Summer School on Scientific Visualization and Presentation

Falun
June 16-18, 2014

Welcome to The Cramér Society Summer School 2014

It gives us great pleasure to welcome you to the The Cramér Society Summer School on Scientific Visualization and Presentation, which is being held on the Campus Lugnet of Dalarna University in Falun, 16-18 June 2014.

On Tuesday you are all welcome to listen to the invited speakers, Jo Røislien and Hadley Wickham; both experts at presentation and visualization. An important, and increasingly so, field to master if you are a statistician.

Moreover, one of the major objectives of the Summer School is for doctoral students in statistics/mathematic statistics and related subjects to get to know each other and become acquainted with ongoing doctoral projects. Therefore, the program on Monday and Wednesday and parts of the program on Tuesday, are for doctoral students only and several social activities are included in the program.

We hope that you will enjoy and actively take part in the scientific and social program of the Summer School. Thank you for attending and for bringing your expertise to our Summer School.

The Organizing Committee

Cramérsällskapet

Contents

Program

Monday June 16

10:45-11:30	**Registration** (Scandic Hotel Lugnet)	
11:30-12:15	**— Lunch —**	
12:15-12:20	**Welcome to the Summer School**	**Lars Rönnegård**
12:20-14:00	**Writing and presenting mathematics and statistics**	**Tom Britton**
14:00-14:30	**— Coffee break —**	
14:30-16:30	**Presentations by PhD students** Chair: Xiangli Meng	
14:30-14:50	Estimated Lifetimes of Road Pavements in Sweden Using Time-To-Event Analysis	Kristin Svenson
14:50-15:10	Using Paradata to Monitor Quality and Costs in Survey Production	Anton Johansson
15:10-15:30	PANIC residuals based pooled tests with contemporaneous correlated errors	Xijia Liu
15:30-15:50	Triggering Solar-Powered Vehicle Activated Signs using Self Organizing Maps with K-Means	Diala Jomaa
15:50-16:10	Patterns of changing cancer risks with time since diagnosis of a sibling	Myeongjee Lee
16:10-16:30	Coarsening to improve balance on key covariates in matching	Philip Fowler
16:30	**— Dinner —**	
18:30	**Visit to Falun Copper Mine**	

Tuesday June 17

09:00-10:30	**Presentations by PhD students** Chair: Yujiao Li	
9:00-9:20	Doubly robust estimation generalized estimating equations - the R package drgee	Johan Zetterqvist
9:20-9:40	Process GPS data on car-movements in Dalarna, Sweden	Xiaoyun Zhao
9:40-10:00	Stationary generalized asymmetric Laplace models	Nima Shariati
10:00-10:20	Firm relocation and firm performance - evidence from the Swedish wholesale and retail sector	Zuzana Macuchova
10:00-11:00	**Registration** & **Coffee** (Dalarna University)	
11:00-12:00	**The art of complex communcation**	**Jo Røislien**
12:00-13:00	**— Lunch —**	Sponsored by: SAS THE POWER TO KNOW.
13:00-14:30	**Data manipulation**	**Hadley Wickham**
	Creating layered graphics with ggplot2	**Hadley Wickham**
14:30-15:00	**— Coffee break —**	
15:00-16:30	**Cognitive psychology for visualisation**	**Hadley Wickham**
	ggvis sneak peek	**Hadley Wickham**
17:00	**— Dinner —**	
18:30	**Bowling**	

Note: The events between 11.00 and 16.30 on Tuesday June 17 are open to all participants; the other events are only for PhD students. You will find the registration in the library at the main entrance of Dalarna University.

Wednesday June 18

08:00-09:50	**Presentations by PhD students** Chair: Xiaoyun Zhao	
8:30-8:50	An Inferential Framework for Domain Selection in Functional ANOVA	Johan Strandberg
8:50-9:10	Confidence of the heuristic solutions	Xiangli Meng
9:10-9:30	Comparing idiosyncratic unit root tests based on the residuals estimated from ML and PC in a dynamic factor model	Xingwu Zhou
9:30-9:50	Spatial analysis of Swedish business	Yujiao Li
09:50-10:20	**— Coffee break —**	
10:20-11:20	**Presentations by PhD students** Chair: Kristin Svenson	
10:20-10:40	Statistical methods for the detection, analyses and integration of biomarkers in the human genome and transcriptome	Chen Suo
10:40-11:00	Tackle 'em bugs! Managing the issue overflow in software engineering	Markus Borg
11:00-11:20	Less is more − how to communicate simple but distinct	Magnus Fahlström
11:30-12:30	**What do all the numbers really mean?**	**Allan Gut**
12:30-13:30	**— Lunch —**	

Cramérsällskapet

Organizing Committee

The Organizing Committee for The Cramér Society Summer School 2014 has consisted of

Lars Rönnegård, Dalarna University (chair)

Maria Karlsson, Umeå University

Leif Ruckman, Karlstad University

The Organizing Committee had great assistance from members of the Cramér Society boards for years 2013 and 2014 and from PhD students at Dalarna University.

Sponsors

The following companies and organizations are sponsoring The Cramér Society Summer School 2014

Swedish Statistical Society , *http* : *//www.statistikframjandet.se*

GRAPES Swedish network for Graduate and Postgraduate Educations in Statistics, *http* : *//www.grapestat.se*

SAS Institute, *http* : *//www.sas.com/en_us/insights/big−data/data−visualization.html*

Statisticon, *http* : *//www.statisticon.se/*

Dalarna University, *http* : *//www.du.se/en*

Participants

Invited Speakers	
Allan Gut	Uppsala University
Hadley Wickham	Rice University, Texas, USA
Jo Røislien	University of Stavanger, Norway
Tom Britton	Stockholm University
Contributed Speakers (PhD students)	
Anton Johansson	Stockholm University
Chen Suo	Karolinska Institutet
Diala Jomaa	Dalarna University
Johan Strandberg	Umeå University
Johan Zetterqvist	Karolinska Institutet
Kristin Svenson	Dalarna University
Magnus Fahlström	Dalarna University
Markus Borg	Lund University
Myeongjee Lee	Karolinska Institutet
Nima Shariati	Lund University
Philip Fowler	Umeå University
Xiangli Meng	Dalarna University
Xiaoyun Zhao	Dalarna University
Xijia Liu	Uppsala University
Xingwu Zhou	Uppsala University
Yujiao Li	Dalarna University
Zuzana Macuchova	Dalarna University
Organizers	
Lars Rönnegård	Dalarna University
Leif Ruckman	Karlstad University
Maria Karlsson	Umeå University
Other Participants	
Angélica Naesman	Swedish Pensions Agency
Arianna Comin	The National Veterinary Institute
Bengt Norrby	Swedish Pensions Agency
Céderic Perriard	Swedish Pensions Agency
Danne Mikula	Swedish Pensions Agency
Estelle Ågren	The National Veterinary Institute
Estrella Zarate	Swedish Pensions Agency
Fredrik Johansson	Uppsala University
Hanna Karlsson Ruiz	Swedish Pensions Agency
Hanna Linnér	Swedish Pensions Agency
Henrik Renlund	Uppsala Clinical Research Centre

Ingemar Svensson	Swedish Pensions Agency
Johan Lyhagen	Uppsala University
Johan Westerbergh	
Karl Birkholz	Swedish Pensions Agency
Kenneth Carling	Dalarna University
Kristina Juhlin	Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring
Lisbeth Hansson	Uppsala University
Love Hansson	Swedish Pensions Agency
Maria Nöremark	The National Veterinary Institute
Martin Bergström	The National Veterinary Institute
Mikael Elenius	Swedish Pensions Agency
Mikael Johansson	SAS
Nils Holmgren	Swedish Pensions Agency
Ronnie Pingel	Uppsala University
Thomas Rosendal	The National Veterinary Institute
Thommy Perlinger	Uppsala University
Tommy Lowén	Swedish Pensions Agency
Tuija Sonkkila	Aalto University, Finland

Abstracts - Invited Speakers

Data manipulation

Hadley Wickham

Abstract

R's built in subsetting is powerful, but can be verbose. The '**dplyr**' package is an powerful alternative to express most data manipulations, using a consistent family of functions. You'll learn:

- how to use the filter, select, arrange, mutate, and summarise functions
- how missing values work in R
- combine multiple data frames with joins.

Creating layered graphics with ggplot2

Hadley Wickham

Abstract

The key to creating rich, informative graphics in **ggplot2** is to use multiple layers. Layers can display different datasets, use different mapping between variables and aesthetics, use different transformations or different geoms. You'll learn how to use layers effectively to create rich graphics that combine raw data, context and summaries.

Cognitive psychology for visualisation

Hadley Wickham

Abstract

To design and critique visualisations you need to know a little bit about how the brain works. This talk will give you the basics of perception as it applies to visualisation.

I'll cover four main principles:

1. Match perceptual and data topology
2. Make important comparisons easy
3. Visual connections should reflect real connections
4. Beware of animation!

Each topic will be illustrated with real examples from around the web, and you'll be able to put the principles to work right away in your own visualisations. I'll also show some optical illusions, cases where our visual system fails us, and show how some common visualisation techniques can be extremely misleading.

ggvis sneak peek

Hadley Wickham

Abstract

I'll give you a sneak peek at **ggvis**, the successor to **ggplot2**. Like **ggplot2**, **ggvis** allows you to describe visualisations declaratively. Unlike **ggplot2**, **ggvis** graphics are fundamentally of the web: they're built using `html`, `js`, and `css`. More importantly, **ggvis** graphics are fundamentally reactive. You can bind plot parameters to sliders and dropdowns, and visualise streaming data as it comes in.

The art of complex communication

Jo Røislien

Abstract

How do you communicate complex scientific knowledge beyond the congregation? Based on the personal experience of being a university researcher who suddenly finds himself at the center of the production of a large scale TV-series about mathematics, statistics and numbers for the general public, aimed at attracting hundreds of thousands of viewers, week after week, this talk on communication in general, and statistics in particular, points to common mistakes often made in science communication, and offers advice on how to solve the problem, using anecdotes and concrete examples and film clips from TV-series Siffror and the mathematics short movie Chasing the worldâs largest number. Learning from popular culture and psychological studies, the importance of both metaphors and visuals images is highlighted. Seeing is believing.

Abstracts - Contributed Speakers (PhD students)

Using Paradata to Monitor Quality and Costs in Survey Production

Anton Johansson

Increasing costs and declining response rates are a major challenge for large scale survey production. In the Swedish Labour Force Survey (LFS), conducted by Statistics Sweden, a lot of effort is put into controlling the survey processes and survey costs.

Paradata (i.e. data about the data collection process itself) can be used in many different ways to monitor different quality aspects during data collection. There are various ways of how paradata can be used in a survey [1].

Still, finding quality indicators that serve both for the purpose of monitoring overall quality of the survey and the purpose of giving survey managers guidelines how to steer data collection efforts is not easy. Paradata (and quality indicators that make use of paradata) need to be simple enough to be interpretable, but also detailed enough to contain the most relevant quality aspects of the survey.

My future work will therefore focus on how to describe, analyze and interpret paradata from a statistical production perspective.

References

[1] Kreuter, F. (2013). Improving Sruveys with Paradata. *Wiley Series in Survey Methodology* .

Anton Johansson, Department of Statistics, Stockholm University, Stockholm

Integration of Somatic Mutation, Gene Expression and Functional Data in Predicting Breast Cancer Survival

Chen Suo, Donghwan Lee, Dhany Saputra, Stefano Calza and Yudi Pawitan

Most carcinomas are driven to develop by a few genetic alterations [1]. In the central dogma of biology, deoxyribonucleic acid (DNA), which contains genes, is transcribed to messenger ribonucleic acid (mRNA), and then translated to proteins. Irregularities in any of these processes may contribute to the development of disease. These potential irregularities include point mutations in DNA sequences, abnormal mRNA expression level which can be quantified by copies of mRNA produced from a gene, etc. High throughput sequencing technologies provide a powerful tool to measure simultaneously thousands of mutation and gene expression level in genome and transcriptome, respectively. Sequencing experiments can be used for comprehensive molecular studies of human cancers by comparing the sequences of DNA/RNA between normal and tumor tissues. But it is not immediately obvious how to subsequently integrate the complex information across the different types of molecular data, owing to a lack of mature statistical tools to identify potential driver genes in cancer progression. Fundamental challenges also lie in identifying patient-specific mutational event contributing to the heterogeneity between tumors and in translating the findings to the clinic.

We build an analytic pipeline using existing bioinformatics tools and propose a novel method to integrate genomic and transcriptomic profiles based on network enrichment analyses [2], revealing statistical evidence in the functional implications of mutated driver genes found inter- and intra-patients. We develop a driver gene score to capture the accumulative effect of driver genes. To contribute to the score, a gene has to be frequently mutated, with high or moderate mutational impact, exhibiting an extreme expression and linked to a large number of differentially expressed neighbors in the functional network.

We apply the pipeline to 60 matched tumor and normal samples of the same patient from The Cancer Genome Atlas breast cancer project [3]. We show that breast cancer patients carrying more mutated driver genes with functional implications and extreme expression pattern have worse survival than those with less mutated driver genes. Two validation data, a set of additional 671 TCGA samples and a Swedish microarray dataset, are tested as negative controls for evaluating the performance of driver gene score. In conclusion, integration of somatic mutation, expression and knowledge-based functional data allows identification of potential clinically relevant driver genes in cancer.

Chen Suo, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm

References

[1] Futreal P.A., Coin L., Marshall M., Down T., Hubbard T., Wooster R., Rahman N. and Stratton M.R. (2004). A cencus of human cancer genes. *Nature Reviews Cancer*, 4, 177.183.

[2] Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, Lehtiö J, Pawitan Y. (2012). Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinforma*, 13:226.

[3] Cancer Genome Atlas N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490: 61-70.

Triggering Solar-Powered Vehicle Activated Signs using Self Organizing Maps with K-Means

Diala Jomaa, Siril Yella and Mark Dougherty

Solar-powered vehicle activated signs (VAS) are speed warning signs powered by batteries recharged by solar panels. These signs are more desirable than other active warning signs due to low installation costs and minimal maintenance requirements. However solar-powered VASâs are often challenged by the limited power capacity available to keep the sign operational. In order to be able to operate the sign more efficiently, it is proposed that the sign be triggered appropriately by taking into account the prevalent conditions. Triggering the sign depends on many factors such as the prevailing speed limit, road geometry, traffic behaviour, the weather and the number of hours of daylight. The main goal of this work is to therefore develop an intelligent algorithm that would help optimise the trigger point to give the best compromise between speed reduction and power consumption. A systematic data collection has been carried out whereby vehicle speed data were gathered whilst varying the value of trigger speed threshold. A two stage algorithm is then utilised to extract the trigger speed value. The algorithm is first using Self Organizing Map (SOM), to effectively visualize and explore properties of the data that is then clustered in the second stage using K-means. Preliminary results achieved in the study indicate that using SOM in conjunction with K-means is found to perform well as opposed to direct clustering of the data by K-means alone. Using SOM in the current case has helped the algorithm determine the number of clusters in the data set, which is a frequent problem in data clustering.

References

[1] Jomaa D, Yella S and Dougherty M. (2013). Review of the effectiveness of vehicle activated signs. *Journal of Transportation Technologies*,3:123-130.

[2] Shukla S K, Rungta S and Sharma L K. (2012). Self-Organizing Map based Clustering Approach for Trajectory Data. *International Journal of Computer Trends and Technology*,3:321-324.

[3] Watts M J and Worner S P. (2009). Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering. *Ecological Modelling*,6:821:829.

Diala Jomaa, Department of Computer Engineering, Dalarna University, Borlange

An Inferential Framework for Domain Selection in Functional ANOVA

Johan Strandberg

Suppose that we observe a collection of functions from different populations. Our aim is to test the hypothesis that the distributions of all functional populations are equal against the alternative that at least one of the population differ from the other. We present a procedure for performing an ANOVA test on functional data, including pairwise group comparisons that is based on the Interval Testing Procedure (ITP) presented in [1]. The ITP is a non-parametric procedure that selects subintervals of the domain wherein data from different groups statistically differ. To illustrate the methodology we also present a real case study where the 3D kinematic motion data of the knee joint is analysed.

References

[1] Pini A. and Vantini S. (2013). The Interval Testing Procedure: Inference for Functional Data Controlling the Family Wise Error Rate on Intervals. *Tech. Rep. MOX, 13/2013*.

Johan Strandberg, Department of Mathematics and Mathematical Statistics, Umeå University, 90187 Umeå, Email: johan.strandberg@math.umu.se

Doubly robust estimation generalized estimating equations - the R package drgee

Johan Zetterqvist

The R package drgee provides functions to estimate parameters in GEE models with independent working covariance when the covariates can be divided into one exposure and a set of nuisance variables. In this case we are only interested in parameters quantifying the association between the exposure and the outcome, conditional on nuisance variables. Standard GEE estimators of such parameters require a correctly specified outcome nuisance model describing how the expected outcome depends on the nuisance variables. Alternative GEE-like estimators can be constructed such that they instead require a correctly specified exposure nuisance model describing the expected exposure conditional on the nuisance variables, e.g., G-estimators. By combining an outcome nuisance model and an exposure nuisance model, we can construct estimators that only require that one of the nuisance models is correctly specified, so called "doubly robust estimators" [1, 2]. The drgee package provides estimators of all the three types with link functions identity, log and logit. In my presentation I will give an overview of the theory behind the three estimation methods and describe the functionality of the package drgee. I will also talk about ongoing work on an extension to conditional GEEs [3], i.e., doubly robust estimators which combines two GEE models with cluster-specific intercepts.

References

[1] Robins J.M. (2000). Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models. *Proceedings of the American Statistical Association*, 1999:6–10.

[2] Tchetgen E.J.T., Robins J.M. and Rotnitzky A. (2010). On Doubly Robust Estimation in a Semiparametric Odds Ratio Model. *Biometrika*, 97(1):171–180.

[3] Goetgeluk S. and Vansteelandt S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780

Johan Zetterqvist, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna

Estimated Lifetimes of Road Pavements in Sweden Using Time-To-Event Analysis

Kristin Svenson

The aim of my research is to estimate lifetimes of road pavement in Sweden with time-to-event analysis. Applications of reliable lifetime estimates can be found in maintenance planning, life-cycle cost (LCC) analyses and marginal cost estimations.

In the settings of maintenance planning and LCC analyses it is also of interest to analyze the impact of different variables on the road's expected lifetime; such variables are pavement type, road type, bearing capacity, road width, speed limit, stone size and climate zone. This is obtained by fitting a Cox proportional hazards model to road data from the Swedish Traffic Agency's Pavement Management Systems. Results so far show that among the nine analyzed pavement types, stone mastic had the longest expected lifetime with a hazard ratio (risk of needing maintenance) estimated to be 36 percent lower than asphalt concrete. Among road types, 2+1 roads had 22 percent higher hazard ratio than ordinary roads indicating significantly lower lifetimes. Increased speed lowered the lifetime, while increased stone size (up to 20 mm) and increased road width lengthened the lifetime.[1]

To develop the present model, further research topics include adding a frailty (random effect) to the Cox model. This frailty can account for heterogeneities in the data which are uncovered by the fixed effects, e.g. heavy traffic (of which measures are uncertain) and different underlying road constructions. The frailty can also account for the spatial correlation in the road data, i.e. that the road network is spatially connected.

References

[1] Svenson, Kristin (2014). Estimated Lifetimes of Road Pavements in Sweden Using Time-To-Event Analysis. *Journal of Transportation Engineering*, (accepted).

Kristin Svenson, Dept. of Statistics, School of Technology and Business Studies, Dalarna University, Borlänge. E-mail: kss@du.se

Less is more - how to communicate simple but distinct

Magnus Fahlström

When you present something you want the recipients to perceive the content they way you intend without ambiguity. *A picture is worth a thousand words* is a famous phrase. If this phrase is true - how do one control the thousand words? In this session I will present some ideas for my research in the light of the theme - *Scientific Visualization and Presentation.* My research is about class room noise and the impact on students chance of learning as intended. I will address questions such as: In what way is it useful to transform future results to a form of: *Wasted Learning Units?*

Magnus Fahlström, Department of Mathematics Education, Dalarna University, Falun

Tackle 'em bugs! Managing the issue overflow in software engineering

Markus Borg

Software maintenance is one of the most expensive phases in software engineering, as shown by studies from the 1970s until today. Some practitioners consider operational and maintenance costs to require as much as 85-90% of their project budgets [3]. Thus, making software maintenance more efficient has potential to considerably reduce overall development costs. Issue management is a central part of software maintenance, revolving around issue reports collected in a central repository. Large organizations receive 10,000s of issue reports yearly, resulting in challenging prioritization decisions and work allocation.

Our work aims to support issue management by providing automated decision support [2]. A fundamental approach in our work is to make use of the rich information in issue repositories, i.e., to let previous work guide future decisions in issue management. Issue reports typically contain both text describing the experienced problem and other pieces of (often nominal or ordinal) information such as the version of the faulty software, the severity of the issue, and details on the execution environment. Also, issue reports are not independent, but are parts of complex networks of related problems [1].

Our contributions address several parts of issue management. First, we have used Information Retrieval (IR) techniques to detect duplicated issue reports. Second, we have applied machine learning to train classifiers (ensemble learning under stacked generalization) to automate assignment of issue reports to appropriate development teams, and to prioritize among issues. Third, we have replicated a study showing that (k-means) clustering issue reports based on their textual content can lead to clusters of issues with significantly different average resolution times. Fourth, we are currently evaluating support for change impact analysis by a recommendation system combining IR and network analysis.

References

[1] Borg, M., Pfahl, D., and Runeson, P. Analyzing Networks of Issue Reports (2013) In *Proc. of the Eur. Conf. on SW Maint. and Reeng.*, pp. 79-88.

[2] Borg, M. and Runeson, P. Changes, Evolution and Bugs: Recommendation Systems for Issue Management, (2014) *Recommendation Systems for Software Engineering*, Springer, pp. 477-509.

[3] Erlikh, L. Leveraging Legacy System Dollars for E-Business, . (2000) *IT Professional*, 2(3), pp. 17-23.

Markus Borg, Dept. of Computer Science, Lund University, markus.borg@cs.lth.se

Patterns of changing cancer risks with time since diagnosis of a sibling

Myeongjee Lee, Kamila Czene, Paola Rebora and Marie Reilly

It is well accepted that the diagnosis of cancer confers an increased risk on family members [1], but there are many unanswered questions concerning how this increased risk depends on the age at diagnosis of the first cancer in the family (index person), the age of the relative(s) at risk, and the time since the index diagnosis. Using the Swedish cancer register, we identified patients diagnosed with one of four major cancers (colorectal, breast, prostate and melanoma) as the first cancer. Using the Swedish Multi-Generation register, we extracted data on siblings of these patients (case siblings) and siblings of matched controls who were free of cancer on the date of diagnosis of the index persons (control siblings). We followed these siblings from the date of cancer diagnosis of the index person to diagnosis of the same cancer, censoring at death, emigration, end of study or diagnosis of a different cancer. We modeled the number of cancer diagnoses with Poisson regression, with follow-up time (person-years) as an offset and case/control status, age and time since diagnosis as predictors. The coefficient of status represents the log of the incidence rate ratio (IRR), comparing incidence rates (number of events per person-years of follow-up) in case siblings and in control siblings. We also presented the IRR graphically from flexible parametric survival models [2]. The overall familial risk estimates confirmed the published values for the four cancers. Higher cancer incidence was found in case siblings than control siblings for all cancers at all ages. The risk profile for case siblings was found to be approximately constant for up to 20 years for colorectal, breast and melanoma, but there was evidence of a sharp decline for prostate cancer, consistent with a lead-in bias from screening of family members. These results can contribute to the genetic counseling and optimal screening of family members of cancer patients.

References

[1] Hemminki K, Li X. (2004). Familial risks of cancer as a guide to gene identification and mode of inheritance. *Int J Cancer*, 110: 291-4.

[2] Royston P, Lambert PC. (2011). Flexible parametric survival analysis using Stata : beyond the Cox modeled. College Station, *TX: Stata Press*, xxvi, 347 p.

Myeongjee Lee, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm

Stationary generalized asymmetric Laplace models

Nima Shariati

In engineering and financial applications, one frequently encounters data that have considerable modulation in their variation in long run while behaving like a stationary Gaussian process in short run. Various at hoc data driven approaches have been applied to address this situation, but it seems more appropriate to find a model that captures this particular behavior. A suitable simple model is proposed which features generalized asymmetric Laplace distribution as its marginal. This model is obtained by an autoregressive type stationary process with gamma distributed marginal as variance-mean mixture scaling of a stationary Gaussian process. An estimation method for both model and distributional parameters is developed and used to fit some real data.

References

[1] Johannesson, P., Podgórski, K. and Rychlik, I. (2014). Modelling roughness of road profiles on parallel tracks using roughness indicators. *TechReport 4, MV Chalmers reports.*

[2] Kozubowski, T. J. and Podgórski, K. (2007). Invariance properties of the negative binomial Levy process and stochastic self-similarity. *International Mathematical Forum*, 2:1457-1468.

[3] Sim, C. H. (1971). First-order autoregressive models for gamma and exponential processes. *Journal of Applied Probability*, 27:325-332.

Nima Shariati, Department of Statistics, Lunds University, SE-220 07 Lund, Sweden

Coarsening to improve balance on key covariates in matching

Philip Fowler, Xavier de Luna and Ingeborg Waernbaum

In observational studies a researcher often needs to control for a large amount of confounding variables in order to not get biased estimates of average treatment effects. One commonly used approach to controlling for covariates is matching, where the researcher tries to compare units that have similar values of the confounding variables. However, as the number of covariates grow, it becomes increasingly more difficult to find good matches. Propensity score matching [1], that is matching on the probability of receiving treatment, is often used to get around this problem but can sometimes result in matches having rather different values on some important covariates. Here we discuss the case where a few variables can be coarsened, that is into discretized by grouping them into intervals, and still retain all their information regarding treatment assignment. We suggest that a data driven algorithm could be used to determine the level of coarsening each key covariate gets, which in turn can be matched exactly on, guaranteeing that the matches will have similar values on said covariates. This could then be combined with further, less exact, matching on the other covariates.

References

[1] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41-55.

Philip Fowler, Department of Statistics, Umeå University, Umeå

Confidence of the heuristic solutions

Kenneth Carling and Xiangli Meng

Solutions to combinatorial optimization, such as p-median problems of locating facilities, frequently rely on heuristics to minimize the objective function. The minimum is sought iteratively and a criterion is needed to decide when the procedure (almost) attains it. However, pre-setting the number of iterations dominates in operational research applications, which implies that the quality of the solution cannot be ascertained. A small branch of the literature suggests using statistical principles to estimate the minimum and use the estimate for either stopping or evaluating the quality of the solution. In this paper we use test-problems taken from Baesley OR library [1] and apply Simulated Annealing on these p-median problems. We do this for the purpose of comparing suggested methods of minimum estimation and, eventually, provide a recommendation for practioners. An illustration ends the paper being a problem of locating some 70 distribution centers of the Swedish Post in a region.

References

[1] Beasley, J.E., (1990). OR library: Distributing test problems by electronic mail. *Journal of Operational Research Society*,41(11): 1067-1072.

School of Technology and Business Studies, Dalarna university, SE 791 88 Falun, Sweden

Process GPS data on car-movements in Dalarna, Sweden

Xiaoyun Zhao, Kenneth Carling, Johan Håkansson

The advancement of GPS technology enables the GPS devices not only to be used as orientation and navigation tools, but also as instruments to capture travelled routes. GPS vehicle tracking data provides an essential unprocessed material for a broad range of applications such as traffic management and control, transportation routing and planning. The unprocessed GPS data can be directly used on very few aspects such as simply extracting the velocity and coordinates. To become more useful, it has to be related to the underlying road network by means of map matching algorithms. This paper processes a collected GPS data of seven-week positional recordings of 316 volunteers. It extracts the detailed information of the GPS data and investigates the value of tracking the movements of cars. The technical process from viewing the original unprocessed data to developing visualized maps in relation to the road network are presented. The visualized maps captures the complexity of the real travel trips and shows that the trips are neither totally repetitious nor totally various considering the accuracy of the data. The processed GPS data enables the evaluation of the pollutants emissions from car movements and provides the access to more related urban planning studies.

References

[1] Ashbrook D and Starner T. (2003). Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275-286.

[2] Jia T, Carling K and Håkansson J. (2013). Trips and their CO2 emissions to and from a shopping center. *Journal of Transport Geography*, 33:135-145.

[3] Shoval N. Tracking technologies and urban analysis. *Cities 2008*, 25:21-28.

[4] Wolf J, Guensler R and Bachman W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board*, 1768(1):125-134.

Xiaoyun Zhao, School of Technology and Business Studies, Dalarna University, Sweden

PANIC residuals based pooled tests with contemporaneous correlated errors

Xijia Liu

Under a rather general model and some less restrictive assumptions, PANIC procedure proposed by Bai and Ng (2004) [1], can provide a consistent estimation for both common factors and idiosyncratic errors without considering the integrated order such that the nonstationarity can be tested for common factors and idiosyncratic errors respectively. However, a further restrictive assumption, the idiosyncratic errors should be independent among all cross-section units, must be made when one want to do a panel unit root test on the idiosyncratic errors. In this study, in order to fill in this gap, I propose two pooled tests for idiosyncratic errors. First, following the spirit of Breitung and Das (2005) [2], do robust OLS t test on the estimation of the idiosyncratic errors. Second, I consider doing the bootstrap tests, for example Chang (2004) [3] and Palm et al. (2011) [4], on the estimation of the idiosyncratic errors. For both of them, I do theoretical analysis first and investigate the small sample properties by Monte Carlo simulations.

References

[1] Bai, J. and Ng, S. (2004). A Panic Attack on Unit Root and Cointegration, *Econometrica*, 72:1127-177.

[2] Breitung, J. and Das, S. (2005). Panel unit roots under cross-sectional dependence, *Statistica Neerlandica*, 4:414-433.

[3] Chang, Y. (2005). Bootstrap unit root tests in panels with cross-sectional dependency, *Journal of Econometrics*, 120:263-293.

[4] Palm, F. C., Smeekes, S. and Urbain, J. P. (2011). Cross-sectional dependence robust block bootstrap panel unit root tests, *Journal of Econometrics* 163:85-104

Address correspondence to Xijia Liu, Department of Statistics, Uppsala University, Box 513, 751 20 Uppsala, Sweden; e-mail: xijia.liu@statistik.uu.se.

Comparing idiosyncratic unit root tests based on the residuals estimated from ML and PC in a dynamic factor model

Xingwu Zhou

Dynamic factor models are widely used in econometric and financial area. For $i = 1, \cdots, N$ and $t = 1, \cdots, T$, consider a dynamic factor model $x_{i,t} = \boldsymbol{\lambda}_i' \boldsymbol{f}_t + u_{i,t}$, where $x_{i,t}$ is the observed data, $\boldsymbol{f}_t$ is a $r \times 1$ vector of the unobserved dynamic factors, $\boldsymbol{\lambda}_i$ is a $r \times 1$ vector of the corresponding factor loadings and $u_{i,t}$ is the unobserved dynamic idiosyncratic error term.

Testing the stationarity of the idiosyncratic error term $u_{i,t}$ is important, e.g., to test if the coefficient ρ_i in $u_{i,t} = \rho_i u_{i,t-1} + \varepsilon_{i,t}$ is less than 1. The idiosyncratic error term can be estimated either from the method of principal components (PC) or the method of (quasi) maximum likelihood (ML). In this paper, we compare the size and power properties of some commonly used unit root tests under the two sets of residuals. We also compare the estimation efficiency of the common factors, factor loadings and the idiosyncratic components across the two methods when the factor model is either stationary or non-stationary. The simulation results show that when N is small and T is large, ML dominates PC. When N and T are both large, the difference is small.

Department of Statistics, Uppsala University,
E-mail address: xingwu.zhou@statistics.uu.se

Spatial analysis of Swedish business

Yujiao Li

Based on the theory of economic geography and network society, we focused on the spatial correlation of the business of Swedish firms. Through the geographic coordinates and their financial information, spatial competition and cooperation among firms were shown by sets of distance-dependent variogram. We separately studied different cities and industries so as to gain more detailed feature of economic agglomeration. We discussed the traditional manufacture, wholesale and retailing as well as the creative industry including the design,advertisement,high-technology innovative industry which are more geographical sticky due to their knowledge spillover in current digital age. We examined the assumption that creative industry prefer to cluster and thrive economically, attract other firms and investment simultaneously. Thus spatial dependence within and between industries were also taken into consideration. The spatial dependence in the same distance also varied year by year, which indicated that the extent of economic spatial dependence in the same distance had been changing.

References

[1] Coe, N.M., Kelly, P.F. & Yeung, H.W.C. (2013). *Economic Geography (2rd ed)*. 294-365, NY: Wiley.

[2] Daunfeldt, S-O., Lang, A., Macuchova, Z. & Rudholm, N. (2013). Firm growth in the Swedish retail and wholesale industries. *The Service Industries Journal*, 10-11.

[3] Hartley, J., Potts, J., Cunningham, S., Flew, T., Keane, M. & Banks, J. (2013). *Key concepts in Creative Industries*. 14-52, SAGE Publications Ltd.

[4] Peter Maskell (2014). The Firm in Economic Geography. *Economic Geography Journal*, 77:329-344.

Yujiao Li, Department of Microdata Analysis, Dalarna University, Dalarna

Firm relocation and firm performance - evidence from the Swedish wholesale and retail sector

Zuzana Macuchova

This study analyses the effects of firm relocation on firm performance within the Swedish wholesale sector. Firm relocation is in general not as extensively studied as other events of the firm demography, as for example firm entry. This is reflected in the theory, where firm relocation theories are seen mainly as specific application of location theories. However, in recent decades there has been a renewed interest in this topic, which is strongly influenced by the seminal work of Krugman and the 'new economic geography', as well as by better accessibility of firm-level data. Size of the firm, growth in previous period, or past profitability are some of the commonly mentioned characteristics influencing firm relocation. In our study, using micro-level data on Swedish limited liability firms within the wholesale trade sector in the period 1998-2010, we seek to analyse the firm's performance in the pre- and post-relocation period and in this way answer the research question: Does firm's relocation contributes to higher firm's performance? It is striking, that such empirical evidence is in the current literature on firm relocation clearly missing. The descriptive statistics show that those firms in our dataset that have relocated, were on average smaller and younger, and had lower profits in the period prior to relocation. To investigate the causal effects of relocation, we use three different models of propensity score matching. Propensity score matching method has a large advantage, it enables to compare 'comparable' units and thus, allows us to analyse the causal effects of the firm's relocation on firm's performance, taking into consideration the possible self-selection bias. To investigate the causal effects of relocation, the relocating firms were matched with firms, having almost identical characteristics in the period prior to relocating, however, different in the extent that they haven't relocated their activities. The results from the matched sample indicate that for an average firm, the relocation leads to higher profits in the post-relocation period. Our conclusion is that relocation, indeed, on average contributes to increasing profitability of relocating firms.

Zuzana Macuchova, Dalarna University, SE- 791 88 Falun, Sweden.