field
It is known that a scalar random field is the random scalar function \( \xi(\omega, z) = \xi(z) \in S \), defined on the probability space \( \{\Omega, F, P\} \), \( \omega \in \Omega \), and measurable for every value of
vector argument \( z = (z_1, z_2, ..., z_n) \in Z \subseteq \mathbb{R}^n \). A discrete scalar random field with a
continuous argument is the scalar random field \( \xi(z) \in S \) with finite or countable set of
values \( S = (s_i) \) and the continuum set \( Z \) of the argument value \( z \in Z \subseteq \mathbb{R}^n \).

We define scalar random field by scalar random function of scalar variable \( \xi(t) \in S \)
at \( t = \phi(z) \), where \( z = (z_1, z_2, ..., z_n) \in Z \subseteq \mathbb{R}^n \), \( \phi(z) \) is nonnegative and nondecreasing
function on every variable \( z_i \). In this definition, one of scalar variables \( z_1, z_2, ..., z_n \) can
be the time, and the other variables are considered as spatial variables. The value
\( z_2 = (z_{1,2}, z_{2,2}, ..., z_{n,2}) \) of the argument \( z = (z_1, z_2, ..., z_n) \) will be called "next" one
with regard to the value \( z_1 = (z_{1,1}, z_{2,1}, ..., z_{n,1}) \). If \( \tau = \phi(z_2 - z_1) \) is nonnegative, i.e., if
\( \tau = \phi(z_2 - z_1) > 0 \), then \( z_2 \geq z_1 \). The argument \( t = \phi(z) \) will be called the generalized
time and the function \( \phi(z) \) will be called the generating function, \( \tau = \phi(z_2 - z_1) \) is the
interval of generalized time, and \( \tau_i = \phi(z_{1,1}, ..., z_{i,2} - z_{i,1}, ..., z_{n,1}) \) is the interval on the
\( i^{th} \) argument, \( i = 1, \ldots, n \).

The random field defined as above will be called the scalar random field generated
by scalar random process [3]. It will be denoted as \( \xi(z) \in S, z \in Z \subseteq \mathbb{R}^n \).

Let assume that the generating random process \( \xi(t) \) is a Poisson one with the
parameter \( \lambda \). With regard to the so generated random field, the following theorem is
valid.

**Theorem 1** Let \( \xi(z), z = (z_1, z_2, ..., z_n) \in Z \subseteq \mathbb{R}^n \), be the scalar random field generated
by the Poisson random process \( \xi(t) \) with the parameter \( \lambda \) and nonnegative nondecreasing
function \( t = \phi(z) \) on every argument \( z_1, z_2, ..., z_n \). The increment \( \eta = \xi(z_2 - z_1) \)
of this field for the generalized time \( \tau = \phi(z_2 - z_1) \) is represented as weighted sum

\[
\eta = \sum_{i=1}^{n} \lambda \lambda_i \eta_i
\]

of independent increments \( \eta_i = \xi(z_{1,1}, ..., z_{i,2} - z_{i,1}, ..., z_{n,1}) \) on every coordinate \( z_i \) dis-
tributed under the Poisson law

\[
P(\eta_i = k/\tau_i) = \frac{\lambda \lambda_i \tau_i}{k!} e^{-\lambda \lambda_i \tau_i}, i = 1, n, k = 0, 1, 2, ...
\]

then and only then when the function \( \tau = \phi(z) = \phi(z_1, z_2, ..., z_n) \) is a linear function in
the form of

\[
t = \phi(z) = \sum_{i=1}^{n} \lambda_i z_i
\]

where \( \lambda_i \) are some non-negative constants. The increment \( \eta = \xi(z_2 - z_1) \) of a random
field for the generalized time \( \tau = \phi(z_2 - z_1) \) is distributed under the Poisson law with
the parameter \( \lambda \tau \).

As it follows from the theory of Poisson random processes, the parameter \( \lambda \) in the
theorem represents the average number of jumps of the random field in a generalized
interval unit of the generalized time \( \tau \). The parameter \( \lambda_i^* = \lambda \lambda_i, i = 1, n \), represents
the average number of jumps of random field in the interval unit \( \tau_i \) on \( i^{th} \) coordinate.

A random field, which increment for the generalized time \( \tau = \phi(z_2 - z_1) \) is distributed
under the Poisson law, will be called the Poisson random field. It follows from the
theorem, the Poisson random field can be generated by Poisson random process
only with linear generating function (2).
3 Computer simulation

The above theorem defines the following algorithm for modeling a single trajectory of points of increase of the proposed random field.

1. Let set the initial point $z_0 = (z_{1,0}, z_{2,0}, ..., z_{n,0})$ of the random field and model the value $\xi(z_0)$ of the random field in this point as a random number from the probability distribution of the initial state $a_i(z_0) = P(\xi(z_0) = s_i), i = 1, 2, 3, ..., s_i = i - 1$. This can be, for example, the Poisson distribution $Po(c)$:

$$a_i(z_0) = P(\xi(z_0) = s_i) = \frac{e^{i-1}}{(i-1)!} c^{-e}, i = 1, 2, 3, ..., s_i = i - 1. \quad (3)$$

2. Let model the intervals on individual coordinates between neighboring state transitions of field as random numbers $\gamma_i$ from exponential distributions $f_{\gamma_i}(\tau_i) = \lambda^* e^{-\lambda^* \tau_i}, \lambda^* = \lambda \lambda_i, i = 1, n$.

3. Let expect the coordinates of the point of increase of the Poisson field $z_{2,i} = z_{1,i} + \gamma_i, i = 1, n$.

4. Let return to the step 2 of the algorithm.

![Figure 1: Implementation of curve of points of increase of Poisson random field](image)

In accordance with the above algorithm, the numerical modeling of a random field on the plane $z_1, z_2$ with $\lambda_1 = 3, \lambda_2 = 1, \lambda = 1$ is defined, with the initial argument value $z_0 = (6, 4)$ ($t_0 = 3 \cdot 6 + 1 \cdot 4 = 22$), the final value $z = (20, 20)$ ($t = 3z_1 + 20$) and the parameter $c = 3$ in the formula (3).

Fig. 1 presents a separate implementation of the curve of points of increase of the random field. On the coordinate axes $z_1$ and $z_2$, the points of increase of the field
Figure 2: Set of implementations of curves of points of increase of Poisson random field are indicated on the coordinates. Fig. 1 shows that the field points are not evenly distributed in the rectangle $[6, 20] \times [4, 20]$, but are subjected to certain regularity. Fig. 2 presents a set of implementations of the curves of points of increase of the Poisson random field (dispersion field).

References


The main goal of the work is to study the limit behavior of optimal stopping and exit times for some classes of random processes, in particular Ito’s diffusion, random walk and diffusion process with non-Lipschitz diffusion coefficient.

1 Limit behavior of optimal stopping times for Ito’s diffusion

We consider an asset whose price is the solution of the following linear stochastic differential equation
\[ dX_n(t) = r_n X_n(t) dt + \alpha_n X_n(t) dW(t), X_n(0) = x > 0, n \geq 0. \]

Definition 1. The discounted profit is given by
\[ g_n(s, x) = e^{-\rho_n s} \cdot (x - \alpha_n). \]

Definition 2. The optimal stopping time is defined as follows
\[ \tau^*_n = \argmax_{\tau \in \Gamma} \left[ E^{(s, x)} e^{-\rho_n \tau} \cdot (X_n(\tau) - \alpha_n) \right], \]
where \( \Gamma \) is the set of all stopping times.

Definition 3. Optimal discounted profit is of the following form
\[ g^*_n(s, x) = E^{(s, x)}[g_n(\tau^*_n, X_n(\tau^*_n))]. \]

We consider the case when \( r_n < \rho_n \). Explicit form of \( \tau^*_n \) was established, e.g. by the Oksendal in [1].

Theorem 1. Let \( r_n, \alpha_n, \rho_n, \alpha_n \) converge to \( r_0, \alpha_0, \rho_0, \alpha_0 \). Then the following convergence holds for every \( \epsilon > 0 \)
\[ P(|\tau^*_n - \tau^*_0| > \epsilon) \longrightarrow 0. \]
Moreover the stopping time \( \tau^*_0 \) is optimal for the limit process \( X_0(t) \) with the limit parameters \( r_0, \alpha_0, \rho_0, \alpha_0 \).

The convergence of optimal discounted profit \( g^*_n \) to the optimal discounted profit of the limit process also holds.
2 Optimal stopping problem for a random walk with polynomial reward function

For a random walk $X_t$, $t \in N^+$ with a drift to the left and polynomial reward function of the following form

$$g(x) = \sum_{k=1}^n C_k \cdot (x^+)^k, C_k \in \mathbb{R}$$

we study the optimal stopping problem of finding the optimal stopping time

$$\tau^* = \arg\max_{\tau \in \Gamma_{0}} E_x g(X_\tau) I\{\tau < \infty\}$$

using the Appel polynomials $Q_k(y)$ of the random variable $M = \sup_{k \geq 0}(X_k - X_0)$. Here $\Gamma_{0}^\infty$ is the set of all Markov times on $[0, \infty]$. This method was first proposed by Novikov and Shiryaev in the work [2] for the particular case $n = 1$.

**Definition 4.** Appel polynomials $Q_k$ of order $k \geq 0$ for random variable $M$ are defined through the expansion

$$\frac{\exp(u \cdot y)}{\mathbb{E} \exp(u \cdot M)} = \sum_{k=0}^{\infty} \frac{u^k}{k!} Q_k(y).$$

We establish the necessary and sufficient conditions for $C_k$ under which the linear combination of Appel polynomials associated with reward function has unique positive root $x_n$.

**Theorem 2.** Under the conditions mentioned above the optimal stopping time is the first exit time of the random walk $X_t$ from the interval $[-\infty, x_n]$.

Examples of random walks and reward functions which satisfy the condition of uniqueness of the positive root $x_n$ are constructed.

3 Asymptotics of exit times for diffusion processes

We study the process that is the solution of the following stochastic differential equation

$$X_n(t) = X_n(0) + \int_0^t b_n(s, X_n(s))ds + \int_0^t \sigma_n(s, X_n(s))dW(s), n \geq 0, t \geq 0$$

with non-random initial conditions and coefficients satisfying the following assumptions

**(a)** coefficients $b_n(s, x)$ and $\sigma_n(s, x)$ are continuous in $s$ and $x$;

**(b)** coefficients $b_n(s, x)$ and $\sigma_n(s, x)$ are of the linear growth;

**(c)** coefficient $b_n(s, x)$ satisfies the Lipschitz condition in space;

**(d)** coefficient $\sigma_n(s, x)$ satisfies the Yamada condition in space;
It was proved by Yamada [3] that under conditions (a) – (d) the stochastic differential equation (1) has unique strong solution $X_n$.

The following pointwise convergence of coefficients and initial condition is assumed

$$b_n(t, x) \to b_0(t, x), \sigma_n(t, x) \to \sigma_0(t, x), n \to \infty,$$

and $X_n(0) \to X_0(0), n \to \infty$.

**Theorem 3.** Under convergence (2) the process $X_n$ converges uniformly in probability to the process $X_0$ on the arbitrary interval $[0, T]$.

The above theorem is the generalization of result obtained in [4] on the weaker Yamada conditions.

Let $\tau_n$ and $\tau_0$ be the first exit times of the processes $X_n$ and $X_0$ from the interval $[l, r]$ where $l < X_n(0) < r, n \geq 0$.

**Theorem 4.** Under convergence (2) the following convergence of exit times take place for every $\epsilon > 0$

$$P(|\tau_n - \tau_0| > \epsilon) \longrightarrow 0.$$

The obtained result is useful in the investigation of the optimal stopping times convergence for diffusion processes with non-Lipschitz diffusion.

**References**


ON COMPARISON OF CLASSICAL AND CEPSTRUM-BASED FORECASTS

V.A. Valoshka*, Yu.S. Kharin**
Research Institute for Applied Problems of Mathematics and Informatics
Minsk, BELARUS
e-mail: *valoshka@bsu.by, **kharin@bsu.by

Abstract

The alternative cepstrum-based approach to forecasting is experimentally compared with the classical correlation-based one. For the class of time series with the polymodal spectral densities the alternative forecast is shown to be sufficiently more accurate then the classical forecast, when both forecasts use the same small number of estimated parameters.

1 Introduction

The usual approach in linear forecasting of a wide-sense stationary time series with unknown spectral density is to estimate some correlation coefficients and to compute estimates of the optimal forecasting coefficients. When the correlation function vanishes fast, its enough to estimate a small number of correlation coefficients, and the traditional methods bring a good result. But for the long-memory processes, when the large number of correlations should be estimated, it is useful to have some suitable parsimonious model of the spectral density. Then we can accurately estimate a few parameters of this model and transform it to the estimates of a necessary number of correlations. This paper presents some experimental results, demonstrating the benefit of using the parsimonious Bloomfield model [2] for forecasting time series with polymodal spectral density.

2 Mathematical model and algorithms

Let \( \{x_t\}_{t \in \mathbb{Z}} \) be the Gaussian stationary time series with zero mean \( E\{x_t\} = 0 \), covariance function \( \sigma_\tau = E\{x_t x_{t+\tau}\}, \tau \in \mathbb{Z} \), and the spectral density \( S(\lambda) = \sum_{\tau \in \mathbb{Z}} \sigma_\tau \cos(\tau \lambda), \lambda \in \Pi = [-\pi, \pi] \), which is known to be non negative. The correlation coefficients (or correlations) are \( \theta_\tau = \sigma_\tau / \sigma_0 \). The theoretical cepstral coefficient [2] is the fourier coefficient of the logarithm of spectral density:

\[
l_\tau = \frac{1}{2\pi} \int_\Pi \ln S(\lambda) \cos(\tau \lambda) d\lambda, \ \tau \in \mathbb{Z}.
\]

For the time series \( x_1, \ldots, x_T \) of the length \( T \in \mathbb{N} \) we should fix some smooth parameter \( p \in \mathbb{N} \) and compare two approaches of one step forward forecasting for \( x_{T+1} \) on the base of the mean squared error, or risk:

\[
R = E \left\{ (x_{T+1} - \hat{x}_{T+1})^2 \right\}.
\]
In the traditional approach the first \( p \) correlations \( \theta_1, \ldots, \theta_p \) are straightly estimated by the sample covariances, and then the linear forecast of depth \( p \) is computed. In the alternative approach the first \( p \) cepstrum coefficients \( l_1, \ldots, l_p \) are estimated, and then the estimates of the first \( T \) correlations \( \hat{\theta}_1, \ldots, \hat{\theta}_T \) and finally the linear forecast of depth \( T \) are computed. In [2] it was suggested that the alternative approach reaches a sufficient gain w.r.t. the traditional one for the time series with polymodal spectral densities, which are concentrated in the thin neighborhoods of a few values of \( \lambda \in \Pi \). Here some experimental results supporting this hypothesis are presented.

Introduce the algorithms. The biased sample covariances and classical estimates of correlations [1]:
\[
\hat{\sigma}_\tau = \hat{\sigma}_{-\tau} = \frac{1}{T} \sum_{t=1}^{T-\tau} x_t x_{t+\tau}, \quad \hat{\theta}_\tau^C = \hat{\sigma}_\tau / \hat{\sigma}_0, \quad \tau = 0, 1, \ldots, T - 1.
\]
The smoothed periodogram and the separated from zero one:
\[
\hat{S}(\lambda) = \sum_{\tau=1-K}^{K-1} \hat{\sigma}_\tau f(\tau/K) \cos(\tau \lambda), \quad \hat{S}^+(\lambda) = \max \left\{ \hat{S}(\lambda), \varepsilon \hat{\sigma}_0 \right\}, \quad \lambda \in \mathbb{R}.
\]
Here \( f : \mathbb{R} \to \mathbb{R}, \quad K \in \mathbb{N} \) and \( \varepsilon > 0 \) are respectively the kernel of smoothing, the width of the smoothing window and zero separator. It will be clear from the results of experiments that varying these three parameters can sufficiently increase the accuracy of the alternative (cepstrum-based) forecast. Introduce the plug-in cepstrum estimates [1] and the alternative estimates of the correlations:
\[
\hat{\theta}_\tau^A = \frac{\Theta_\tau}{\Theta_0}, \quad \Theta_\tau = \frac{1}{2\pi} \int_\Pi \exp \left( 2 \sum_{j=1}^{p} \hat{l}_j \cos(j \lambda) \right) \cos(\tau \lambda) d\lambda,
\]
where the integrals for \( \Theta_\tau \) in the algorithm are computed approximately by numerical methods. Finally using Durbin-Levinson formulas [2] the classical estimates of the first \( p \) correlations \( \hat{\theta}_1^C, \ldots, \hat{\theta}_p^C \) are transformed into the estimates of the optimal forecasting coefficients \( \hat{a}_1^C, \ldots, \hat{a}_p^C \) at the depth of \( p \). Similarly, the alternative estimates of the first \( T \) correlations \( \hat{\theta}_1^A, \ldots, \hat{\theta}_T^A \) are transformed into the estimates of the optimal forecasting coefficients \( \hat{a}_1^A, \ldots, \hat{a}_T^A \) at the depth of \( T \). The classical and the alternative forecasting statistics for \( x_{T+1} \) are:
\[
\hat{x}_{T+1}^C = \sum_{t=T-p+1}^{T} x_t \hat{a}_{T+1-t}^C, \quad \hat{x}_{T+1}^A = \sum_{t=1}^{T} x_t \hat{a}_{T+1-t}^A,
\]
and the corresponding risks to be examined:
\[
\mathcal{R}_C(S, T, p) = E \left\{ (\hat{x}_{T+1}^C - x_{T+1})^2 \right\}, \quad (2)
\]
\[
\mathcal{R}_A(S, T, p; f, K, \varepsilon) = E \left\{ (\hat{x}_{T+1}^A - x_{T+1})^2 \right\}. \quad (3)
\]
3 Computer experiments

For illustration of the gain of alternative forecast w.r.t. the classical one we used the Gaussian time series with the polymodal spectrum specified by the Bloomfield model of fifth order:

\[ S(\lambda) = \exp \left( 2 \sum_{j=1}^{5} l_j \cos(j\lambda) \right), \] (4)

\[ (l_1, \ldots, l_5) = (-0.00864, -1.15869, -1.97555, 0.06574, -3.36081). \]

The time series \( x_t \) was generated as \( x_t = \sum_{j=0}^{123} c_j \xi_{t+j} \), using the standard discrete Gaussian white noise \( \xi_t \) and 124 pre-calculated moving average coefficients \( c_j \), associated with the spectral density (4): \( S(\lambda) = \left| \sum_{j \geq 0} c_j e^{ij\lambda} \right|^2, \lambda \in \mathbb{R}, i = \sqrt{-1}. \) The plots of moving average coefficients \( c_j \), spectral density (4) and time series realization \( x_t \) are presented in Figure 1.

![Figure 1](image_url)

Figure 1: Moving average coefficients (A), spectral density (B) and time series realization (C)

The risks (2) and (3) was computed by Monte-Carlo method with 3000 replications. The number \( p \) of estimated parameters was set to five for both classical and alternative forecasts. In (1) the following kernels of smoothing were used [1]:

\[ f_{\cos}(x) = \frac{1 + \cos(\pi x)}{2}, \quad f_{\text{const}}(x) \equiv 1, \quad f_{\text{Parzen}(q)}(x) = 1 - |x|^q. \]
The optimal widths $K$ of smoothing windows was manually selected for every kernel. The value $\varepsilon = 10^{-11}$ was chosen in a similar way. The plots of the Monte-Carlo estimates of the risks (2) and (3) on a logarithmic scale are presented in Figure 2; they illustrate significant gain of the risk (3) for the Bloomfield forecasting statistic w.r.t. the risk (2) for the classical one. Moreover, the best of alternative forecasts has the smallest optimal width of the smoothing window ($K = 12$) and the kernel $f(x)$, smooth both at $x = 0$ and at $x = 1$. The average two of alternative forecasts have close optimal widths of the smoothing windows ($K = 17$ and $K = 20$) and the kernels $f(x)$, smooth at $x = 0$ and nonsmooth at $x = 1$. Finally, the worst one of alternative forecasts have the largest optimal width of the smoothing window ($K = 77$) and the kernel $f(x)$, nonsmooth both at $x = 0$ and at $x = 1$. This trend suggests an obvious assumption on preferability of using kernels $f(x)$, totally smooth on $\mathbb{R}$.

![Figure 2: Risks of classical and alternative forecasts (logarithmic scale)](image)

**References**


ON FORMULAS OF LINEAR INTERPOLATION FOR OPERATORS GENERATED BY SOLUTIONS OF STOCHASTIC DIFFERENTIAL EQUATIONS

L.A. YANOVIICH

National Academy of Sciences, Institute of Mathematics
Minsk, BELARUS

e-mail: yanovich@im.bas-net.by

Abstract

Formulas of linear interpolation with respect to two or three nodes for the operator, which gives a solution of a linear stochastic differential equation, are constructed.

1 Introduction

The solutions of stochastic differential equations and their probabilistic characteristics are operators depending on the input random or deterministic functions and numerical parameters. These operators have a special structure and in the vast most cases are nonlinear [1-3]. For their approximation methods of operator interpolation can be used. One such approach is illustrated below in the example of linear stochastic equations.

2 The results

Consider the scalar linear stochastic differential equation of the form

\[ dX(t) = [a_1(t)X(t) + a_2(t)]dt + [a_3(t)X(t) + a_4(t)]dW(t) \]  

with the initial condition \( X(0) = X_0 \), where \( X_0 \) and deterministic continuous on \( T \subseteq \mathbb{R}_+ \) vector-function \( a(t) = (a_1(t), a_2(t), a_3(t), a_4(t)) \) are given; \( W(t) = W(t, \omega) \) is standard Wiener process defined by a complete probabilistic space \( \{\Omega, F, P\} \) with a distinguished on it right continuous flow of \( \sigma \)-algebras \( F_t (t \geq 0, \omega \in \Omega) \).

The solution \( X(t) = X(t, \omega) \) of the problem can be written as

\[ X(t) = F(a(t)) = F(a_1, a_2, a_3, a_4), \]

where the random operator \( F(a) \) dependent on the coefficients \( a_i(t) \) \( (i = 1, 2, 3, 4) \) of the equation (1), has the form

\[
F(a(t)) = p_0(t) \left[ X_0 \exp\{\int_0^t a_3(\tau)dW(\tau)\} + \int_0^t p_1(s) \exp\{\int_s^t a_3(\tau)dW(\tau)\}ds + \int_0^t p_2(s) \exp\{\int_s^t a_3(\tau)dW(\tau)\}dW(s) \right].
\]
In the formula (2) functions \( p_i(t) \) \( (i = 0, 1, 2) \) are given by

\[
p_0(t) = \exp\left\{ \int_0^t \left[ a_1(\tau) - \frac{1}{2}a_3^2(\tau) \right] d\tau \right\},
\]

\[
p_1(t) = \frac{a_2(t) - a_3(t)a_4(t)}{p_0(t)}, \quad p_2(t) = \frac{a_4(t)}{p_0(t)}.
\]

(3)

The linear interpolation problem for operators of such form is construction of linear with respect to the coordinates \( a_i(t) \) \( (i = 1, 2, 3, 4) \) of vector \( a(t) \) of operator \( L_1(F; a) \) such that at points \( \tilde{a}_i = \tilde{a}_i(t) = (a_{1i}(t), a_{2i}(t), a_{3i}(t), a_{4i}(t)) \) \( (i = 0, 1) \), the equalities \( L_1(F; \tilde{a}_i) = F(\tilde{a}_i) \) \( (i = 0, 1) \) hold true. Two types of such interpolation polynomials are studied. The first of them contains the Gateaux differential of interpolated operator \( F(a) \) and has the form

\[
L_1(F; a) = F(\tilde{a}_0) + \int_0^1 \delta F \left[ \tilde{a}_0 + \tau(\tilde{a}_1 - \tilde{a}_0); a - \tilde{a}_0 \right] d\tau,
\]

(4)

where Gateaux differential of operator \( F(a) \) at the point \( a = \tilde{a}_0 + \tau(\tilde{a}_1 - \tilde{a}_0) \) in the direction \( a - \tilde{a}_0 \) in (4) is a total differential of operator \( F(a) = F(a_1, a_2, a_3, a_4) \), which is defined as the sum of partial differentials

\[
\delta F [a; h] = \sum_{i=1}^{4} \delta a_i F [a_1, a_2, a_3, a_4; h_i]
\]

respectively in the directions \( h_i = a_i - a_{i0} \).

More visual version of the formula (4) is as follows

\[
L_1(F; a) = F(a_{10}, a_{20}, a_{30}, a_{40}) + \sum_{i=1}^{4} \int_0^1 \delta a_i F[a_{10} + \tau(a_{11} - a_{10}), a_{20} + \tau(a_{21} - a_{20}),
\]

\[a_{30} + \tau(a_{31} - a_{30}), a_{40} + \tau(a_{41} - a_{40}); a_i - a_{i0}] d\tau.
\]

Condition \( L_1(F; \tilde{a}_i) = F(\tilde{a}_i) \) for \( i = 1 \) is easy to check, using the well-known [4] ratio

\[
\frac{d}{d\tau} F(\tilde{a}_0 + \tau(\tilde{a}_1 - \tilde{a}_0)) = \delta F [\tilde{a}_0 + \tau(\tilde{a}_1 - \tilde{a}_0); \tilde{a}_1 - \tilde{a}_0].
\]

The validity of \( L_1(F; \tilde{a}_0) = F(\tilde{a}_0) \) is obvious.

If \( a_3(t) \equiv 0 \) then the operator \( F(a(t)) \) is transformed to

\[
F(a(t)) = \exp\{ \int_0^t a_1(\tau)d\tau \} X_0 +
\int_0^t a_2(s) \exp\{ \int_s^t a_1(\tau)d\tau \} ds + \int_0^t a_4(s) \exp\{ \int_0^t a_1(\tau)d\tau \} dW(s).
\]

The mathematical expectation of the solution \( X(t) = F(a(t)) \) in this case has the form

\[
E\{X(t)\} = \exp\{ \int_0^t a_1(\tau)d\tau \} \left[ X_0 + \int_0^t a_2(s) \exp\{ \int_0^s a_1(\tau)d\tau \} ds \right].
\]

(5)
Another version of the formula of linear interpolation is given by special form of the Stieltjes integral. Let us assume that the function \( F[\tilde{a}_1 + \tau(\tilde{a}_1 - \tilde{a}_0)] \) has a bounded variation on the segment \([0, 1]\) with respect to numerical parameter \( \tau \). We now consider the formula of the form

\[
L_1(F; a) = F(\tilde{a}_0) + \int_0^1 \Omega(a(\tau), \tilde{a}_0(\tau), \tilde{a}_1(\tau))d\tau F(\tilde{a}_0 + \tau(\tilde{a}_1 - \tilde{a}_0)),
\]

where

\[
\Omega(a(\tau), \tilde{a}_0(\tau), \tilde{a}_1(\tau)) = \frac{1}{4} \sum_{i=1}^{4} \frac{a_i(\tau) - a_{i0}(\tau)}{a_{i1}(\tau) - a_{i0}(\tau)},
\]

and the integral in (6) is understood as a Stieltjes integral for each fixed trajectory of the Wiener process \( W(t) \). We must require \( a_{i1}(\tau) \neq a_{i0}(\tau) \) for \( \tau \in [0, 1] \). It is easy to verify that for the formula (6) the interpolation conditions \( L_1(F; \tilde{a}_i) = F(\tilde{a}_i) \) for \( i = 0, 1 \) also hold true.

Interpolation formulas (4) and (6) coincide with the value of operator \( F(a) \) at the two nodes \( \tilde{a}_0(t) \) and \( \tilde{a}_1(t) \). Based on these formulas the linear operator polynomials, for which the interpolation conditions hold true not only at these nodes, but also at the third node specially picked up in addition, can also be constructed.

Consider the vector \( p(t) = (p_1(t), p_2(t), p_3(t), p_4(t)) \) with mutually orthonormal on the segment \([0, 1]\) with respect to weight \( \rho(t) \) coordinates \( p_i(t) \), i.e.

\[
\int_0^1 \rho(t)p_i(t)p_j(t)dt = \delta_{ij} \quad (1 \leq i, j \leq 4),
\]

with \( \delta_{ij} \) being the Kronecker symbol. By \((a(t), p(t))\) we denote the scalar product of the vectors \( a(t) \) and \( p(t) \). Then for the operator first-degree polynomial with respect to variables \( a_i(t) \) \( (i = 1, 2, 3, 4) \) of the form

\[
L_{11}(F; a) = L_1(F; a) + \frac{1}{4} r_1(a) \int_0^1 \rho(\tau)(a(\tau), p(\tau))d\tau,
\]

where \( r_1(a) = F(a) - L_1(F; a) \), the equalities \( L_1(F; \tilde{a}_i) = F(\tilde{a}_i) \) \( (i = 0, 1) \), \( L(F; p) = F(p) \) hold true. It is true, since \( r_1(\tilde{a}_0) = r_1(\tilde{a}_1) = 0 \), and the integral in equation (7) when \( a(\tau) = p(\tau) \) is equal 4.

The interpolation formulas (4), (6) and (7) presented here can be used not only to random operators, which define an explicit solution of the equation, but also to the mathematical expectations of these solutions, in particular, to the operator (5), which depends on function variables \( a_1(\tau) \ a_2(\tau) \), and also to other classes of operators.

For example, if the random operator \( F(a(t)) \) depending on the vector \( a(t) = (a_1(t), a_2(t)) \) has the form

\[
F(a(t)) = \sin(a_1(t) + W(t)) + \cos(a_2(t) + W(t)),
\]

two interpolation nodes \( \tilde{a}_1(t) = (a_{10}(t), a_{11}(t)) \) and \( \tilde{a}_2(t) = (a_{20}(t), a_{21}(t)) \) are given, then the integral in (4) can be exactly calculated and so this interpolation formula takes the form generally accepted for the case of scalar functions.
\[ L_1(F; a) = \frac{a_1(t) - a_{10}(t)}{a_{11}(t) - a_{10}(t)} F(\bar{a}_1(t)) + \frac{a_2(t) - a_{20}(t)}{a_{21}(t) - a_{20}(t)} F(\bar{a}_2(t)). \] (9)

For the operator (8) we present another version of the formula using the form (7):

\[ L_{11}(F; a) = L_1(F; a) + \frac{1}{2} r_1(a) \int_0^1 \left[ \frac{1}{\alpha + 1} a_1(\tau) \tau^\alpha + \frac{1}{\beta + 1} a_2(\tau) \tau^\beta \right] d\tau, \]

where \( r_1(a) = F(a) - L_1(F; a) \), \( L_1(F; a) \) is given by (9), \( \alpha, \beta \) are given numbers \( \alpha \geq 0, \beta \geq 0 \). For the formula (10) the interpolation conditions at the points \( \bar{a}_i \) \( (i = 0, 1) \) \( \bar{a}_2 = (1, 1) \) hold true.

Operator interpolation, as one of methods of approximation for operators can be used for approximate calculation of both deterministic and stochastic linear and non-linear operators. Numerous interpolation formulas for the operators defined in the general linear, Hilbert and function spaces are constructed in [5,6].

**References**


Научное издание

КОМПЬЮТЕРНЫЙ АНАЛИЗ ДАННЫХ И МОДЕЛИРОВАНИЕ: ТЕОРЕТИЧЕСКАЯ И ПРИКЛАДНАЯ СТОХАСТИКА

ТРУДЫ
Десятой Международной научной конференции
Минск, 10—14 сентября 2013 г.

В двух томах
ТОМ 1

На английском языке

Ответственный за выпуск В. С. Повколас

Подписано в печать 08.08.2013. Формат 60×84 1/8. Бумага офсетная.
Ризография. Усл. печ. л. 26,97. Уч.-изд. л. 23,36.
Тираж 130 экз. Заказ 511.

Республиканское унитарное предприятие
«Издательский центр Белорусского государственного университета».
ЛИ № 02330/0494361 от 16.03.2009. Ул. Красноармейская, 6, 220030, Минск.

Отпечатано с оригинала-макета заказчика
в республиканском унитарном предприятии
«Издательский центр Белорусского государственного университета».
ЛП № 02330/0494178 от 03.04.2009. Ул. Красноармейская, 6, 220030, Минск.
ДЛЯ ЗАМЕТОК
ДЛЯ ЗАМЕТОК
ДЛЯ ЗАМЕТОК