Berlin Symposium December 2017 Abstracts

Malgorzata Bogdan
University of Wroclaw

Title: Sorted L-One Penalized Estimation.

Abstract: Sorted L-One Penalized Estimation is a relatively new convex optimization method for identifying important predictors in large data bases. In a variety of settings it allows to control False Discovery Rate and enhances estimation properties of constructed models. In this talk we will present current theoretical results on SLOPE's properties and show a variety of applications, including Genome Wide Association Studies and construction of sparse portfolios.

Przemysław Biecek
Warsaw University of Technology, Poland

Title: Integrated Machine Learning Genetic Signatures (MLGenSig) based on gene expression and DNA methylation with applications to The Cancer Genome Atlas project

Abstract: Genetic signatures are frequently used in the cancer research. Predictive signatures, screening signatures, targeted signatures and many others are being developed to model patient survival, drug-response or other clinically important factors. This trend follows the rapid development of *-seq platforms for measuring gene expression, DNA methylation, CNV alterations etc. Yet, despite the large variety of different platforms, most genetic signatures are based on a single type of data. In this talk I will present MLGenSig: methodology and tools (R packages) that facilitate the development, maintenance and validation of integrated genetic signatures. Signatures that utilize clinical features, genes expression and methylation based biomarkers in an integrated signature. Among all, following tools will be presented: archivist – for maintenance, randomForestExplainer for validation and MLExpResso for identification of important biomarkers. All presented examples are based on The Cancer Genome Atlas data.

References:
MLExpResso: An R package for joint modelling of genes expression and CpG probes methylation. Aleksandra Dąbrowska, Alicja Gosiewska, Przemysław Biecek (2017), https://github.com/geneticsMiNIng/MLGenSig
randomForestExplainer: A set of tools to understand what is happening inside a Random Forest Aleksandra Paluszyńska, Przemysław Biecek (2017) https://github.com/MI2DataLab/randomForestExplainer
An R package for accessing data from The Cancer Genome Atlas Data Marcin Kosinski, Przemyslaw Biecek, Witold Chodor (2016) https://github.com/RTCGA/RTCGA
archivist: An R Package for Managing, Recording and Restoring Data Analysis Results Przemyslaw Biecek, Marcin Kosinski (2017) URL: https://github.com/pbiecek/archivist

MLExpResso: a tool for integrative analyses and visualization of gene expression and DNA methylation data

Aleksandra Dąbrowska[1], Alicja Gosiewska[2], Przemysław Biecek[3]

[1]University of Warsaw, Warsaw, Poland
[2]Warsaw University of Technology, Warsaw, Poland
[3]Warsaw University of Technology, Warsaw, Poland

Abstract: This poster presents MLExpResso - an R package for integrative analyses and visualization of gene expression and DNA methylation data. Key functions of this package:

1. identification of differentially methylated regions based on RRBS data,
2. identification of differentially expressed genes based on RNAseq data,
3. identification of regions with changes in expression and methylation between two conditions,
4. visualization of identified regions.

The joint modeling and visualization of genes expression and methylation improves interpretability of identified signals.

MLExpResso uses methods implemented in various R packages available in Bioconductor for analysis of expression (i.e. DESeq2 [2], edgeR [3]) and methylation (i.e. methyAnalysis [4]). In addition, it supports visualization of identified regions. The developed solution can be used to better understand the interdependence of expression and methylation and their joint effect on the selected feature.

The methodology is supplemented with the example applications to Breast Cancer data from The Cancer Genome Atlas project [5].

This work was supported by the National Science Centre (Opus grant 2016/21/B/ST6/02176).

References

[1] MLExpResso, „Machine Learning for Genetic Signatures" R package [https://github.com/geneticsMiNIng/MLGenSig].

[2] Love M, Huber W, Anders S. *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology 2014 15:550.

[3] Robinson M, McCarthy D, Smyth G. *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics 2010 Jan 1;26(1):139–140.

[4] Du P, Bourgon R. *methyAnalysis: an R package for DNA methylation data analysis and visualization.*

[5] Tomczak K, Czerwińska P, Wiznerowicz M. *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.* Contemporary Oncology 2015;19(1A):A68-A77.

Large-scale predictive modeling for lung cancer

Aleksandra Dąbrowska, University of Warsaw
Mariusz Adamek, Medical University of Silesia
Tadeusz Orłowski, Institute of Tuberculosis and Lung Diseases
Przemysław Biecek, Warsaw University of Technology

The Goal

The main goal of this study is to identify key factors that could be used to predict future patient outcome after surgical treatment of lung cancer. In this study, we use 5-years survival as a primary outcome.

The Data

All analyses were conducted on data from polish monitoring study. Clinical, genetic and demographical characteristics are gathered for over 35,000 patients after surgical treatment for lung cancer in years 2005-2016. Data is available due to the collaboration with the Institute of Tuberculosis and Lung Diseases, the Maria Sklodowska-Curie Oncology Center - Institute. It is a unique dataset both in terms of the number of patients and in terms of the number of features gathered for each patient.

Tools

Our analyses were carried out using R [1]. Results are embedded in a shiny application [2]. The application is available here: http://52.31.27.158/pluco/. Statistical modeling was performed with survival package [3] while visualizations were created with survminer package [4]. For reporting, we have used the tableone package [5].

Results

Based on the available data we have created a list of factors that are linked with patient survival. Due to significant differences, a separate model is built for each gender. Among significant factors, we have identified: cancer center, uicc score, smoking, family history, and age.

Acknowledgments

This work was financially supported by *NCN Opus grant 2016/21/B/ST6/02176*.

Literature

[1] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[2] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.3. https://CRAN.R-project.org/package=shiny

[3] Therneau T (2015). _A Package for Survival Analysis in S_. version 2.38. https://CRAN.R-project.org/package=survival
[4] Alboukadel Kassambara and Marcin Kosinski (2017). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.0. https://CRAN.R-project.org/package=survminer
[5] Kazuki Yoshida and Justin Bohn. (2017). tableone: Create 'Table 1' to Describe Baseline Characteristics. R package version 0.8.1. https://CRAN.R-project.org/package=tableone

A time-varying AR coefficient model of functional near-infrared spectroscopy data

Timothy D. Johnson
Department of Biostatistics
University of Michigan

Abstract

Functional near-infrared spectroscopy (fNIRS) is a relatively new neuroimaging technique. It is a low cost, portable, and non-invasive method to monitor brain activity. Similar to fMRI, it measures changes in the level of blood oxygen in the brain. Its time resolution is much finer than fMRI, however its spatial resolution is much courser—similar to EEG or MEG. fNIRS is finding widespread use on young children whom cannot remain still in the MRI magnet and it can be used in situations where fMRI is contraindicated—such as with patients whom have cochlear implants. In this talk, I propose a fully Bayesian time-varying autoregressive model to analyze fNIRS data. The hemodynamic response function is modeled with the canonical HRF and the low frequency drift with a variable B-spline model (both locations and number of knots are allowed to vary). Both the model error and the auto-regressive process vary with time. Via a simulation studies, I show that this model naturally handles motion artifacts and gives good statistical properties. The model is then apply to a fNIRS study.

Joerg Polzehl

Title: Towards in-vivo histology of the brain

Abstract: Recent advances in neuro-imaging attempt to enable in-vivo histology of the brain. Doing so requires increased spatial resolution up to a situation where the signal meets the noise floor. The talk will cover research conducted at WIAS, in collaboration with MR physicists, on statistical issues in modeling imaging data characterized by low signal-to-noise ratio (SNR). I'll cover several specific, but interrelated problems:
- characterization of the signal distribution in MR experiments,
- effects of preprocessing on the signal distribution,
- estimation of the noise profile in MR images,
- use of spatial information for variance reduction in (collections of) MR images,
- bias due to incorrect modeling in MR experiments.
I'll consider two specific imaging experiments to illustrate problems, effects and solutions:
- diffusion weighted MR, with an analysis based on data of the Human Connectome Project,
- multi-parameter mapping, using data measured at the Wellcome Trust Center for Neuroimaging, London.

Literature:
- S. Becker, K. Tabelow, S. Mohammadi, N. Weiskopf and J. Polzehl, Adaptive smoothing of multi-shell diffusion-weighted MR data by msPOAS, NeuroImage, 95 (2014) pp. 90--105.
- K. Tabelow, H.U. Voss and J. Polzehl, Local estimation of the noise level in MRI using structural adaptation, Medical Image Analysis, 20 (2015) pp. 76--86.
- J. Polzehl and K. Tabelow, Low SNR in dMRI models, JASA, 11 (2016) pp. 1480--1490.
- K. Tabelow, Ch. D'Alonzo, L. Ruthotto, M. F. Callaghan, N. Weiskopf, J. Polzehl and S. Mohammadi, Removing the estimation bias due to the noise floor in multi-parameter maps, accepted for ISMRM annual meeting 2017.

Damian Brzyski
Title: "Finding graphs in the brain"

Abstract: 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 images 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 -- the tensor regression -- which we use to investigate associations between brain's structural connections and HIV disease-related outcomes. Specifically, we utilize a collection of DTI-derived subject-specific structural and functional connectivity matrices to represent a graph of connections between brain regions. Each matrix is treated as a subject's covariate in a tensor regression modeling approach. The corresponding regression coefficients are of the form of a matrix as well. The task is to estimate this matrix, taking under account not only the considered model but also the complicated brain structure revealed by subjects' connectivity matrices.

Karsten Tabelow

Title: Adaptive smoothing of Multi Parameter Maps

Abstract: Low signal-to-noise ratio (SNR) renders the interpretation and analysis of images from quantitative magnetic resonance imaging (qMRI) of the human brain difficult if not impossible. The drop in SNR inherently comes with the increase of spatial resolution, which is required to image fine details of the tissue. Here, we present a novel adaptive smoothing method for qMRI in the framework of multi-parameter mapping that is able to reduce variability of the data and hence increase the SNR without introducing a bias especially at tissue borders which is often observed for non-adaptive smoothing methods.

Mark Fiecas

Title: A longitudinal Bayesian model for spectral analysis of neuroimaging time series data

Abstract: In this talk, we will give an overview of statistical methodologies for spectral analysis of time series data. We will briefly discuss the common approaches for spectral analysis, and discuss their limitations for analyzing data whenever the study has a longitudinal experimental design. To address the limitations, we propose a Bayesian model for spectral analysis that accounts for the covariation within a subject. Our proposed model using smoothing splines to estimate the spectra of the time series, and we induce correlation across visits through the prior distributions of the model parameters. We discuss the merits of our proposed model in the context of a longitudinal fMRI experiment, and we illustrate its utility using simulated and empirical data.

M. Staniak

Title: Local interpretability of machine learning models

Abstract: In [1] authors proposed a method (LIME) of explaining predictions generated by black boxes (complex machine learning models) byfi tting a simple, interpretable model locally around an observation of interest. This is done by simulating a set of observations that are similar (close) to the given observation for so called local exploration, and then fitting a chosen interpretable model to this new dataset. To ensure the simplicity and interpretability of the model, a penalty term can be added to the fitting procedure. In this talk I will show how we adapted this idea to numerical data and regression problems, as opposed to classification problems from image and text data analysis which were the focus of the original article. To aid understanding of a model, visualization techniques can be applied. We implemented some of them in an R package live, which I will use to present examples. Model visualization is a growing field in statistics as exemplified by [2]. Use of its ideas boosts the advantages of LIME methodology.

References

[1] Tulio Ribeiro, Singh, Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[2] Wickham, Cook, Hofmann, Visualizing Statistical Models: Removing the Blindfold. Statistical Analysis; Data Mining 8(4), 2015.
[3] Biecek, Staniak, live: Locally Interpretable Visual Explanations, in preparation, 2017.

Pawel Drozd
Title: Agent-based modeling of spread of public opinion".

Abstract: Subject of my thesis was agent base modeling of spread of public opinion. Using modified q-voter model I carried out research on impact of three factors on results of simulation. Those factors were: network density, type of random sampling used for picking neighbours of selected agent and number q of said neighbours. Results show that modifying the type of random sampling and value of coefficient q have impact on model.
Keywords: stochastic simulation, Monte Carlo methods, complex networks, opinion dynamics

Marcin Halupka
Title: How much do we know about our clients? A short story on recommendation systems.

Abstract: Every day we gather a lot of data about sets of products clients buy from shop shelves. Can we create business value for shops by extracting knowledge from that data and revealing, maybe hidden so far, clients preferences and as a result helping them with crafting better product offerings for their clients? The answer obviously is yes, that is the well-known purpose behind building recommendation systems. One of the common techniques is called collaborative filtering – method of making predictions about the interests of a given client by collecting preferences from many other clients. In presentation, I will quickly present a variation of technique mentioned above using item-item matrix and explain it in context of project I was working on - creating a simple recommendation engine for e-commerce platform.

Ali Shojaie
Title: Analyzing Non-Stationary High-Dimensional Time Series: Structural Break Detection and Parameter Estimation

Abstract: Assuming stationarity is unrealistic in many time series applications, including neuroscience. A more realistic alternative is to assume piecewise stationarity, where the model is allowed to change at potentially many time points. We propose a three-stage procedure for consistent estimation of both structural change points and parameters of high-dimensional piecewise vector autoregressive (VAR) models. In the first step, we reformulate the change point detection problem as a high dimensional variable selection one, and solve it using a penalized least square estimator with a total variation penalty. We show that the proposed penalized estimation method over-estimates the number of change points. We then propose a backward selection criterion to identify the change points. In the last step of our procedure, we estimate the VAR parameters in each of the segments. We show that the proposed procedure consistently detects the number of change points and their locations. We also show that the procedure consistently estimates the VAR parameters. The performance of the method is illustrated through several simulation studies, as well as an analysis of EEG data.

Jaroslaw Harezlak
Title: What can raw accelerometry data tell us about human activity? Walking vs. stair climbing

Abstract: Wearable accelerometers offer a noninvasive measure of physical activity (PA). They record high frequency three-dimensional time series data. Among many human activities, walking is the most common moderate level PA. Our work addresses the classification of walking into level walking, descending stairs and ascending stairs. We apply our method, based on the extracted short-time interpretable features arising from the Fourier and wavelet transforms, to data collected on 32 middle-age participants. We build subject-specific and group-level classification models utilizing a tree-based classifier. We evaluate the effects of sensor location and tuning parameters on the classification accuracy of these models. In the group-level classification setting, we propose a robust feature normalization approach and evaluate its performance. In summary, our work provides a framework for better extraction and use of the raw accelerometry data to differentiate among different walking modalities. We show that both at a subject-specific level and at a group level, overall classification accuracy is above 80% indicating excellent performance of our method.

Marcin Straczkiewicz
Title: The challenges in walking recognition using accelerometers at various body locations

Abstract: Accelerometers are frequently used to measure physical activity in large observational studies (e.g. National Health and Nutritional Examination Survey (NHANES), UK Biobank, Women Health Initiative (WHI), etc.) because they are convenient to wear, cheap and provide objective and reproducible proxy measurements of physical activity. The outcomes of activity measurements are often provided as 24-hour activity cycle summaries, such as activity counts, vector magnitude or number of steps. However, the collected data contain a lot more information that is not presently used. This is due to the size and complexity of the data as well as to the methodological gap between scientific questions and what can reliably be estimated from the data. One of such divergences is the accuracy of walking recognition algorithms which is heavily dependent on the selection of sensor placement on a human body. Typical body places used in research studies, such as thigh or hip, show the highest accuracy of walking detection. However as noticed they provide a significantly lower level of compliance compared to e.g. a wrist. In fact a level of wear time is a vital factor for a body placement selection in a large scale free-living studies. Unfortunately, wrists are more prone to distortions that hinder the process of detection of salient walking features, such as periodicity, duration, speed and intensity. The author describes challenges related to the walking recognition using the algorithm based on Continuous Wavelet Transform. The discussion is conducted based on the results of free-living experiment on a cohort of older Americans (N=51) equipped with wearable accelerometers on hip, waist and both wrists.