

# 1 Geometric inference and robustness

Catherine Aaron

*Université Clermont Auvergne, France*

When observing an iid sample in  $\mathbb{R}^D$ , drawn according to a distribution  $\mathbb{P}_X$  supported by a compact  $d$  dimensional set  $S$  we aim at estimating  $S$  (when  $d = D$  we will speak of set estimation, when  $d < D$  we will speak of manifold estimation) its levels sets  $L_t = \{x, f(x) \geq t\}$ , its possible boundary  $\partial S$  and related quantities such as its volume  $|S|_d$  or its perimeter  $|\partial S|_{d-1}$ . A paradoxal point being that geometric inference can be used to detect outliers. It was the point of one of the first papers in the topic : ("Detection of abnormal behaviour via nonparametric estimation of the support" by Luc Devroye and Gary L Wise in 1980). But geometric inference method are, mostly, not robust. The aim of the talk consists in presenting some tools related to this topics, to discuss there (mostly lack of) robustness and, as perspective propose some clues to deal with robustness.

RNS1

# 2 Testing many restrictions under heteroskedasticity

Stanislav Anatolyev<sup>1</sup>, Mikkel Mikkelsen<sup>2</sup>

*1 – CERGE-EI, Czech Republic; 2 – University of Wisconsin-Madison, USA*

We propose a hypothesis test that allows for many tested restrictions in a heteroskedastic linear regression model. The test compares the conventional F statistic to a critical value that corrects for many restrictions and conditional heteroskedasticity. The correction utilizes leave-one-out estimation to correctly center the critical value and leave-three-out estimation to appropriately scale it. Large sample properties of the test are established in an asymptotic framework where the number of tested restrictions may be fixed or may grow with the sample size and can even be proportional to the number of observations. We show that the test is asymptotically valid and has non-trivial asymptotic power against the same local alternatives as the exact F test when the latter is valid. Simulations corroborate the relevance of these theoretical findings and suggest excellent size control in moderately small samples also under strong heteroskedasticity.

SHDA1

# 3 Retrieving grouped LATEs via Classifier-Lasso

Nicolas Apfel<sup>1</sup>, Helmut Farbmacher<sup>2</sup>, Martin Huber<sup>3</sup>, Henrika Langen<sup>3</sup>

*1 – University of Surrey, UK; 2 – TU Munich, Germany; 3 – University of Fribourg, Switzerland*

In the context of an endogenous binary treatment with heterogeneous effects and multiple instruments, we propose to apply a classifier-Lasso (C-Lasso, Su, Shi, Phil-

ipps, 2016) and an Agglomerative Hierarchical Clustering (AHC, Ward, 1963) procedure to identify complier groups with identical local average treatment effects (LATE), in spite of relying on distinct instruments. Our procedure is based on the fact that the LATE needs to be homogeneous for any two or multiple instruments that (i) satisfy the LATE assumptions and (ii) generate identical complier groups in terms of treatment probabilities given the respective instruments. Under the assumption that a majority or plurality of instruments with identical complier groups satisfies the LATE assumptions, our procedure permits identifying the valid instruments (in the sense that they satisfy the exclusion restriction) in a data driven way. The choice of tuning parameters is discussed and for this we propose the use of an information criterion. We also provide a simulation study investigating the finite sample properties of our LATE C-Lasso and AHC approach and an empirical application investigating the effect of incarceration on recidivism in the US with judge assignments serving as instruments.

CI

## 4 Agglomerative Hierarchical Clustering for Selecting Valid Instrumental Variables

Nicolas Apfel<sup>1</sup>, Xiaoran Liang<sup>2</sup>

1 – University of Surrey, UK; 2 – University of Bristol, UK

We propose a novel procedure for Mendelian Randomization (MR) with pleiotropic SNPs which combines agglomerative hierarchical clustering and a test of overidentifying restrictions for selecting valid instrumental variables (IV) from a large set of candidate IVs. Some of these IVs may be invalid in the sense that they fail the exclusion restriction. We show that if the largest group of IVs is valid, our method achieves oracle properties. Unlike existing techniques, our method can deal with weak instruments, multiple endogenous regressors as in Multivariable MR and heterogeneous causal pathways. In simulations, we show that our procedure outperforms the two closest methods, the Hard Thresholding and the Confidence Interval method.

SME3

## 5 Evaluation of Misspecified Linear Regression Models for Subgroup Analysis

Saïde Atmaca

*Ulm University of Applied Sciences in Cooperation with Boehringer Ingelheim Pharma GmbH & Co. KG, Germany*

In clinical trials, subgroup analyses are routinely performed to assess the treatment effect in certain subpopulations. For linear models with continuous endpoints, the use of linear regression represents a classical approach to determine the influence

of subgroup status on treatment effect. For a valid interpretation of the resulting regression estimates, correct model specification is a fundamental assumption. One source of model misspecifications is that the choice of included variables does not represent the true underlying model, i.e. the omission of relevant (underfitting) or the inclusion of irrelevant (overfitting) variables.

We investigate model misspecification in the context of subgroup analyses by comparing a full population model, including interaction terms for the subgroup status, with a subgroup model that represents a separate analysis within each subgroup category. A full population model that is not fully interacted with the subgroup status may be underfitted compared to the subgroup model, which can cause bias to the parameter estimates (the so-called omitted variable bias). In this work, we assess the estimators in underfitted full population analysis models. Using a simulation study the relationship between the two types of models is examined under different parameter settings. In a second step, an analytical approach is taken to better understand which question the estimators may answer. The simulation results indicate that the absence of an interaction term in the full population model only leads to underfitting when the effect of the variables is different across the subgroup categories. Moreover, the analytical results underpin that by adding interaction terms between the variables and the subgroup status, the estimators consider the observations of the included variables for the respective groups separately. In the absence of the grouping variable or the corresponding interactions, the consideration of observations is expanded across groups.

IBSDRP

## 6 General-purpose imputation of planned missing data in social surveys: different strategies and their effect on correlations

Julian B. Axenfeld<sup>1</sup>, Christian Bruch<sup>2</sup>, Christof Wolf<sup>2</sup>

*1 – University of Mannheim, Germany; 2 – GESIS Leibniz Institute for the Social Sciences, Germany*

Planned missing survey data, for example stemming from split questionnaire designs, are becoming more and more common in large-scale social surveys, making imputation indispensable to obtain reasonably analyzable data. This is especially because surveys are facing pressures to shorten questionnaires: Long questionnaires are associated with low response rates, poor response quality, and are particularly considered inappropriate for the increasingly popular online mode. However, these data can be difficult to impute due to common features of social survey data, such as low correlations, predominantly categorical data, and relatively small sample sizes available to support imputation models with many potential predictor variables.

In this presentation, we discuss findings from a Monte Carlo simulation in which we simulate split questionnaire designs, evaluating different imputation methods based on data from the German Internet Panel (GIP). In this simulation, we also experiment with predictor set specifications in which imputation models are restricted exclusively to variables that have correlations to the imputed variable clearly larger than zero. Our results show that strategies that simplify the imputation exercise

(for instance, predictive mean matching procedures with restricted predictor sets) perform well, while some established strategies lead to strong biases.

SM2

## 7 Posterior Concentration Rates for Bayesian Penalized Splines

**Paul Bach, Nadja Klein**

*Humboldt-Universität zu Berlin, Germany*

Despite their widespread use in practice, the asymptotic properties of Bayesian penalized splines have not been investigated so far. We close this gap and study posterior concentration rates for Bayesian penalized splines in a Gaussian nonparametric regression model. A key feature of the approach is the hyperprior on the smoothing variance, which allows for adaptive smoothing in practice but complicates the theoretical analysis considerably. Our main tool for the derivation of posterior concentration rates with a general hyperprior on the smoothing variance is a novel spline estimator that projects the observations onto the first basis functions of a Demmler-Reinsch basis. Our results show that posterior concentration at near optimal rate can be achieved if the hyperprior on the smoothing variance strikes a fine balance between oversmoothing and undersmoothing. Another interesting finding is that the order of the roughness penalty must exactly match the regularity of the unknown regression function in order to achieve posterior concentration at near optimal rate. Overall, our results are the first posterior concentration results for Bayesian penalized splines and can be generalized in many directions.

MS2

## 8 DoubleML - An Object-Oriented Implementation of Double Machine Learning in R

**Philipp Bach<sup>1</sup>, Victor Chernozhukov<sup>2</sup>, Malte Kurz<sup>1</sup>, Martin Spindler<sup>1</sup>**

*1 – University of Hamburg, Germany; 2 – MIT, USA*

The R package DoubleML implements the double/debiased machine learning framework of Chernozhukov et al. (2018). It provides functionalities to estimate parameters in causal models based on machine learning methods. The double machine learning framework consist of three key ingredients: Neyman orthogonality, high-quality machine learning estimation and sample splitting. Estimation of nuisance components can be performed by various state-of-the-art machine learning methods that are available in the mlr3 ecosystem (Lang et al, 2019). DoubleML makes it possible to perform inference in a variety of causal models, including partially linear and interactive regression models and their extensions to instrumental variable estimation. The object-oriented implementation of DoubleML enables a high flexibility for the model specification and makes it easily extendable. This presentation serves as an introduction to the double machine learning framework and the R package DoubleML. In reproducible code examples with simulated and real data sets, we

demonstrate how DoubleML users can perform valid inference based on machine learning methods.

## References

- [1] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21: C1-C68, <https://doi.org/10.1111/ectj.12097>.
- [2] Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L. and Bischl, B. (2019), mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, <https://doi.org/10.21105/joss.01903>.

CSASOFT2

## 9 A game-theoretic approach for unsupervised ranking and selection of pathways in gene sets

**Chiara Balestra<sup>1,3</sup>, Carlo Maj<sup>2</sup>, Emmanuel Müller<sup>1</sup>, Andreas Mayr<sup>3</sup>**

*1 – Technical University of Dortmund, Germany; 2 – Institute for Genomic Statistics and Bioinformatics IGSB, University Hospital Bonn, Germany; 3 – Institut für Medizinische Biometrie, Informatik und Epidemiologie, University Hospital Bonn, Germany*

Shapley values provide a way to assign fair values to players within a cooperative game. We interpret sets within families of sets as players of a game and extend the application of game theory to unsupervised rankings of pathways in gene sets; each pathway, i.e., each set, has a value dependant on the distribution of its elements within the family of sets. However, the Shapley values do not consider possible overlaps among sets and possibly rank overlapping sets in similar positions.

We propose Shapley values-based feature importance scores aware of redundancy for sets. Inspired by the necessity of reducing the overlap among pathways while still retaining a high coverage of the gene sets, our method assigns an importance score to each set. We apply the obtained ranking to reduce the dimension of families of sets. Our case study for pathways and gene sets shows that the obtained selection of pathways contains lower redundancy while still keeping a high coverage of the genes when compared to naive Shapley values. Moreover, a naive application of Shapley values leads to a positive correlation of the position in the ranking and the pathways' size; with our method, instead, the aforementioned correlation is reduced.

Furthermore, we paired the Shapley values-based feature importance scores with gene set enrichment analysis where multiple hypothesis testing for association with phenotypic traits is an issue. We use our ranking for unsupervised pre-selection of pathways and study the effect on the number of significant pathways found for the single association traits.

BSB2

# 10 Calibration of wind speed ensemble forecasts for power generation

Ágnes Baran, Sándor Baran

*University of Debrecen, Hungary*

In 2020, 36.6% of the total electricity demand of the world was covered by renewable sources, whereas in the EU (UK included) this share reached 49.3%. A substantial part of green energy is produced by wind farms, where accurate short range power predictions are required for successful integration of wind energy into the electrical grid. Accurate predictions of the produced electricity require accurate forecasts of the corresponding weather quantity, where the state-of-the-art method is the probabilistic approach based on ensemble forecasts. However, ensemble forecasts are often uncalibrated and might also be biased, thus require some form of post-processing to improve their predictive performance.

To calibrate (hub height) wind speed ensemble forecasts we propose a novel flexible machine learning approach, which results either in a truncated normal or a log-normal predictive distribution [1]. In a case study based on 100m wind speed forecasts of the operational AROME-EPS of the Hungarian Meteorological Service, the forecast skill of this method is compared with the predictive performance of three different ensemble model output statistics approaches and the raw ensemble predictions. We show that compared with the raw ensemble, post-processing always improves the calibration of probabilistic and accuracy of point forecasts, and from the five competing methods the novel machine learning based approaches result in the best overall performance.

## Reference

[1] Baran, S., Baran, Á., Calibration of wind speed ensemble forecasts for power generation. *Időjárás* 125 (2021), 609–624.

SAES2

# 11 APCtools: An R Package for Descriptive and Model-based Age-Period-Cohort Analysis

Alexander Bauer<sup>1</sup>, Maximilian Weigert<sup>1</sup>, Hawre Jalal<sup>2</sup>

*1 – Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany; 2 – Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, USA*

Age-Period-Cohort (APC) analysis aims to differentiate relevant drivers for long-term developments and is used in many fields of science. The critical challenge in APC analysis is that the main components age, period and cohort are linearly dependent (e.g., cohort = period – age). Accordingly, flexible methods and visualization techniques are needed to properly disentangle observed temporal association structures. We introduce APCtools, the first sophisticated R package that covers all aspects of APC analysis. The package comprises functions both for descriptive

and regression model-based analyses. For the former, we use density (or ridgeline) matrices and adapted heatmaps as innovative visualization techniques that build on the concept of Lexis diagrams. Since Lexis diagrams and derived visualizations share the problem of visually underrepresented diagonals we offer hexagonally binned heatmaps (hexamaps) as an alternative in which all three axes have similar visual weight. Model-based analyses build on the separation of the temporal dimensions based on generalized additive models and can be used to simultaneously account for further control variables. The main temporal effect is estimated based on a tensor product interaction surface between two dimensions, representing the third dimension on its diagonal. Additional to heatmap and hexamap visualizations of the resulting tensor product estimate, partial APC plots allow for detailed insights in the dependency structure among the temporal dimensions. We showcase the functionality of the package by analyzing how the travel behavior of German travelers changed over the last decades.

CSASOFT1

## 12 Quantile-based MANOVA: A new tool for inferring multivariate data in factorial designs

**Marléne Baumeister, Marc Ditzhaus, Markus Pauly**

*TU Dortmund University, Germany*

Multivariate Analysis of Variance (MANOVA) allows for the joint comparison of multiple responses and is commonly used in fields such as medicine, ecology or psychology. However, the classical MANOVA relies in assumptions, e.g. normality or homogeneous covariance matrices, which are often difficult to justify. To overcome these difficulties there are less restrictive mean-based MANOVA concepts proposed for testing global hypothesis about multivariate expectations, e.g. (Friedrich and Pauly, 2018). In case of outliers or distributions with larger tails, however, non-robust estimators like the mean can have some drawbacks. Despite the usage of quantiles is intuitive in that case and often applied in descriptive statistics, e.g. boxplots, quantiles "[appear] to be quite underused in medical research" (Beyerlein, 2014). Therefore we developed a flexible quantile-based MANOVA method. The approach is adaptable to general factorial designs and has the advantage that it fits to median and other quantile-based statistical methods. To achieve this, we considered two quadratic-form type test statistics and three different strategies for estimating the covariance. The test statistics' distribution is approximated via resampling. We prove by empirical processes that our method is valid in theory and even works in case of general heterogeneous or heteroscedastic data beyond normality. In a simulation study, we compare the novel procedures with state-of-the-art mean-based approaches.

### References

- [1] Beyerlein, A. (2014). Quantile Regression—Opportunities and Challenges From a User's Perspective. *American Journal of Epidemiology*, 180(3):330–331.
- [2] Friedrich, S. and Pauly, M. (2018). MATS: Inference for potentially singular and

## 13 mlr3tuning: A general framework for ML hyperparameter tuning

Marc Becker<sup>1</sup>, Michel Lang<sup>2</sup>, Bernd Bischl<sup>1</sup>

*1 – Ludwig-Maximilians-Universität München, Germany; 2 – Technische Universität Dortmund, Germany*

The performance of modern ML algorithms depends on their hyperparameters. While R offers packages to solve general optimization problems, it lacks a common interface for hyperparameter tuning.

mlr3tuning offers exactly that and already includes these tuners: random search, grid search, iterated racing, GenSA and nloptr. Packages mlr3hyperband and mlr3mbo extend this list with multifidelity approaches and Bayesian optimization. Furthermore, mlr3tuning offers a flexible system for termination, a generic archive for logging of evaluations and the capacity for nested resampling. Search space parameters can be of different types, including mixed categorical-numerical spaces, including hierarchical dependencies. These can be optimized on arbitrary scales via transformations. This is further simplified by mlr3tuningspaces which offers a extensible collection of common search spaces for learners.

An important aspect of mlr3tuning is its modularity integration into the mlr3 ecosystem (Lang et al., 2019). Learners can be tuned for a large number of metrics and resampling methods and the tuning process can be parallelized via different backends. Furthermore, arbitrary ML pipelines from mlr3pipelines (Binder et al., 2021) can be jointly optimized to handle the combined algorithm selection and hyperparameter optimization (CASH) problem.

### References

- [1] Binder et al. (2021). mlr3pipelines-Flexible Machine Learning Pipelines in R. Journal of Machine Learning Research, 22(184), 1–7.
- [2] Lang et al. (2019). mlr3: A modern object-oriented machine learning framework in R. Journal of Open Source Software, 4(44), 1903.

CSASOFT2

## 14 Firth correction in Cox proportional hazard analysis in the presence of zero events

Lars Beckmann, Anke Schulz, Guido Skipk

*IQWiG, Germany*

For the early benefit assessment of drugs in Germany, the pharmaceutical company



must describe the extent of an added benefit of the drug to be assessed compared with an appropriate comparator therapy [1]. The confidence interval of a significant effect must lie completely outside a certain corridor around the null effect for the extent of the effect to be regarded as minor, considerable or major. The corridors are defined by different thresholds depending on outcome category. For time to event analyses, the 95% confidence interval for the hazard ratio is required to determine the extent.

If no events are observed in one of the study arms, the standard Cox proportional hazard regression does not provide effect estimates with corresponding confidence intervals, while the log rank test provides appropriate p-values. Thus, in the case of a statistically significant effect, the extent cannot be determined and the overall assessment of the added benefit might be hampered.

Heinze and Schemper proposed an adaption of the Firth correction to reduce bias from maximum likelihood estimation for the Cox proportional hazard [2]. To assess the applicability of this approach, we performed a simulation study of time to event analyses with zero events. We will present results from this study and discuss the situations, in which the application of the Firth correction provides reliable estimates that can be used for the assessment of the extent of an added benefit.

## References

[1] IQWiG. General Methods 6.0 [online]. 2020. URL: <https://www.iqwig.de/en/about-us/methods/methods-paper/>

[2] Heinze and Schemper (2001). *Biometrics* 57(1):114–119.

SEHA2

# **15 A Test of Independence for Locally Stationary Processes Using a Weighted Characteristic Function-based Distance**

**Carina Beering**

*Helmut Schmidt University Hamburg, Germany*

We propose a testing procedure for independence of locally stationary processes based on a weighted distance composed of characteristic functions (CF) and its empirical version. The essential idea of this concept is inspired by the distance covariance defined by Székely et al. (2007) and was taken up by Jentsch et al. (2020). For the purpose of compiling a testing procedure at the end, we provide the needed results with the notion of the beneficial effects of a bootstrap analogue. Therefore, we establish the bootstrap versions of the previously presented findings. Beforehand, we transfer the concept of empirical weighted CF-distance to the bootstrap world. Finally, a simulation study is performed using our testing procedure to detect dependence of different forms.

## References

[1] Jentsch, C., Leucht, A., Meyer, M. and Beering, C. (2020). Empirical characteristic functions-based estimation and distance correlation for locally stationary

processes. *Journal of Time Series Analysis* 41, 110-133.

[2] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35, 2769-2794.

TSA3

## 16 Modeling trajectories of slowly progressing diseases: A mixed-model-based algorithm for variable transformation, prediction and age-of-onset estimation

**Charlotte C. Behning<sup>1</sup>, Monika Fleckenstein<sup>2</sup>, Maximilian Pfau<sup>3</sup>, Christine Adrion<sup>4</sup>, Lukas Goerdts<sup>5</sup>, Moritz Lindner<sup>5</sup>, Steffen Schmitz-Valckenberg<sup>2</sup>, Frank G. Holz<sup>5</sup>, Matthias C. Schmid<sup>1</sup>**

*1 – Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Germany; 2 – John A. Moran Eye Center, University of Utah, Salt Lake City, USA; 3 – Ophthalmic Genetics and Visual Function Branch, National Eye Institute, Bethesda, MD, USA; 4 – Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University, Munich, Germany; 5 – Department of Ophthalmology, University Hospital Bonn, Bonn, Germany*

Modeling trajectories of slowly progressing diseases is essential to understand the course of the disease, to identify associated risk factors and to design clinical (interventional) trials. Determining the age of disease also remains a largely unsolved problem. Here, we aim to model the progression of geographic atrophy (GA) in patients with age-related macular degeneration (AMD) by building a suitable statistical regression model. Based on theoretical considerations, we develop a linear mixed-effects model for GA size progression that incorporates covariable-dependent enlargement rates as well as correlations between longitudinally collected GA size measurements. Instead of simply including a patient specific intercept and slope, we directly incorporate the (risk-factor dependent) age of disease at study entry into our model. This way, we account for the fact, that patients enter a clinical study at different stages in their disease history. To capture nonlinear progression in a flexible way, we systematically assess Box-Cox transformations with different transformation parameters  $\lambda$ . The selection of an optimal transformation parameter is based on Akaike information criterion (AIC). We evaluate our model using data collected for two longitudinal, prospective, multicenter cohort studies of GA size progression that included 150 eyes from 101 patients. The model allows for age-of-onset estimation, identification of risk factors and prediction of future GA size. Based on the assessment of the transformation approach, we recommend a square-root transformation of atrophy size before model fitting.

### Reference

[1] <https://doi.org/10.1186/s12874-021-01356-0>

ARM3