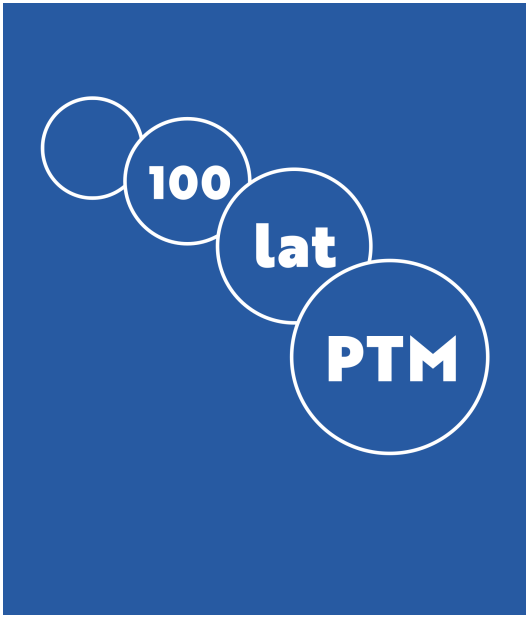

statystyka

patron sesji: Jerzy Spława-Neyman

ptm rok zał. 1919

Jubileuszowy Zjazd Matematyków Polskich

w stulecie

Polskiego Towarzystwa Matematycznego

Kraków 3 -7 września 2019

Spis treści

The cooperation of nuclear and ℓ_1 norms in revealing the association between HIV disease and brain connectivity structure

Damian Brzyski
damian.brzyski@pwr.edu.pl
Politechnika Wrocławska

Classical regression methods treat covariates as a vector and estimate a corresponding vector of regression coefficients. In medical applications, however, regressors in a form of multidimensional arrays can be often met. For example, one may be interested in identifying regions of the brain associated with an outcome of interest based on MRI images. Turning such image array into a vector is an unsatisfactory solution, since it destroys the inherent spatial structure of the image and could be very challenging from the computational point of view. In my talk, I will present an alternative approach, where the whole matrix of regression coefficients is estimated. The method we propose, called Sparsity Inducing Nuclear Norm Estimator (SpINNEr), simultaneously imposes two types of penalties on the matrix – the nuclear and ℓ_1 norms – to encourage the low rank of the solution and its entry-wise sparsity. Our software allows for the automatic selection of the weights defining the optimal trade-off between two considered types of penalties. SPINNER has been applied to investigate associations between brain's structural connections and HIV disease-related outcomes.

References

[1] H. Zhou, Lexin Li *Regularized matrix regression*, JRSS-B, 76(2), 2013.

● Powrót do indeksu abstraktów sekcji

Estimation of treatment effects in cancer clinical trials when the proportional hazard assumption is not fulfilled

Tomasz Burzykowski
tomasz.burzykowski@uhasselt.be
Hasselt University, Belgia

Currently, treatment effects in cancer clinical trials with time-to-event endpoints are estimated almost exclusively by using the proportional-hazard (PH) model. The PH assumption is, however, very restrictive and is almost surely not fulfilled in the case of, for instance, modern immunotherapies. Thus, there is a need to develop and use other methods of the estimation of treatment effects. In the presentation several such methods will be reviewed, including the accelerated failure-time model, the restricted mean survival time, and the cure-fraction model.

● Powrót do indeksu abstraktów sekcji

Efektywność pośrednia w nieparametrycznych problemach testowania

Bogdan Ćmiel
cmielbog@mat.agh.edu.pl
Akademia Górnicz-Hutnicza

Zdefiniujemy efektywność pośrednią, omówimy jej własności oraz przedyskutujemy jej praktyczną interpretację. Pokażemy, że efektywność pośrednia jest dobrym narzędziem do teoretycznego porównywania testów nieparametrycznych. Przedstawimy twierdzenia o efektywności pośredniej oraz symulacje numeryczne potwierdzające użyteczność tych twierdzeń na przykładach problemów testowania stochastycznego uporządkowania oraz testowania zgodności. Twierdzenia te oraz symulacje będą dotyczyły testów typu Kołmogorowa-Smirnowa i Andersona-Darlinga. Przedstawiony referat będzie streszczeniem wyników z pracy Inglot et al. (2019) i Ćmiel et al. (2019).

Bibliografia

[1] T. Inglot, T. Ledwina, B. Ćmiel, *Intermediate efficiency in nonparametric testing problems with an application to some weighted statistics*, ESAIM: Probability and Statistics, https://doi.org/10.1051/ps/2018022 (2019)

[2] B. Ćmiel, T. Inglot, T. Ledwina, *Intermediate effciency of some weighted goodness-of-fit statistics*, http://arxiv.org/abs/1906.09143 (2019)

• Powrót do indeksu abstraktów sekcji

(dual-frequency)-dependent dynamic functional connectivity analysis of visual working memory capacity

Anna Dudek
aedudek@agh.edu.pl
Akademia Górniczo-Hutnicza

We develop a novel methodology for studying the dynamic functional connectivity within the brain from EEG traces. Our observations consist of replicated realizations of spatio-temporal processes that are locally time-harmonizable. Our bootstrap-based methodology estimates confidence intervals for both the spatial time-varying Loève-spectrum and the spatial time-varying dual-frequency coherence functions under realistic modeling assumptions. We illustrate the application of this methodology on a data set arising from an experiment designed to assess the capacity of the visual working memory. Our real data analysis pipeline starts with the clustering of our replicated time series obtained from toroidal mixture modeling of the corresponding response variables which describe the quality of memorization. Then we estimate the spatial time-varying dual-frequency coherence functions and the corresponding connectivity matrices within each cluster. This procedure allows us to potentially identify specific patterns in the dynamic functional connectivity characterizing each cluster. More specifically we reveal that better visual working memory performance is apparently associated to longer connectivity period within the prefrontal cortex between the alpha-beta frequency bands during the

memorization task.

Joint work with J. Aston, D. Dehay, J-M. Freyermuth, D. Scucs, L. Colling

References

[1] J. Aston, D. Dehay, A.E. Dudek, J-M. Freyermuth, D. Scucs, L. Colling, *(dual-frequency)-dependent dynamic functional connectivity analysis of visual working memory capacity*, submitted.

● Powrót do indeksu abstraktów sekcji

Analysis of Langevin Monte-Carlo via convex optimization

Alain Durmus
alain.dirmus@cmla.ens-cachan.fr
ENS Paris-Saclay

We provide new insights on the Unadjusted Langevin Algorithm. We show that this method can be formulated as a first order optimization algorithm of an objective functional defined on the Wasserstein space of order 2. Using this interpretation and techniques borrowed from convex optimization, we give a non-asymptotic analysis of this method to sample from log concave smooth target distribution. Our proofs are then easily extended to the Stochastic Gradient Langevin Dynamics, which is a popular extension of the Unadjusted Langevin Algorithm. Finally, this interpretation leads to a new methodology to sample from a non-smooth target distribution, for which a similar study is done. This is a joint work with Szymon Majewski and Błażej Miasojedow. • Powrót do indeksu abstraktów sekcji

Testing covariance structure and estimation of unknown parameters under doubly multivariate models

Katarzyna Filipiak
katarzyna.filipiak@put.poznan.pl
Politechnika Poznańska

The covariance matrix of doubly multivariate data has often a separable structure, that is it can be presented as a Kronecker product of two positive definite matrices. In particular, one of the separability components can be further specified, as e.g. compound symmetry or autoregression of order one. In this talk two methods of covariance structure fitting will be presented, as well as two testing procedures based on the likelihood ratio and Rao score test will be developed. Using simulation studies it will be shown that the Rao score test outperforms the likelihood ratio test in the number of contexts. Finally, the estimators of unknown parameters under the tri-linear multivariate model will be given.

All the results will be illustrated by real data examples.

References

[1] K. Filipiak and D. Klein, *Approximation with a Kronecker product structure with one component as compound symmetry or autoregression*, Linear Algebra Appl. **559**: 11–33 (2018).

[2] K. Filipiak and D. Klein, *Estimation of parameters under a generalized growth curve model*, J. Multivariate Anal. **158**: 73–86 (2017).

[3] K. Filipiak, D. Klein, and A. Roy, *A comparison of likelihood ratio tests and Rao's score test for three separable covariance matrix structures*, Biometrical J. **59**: 192–215 (2017).

[4] K. Filipiak, D. Klein, and A. Roy, *Score test for a separable covariance structure with the first component as compound symmetric correlation matrix*, J. Multivariate Anal. **150**: 105-124 (2016).

● Powrót do indeksu abstraktów sekcji

Szacowanie błędu klasyfikacji w źle wyspecyfikowanym modelu regresji binarnej

Konrad Furmańczyk
konfur@wp.pl
Wydział Zastosowań Informatyki i Matematyki SGGW

W referacie zostanie przedstawione oszacowanie ekscesu ryzyka dla błędu klasyfikacji w źle wyspecyfikowanym modelu regresji binarnej. Podamy wyniki dla klasyfikacji opartej o regresje logistyczną oraz klasyfikacje liniową. Uzyskane wyniki są oparte na pracach [1]–[3].

Bibliografia

[1] M. Kubkowski, *Misspecification of binary regression model: properties and inferential procedures. Ph.D. Thesis*, Warsaw University of Technology, Warsaw, 2019.

[2] M. Kubkowski and J. Mielniczuk, *Selection consistency of two-step selection method for misspecified logistic model*, Submitted

[3] T. Zhang, *Statistical behavior and consistency of classification methods based on convex risk minimization*, Ann. Stat. **32**: 56–134 (2004).

● Powrót do indeksu abstraktów sekcji

Nieparametryczna estymacja prawdopodobieństw ruiny w modelach przełącznikowych

Lesław Gajek
leslaw.gajek@p.lodz.pl
Politechnika Łódzka

Modele przełącznikowe uogólniają wiele znanych modeli ryzyka w ubezpieczeniach (zobacz na przykład Lu (2006)). Wykorzystują one łańcuch Markowa, który zmienia rozkład wielkości szkody lub/i czasu oczekiwania na nią, natomiast ubezpieczyciel może dostosować swoją strategię zmieniając wysokość składki. W pracy Gajek, L. , Rudź, M. (2018) pokazano, że wektor prawdopodobieństw ruiny w nieskończonym horyzoncie czasu, Ψ, jest punktem stałym wektorowego operatora ryzyka w tym modelu. Operator ryzyka jest kontrakcją w odpowiedniej przestrzeni metrycznej, co ma istotne znaczenie dla naszej metody estymacji. Konstruując odpowiedni empiryczny operator ryzyka, definiujemy jego punkt stały Ψ_n jako estymator wektora Ψ. Pokażemy, że estymator Ψ_n jest zbieżny do Ψ w odpowiedniej ważonej metryce średniokwadratowej.

Bibliografia

[1] Y. Lu, *On the severity of ruin in a Markov-modulated risk model*, Scandinavian Actuarial Journal, **4**: 183–202 (2006).

[2] L. Gajek and M. Rudź, *Banach Contraction Principle and ruin probabilities in regime switching-models*, Insurance: Mathematics and Economics, **80**: 45–53 (2018).

Powrót do indeksu abstraktów sekcji

Model selection in the space of coloured Gaussian models

Piotr Graczyk
graczyk@univ-angers.fr
Larema Université d'Angers, Francja

In order to make Graphical Gaussian Models a viable modeling tool in the modern Big Data Science, i.e. when the number of variables outgrows the number of observations, Højsgaard and Lauritzen introduced in 2008 model classes which set equality restrictions on certain entries of covariance matrix or precision matrix. Such models can be represented by **coloured** graphs.
The estimation theory for Coloured Graphical Models is well established, whereas the Model Selection within the Coloured Graphical Models class is still not satisfactory. We consider multivariate Gaussian models for the random variable $X = (X_1, \ldots, X_p)$, invariant under the action of a subgroup of the group of permutations on $\{1, \ldots, p\}$. Using the representations of the symmetric group on the field of reals, we derive the analytic expression of the normalizing constant of the Diaconis-Ylvisaker conjugate prior for the precision parameter $K = \Sigma^{-1}$. We can thus perform Bayesian model selection in the class of complete Gaussian models invariant (coloured) by the action of a subgroup of the symmetric group, which we could also call complete RCOP models. We illustrate our results with Frets' Heads example of dimension 4, several simulated examples of dimension $p < 9$ and a high-dimensional example with $p = 100$ in the case of a cyclic

group generated by one permutation.

This is a joint work with H. Ishi(Nagoya), B. Kołodziejek(Pol. Warszawska) and H. Massam(Toronto).

References

[1] Hojsgaard, S.; Lauritzen, S. L. Graphical *Gaussian models with edge and vertex symmetries*. J. R. Stat. Soc. Ser. B, 70:1005–1027, 2008.

[2] Maathuis, M.,Drton, M., Lauritzen, S. and Wainwright, M. Editors , *Handbook of Graphical Models*, Chapman and Hall - CRC Handbooks of Modern Statistical Methods, Chapman and Hall - CRC, 2018, 536 p.

● Powrót do indeksu abstraktów sekcji

Brain Connectivity-Informed Adaptive Regularization for Generalized Outcomes

Jaroslaw Harezlak
harezlak@iu.edu
Indiana University, USA

A challenging problem in the brain imaging research is a principled incorporation of information from different imaging modalities in regression models. Frequently, data from each modality is analyzed separately using, for instance, dimensionality reduction techniques, which result in a loss of information. We propose a novel regularization method, griPEER (generalized ridgified Partially Empirical Eigenvectors for Regression) to estimate the association between the brain structure features and a scalar outcome within the generalized linear regression framework. griPEER provides a principled approach to use external information from the structural brain connectivity to improve the regression coefficient estimation. Our proposal incorporates a penalty term, derived from the structural connectivity Laplacian matrix, in the penalized generalized linear regression. We address both theoretical and computational issues and show that our method is robust to the incomplete structural brain connectivity information. griPEER is evaluated via extensive simulation studies and it is applied in classification of the HIV+ and HIV- individuals.

● Powrót do indeksu abstraktów sekcji

Heavy-Tailed Distributions in Models of Secondary Tumors

Marek Kimmel
kimmel@rice.edu
Rice University, USA

Recent progress in microdissection and in DNA sequencing has facilitated the subsampling of multi-focal cancers in organs such as the liver in several hundred spots, helping to determine the pattern of mutations in each of these spots. These studies have led to diverse conclusions concerning the Darwinian (selective) or neutral evolution in cancer. Mathematical models of the development of multi-focal tumors have been devised to support these claims. We offer a model for the development of a multifocal tumor: it is a mathematically rigorous refinement of a model of Ling et al. (2015). Guided by numerical studies and simulations, we show that the rigorous model, in the form of an infinite-type branching process, displays distributions of tumor size which have heavy tails and moments that become infinite in finite time. To demonstrate these points, we obtain bounds on the tails of the distributions of the process and an infinite series expansion for the first moments. In addition to its inherent mathematical interest, the model is corroborated by recent literature on apparent super-exponential growth in cancer metastases.

Joint work with Philip Ernst, Monika Kurpas and Quan Zhou

References

[1] P. Ernst, M. Kimmel, M. Kurpas and Q. Zhou, *Heavy-tailed distributions in branching process models of secondary cancerous tumors*, Adv. Appl. Prob. **50A** 99 – 114 (2018).

● Powrót do indeksu abstraktów sekcji

Bayesian image analysis in transformed spaces

John Kornak
john.kornak@ucsf.edu
University of California, USA

Bayesian image analysis can improve image quality, by balancing a priori expectations of image characteristics, with a model for the noise process via Bayes Theorem. We will give a reformulation of the conventional Bayesian image analysis paradigm in Fourier and wavelet spaces, e.g. for Fourier space the prior and likelihood are given in terms of spatial frequency signals. By specifying the Bayesian model in transformed spaces, spatially correlated priors, that are relatively difficult to model and compute in conventional image space, can be efficiently modeled as a set of independent processes across; the priors are modeled as independent over the transformed space, but tied together by defining a parameter function over the space for the values of the pdf parameters. The originally inter-correlated and high-dimensional problem in image space is thereby broken down into a series of (trivially parallelizable) independent one-dimensional problems. We will describe the Bayesian image analysis in transformed space modeling approach, illustrate its computational efficiency and speed, and demonstrate useful properties such as isotropy and resolution invariance to model specification which are difficult to obtain in the conventional formulation. We will describe applications in medical imaging, and contrast with results using conventional Bayesian image analysis models. Finally, we will showcase a Python

package that is under development to make the approach widely accessible.

● Powrót do indeksu abstraktów sekcji

Odporna estymacja 'szkieletu' rozkładu wielowymiarowego

Andrzej Kozek
Andrzej.Kozek@mq.edu.au
Macquarie University, Australia

Rozważany jest problem estymacji 'szkieletu' wielowymiarowego rozkładu P w R^k rozumianego jako zbiór S_n złożony z n punktów w R^k dobranych tak, by dyskretny rozkład jednostajny na S_n możliwie jak najlepiej przybliżał oryginalny i nieznany rozkład P. Estymacja jest oparta na N niezależnych obserwacjach $X_1, X_2, \ldots, X_N$ o tym samym rozkładzie P gdzie $n << N$. Zakładamy, że ilość punktów n w S_n jest znana. Proponowana metoda jest na tyle elastyczna, że zwiększenie n wykorzystuje wcześniej dokonane obliczenia. Proponowane jest łączne zastosowanie kilku znanych teorii i metod: metody wielowymiarowych kwantyli (Koltchinskii, Chaudhuri), odpornej metody ich estymacji używając M-estymatorów (Huber), interpretacji M-funkcjonałów jako kwantyli rozkładów zakłóconych (Pawlak, Kozek) oraz teorii równomiernych sekwencyjnych rozkładów punktów (low discrepancy points (ldp), Woźniakowski, Chen i inni). Startując z n punktami ldp w kuli jednostkowej w R^k wybieramy do estymacji odpowiadające tym punktom n k-wymiarowe kwantyle. Kwantyle te estymujemy odpornie (robust) przy pomocy specjalnie dobranych M-estymatorów. Rezultatem jest estymator 'szkieletu' składający się z n punktów oszczędnie aproksymujący wyjściowy k-wymiarowy rozkład. W pracy badamy asymptotyczne własności proponowanej metody.

Powrót do indeksu abstraktów sekcji

Two-step selection method for misspecified binary regression

Mariusz Kubkowski
m.kubkowski@mini.pw.edu.pl
Politechnika Warszawska

Współautor:
Jan Mielniczuk
j.mielniczuk@ipipan.waw.pl
Politechnika Warszawska

Rozważamy dwuetapową procedurę selekcji predyktorów w sytuacji, gdy model regresji binarnej:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q(\mathbf{x}) \tag{1}$$

jest źle wyspecyfikowany, tj. $q(\mathbf{x}) = q(\boldsymbol{\beta}^T\mathbf{x})$ i odpowiadająca funkcja minus log-wiarogodności nie jest równa założonej funkcji straty ρ. Badamy problem znalezienia zgodnego estymatora $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}^*$ minimalizującego funkcję ryzyka:

$$R(\mathbf{b}) = \mathbb{E}\rho(\mathbf{b}^T\mathbf{X}, Y)$$

dla $\mathbf{b} \in \mathbb{R}^p$. Celem selekcji w rozważanym przez nas problemie jest znalezienie takiego wektora $\hat{\beta}$, który ma taki sam nośnik jak β^* z wysokim prawdopodobieństwem. Proponowana procedura składa się z odsiewania i porządkowania predyktorów przy użyciu metody Lasso w pierwszym kroku, a następnie wybrania zbioru minimalizującego Uogólnione Kryterium Informacyjne w powstałej rodzinie modeli. Omawiamy warunki dostateczne zgodności wyboru predyktorów w tej procedurze.

We consider two-stage selection procedure when the underlying binary regression model:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q(\mathbf{x}) \tag{2}$$

is misspecified, ie. $q(\mathbf{x}) = q(\boldsymbol{\beta}^T\mathbf{x})$ and the corresponding minus log-likelihood function is not equal considered loss function ρ. We study a problem of finding consistent estimator $\hat{\boldsymbol{\beta}}$ of the parameter $\boldsymbol{\beta}^*$, which minimizes risk function:

$$R(\mathbf{b}) = \mathbb{E}\rho(\mathbf{b}^T\mathbf{X}, Y)$$

for $\mathbf{b} \in \mathbb{R}^p$. The main aim of selection in the considered problem is to find vector $\hat{\beta}$ which recovers the support of β^* with high probability. Proposed procedure consists of screening and ordering of predictors using Lasso method and then selecting a subset of predictors which minimizes Generalized Information Criterion on the corresponding nested family. We discuss sufficient conditions for consistent selection of predictors in this procedure.

Bibliografia

[1] M. Kubkowski, J. Mielniczuk *Selection consistency of Lasso-based procedures for misspecified high-dimensional binary model and random regressors*, arXiv preprint arXiv:1906.04175, 2019.

[2] M. Kubkowski, *Misspecification of binary regression model: properties and inferential procedures*, PhD thesis (under review), 2019.

• Powrót do indeksu abstraktów sekcji

Limit theorems for empirical cluster functionals with applications to statistical inference

Rafał Kulik
rkulik@uottawa.ca
University of Ottawa, Canada

Limit theorems for empirical cluster functionals are discussed. Conditions for weak convergence are provided in terms of tail and spectral tail processes and can be verified for a large class of multivariate time series, including geometrically ergodic Markov chains. Applications include asymptotic normality of blocks and runs estimators for the extremal index and other cluster indices. Results for multiplier bootstrap processes are also provided.

References

[1] H. Drees, *Limit theorems for empirical cluster functionals*, Annals of Statistics **38**: 2145–2186 (2010).

[2] R. Kulik, P. Soulier, O. Wintenberger, *The tail empirical process of regularly varying functions of geometrically ergodic Markov chains*, Stochastic Processes and their Applications (2019).
DOI: https://doi.org/10.1016/j.spa.2018.11.014.

[3] R. Kulik, P. Soulier, *Heavy Tailed Time Series*, Springer, 2020.

● Powrót do indeksu abstraktów sekcji

Non-asymptotic Analysis of Biased Stochastic Approximation Schemes

Błażej Miasojedow
b.miasojedow@mimuw.edu.pl
Uniwersytet Warszawski

Stochastic approximation (SA) is a key method used in statistical learning. Recently, its non- asymptotic convergence analysis has been a fundamental issue considered in many papers. However, most of these analyses are made under restrictive assumptions such as unbiased gradient estimates and convex objective function, which significantly limit their applications to sophisticated tasks such as online and reinforcement learning. These restrictions are all essentially relaxed in this work. In particular, we consider two general SA schemes to minimize a non-convex objective function. We consider update procedure whose mean field is not necessarily of gradient-type, covering in particular approximate second-order method and allow the one-step update to be a biased estimator of the target mean-field. We illustrate these settings with the online EM algorithm and the policy-gradient method for average reward maximization in reinforcement learning.

References

[1] B. Karimi, B. Miasojedow, E. Moulines, H.-T. Wai, *Non-asymptotic Analysis of Biased Stochastic Approximation Scheme*, Proceedings of the Thirty-Second Conference on Learning Theory, 1944–1974, 2019.

● Powrót do indeksu abstraktów sekcji

Poisson Tree MCMC

Wojciech Niemiro
W.Niemiro@mimuw.edu.pl.com
Uniwersytet im. Mikołaja Kopernika w Toruniu

Co-authors:
Tomasz Cąkała
tc360950@mimuw.edu.pl.com
Uniwersytet Warszawski

Błażej Miasojedow
B.Miasojedow@mimuw.edu.pl.com
Uniwersytet Warszawski

Poisson Tree MCMC algorithms belong to the family of "particle MCMC" (pMCMC) algorithms introduced by Andrieu et al. (JRSS (B) 2010), including particle Metropolis-Hastings and particle Gibbs Sampler. We introduced versions of these algorithms in which the number of "children" of a particle at a given time has a Poisson distribution. In contrast with the classical versions with deterministic number of particles, Poisson Tree MCMC can be directly applied to compute posterior distributions for continuous time semi-Markov processes. We show that Poisson Tree MCMC algorithms preserve the target (posterior) distribution on the space of trajectories of the hidden process. We also prove that for discrete time models, Poisson Tree Gibbs Sampler is uniformly ergodic.

Another advantage of our scheme is that descendants of different particles can evolve independently. This makes it easy to parallelize computations. Simulations show that this

leads to a substantial gain in efficiency.

References

[1] C. Andrieu, A. Doucet, R. Holenstein, *Particle Markov chain Monte Carlo methods*. Journal of the Royal Statistical Society B, 2010

[2] F. Lindsten, M.I. Jordan, T.B. Schön, *Particle Gibbs with Ancestor Sampling*, Journal of Machine Learning Research, 2014

● Powrót do indeksu abstraktów sekcji

Statistical Real-Time Tools for Exploring Dependence in Multivariate Time Series

Hernando Ombao
ombaostat@gmail.com
King Abdullah University of Science and Technology, Saudi Arabia

This work is motivated by the problem of characterizing multi-scale changes in multivariate time series resulting from an external stimulus or shock to the system. One particular goal is to develop a method that can track real-time changes in dependence. This talk will cover a quick overview of the classical measures: coherence, partial coherence and dual-frequency coherence and then introduce some non-stationary generalizations of these (in particular, the evolutionary dual-frequency coherence). We then discuss partial directed coherence which, unlike the previously mentioned measures, can capture directionality between components under the framework of vector autoregressive processes. The latter part of the talk will cover some of the real-time techniques for estimating the different measures of connectivity and for extracting low-dimensional signal summaries. These methods will be critical for understanding biofeedback and adjusting the stimuli adaptively during the experiment. These methods will be applied to various brain signals to track dynamic changes in connectivity in an experiment that seeks to find associations between brain physiological signals and creative thinking. This is joint work with the Biostatistics Research Group at KAUST (Dr. Chee-Ming Ting and Marco Pinto).

Powrót do indeksu abstraktów sekcji

Metody najbliższego sąsiada w modelowaniu predykcyjnym

Mirosław Pawlak
mirek.pawlakk@gmail.com
University of Manitoba i Akademia Górniczo-Hutnicza

Metody najbliższego sąsiada są jednymi z najprostszych oraz intuicyjnie atrakcyjnych procedur stosowanych w nieparametrycznym uczeniu oraz predykcji. Są one szczególnie przydatne w sytuacjach, w których nieznana jest struktura danych, a dane bezpośrednio decydują o jakości predykcji. Metody najbliższego sąsiada wymagają wyboru miary odległości i schematu ważenia. Wiele znanych algorytmów uczenia takich jak losowe lasy, *AdaBoost* czy *gradient boosting* mogą być postrzegane jako ważone metody najbliższego sąsiada z prawidłowo wyuczoną funkcją odległości.
W referacie omówię wykorzystanie metod najbliższego sąsiada w kontekście nieparametrycznej analizy regresyjnej i szeregów czasowych. W szczególności, przedstawię wybrane klasy modeli nieparametrycznych i semi-parametrycznych dla szeregów czasowych. Ponadto omawiany będzie problem nieparametrycznego testowania modelu i przekleństwo wymiaru w przypadku danych wielowymiarowych.

Bibliografia

[1] G. Biau and L. Devroye, *Lectures on the Nearest Neighbour Method*, Springer-Verlag, Berlin, 2015.
[2] W. Greblicki and M. Pawlak, *Nonparametric System Identification*, Cambridge University Press, Cambridge, 2008.

Powrót do indeksu abstraktów sekcji

A novel weighted likelihood estimation with empirical Bayes flavor

Krzysztof Podgórski
Krzysztof.Podgorski@stat.lu.se
Lund University, Szwecja

We propose a novel approach to estimation, where a set of estimators of a parameter is combined into a weighted average to produce the final estimator. The weights are chosen to be proportional to the likelihood evaluated at the estimators. We investigate the method for a set of estimators obtained by using the maximum likelihood principle applied to each individual observation. The method can be viewed as a Bayesian approach with a data driven prior distribution. We provide several examples illustrating the new method, and argue for its consistency, asymptotic normality, and efficiency. We also conduct simulation studies to assess the performance of the estimators. This straightforward methodology produces consistent estimators comparable with those obtained by the maximum likelihood method. The method also approximates the distribution of the estimator through the 'posterior' distribution. Many straightforward generalizations are suggested and can be subject future studies. The talk is based on (Hossain et al. 2018) and provides an alternate estimation in the singular cases of likelihood discussed in (Podgórski and Wallin 2015).

References

[1] Podgórski, K., Wallin, J. *Maximizing leave-one-out likelihood for the location parameter of unbounded den-*

sities, Annals of the Institute of Statistical Mathematics **67**(1):19–38 (2015).

[2] Hossain, M., Kozubowski, T.J., Podgórski, *K. A novel weighted likelihood estimation with empirical Bayes flavor*, Communications in Statistics – Simulation and Computation **47**(2): 392–412 (2018).

● Powrót do indeksu abstraktów sekcji

Geometria rozbicia MAP w bayesowskich modelach mieszanek

Łukasz Rajkowski
l.rajkowski@mimuw.edu.pl
Uniwersytet Warszawski

Współautor: John Noble

Bayesowskie modele mieszanek stanowią jedno z narzędzi do przeprowadzania analizy skupień. Ich istotą jest nałożenie pewnego rozkładu prawdopodobieństwa (*prawdopodobieństwo a priori*) na możliwe rozbicia danych na grupy i oparciu wnioskowania o rozkład warunkowy (*a posteriori*), wykorzystując rozkład prawdopodobieństwa danych pod warunkiem rozbicia (rozkład próbkowy) oraz twierdzenie Bayesa. W moich badaniach analizuję własności rozbicia maksymalizującego prawdopodobieństwo a posteriori (tzw. rozbicia MAP). W sytuacji, gdy rozkład próbkowy jest skonstruowany przy użyciu sprzężonych rodzin wykładniczych, dowolne dwie grupy w estymatorze MAP są rozdzielone przez poziomnice liniowego funkcjonału statystyki dostatecznej. W szczególności, gdy w ramach danej grupy dane losowane są z rozkładu gaussowskiego o nieznanej średniej i macierzy kowariancji (losowanymi odpowiednio z rozkładów gaussowskiego i Wisharta), oznacza to możliwość rozdzielenia dowolnych dwóch grup przez kwadrykę (powierzchnię określoną równaniem drugiego stopnia). W przypadku, gdy wewnątrzgrupowa macierz kowariancji jest ustalona, grupy są rozdzielone przez hiperpłaszczyzny, co stanowi elegancki odpowiednik własności regionów decyzyjnych w analizie dyskryminacyjnej Fi-

shera.

Bibliografia

[1] Ł. Rajkowski and J. Noble, *A note on the geometry of the MAP partition in Conjugate Exponential Bayesian Mixture Models*,
arXiv preprint, arXiv:1902.01141v2 (2019).

● Powrót do indeksu abstraktów sekcji

A regression framework for multi-view analysis of high-dimensional structured data

Timothy Randolph
trandolp@fredhutch.org
Fred Hutchinson Cancer Center, Seattle, USA

The analysis of genomic and other 'omics data is often carried out from multiple perspectives. For example, one sample from a microbiome study involves many microbial measurements, but these measures may take multiple forms (species abundance, species presence, microbe gene counts, etc) and have auxiliary information (phylogenetic relatedness, metabolic potential, etc). These multiple "viewsóf each sample don't fit neatly into standard statistical analyses or dimension-reduction methods. Here we present a framework of incorporating more than one view of such data into both dimension-reduction methods and penalized regression models.

• Powrót do indeksu abstraktów sekcji

Szybka i odporna selekcja cech w modelach regresyjnych

Wojciech Rejchel
wrejchel@gmail.com
Uniwersytet im. Mikołaja Kopernika w Toruniu

Współautorka: Małgorzatą Bogdan (Uniwersytet Wrocławski)

Selekcja cech jest zagadnieniem ważnym, zwłaszcza gdy badamy wysokowymiarowe zbiory danych, w których liczba cech znacząco przekracza liczbę obserwacji. W wielu przypadkach znalezienie małego zbioru złożonego z cech istotnych jest równie ważne, bądź ważniejsze, jak poprawna estymacja czy predykcja.

Rozważamy problem selekcji cech w modelu $Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n$, gdzie $Y_i \in \mathbb{R}$ jest zmienną zależną, $X_i \in \mathbb{R}^p$ wektorem cech, β prawdziwym parametrem oraz ε_i błędem losowym. Zakładamy, że nieznana funkcja g jest rosnąca względem pierwszej zmiennej. Rozkład błędów ε_i jest dowolny, w szczególności nie wymagamy istnienia jego momentów.

Proponujemy prostą i obliczeniowo szybką procedurę selekcji cech, która oparta jest na standardowym algorytmie Lasso ze zmiennymi Y_i zastąpionymi przez ich rangi R_i. Przedstawimy teoretyczne i numeryczne wyniki dotyczące zgodności w wyborze modelu naszych metod. Zaproponowane rozwiązania porównamy z procedurą LADLasso [1], która jest często używanym narzędziem w odpornej selekcji cech.

Bibliografia

[1] J. Fan, Y. Fan and E. Barut, *Adaptive robust variable selection*, Ann. Statist., **42**: 324–351 (2014).

● Powrót do indeksu abstraktów sekcji

Własności estymatorów w modelowaniu przyczynowości

Krzysztof Rudaś
k.rudas@mini.pw.edu.pl
Politechnika Warszawska

Od kilku lat modelowanie przyczynowości staje się coraz istotniejszą gałęzią statystki, wykorzystywaną między innymi w medycynie czy kampaniach marketingowych. Celem modelowania przyczynowości jest wskazanie elementów populacji (np. pacjentów) dla których nasze działanie (np. nowa terapia) daje pozytywne rezultaty. Aby uzyskać tę informację, porównuje się dwie sytuacje. Pierwsza, kiedy dany element populacji został poddany działaniu i druga, kiedy nie został. Niestety nie posiadamy tych dwóch informacji jednocześnie. Problem ten rozwiązuje się, stosując podział całej populacji na zbiór eksperymentalny i kontrolny, a następnie estymując różnicę zysku w tych dwóch sytuacjach.

W swojej prezentacji skupię się na założeniu liniowości odpowiedzi w obydwu grupach. Wówczas najpopularniejszą metodą estymacji jest model podwójny. Metoda ta polega na konstrukcji dwóch estymatorów liniowych, odpowiednio na grupie eksperymentalnej i kontrolnej i policzenia różnicy między nimi. W swoim wystąpieniu zaprezentuję metody alternatywne, a także porównam ich własności z modelem podwójnym.

Bibliografia

[1] K. Rudaś and S. Jaroszewicz, *Linear regression for uplift modeling*, Data Mining and Knowledge Disco-

very **vol.32, num.5**: 1275–1305 (2018).

● Powrót do indeksu abstraktów sekcji

Zmienność średnich i kwantyli mieszanek uporządkowanych rozkładów przy niedokładnym wyborze rozkładu a priori

Tomasz Rychlik
trychlik@impan.pl
Polska akademia Nauk

Rozważamy rodziny rozkładów prawdopodobieństwa indeksowane parametrami z przedziału osi rzeczywistej. Zakładamy jedynie, że wzrost wartości parametru powoduje wzrost rozkładu względem standardowego porządku stochastycznego. Badamy konsekwencje nieprecyzyjnego wyboru rozkładu a priori na zbiorze parametrów modelu. Dokładniej, wybieramy ustalony rozkład a priori, podczas gdy rzeczywisty rozkład a priori należy do jego nieparametrycznego otoczenia. Wówczas wyznaczamy optymalne dolne i górne oszacowania różnic między wartościami oczekiwanymi i kwantylami mieszanek względem prawdziwego i założonego rozkładu mieszającego w różnych jednostkach skali.

● Powrót do indeksu abstraktów sekcji

Zasada Morozowa dla poissonowskich problemów odwrotnych

Zbigniew Szkutnik
szkutnik@agh.edu.pl
Akademia Górniczo-Hutnicza

Rozważymy problem odwrotnej estymacji funkcji intensywności procesu Poissona. Dla zaobserwowanego procesu Poissona (tzn. losowej miary punktowej) o funkcji intensywności $g = \mathcal{K}f$, gdzie $\mathcal{K}$ jest pewnym zwartym operatorem działającym między pewnymi ośrodkowymi przestrzeniami Hilberta, chcemy wyestymować funkcję f. Jest to ogólna postać tzw. poissonowskiego problemu odwrotnego, zwykle źle postawionego w sensie Hadamarda. Jego rozwiązanie wymaga zastosowania jakiejś formy regularyzacji i związanej z nią, opartej na danych metody wyboru parametru wygładzającego. Przedstawimy pewną ogólną konstrukcję metody wyboru parametru regularyzacji odpowiednią dla tego typu problemów, a opartą na tzw. zasadzie rozbieżności Morozowa. W odróżnieniu od podobnych metod opisanych w literaturze, nie będzie potrzebna dyskretyzacja rozważanego problemu. Omówimy pewne własności asymptotyczne otrzymanych estymatorów i przykładowe zastosowania do problemów stereologicznych sformułowanych jako poissonowskie problemy odwrotne.

• Powrót do indeksu abstraktów sekcji

Classifier chains for multi-label classification

Paweł Teisseyre
Pawel.Teisseyre@ipipan.waw.pl
Polska Akademia Nauk

Multi-label classification (MLC) has received increasing attention in recent years, motivated by a large number of new applications. In MLC each object of our interest is described by a vector of features and a vector of binary labels. The main objective is to build a model which predicts labels using features. For example in medical diagnosis the goal is to predict occurrences of diseases using some information about a patient. Classifier chains (CC) are among the most popular and successful methods used in MLC. The goal of the talk is to present various modifications of standard CC method, e.g. adaptive classifier chains (ACC) and parsimonious classifier chains (parCC). The modifications aim to build powerful models under a constraint on the total number of features. Reducing the number of features is crucial in the domains where the acquisition of the feature values is costly, e.g. in medical diagnosis where each diagnostic test is associated with its cost.

References

[1] P. Teisseyre, *CCnet: Joint multi-label classification and feature selection using classifier chains and elastic net regularization*, *Neurocomputing* **235**: 98–111 (2017).

[2] P. Teisseyre and D. Zufferey and M. Slomka, *Cost-sensitive classifier chains: selecting low-cost features in multi-label classification*, *Pattern Recognition* **86**:

290–319 (2019).

● Powrót do indeksu abstraktów sekcji

Asymptotics of the overflow in urn models

Jacek Wesołowski
wesolo@mini.pw.edu.pl
Politechnika Warszawska

The number of occupied urns after n balls have been thrown in is often interpreted as a measure of richness. In diversity analysis, the number M_k of urns with exactly k balls, is called abundance count of order k and a popular estimator of species richness, called Chao estimator, is based on M_1 and M_2, Chao and Chiu (2016).

Balls (n of them) are to be placed into urns, each urn of capacity r. If the urn selected for the given ball is already full, the ball lands in the overflow. We study the size of the overflow, $V_{n,r}$, when $n \to \infty$. Hwang and Janson (2008) covered the Poissonian asymptotic of $V_{n,1}$. We establish Poissonian and Gaussian asymptotics of $V_{n,r}$ for any $r \geqslant 1$. Note that the abundance count, $M_{n,r}$, is the *second discrete derivative* of $V_{n,r}$, i.e. its asymptotics can be read out from that of $V_{n,r}$.

The talk is based on Gouet, Hitczenko and Wesołowski (2019).

References

[1] A. Chao , C.-H. Chiu, *Species richness: estimation and comparison*, WileyStatsRef: Statistics Reference Online: 126 (2016).

[2] R. Gouet, P. Hitczenko, J. Wesołowski, *Asymptotics of the overflow in urn models*, arXiv **1905.06663**: 1–23 (2019).

[3] H.K. Hwang, S. Janson, *Local limit theorems for finite*

and infinite urn models, Ann. Probab. **36(3)**: 992–1022 (2008).

● Powrót do indeksu abstraktów sekcji

Adaptacyjny jednostronny dwupróbkowy test Kaplana-Meiera

Grzegorz Wyłupek
wylupek@math.uni.wroc.pl
Uniwersytet Wrocławski

W referacie przedyskutujemy istniejące podejścia, znane z literatury, do detekcji stochastycznego uporządkowania dwóch funkcji przeżycia jak również postawimy i rozwiążemy nowatorski problem testowania dotyczący jego istnienia. Dokładniej, hipoteza zerowa orzeka brak uporządkowania, natomiast alternatywa wyraża jego istnienie. Wprowadzona statystyka testowa jest pewnym funkcjonałem standaryzowanego dwupróbkowego procesu Kaplana-Meiera próbkowanego w losowej liczbie losowych punktów będących obserwowanymi czasami przeżycia w połączonych próbach oraz wykorzystuje informację zawartą w specjalnie do tego celu zdefiniowanej jednostronnej ważonej statystyce log-rank. Statystyka testowa automatycznie waży wielkość i znak bloków ją budujących stając się przez to czułą procedurą w rozważanym problemie testowania. Adaptacyjna konstrukcja sprawia, że odpowiadający jej test kontroluje asymptotyczne błędy obu rodzajów na ustalonym poziomie istotności α. Przeprowadzone badania symulacyjne pokazują, że błędy te są kontrolowane w satysfakcjonującym stopniu również gdy liczba obserwacji jest skończona. Na tle najlepszych i najpopularniejszych testów nowe rozwiązanie wypada bardzo dobrze. Analiza zbioru danych rzeczywistych potwierdza tę konkluzję.

Bibliografia

[1] G. Wyłupek, Data-driven Kaplan-Meier one-sided two-sample tests, *Under review*, (2019)

● Powrót do indeksu abstraktów sekcji