ENSAI CREST CENTER FOR RESEARCH IN ECONOMICS AND STATISTICS UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR cnrs Mediametrie NWU®

The 5th workshop on *Goodness-of-fit, Change Point and related problems (GOFCP)* is held in Rennes, France, at Ecole Nationale de Statistique et Analyse de l'Information (ENSAI). The workshop follows those of Trento (2019), Bad Herrenalb (2017), and Athens (2015).

This edition counts 28 invited speakers and 13 invited poster presentations. The variety of topics and the content of the presentations show the importance of the workshop themes in the modern developments of statistical methodology and applications. Talented early-stage and widely recognized confirmed researchers, who have accepted the invitation for attending the workshop, are crucial to the success of the event. I thank them all sincerely for their contributions.

I would also like to mention the fundamental role played by all members of the Scientific Program Committee and Simos Meintanis in designing the final program of the workshop. Local details and logistics of this event were carried out by the Organizing Committee. Special thanks go to ENSAI PhD students involved in the organization, and the ENSAI staff who efficiently and enthusiastically provided all the needed support.

Valentin Patilea

Contents

Workshop Program

Friday, Sept. 2

Registration and opening address: *9:15 - 10:00*

IS–1: *10:00 - 12:00* – Chair Ingrid Van Keilegom

- Alexandre Tsybakov: *Statistical decisions for variable selection*
- Lajos Horváth: *Change point detection in high dimensional data with $U-$statistics*
- Jean-Marc Bardet: *Data-driven semi-parametric detection of multiple changes in long-range dependent processes*
- Richard J. Samworth: *High-dimensional changepoint estimation with heterogeneous missingness*

Lunch

IS–2: *13:30 - 15:30* – Chair François Portier

- Jad Beyhum: *Testing for error invariance in separable instrumental variable models*
- Arthur Gretton: *A Kernel Stein Test for Comparing Latent Variable Models*
- Ingrid Van Keilegom: *Instrumental variable quantile regression under random right censoring*
- Tom Berrett: *Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility*

Coffee/Tea

IS–3: *16:00 - 18:00* – Chair Simos G. Meintanis

- Elia Lapenta: *Partly Linear Instrumental Variables Regression Without Smoothing on the Instruments*
- François Portier: *Density model checks via the lack-of-fitness*
- Wenceslao González-Manteiga: *Testing specification in some complex models using correlation distances*
- Abderrahim Taamouti: *Copula-Based Estimation of Health Inequality Measures With an Application to COVID-19*

Dinner: from 19:30

Saturday, Sept. 3

IS–4: *9:00 - 11:00* – Chair Richard J. Samworth

- Bruno Ebner: *A general maximal projection approach to uniformity testing on the hypersphere*
- Thomas Verdebout: *Asymptotic power of Sobolev tests for uniformity on hyperspheres*
- Eduardo García-Portugués: *Some new tests of independence for circular data*
- Marc Hallin: *A Measure-Transportation-Based GOF Test for Directional Data*

Coffee/Tea

Poster Session: *11:15 - 12:15* – Chair Valentin Patilea

Short oral presentations of the posters (by alphabetical order of the presenters)

Lunch

IS–5: *13:15 - 14:45* – Chair Natalie Neumeyer

- Simos G. Meintanis *Tests for the stable paretian hypothesis for i.i.d. data and for innovations in multivariate GARCH models*
- Maria Dolores Jiménez-Gamero: *Testing Poissonity of a large number of populations*
- Yi Yu: *Change point inference in high-dimensional regression models under temporal dependence*

Coffee/Tea

IS–6: *15:00 - 16:30* – Chair Juan Carlos Escanciano

- Denis Belomestny: *Semiparametric estimation of McKean-Vlasov SDEs*
- Natalie Neumeyer: *Semiparametric transformation models: mean and boundary regression*
- Pedro H. C. Sant'Anna: *Testing Instrument Validity in Marginal Treatment Effects Models*

Departure for Saint-Malo: 16:35

List of posters (alphabetical order)

1. Facundo Argañaraz
   - *Necessary and Sufficient Conditions for Existence of Locally Robust/Orthogonal Moments For Inference*
2. Mercedes Conde-Amboage
   - *Testing quantile regression models with censored data and high-dimensional covariates*
3. Gilles Crommen
   - *A Gaussian model for survival data subject to dependent censoring and confounding*
4. Heishiro Kanagawa
   - *When does kernel Stein discrepancy detect (non)convergence of moments?*
5. Adrián Lago
   - *The two-sample problem under random truncation*
6. Hassan Maissoro
   - *Learning the Smoothness of Weakly Dependent Functional Times Series*
7. Andrea Meilán-Vila
   - *Testing a parametric circular regression function*
8. Bojana Milošević
   - *Independence tests for randomly censored data: novel proposal and the review of recent developments*
9. Luis-Alberto Rodríguez
   - *A uniform kernel trick for high-dimensional two-sample problems*
10. Antonin Schrab
    - *KSD Aggregated Goodness-of-fit Tests*
11. Jeremy VanderDoes
    - *The Maximally Selected Likelihood Ratio Test In Random Coefficient Models*
12. Jaco Visagie
    - *On the effect of the Kaplan-Meier estimator's assumed tail behaviour on goodness-of-fit testing*
13. Sunny Wang
    - *Adaptive Functional Principal Components Analysis*

Sunday, Sept. 4

IS–7: *9:00 - 10:30* – Chair Maria Dolores Jiménez-Gamero

- Axel Bücher: *Testing for independence in high dimensions based on empirical copulas*
- Dominic Edelmann: *A regression perspective on generalized distance covariance and its application to genome-wide association studies*
- Ivan Kojadinovic: *Open-end monitoring procedures for multivariate observations that can be senstive to all types of changes in the distribution function*

Coffee/Tea

IS–8: *11:00 - 12:30* – Chair Jad Beyhum

- Weichi Wu: *Confidence surfaces for the mean of locally stationary functional time series*
- Miguel A. Delgado: *Chi-squared Goodness-of-fit Tests for Conditional Distributions*
- Juan Carlos Escanciano: *Quadratic Model Checks for High-Dimensional Models*

Lunch and closing discussions

Invited Talks

Abstracts are listed in alphabetical order by the presenting author

Data-driven semi-parametric detection of multiple changes in long-range dependent processes

Jean-Marc Bardet[1,*], Abdellatif Guenaizi

[1] *SAMM, Université Paris 1 Panthéon-Sorbonne, France, bardet@univ-paris1.fr*

Abstract

This talk is devoted to the offline multiple changes detection for long-range dependent processes. The observations are supposed to satisfy a semi-parametric long-range dependent assumption with distinct memory parameters on each stage. A penalized local Whittle contrast is considered for estimating all the parameters, notably the number of changes. Consistency as well as convergence rates are obtained. Monte-Carlo experiments exhibit the accuracy of the estimators. They also show that the estimation of the number of breaks is improved by using a data-driven slope heuristic procedure of choice of the penalization parameter.

Keywords. Long-range dependence, Local Whittle estimation, Change detection.

References

1. Bardet, J.-M. and Guenaizi, A. (2020) Data-driven semi-parametric detection of multiple changes in long-range dependent processes, *Elec. Journ. Stat.*, 14, 3606-3643.

2. Beran, J. (1994) *Statistics for Long-Memory Processes.* Chapman and Hall, New York.

3. Dahlhaus, R. (1989) Efficient parameter estimation for self-similar processes, *Ann. Statist.*, 17, 1749-1766.

4. Doukhan, P., Oppenheim, G. and Taqqu M.S. (Editors) (2003) *Theory and applications of long-range dependence*, Birkhäuser.

5. Fox, R. and Taqqu, M.S. (1986) Large-sample properties of parameter estimates for strongly dependent Gaussian time series. *Ann. Statist.* 14, 517-532.

6. Giraitis, L., Robinson P.M. and Samarov, A. (1997) Rate optimal semi-parametric estimation of the memory parameter of the Gaussian time series with long range dependence. *J. Time Ser. Anal.*, 18, 49-61.

7. Giraitis, L. and Surgailis, D. (1990) A central limit theorem for quadratic forms in strongly dependent linear variables and its applications to the asymptotic normality of Whittle estimate. *Prob. Th. and Rel. Field.* 86, 87-104.

8. Giraitis, L. and Taqqu, M.S. (1999) Whittle estimator for finite-variance non-Gaussian time series with long memory. *Ann. Statist.* 27, 178-203.

9. Robinson, P.M. (1995) Gaussian semiparametric estimation of long range dependence. *Ann. Statist.*, 23, 1630-1661.

*Presenting author

Semiparametric estimation of McKean-Vlasov SDEs

Denis Belomestny[1,*], Vytautė Pilipauskaitė[2] & Mark Podolskij[2]

[1] *Faculty of Mathematics, University of Duisburg-Essen*
[2] *Department of Mathematics, University of Luxembourg*

Abstract

In this talk, we study the problem of semiparametric estimation for a class of McKean-Vlasov stochastic differential equations. Our aim is to estimate the drift coefficient of a MV-SDE based on observations of the corresponding particle system. We propose a semiparametric estimation procedure and derive the rates of convergence for the resulting estimator. We further prove that the obtained rates are essentially optimal in the minimax sense. Also a goodness-of-fit test for detecting interactions between particles will be presented.

Keywords. McKean-Vlasov stochastic differential equations, particle system, interaction, goodness-of-fit

References

1. Belomestny, Denis, Vytautė Pilipauskaitė, and Mark Podolskij. "Semiparametric estimation of McKean-Vlasov SDEs." arXiv preprint arXiv:2107.00539 (2021).

*Presenting author

Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility

Tom Berrett[1,*] & Richard Samworth[2]

[1] *University of Warwick, tom.berrett@warwick.ac.uk*
[2] *University of Cambridge, r.samworth@statslab.cam.ac.uk*

Abstract

Given a set of incomplete observations, we study the nonparametric problem of testing whether data are Missing Completely At Random (MCAR). Our first contribution is to characterise precisely the set of alternatives that can be distinguished from the MCAR null hypothesis. This reveals interesting and novel links to the theory of Fréchet classes (in particular, compatible distributions) and linear programming, that allow us to propose MCAR tests that are consistent against all detectable alternatives. We define an incompatibility index as a natural measure of ease of detectability, establish its key properties, and show how it can be computed exactly in some cases and bounded in others. Moreover, we prove that our tests can attain the minimax separation rate according to this measure, up to logarithmic factors. Our methodology does not require any complete cases to be effective, and is available in the `R` package `MCARtest`.

Keywords. Missing data, minimax hypothesis testing, compatibility

References

1. Berrett, T. B. and Samworth, R. J. (2022). Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility. *Preprint, available at* `arXiv:2205.08627`.

*Presenting author

Testing for error invariance in separable instrumental variable models

Jad Beyhum[1,*], Jean-pierre Florens[2], Elia Lapenta[3] & Ingrid Van Keilegom[4]

[1] *CREST, ENSAI, jad.beyhum@gmail.com*
[2] *Toulouse School of Economics, jean-pierre.florens@tse-fr.eu*
[3] *CREST, ENSAE, elia.lapenta@gmail.com*
[4] *ORSTAT, KU Leuven, ingrid.vankeilgom@kuleuven.be*

Abstract

The hypothesis of error invariance is central to the instrumental variable literature. It means that the error term of the model is the same across all potential outcomes. In other words, this assumption means that treatment effects are constant across all subjects. It allows to interpret instrumental variable estimates as average treatment effects over the whole population of the study. When this assumption does not hold, the bias of instrumental variable estimators can be larger than that of naive estimators not taking confounding into account. This paper develops two tests for the assumption of error invariance when the treatment suffers from confounding, an instrumental variable is available and the model is separable. The first test assumes that the potential outcomes are linear in the regressors and is computationally simple. The second test is nonparametric and relies on Tikhonov regularization. The treatment can be either discrete or continuous. We show that the tests have asymptotically correct level and asymptotic power equal to one against a range of alternatives. Simulations demonstrate that the proposed tests attain excellent finite sample performances. The methodology is also applied to a model of demand in a fish market.

Keywords. Instrumental variables, Hypothesis testing, Separable model, Nonparametric statistics

[*] Presenting author

Testing for independence in high dimensions based on empirical copulas

Axel Bücher[1,*], Cambyse Pakzad[2]

[1] *Heinrich-Heine-Universität Düsseldorf, axel.buecher@hhu.de*
[2] *Université Côte d'Azur, cambyse.pakzad@univ-cotedazur.fr*

Abstract

Testing for pairwise independence for the case where the number of variables may be of the same size or even larger than the sample size has received increasing attention in the recent years (Leung and Drton, 2018; Han, Chen, Liu, 2017). We contribute to this branch of the literature by considering tests that allow to detect higher-order dependencies. The proposed methods are based on connecting the problem to copulas and making use of the Moebius transformation of the empirical copula process; an approach that has already been used successfully for the case where the number of variables is fixed (Genest and Rémillard, 2004). Based on a martingale central limit theorem, it is shown that respective test statistics converge to the standard normal distribution, allowing for straightforward definition of critical values. The results are illustrated by a Monte Carlo simulation study.

Keywords. Empirical copula process, high dimensional statistics, higher order dependence, Moebius transform, rank based inference.

References

1. Bücher, A. and Cambyse, P. (2022). Testing for independence in high dimensions based on empirical copulas. *ArXiv preprint*, arXiv:2204.01803.

2. Genest, C. and Rémillard, B. (2004). Tests of independence and randomness based on the empirical copula process. *Test* 13(2):335–370.

3. Han, F., Chen, S. and Liu, H. (2017). Distribution-free tests of independence in high dimensions. *Biometrika* 104(4):813–828.

4. Leung, D. and Drton, M. (2018). Testing independence in high dimensions with sums of rank correlations. *Ann. Statist.* 46(1):280–307.

*Presenting author

Chi-squared Goodness-of-fit Tests for Conditional Distributions.

Miguel A. Delgado[1,*] & Julius Vainora[2]

[1] *Universidad Carlos III de Madrid*
[2] *University of Cambridge*

Abstract

We propose Pearson-type goodness-of-fit tests to check for continuous conditional distributions specification. Each observation in the sample is simultaneously cross-classified according to the explanatory variables vector and the Rosenblatt transform of the dependent variable using the specified model. The resulting contingency table forms a basis for the proposed tests. As in the classical case for marginal distributions, the Pearson statistic is identical to the Wald and Lagrange multiplier statistics based on the grouped data likelihood, which are in turn asymptotically equivalent to the likelihood ratio statistic under the null hypothesis. We also provide a Chernoff-Lehmann result for the Pearson statistic using the raw data maximum likelihood estimator, which forms a basis to show that the corresponding Wald statistic is asymptotically distributed as a chi-squared with degrees of freedom invariant to the number of parameters in the model. The asymptotic distribution of the statistics does not change when the explanatory variables classification is data dependent. The finite sample properties of the tests are examined by means of Monte Carlo experiments.

Keywords. Conditional distribution specification testing; Rosenblatt transform; Chi-square tests; Trinity of tests.

*Presenting author

A general maximal projection approach to uniformity testing on the hypersphere

Bruno Ebner[1]

[1] *Karlsruhe Institute of Technology (KIT), Institute of Stochastics, Englerstr. 2, 76128 Karlsruhe, Germany, email: Bruno.Ebner@kit.edu*

Abstract

We propose a novel approach to uniformity testing on the d-dimensional unit hypersphere $\mathcal{S}^{d-1}$ based on maximal projections. This approach gives a unifying view on the classical uniformity tests of Rayleigh and Bingham, as well as links to measures of multivariate skewness and kurtosis. We derive the limiting distribution under the null hypothesis using limit theorems for Banach space valued stochastic processes and present strategies to simulate the limiting processes by applying results on spherical harmonics theory. The behavior under contiguous and fixed alternatives is examined and consistency of the testing procedure is shown for some classes of alternatives. The theoretical findings and empirical powers of the procedures are evaluated in a broad competitive Monte Carlo simulation study.

Keywords. uniformity tests, maximal projections, directional data, stochastic processes in Banach spaces, contiguous alternatives, Monte Carlo simulations

A regression perspective on generalized distance covariance and its application to genome-wide association studies

Dominic Edelmann[1,2,*], Fernando Castro-Prado[3,4], Jelle Goeman[3,4], Wenceslao González-Manteiga[3], Javier Costas [4] & David R. Penas[3]

[1] *German Cancer Research Center, Heidelberg, Germany, dominic.edelmann@dkfz.de.*
[2] *National Center for Tumor Diseases, Heidelberg, Germany.*
[3] *University of Santiago de Compostela, Spain.*
[4] *Health Research Institute of Santiago de Compostela, Spain.*

Abstract

In a seminal paper, Sejdinovic, et al. showed the equivalence of the Hilbert-Schmidt Independence Criterion (HSIC) and a generalization of distance covariance. In this talk, the two notions of dependence are unified with a third prominent concept for independence testing, the "global test" introduced by Goeman, et al. The new viewpoint provides novel in-sights into all three test traditions, as well as a unified overall view of the way all three tests contrast with classical association tests. Moreover, a regression perspective on certain versions of HSIC and generalized distance covariance is obtained, allowing such tests to be used with nuisance covariates or for survival data. Subsequently, we develop distance covariance methods for genome-wide association studies. First, we show how distance covariance can be used for testing between two single nucleotide polymorphisms (SNPs). Second, we develop a family of tests for the association of single nucleotide polymorphisms (SNPs) with quantitative responses. Using the equivalence to the global test framework, we show that certain versions of distance covariance correspond to locally most powerful tests for specific statistical models leading to a theoretically based understanding in which situations these tests perform well. As an example, one version of distance covariance can be interpreted as the locally most powerful test in a dominant-recessive model of genetic association. Closed form expressions for the distributions of the test statistics and corresponding estimators are derived. The performance of the approach is investigated in various simulation studies and a real world example. Extensions to survival data and testing groups of SNPs are discussed.

Keywords. distance covariance, global test, Hilbert-Schmidt Independence Criterion, genome-wide association studies

References

1. Edelmann, D., Goeman, J. J., (2022). A Regression Perspective on Generalized Distance Covariance and the Hilbert-Schmidt Independence Criterion. *Statistical Science*, to appear.
2. Goeman, J. J., Van De Geer, S. A., De Kort, F., Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1), 93–99.
3. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Internat. Conf. Algorithmic Learning Theory, 63-77.*
4. Székely, G. J., Rizzo, M. L., Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann.Statist.*, 35(6), 2769–2794.
5. Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist., 2263-2291.*

*Presenting author

Quadratic Model Checks for High-Dimensional Models

Juan Carlos Escanciano[1]

[1] *Universidad Carlos III de Madrid, jescanci@eco.uc3m.es*

Abstract

This paper proposes a unified approach to model checking based on quadratic tests statistics. Our tests address two important limitations of existing model checks: (i) the lack of robustness to estimated nuisance parameters; and (ii) the limited interpretability in nonparametric settings. These two limitations are particularly important for modern machine learning applications. We generalize the classical model checks literature to a high-dimensional setting and the kernel machine literature to a general supervised setting. The proposed tests are robust to estimated nuisance parameters, in a way that their asymptotic distributions do not depend on that of the estimator or the estimator being asymptotically linear. We provide new tools for the interpretability of our omnibus test, including an asymptotic local power analysis, a new measure of degress of freedom, and a Gaussian Process interpretation. The proposed tests are implemented with the assistance of simple multiplier bootstrap. Several examples illustrate the wide applicability of the procedures. A Monte Carlo experiment shows that the new method presents more accurate size than competing methods and improves the interpretability of empirical tests results.

Keywords. Model checks; High dimensional models; Omnibus tests; Multiplier bootstrap.

References

1. Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics,* 20**,** 105-134.

2. Bravo, F., Escanciano, J.C. and I. Van Keilegom (2020). Two-Step Semiparametric Empirical Likelihood Inference. *The Annals of Statistics*, 48(1), 1-26.

3. Chernozhukov, V., Escanciano, J.C., H. Ichimura, W.K. Newey , and J. Robins (2022). Locally Robust Semiparametric Estimation. Forthcoming in *Econometrica*.

4. Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics,* 25, 613-641.

Some new tests of independence for circular data

Eduardo García-Portugués[1,*], Pierre Lafaye de Micheaux[2,3,4],
Simos G. Meintanis[5,6] & Thomas Verdebout[7,8]

[1] *Department of Statistics, Carlos III University of Madrid; edgarcia@est-econ.uc3m.es.*
[2] *School of Mathematics and Statistics, UNSW Sydney; lafaye@unsw.edu.au.*
[3] *Desbrest Institute of Epidemiology and Public Health, Université de Montpellier.*
[4] *AMIS, Université Paul Valéry Montpellier 3.*
[5] *Department of Economics, National and Kapodistrian University of Athens; simosmei@econ.uoa.gr.*
[6] *Unit for Pure and Applied Analytics, North-West University.*
[7] *Département de Mathématique, Université libre de Bruxelles; tverdebo@ulb.ac.be.*
[8] *ECARES, Université libre de Bruxelles.*

Abstract

We introduce nonparametric tests of independence for bivariate circular data based on trigonometric moments. Our contributions lie in (*i*) proposing nonparametric tests that are locally and asymptotically optimal against bivariate cosine von Mises alternatives and (*ii*) extending these tests, via the empirical characteristic function, to obtain consistent tests against broader sets of alternatives, eventually being omnibus. In particular, one of such omnibus tests is a circular version of the celebrated distance-covariance test. We thus provide a collection of trigonometric-based tests of varying generality and known optimalities. The large-sample behavior of the tests under the null and alternative hypotheses are obtained, while simulations show that the new tests are competitive against previous proposals. Two data applications in astronomy and forest science illustrate the usage of the tests.

The contribution is based on the cited preprint.

Keywords. Characteristic function, directional data, independence, trigonometric moments.

References

1. García-Portugués, E., Lafaye de Micheaux, P., Meintanis, S. G., and Verdebout, T. (2021). Nonparametric tests of independence for circular data based on trigonometric moments. *Under review*. Preprint available at arXiv:2104.14620.

*Presenting author

Testing specification in some complex models using correlation distances

Wenceslao González-Manteiga[1,2,*], Laura Freijeiro-González[1], Alejandra López-Pérez[1] & Manuel Febrero-Bande[1]

[1] *Centre for Mathematical Research and Technology Transfer of Galicia (CITMAga). Department of Statistics, Mathematical Analysis and Optimization. Universidade de Santiago de Compostela, Santiago de Compostela, Spain*
[2] *Presenting author; Email: wenceslao.gonzalez@usc.es*

Abstract

Specification tests in a regression framework refer to the assertion or rejection of some assumption about the model structure. These cover well studied topics as goodness-of-fit or significance tests. In terms of goodness-of-fit tests, these focus on testing if the model belongs to some parametric family, measuring the departure from this hypothesis. Over the years, different methodologies have been developed to perform these. A complete review of this topic can be found in González-Manteiga and Crujeiras (2013, TEST). Once a certain structure of the model is assumed, significance tests can be applied to determine if the model is well specified in terms of covariates: testing if covariates effects are relevant or some can be excluded from the model. Recently, the innovative and well-known distance correlation (Székely et al. (2007, Annals) and Székely and Rizzo. (2017, ARSIA)) as well as its variants: martingale difference divergence (Shao and Zhang (2014, JASA)) and conditional distance correlation (Wang et al. (2015, JASA)), are applied in specification testing. Novel tests for covariates selection in the functional concurrent model jointly with new ones for testing the specification of diffusion models are developed making use of these ideas. As a result, no model estimation is needed and the curse of dimensionality is overcome, respectively. Statistics behaviour is studied in both cases and their good performance is showed by means of complete simulation studies. In practice, wild bootstrap schemes are provided to approximate its p-value. Eventually, they are applied to real datasets.

Keywords. Concurrent model, conditional correlation distance, diffusion model, distance correlation, martingale difference divergence, specification test.

*Authors acknowledge the support from Projects MTM2016-76969-P and PID2020-116587GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" and the Competitive Reference Groups 2021–2024 (ED431C 2021/24) from the Xunta de Galicia through the ERDF. The Galician Supercomputing Center (CESGA) is also acknowledged for providing computational resources for simulation studies.

A Kernel Stein Test for Comparing Latent Variable Models

Heishiro Kanagawa[1], Wittawat Jitkrittum[2,3], Lester Mackey[4], Kenji Fukumizu[5], Arthur Gretton[1,*]

[1] *Gatsby Computational Neuroscience Unit, UCL;* [2] *Max Planck Institute for Intelligent Systems, Tübingen, Germany;* [3] *Google Research, USA;* [4] *Microsoft Research, Cambridge, MA, USA;* [5] *The Institute of Statistical Mathematics, Tokyo, Japan*

Abstract

I will describe a kernel-based nonparametric test of relative goodness of fit, where the goal is to compare two models, both of which may have unobserved latent variables, such that the marginal distribution of the observed variables is intractable. Given the premise that "all models are wrong," the goal of the test is to determine whether one model significantly outperforms the other in respect of a reference data sample. The test generalises earlier kernel Stein discrepancy (KSD) tests to the case of latent variable models, a much more general class than the fully observed models treated previously. The new test, with a properly calibrated threshold, has a well-controlled type-I error. In the case of models with low-dimensional latent structure and high-dimensional observations, our test significantly outperforms the relative maximum mean discrepancy test, which is based on samples from the models, and does not exploit the latent structure. I will illustrate the test on probabilistic topic models of arXiv articles.

Keywords. Kernel Methods, Stein's Method, Nonparametric Goodness-of-fit Test

*Presenting author

A Measure-Transportation-Based GOF Test for Directional Data

Marc Hallin[1,*], Hang Liu[2] & Thomas Verdebout[3]

[1] *Université libre de Bruxelles, mhallin@ulb.ac.be*
[2] *University of Science and Technology of China, hliu01@ustc.edu.cn*
[3] *Université libre de Bruxelles, tverdebo@ulb.ac.be*

Abstract

Based on measure transportation results, we propose new concepts of distribution and quantile functions on the hypersphere. The empirical versions of our distribution functions enjoy the expected Glivenko-Cantelli property of traditional distribution functions and yield fully distribution-free concepts of ranks and signs. They also yield "universally consistent" GOF tests outperforming the projected Cramér-von Mises, Anderson-Darling, and Rothman procedures recently proposed in the literature.

Keywords. Directional data, Measure transportation, Consistent GOF tests.

[*] Presenting author

Change point detection in high dimensional data with $U-$statistics

Lajos Horváth[1]

[1] *University of Utah*

Abstract

We consider the problem of detecting distributional changes in a sequence of high dimensional data. Our proposed methods are nonparametric, suitable for either continuous or discrete data, and are based on weighted cumulative sums of non–degenerate U–statistics. We establish the asymptotic distribution of our proposed test statistics separately in cases of either weakly dependent or strongly dependent coordinates, and provide sufficient conditions for consistency of the proposed test procedures under a general fixed alternative with one change point. The high–dimensional asymptotic setting we consider requires only that $\min\{N, d\} \to \infty$, allowing for substantial flexibility in applications. We further assess finite sample performance of the test procedures through Monte Carlo studies, and conclude with two applications to Twitter data concerning the mentions of U.S. Governors and the frequency of social justice keywords.

Testing Poissonity of a large number of populations

M. Dolores Jiménez-Gamero[1,*] & Jacobo de-Uña-Álvarez[2]

[1] *University of Seville, Spain,* dolores@us.es
[2] *University of Vigo, Spain,* jacobo@uvigo.es

Abstract

Univariate count data appear in many real life situations and the Poisson distribution is frequently used to model this kind of data. Testing the goodness-of-fit of given observations with a probabilistic model is a crucial aspect of data analysis. Because of these reasons, a number of authors have proposed tests for the Poisson law. Most papers on this issue deal with testing Poissonity for a single sample, and the properties of the proposed procedures are studied as the sample size increases. Here we consider the problem of simultaneously testing Poissonity for k samples, where k can increase with the sample sizes. Moreover, k will be allowed to be even larger than the sample sizes. This is important, for instance, in applications with high-dimensional data, as those arising from DNA sequencing experiments. The cases of independent samples and weakly dependent samples are both of them studied. Specifically, a modification of the test statistic in Baringhaus and Henze (1992) is calculated at each sample, and the resulting values are conveniently combined to get a global test statistic which, under the null hypothesis, is asymptotically free distributed, not relying on resampling or Monte-Carlo methods to obtain critical values. Consistency results are also derived. As a real data illustration, we test Poissonity for the read counts along a DNA region that were obtained in a targeted resequencing single-end experiment

Keywords. Goodness-of-fit, count data, Poisson law, k samples, high-dimensional data.

References

1. Baringhaus, L., and Henze, N. (1992), A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statistics & Probability Letters*, 13(4):269–274.

*Presenting author

Open-end monitoring procedures for multivariate observations that can be senstive to all types of changes in the distribution function

Mark Holmes[1], Ivan Kojadinovic[2,*] & Alex Verhoijsen[1,2]

[1] *School of Mathematics & Statistics, The University of Melbourne, Parkville, VIC 3010, Australia*

[2] *CNRS / Université de Pau et des Pays de l'Adour / E2S UPPA, Laboratoire de mathématiques et applications IPRA, UMR 5142, B.P. 1155, 64013 Pau Cedex, France*

Abstract

We propose nonparametric open-end sequential testing procedures based on the empirical distribution function that can detect all types of changes in the contemporary distribution function of multivariate observations. Their asymptotic properties are theoretically investigated under stationarity and under alternatives to stationarity. Monte Carlo experiments reveal their good finite-sample behavior in the case of continuous univariate observations. A short data example concludes the presentation.

Keywords. asymptotic results, change-point detection, open-end monitoring, sequential testing, theoretical quantile estimation.

References

1. J. Gosmann, T. Kley and H. Dette (2021), A new approach for open-end sequential change point monitoring, *Journal of the Time Series Analysis* 42, 63–84.

2. M. Holmes, I. Kojadinovic and A. Verhoijsen (2022), Multi-purpose open-end monitoring procedures for multivariate observations based on the empirical distribution function, *arXiv:2201.10311*, 33 pages.

*Presenting author

Partly Linear Instrumental Variables Regression Without Smoothing on the Instruments

Jean-Pierre Florens[1] & Elia Lapenta[2,*]

[1] *Toulouse School of Economics (jean-pierre.florens@tse-fr.eu)*
[2] *CREST and ENSAE (elia.lapenta@ensae.fr)*

Abstract

We propose a new estimation method for partly linear models identified by Instrumental Variables (IVs). The estimation is based on a class of Generically Comprehensively Revealing functions. Compared to methods available in the literature that smooth on the IVs, our procedure does not smooth on the instruments and thus requires the selection of fewer smoothing parameters. Furthermore, it does not suffer from a curse of dimensionality on the IVs. To deal with the ill-posedness of the inverse problem, we use a Landweber-Friedman regularization scheme. This is a simple iterative method that does not require the inversion of a large matrix whose dimension increases with the sample size. We show that our procedure is equivalent to a classical estimation method that smooths on the IVs but keeps the bandwidth fixed as the sample size increases. We obtain convergence rates for the estimator of the nonparametric part of the model and the asymptotic normality of the estimator of the parametric components. We finally study the implementation of our procedure and propose a data-driven selection of the regularization parameter.

Keywords. Endogeneity, Ill-Posed inverse problem, Instrumental variables, Partly-linear model, Landweber-Friedman regularization.

References

1. Carrasco, M., Florens, J.-P., and Renault, E. (2007). Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization. Chapter 77 in *The Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, Vol. 6: 5633-–5751.

*Presenting author

Tests for the stable paretian hypothesis for i.i.d. data and for innovations in multivariate GARCH models

Simos G. Meintanis[1,*], John P. Nolan[2], Charl Pretorius[3] & Zhou Zhou[4]

[1] *National and Kapodistrian University of Athens, Athens, Greece*
North–West University, Potchefstroom, South Africa
simosmei@econ.uoa.gr
[2] *American University, Washington DC, USA; jpnolan@american.edu*
[3] *North–West University, Potchefstroom, South Africa*
charl.pretorius@nwu.ac.za
[4] *University of Toronto, Ontario, Canada; zhouzhou@utoronto.ca*

Abstract

We propose specification tests for the stable Paretian hypothesis in arbitrary dimension. The methods are based on the empirical characteristic function and can be readily implemented with i.i.d. data as well as with observations from GARCH models. Asymptotic properties of the proposed procedures are investigated, while the finite–sample properties are illustrated by means of an extensive Monte Carlo study. The procedures are also applied on real data from the financial markets.

Keywords. Goodness–of–fit, Empirical characteristic function, Heavy–tailed data

[*] Presenting author

Semiparametric transformation models: mean and boundary regression

Natalie Neumeyer[1]

[1] *University of Hamburg, natalie.neumeyer@uni-hamburg.de*

Abstract

In transformation regression models a sample from (X, Y) is observed and a transformation Λ is applied to the response before fitting a regression model:

$$\Lambda(Y) = m(X) + \varepsilon.$$

The transformation Λ is chosen data-dependently, often with the aim to obtain independence between the covariates X and the errors ε. We give an overview over recent results for mean regression models (i.e. $E[\varepsilon] = 0$) with parametric transformation $\Lambda \in \{\Lambda_\vartheta \mid \vartheta \in \Theta\}$ and nonparametric regression function m, in particular goodness-of-fit tests for the transformation function. We then consider estimation of the transformation parameter in boundary regression models (with one-sided errors $\varepsilon \leq 0$). We discuss some open problems like goodness-of-fit tests for the transformation function in boundary regression.

Keywords. Boundary curve estimation, goodness-of-fit, minimum distance to independence, nonparametric regression, residual-based empirical processes

Density model checks via the lack-of-fitness

Valentin Patilea[1,*] François Portier[2,†]

[1] *CREST, Ensai; valentin.patilea@ensai.fr*
[2] *CREST, Ensai; françois.portier@ensai.fr*

Abstract

Olkin and Spiegelman (1987) introduced a semiparametric estimator of the density defined as a mixture between the maximum likelihood estimator and the kernel density estimator. Mazo and Portier (2021) pointed out that, in that context, the mixture weight associated with the parametric density provides a measure for the goodness-of-fit of the parametric model. They call this mixture weight the *fitness coefficient*, and estimate it by maximum likelihood. Under mild conditions on the, possibly multivariate, density model, Mazo and Portier show that the fitness coefficient converges in probability to 1 if the parametric density model is correct, and zero otherwise. In this contribution, we introduce and investigate the convergence in distribution of the *lack-of-fitness* statistic, defined as the suitably normalized difference between 1 and the fitness coefficient. It is shown that, when the parametric density model is correct, the lack-of-fitness statistic converges in distribution to the positive part of a standard Gaussian variable, regardless the fixed dimension of the i.i.d. observations. Moreover, the new test statistic detects local alternatives at the parametric rate.

Keywords. Concavity, Goodness-of-fit, Leave-one-out density estimator, Pivotalness

References

1. Claeskens, G. and N. L. Hjort (2004). Goodness of fit via non-parametric likelihood ratios. *Scandinavian Journal of Statistics*, 31(4), 487–513.

2. Hjort, N. L. and D. Pollard (2011). Asymptotics for minimisers of convex processes. arXiv preprint arXiv:1107.3806.

3. Mazo, G. and F. Portier (2021). Parametric versus nonparametric: The fitness coefficient. *Scandinavian Journal of Statistics*, 48(4), 1344–1383.

4. Olkin, I. and C. H. Spiegelman (1987). A semiparametric approach to density estimation. *Journal of the American Statistical Association*, 82(399), 858–865.

*V. Patilea acknowledges support from the grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P4-ID-PCE-2020-1112, within PNCDI III.

†Presenting author

High-dimensional changepoint estimation with heterogeneous missingness

Bertille Follain[1], Tengyao Wang[2] & Richard J. Samworth[3,*]

[1] *Statistical Laboratory, University of Cambridge & Ecole Normale Supérieure, PSL Research University, INRIA; bertille.follain@inria.fr*
[2] *Department of Statistics, London School of Economics and Political Science & Department of Statistical Science, University College London; t.wang59@lse.ac.uk*
[3] *Statistical Laboratory, University of Cambridge; r.samworth@statslab.cam.ac.uk*

Abstract

We propose a new method for changepoint estimation in partially-observed, high-dimensional time series that undergo a simultaneous change in mean in a sparse subset of coordinates. Our first methodological contribution is to introduce a 'MissCUSUM' transformation (a generalisation of the popular Cumulative Sum statistics), that captures the interaction between the signal strength and the level of missingness in each coordinate. In order to borrow strength across the coordinates, we propose to project these MissCUSUM statistics along a direction found as the solution to a penalised optimisation problem tailored to the specific sparsity structure. The changepoint can then be estimated as the location of the peak of the absolute value of the projected univariate series. In a model that allows different missingness probabilities in different component series, we identify that the key interaction between the missingness and the signal is a weighted sum of squares of the signal change in each coordinate, with weights given by the observation probabilities. More specifically, we prove that the angle between the estimated and oracle projection directions, as well as the changepoint location error, are controlled with high probability by the sum of two terms, both involving this weighted sum of squares, and representing the error incurred due to noise and the error due to missingness respectively. A lower bound confirms that our changepoint estimator, which we call 'MissInspect' and which is available in the `R` package `InspectChangepoint`, is optimal up to a logarithmic factor. The striking effectiveness of the MissInspect methodology is further demonstrated both on simulated data, and on an oceanographic data set covering the Neogene period.

Keywords. changepoint estimation; missing data; high-dimensional data; segmentation; sparsity

*Presenting author

Testing Instrument Validity in Marginal Treatment Effects Models

Minghai Mao[1] , Pedro H. C. Sant'Anna[2,*], Xiaojun Song[3]

[1] *Liaoning University, maominghai@lnu.edu.cn*
[2] *Microsoft and Vanderbilt University, psantanna@microsoft.com*
[3] *Peking University, sxj@gsm.pku.edu.cn*

Abstract

In this paper, we develop falsification tests for the validity of marginal treatment effects (MTE) model. We first derive the sharp, testable implications of nonparametric and semiparametric MTE models. The nonparametric MTE model requires a continuous instrument but avoids functional form restrictions, whereas the semiparametric model relies on additional modelling assumptions but can be used with discrete instruments. These strongest testable implications are characterized by monotonicity and index sufficiency restrictions on the (conditional) joint distribution of the outcome and treatment on the propensity score. Our test statistics reformulate these testable implications as shape restrictions, and build on the integrated conditional moment approach to assess their plausibility. To establish the validity of our tests, we leverage recent results on the directional differentiability of the least concave majorant operator, and on a numerical bootstrap procedure that remains valid even when one cannot rely on the classical functional delta method for the bootstrap. The finite sample properties of the proposed tests are examined by means of Monte Carlo simulations, and an application on returns to schooling.

Keywords. Causal inference, endogeneity, Hadamard directional differentiability, instrumental variable, least concave majorant, treatment effects.

*Presenting author

Copula-Based Estimation of Health Inequality Measures With an Application to COVID-19

Taoufik Bouezmarni[1], Mohamed Doukali[1,2] & Abderrahim Taamouti[3,*]

[1] *Université de Sherbrooke, Email: Taoufik.Bouezmarni@USherbrooke.ca*
[2] *McGill University, Email: mohamed.doukali@mail.mcgill.ca*
[3] *University of Liverpool Management School, Email: abderrahim.taamouti@liverpool.ac.uk*

Abstract

COVID-19 has created an unprecedented global health crisis that caused millions of infections and deaths worldwide. Many argue that pre-existing social inequalities have led to inequalities in COVID-19's infection and death rates across social classes, with the most-deprived classes are worst hit. In this paper, we derive semi/non-parametric estimators of Health Concentration Curve [hereafter CH] and Gini coefficient for health distribution [hereafter Gini health coefficient] to help quantify inequalities in COVID-19 infections and deaths, and identify the social classes that are most at risk of infection and dying from the virus. We first express CH in terms of copula function, which we use to build our semi/non-parametric estimators. For the semi-parametric estimator, a parametric copula is used to model the dependence between health and socioeconomic variable. The copula function is estimated using maximum pseudo-likelihood estimator after replacing the cumulative distribution of health variable by its empirical analogue. For the non-parametric estimator, we replace the copula function by a Bernstein copula estimator. We establish the consistency and the asymptotic normality of CH's estimators. Thereafter, we use the above estimators of CH to derive copula-based estimators of Gini health coefficient. A Monte-Carlo simulation exercise based on several data-generating processes and sample sizes shows that the semiparametric estimator outperforms the smoothed nonparametric estimator, and that the latter does better than the empirical estimator in terms of Integrated Mean Squared Error. Finally, we run an extensive empirical study to illustrate the importance of CH and Gini health coefficient estimators for investigating inequality in COVID-19's infections and deaths in the U.S. The results show that inequality across U.S. states' socioeconomic variables like income/poverty and race/ethnicity might explain the observed inequalities in the U.S. COVID-19's infections and deaths.

Keywords. Health concentration curve, Gini health coefficient, inequality, copula, semi/non-parametric estimators, COVID-19 infections and deaths.

*Presenting author

Statistical decisions for variable selection

Cristina Butucea[1], Enno Mammen[2], Mohamed Ndaoud[3] & Alexandre Tsybakov[1,*]

[1] *CREST, ENSAE, IP Paris, cristina.butucea@ensae.fr, alexandre.tsybakov@ensae.fr*
[2] *Universität Heidelberg, mammen@math.uni-heidelberg.de*
[3] *ESSEC, ndaoudm@gmail.com*

Abstract

For the core variable selection problem under the Hamming loss, we derive a non-asymptotic exact minimax selector over the class of all s-sparse vectors, which is also the Bayes selector with respect to the uniform prior. While this optimal selector is, in general, not realizable in polynomial time, we show that its tractable counterpart (the scan selector) attains the minimax expected Hamming risk to within factor 2. Moreover it is non-asymptotically exact minimax under the probability of wrong recovery criterion. In the monotone likelihood ratio framework, we establish explicit lower bounds on the minimax risk and provide its tight characterization in terms of the best separable selector risk. As a consequence, we obtain sharp necessary and sufficient conditions of exact and almost full recovery in the location model with light tail distributions and in the problem of group variable selection under Gaussian noise.

Keywords. variable selection, minimax selector, necessary and sufficient conditions of exact and almost full recovery

[*] Presenting author

Instrumental variable quantile regression under random right censoring

Jad Beyhum[1], Lorenzo Tedesco[2] & Ingrid Van Keilegom[3,*]

[1] *CREST, ENSAI & KU Leuven, Belgium, jad.beyhum@gmail.com*
[2] *KU Leuven, Belgium, lorenzo.tedesco@kuleuven.be*
[3] *KU Leuven, Belgium, ingrid.vankeilegom@kuleuven.be*

Abstract

This paper studies a semiparametric quantile regression model with endogenous variables and random right censoring. The endogeneity issue is solved using instrumental variables. It is assumed that the structural quantile of the logarithm of the outcome variable is linear in the covariates and censoring is independent. The regressors and instruments can be either continuous or discrete. The specification generates a continuum of equations of which the quantile regression coefficients are a solution. Identification is obtained when this system of equations has a unique solution. Our estimation procedure solves an empirical analogue of the system of equations. We derive conditions under which the estimator is asymptotically normal and prove the validity of a bootstrap procedure for inference. The finite sample performance of the approach is evaluated through numerical simulations. The method is illustrated by an application to the national Job Training Partnership Act study.

Keywords. Censoring, endogeneity, instrumental variable, quantile regression, semiparametric regression, survival analysis.

References

1. Beyhum, J., Tedesco, L. and Van Keilegom, I. (2022). Instrumental variable quantile regression under random right censoring (submitted).

*Presenting author

Asymptotic power of Sobolev tests for uniformity on hyperspheres

Eduardo García-Portugués[1], Davy Paindaveine[2] and Thomas Verdebout[2,*]

[1] *Department of Statistics, University Carlos III, Mardrid, Spain*
[2] *Department of Mathematics and ECARES, ULB, Belgium*

Abstract

One of the most classical problems in multivariate statistics is considered, namely, the problem of testing isotropy, or equivalently, the problem of testing uniformity on the unit hypersphere. Rather than restricting to tests that can detect specific types of alternatives only, we consider the broad class of Sobolev tests. While these tests are known to allow for omnibus testing of uniformity, their non-null behavior and consistency rates, unexpectedly, remain largely unexplored. To improve on this, we thoroughly study the local asymptotic powers of Sobolev tests under the most classical alternatives to uniformity, namely, under rotationally symmetric alternatives. We show in particular that the consistency rate of Sobolev tests does not only depend on the coefficients defining these tests but also on the derivatives of the underlying angular function at zero.

Keywords. Directional data, testing for uniformity, local powers, Sobolev tests.

References

1. Cutting, C., Paindaveine, D., and Verdebout, Th. (2017). Testing uniformity on high-dimensional spheres against monotone rotationally symmetric alternatives. *Annals of Statistics*, 45(3):1024–1058.

2. García-Portugués, E., Paindaveine, D. and Verdebout, Th. (2022). On the asymptotic power of Sobolev tests of uniformity against rotationally symmetric alternatives, *Submitted*.

*Presenting author

Confidence surfaces for the mean of locally stationary functional time series

Holger Dette[1,] and Weichi Wu[2,*]

[1]*Department of Mathematics, Ruhr University Bochum, Germany, holger.dette@ruhr-uni-bochum.de*
[2]*Center for Statistical Science, Tsinghua University, China, wuweichi@mail.tsinghua.edu.cn*

Abstract

The problem of constructing a simultaneous confidence surface for the 2-dimensional mean function of a non-stationary functional time series is challenging as these bands can not be built on classical limit theory for the maximum absolute deviation between an estimate and the time dependent regression function. In this paper we propose new bootstrap methodology to construct such a region. Our approach is based on a Gaussian approximation for the maximum norm of sparse high-dimensional vectors approximating the maximum absolute deviation. The elimination of the zero entries produces (besides the time dependence) additionally dependencies such that "classical" multiplier bootstrap is not applicable. To solve this issue we develop a novel multiplier bootstrap, where blocks of the coordinates of the vectors are multiplied with random variables, which mimic the specific structure between the vectors appearing in the Gaussian approximation. We prove the validity of our approach by asymptotic theory, demonstrate good finite sample properties by means of a simulation study and illustrate its applicability analyzing a data example.

Keywords. locally stationary time series, functional data, confidence surface, Gaussian approximation, multiplier bootstrap.

*Presenting author

Change point inference in high-dimensional regression models under temporal dependence

Haotian Xu[1], Daren Wang[2], Zifeng Zhao[2] & Yi Yu[3,*]

[1] *Université Catholique de Louvain, haotian.xu@uclouvain.be*
[2] *University of Notre Dame, {dwang24, zzhao2@nd.edu}@nd.edu*
[3] *University of Warwick, yi.yu.2@warwick.ac.uk*

Abstract

This paper concerns about the limiting distributions of change point estimators, in a high-dimensional linear regression time series context, where a regression object $\{y_t, X_t\} \in \mathbb{R} \times \mathbb{R}^p$ is observed at every time point $t \in \{1, \ldots, n\}$. At unknown time points, called change points, the regression coefficients change, with the jump sizes measured in the ℓ_2-norm. We provide limiting distributions of the change point estimators under the regimes where the minimal jump size vanishes and remain a constant. We allow for both the covariate and noise sequences to be temporal dependent, in the functional dependence framework, which is the first time seen in the high-dimensional change point inference literature. We show a block-type long-run variance estimator is consistent under the functional dependence, to enable the practicality of our derived limiting distributions. We also present a feast of byproducts of their own interest, including a novel variant of the dynamic programming algorithm to boost the computational efficiency, change point localisation rates under functional dependence and a new Bernstein inequality for data possessing functional dependence.

Keywords. High-dimensional linear regression; Change point inference; Functional dependence.

*Presenting author

Invited Posters

Abstracts are listed in alphabetical order by the presenting author

Necessary and Sufficient Conditions for Existence of Locally Robust/Orthogonal Moments For Inference

Juan Carlos Escanciano[1] & Facundo Argañaraz[2,*]

[1] *Universidad Carlos III de Madrid, jescanci@eco.uc3m.es*
[2] *Universidad Carlos III de Madrid, farganar@eco.uc3m.es*

Abstract

In this paper we provide a necessary and sufficient condition for the existence of Locally Robust (LR)/Orthogonal/Debiased moments for inference on a parameter of interest in a general semiparametric model. Such orthogonal moments have been proved useful with machine learning or high dimensional first steps. The condition is equivalent to the existence of tests with non-trivial local power for the parameter of interest, and it does not require identification of the parameter of interest or its Fisher information to be positive. We illustrate the result with models with unobserved heterogeneity, for which we provide novel necessary and sufficient conditions for existence of LR moments, check whether the condition holds, and when it does, construct new examples of LR moments. We also show that the existence of LR moments can be achieved through sparsity constraints in cases where it does not hold otherwise.

Keywords. Identification; Locally Robust Inference; Unobserved Heterogeneity; Machine Learning; Sparsity.

References

1. Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, forthcoming.

2. Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

3. van der Vaart, A. (1998). Asymptotic Statistics. *Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.*

*Presenting author.

Testing quantile regression models with censored data and high-dimensional covariates

Mercedes Conde-Amboage[1,*], Ingrid Van Keilegom[2] & Wenceslao González-Manteiga[3]

[1] *Department of Statistics, Mathematical Analysis and Optimization. Universidade de Santiago de Compostela. mercedes.amboage@usc.es.*

[2] *Research Centre for Operations Research and Statistics (ORSTAT). KU Leuven. ingrid.vankeilegom@kuleuven.be.*

[3] *Department of Statistics, Mathematical Analysis and Optimization. Universidade de Santiago de Compostela. wenceslao.gonzalez@usc.es.*

Abstract

Quantile regression, introduced by Koenker and Basset (1978), allows a more detailed description of the behaviour of the response variable, adapts to situations under more general conditions of the error distribution (that is, do not require stringent assumptions, such as homoscedasticity or normality) and enjoys properties of robustness.

In particular, quantile regression provides good results when complex data are considered, for instance, when the response variable is right-censored. Along this poster, a new lack-of-fit test for censored quantile regression models with multiple covariates will be presented.

The test is based on the cumulative sum of residuals with respect to unidimensional linear projections of the covariates. The test is then adapting the ideas of Escanciano (2006) to cope with high-dimensional covariates. It will be shown the limit distribution of the empirical process associated with the test statistic. Furthermore, in order to approximate the critical values of the test, a wild bootstrap mechanism is used. In addition, an extensive simulation study and an interesting application to real data will be developed in order to show the behaviour of the new test in practice.

Keywords. Quantile regression, censored data, lack-of-fit test, bootstrap approach, high-dimensional covariates.

References

1. Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6), 1030–1051.

2. Koenker, R., and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

*Presenting author

A Gaussian model for survival data subject to dependent censoring and confounding

Gilles Crommen[1,*], Jad Beyhum[2] & Ingrid Van Keilegom[3]

[1] *ORSTAT, KU Leuven (gilles.crommen@kuleuven.be)*
[2] *CREST, ENSAI (jad.beyhum@gmail.com)*
[3] *ORSTAT, KU Leuven (ingrid.vankeilegom@kuleuven.be)*

Abstract

This paper considers the problem of inferring the causal effect of a variable Z on a survival time T. The error term of the model and Z are correlated, which leads to a confounding issue. Additionally, T is subject to dependent censoring, that is, T is right censored by a censoring time (C) which is dependent on T. In order to tackle the confounding issue, we leverage a control function approach relying on an instrumental variable $\tilde{W}$. It is assumed that T and C follow a joint regression model with bivariate Gaussian error terms with an unspecified covariance matrix, which allows us to handle dependent censoring in a flexible manner. We derive conditions under which the model is identifiable. A two-step estimation procedure is proposed and we show that the resulting estimator is consistent and asymptotically normal. Simulations are used to confirm the validity and finite-sample performance of the estimation procedure. Finally, the proposed method is used to estimate the effectiveness of JTPA training programs on time until employment.

Keywords: Dependent censoring, causal inference, instrumental variable, control function, survival analysis.

References

1. Deresa, N. W., & Van Keilegom, I. (2020). Flexible parametric model for survival data subject to dependent censoring. *Biometrical Journal*, 62 (1), 136–156.

2. Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT press.

3. Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics, 4*, 2111–2245.

*Presenting author

When does kernel Stein discrepancy detect (non)convergence of moments?

Heishiro Kanagawa[1,*] , Lester Mackey[2] & Arthur Gretton[3]

[1] *Gatsby Computational Neuroscience Unit heishiro.kanagawa@gmail.com*
[2] *Microsoft Research lmackey@microsoft.com*
[3] *Gatsby Computational Neuroscience Unit arthur.gretton@gmail.com*

Abstract

The kernel Stein discrepancy (KSD) is a discrepancy measure between probability distributions. The KSD has been used to measure the quality of samples against a distributon defined by density with a intractable normalising constant, a common setting in approximate posterior inference. Gorham and Mackey (2017) showed that the KSD detects the weak convergence (or non-convergence) under suitable conditions. However, the weak convergence does not imply convergence with respect to growing functions, and thus it remains unclear if the KSD diagnostic may be used for evaluating quantities such as mean and variance. This presentation demonstrates conditions whereby the KSD imply convergence of moments.

Keywords. Bayesian inference, MCMC, Reproducing kernel Hilbert spaces, Stein's method.

References

1. Gorham, J. & Mackey, L. (2017). xMeasuring Sample Quality with Kernels. In *Proceedings of The 34th International Conference on Machine Learning*, 1292-1301

*Presenting author

The two-sample problem under random truncation

Adrián Lago[1*], Juan Carlos Pardo-Fernández[2] & Jacobo de Uña-Álvarez[3]

[1] *Universidade de Vigo, Spain. Email: adrian.lago@uvigo.es*
[2] *Universidade de Vigo, Spain. Email: juancp@uvigo.es*
[3] *Universidade de Vigo, Spain. Email: jacobo@uvigo.es*

Abstract

Truncation is a problem that comes up naturally when collecting data, especially in the Survival Analysis framework. In particular, left truncation induces an observational bias, so large event times are oversampled. Proper estimators of the survival function are therefore needed such as the one introduced in Lynden-Bell (1971).

In this work we propose a Kolmogorov-Smirnov-type test to compare the survival functions of two independent populations with left-truncated data. By means of simulations we first show that the proposed test is not distribution-free. Furthermore, by employing the almost-sure representation of the Lynden-Bell estimator given in Chao and Lo (1988), the null asymptotic distribution of the proposed test statistic is obtained. A bootstrap resampling plan, based on Gross and Lai (1996), is designed to approximate the null distribution of the test statistic.

The practical performance of the proposed test is analysed and compared to the log-rank in a simulation study. We show that the new test reaches reasonable power and can outperform the log-rank test under nonproportional hazards scenarios.

Keywords. Truncation, two-sample problem, Kolmogorov-Smirnov, bootstrap, log-rank

References

1. Chao, M.T., and Lo, S.H. (1988) Some representations of the nonparametric maximum likelihood estimators with truncated data. *The Annals of Statistics* 16(2):661-668.

2. Gross, S.T., and Lai, T.L. (1996) Bootstrap methods for truncated and censored data. *Statistica Sinica* 6(3):509-530.

3. Lynden-Bell, D. (1971) A method of allowing for known observational selection in small samples applied to 3RC quasars. *Monthly Notices of the Royal Astronomical Society* 155(1):95-118.

*Presenting author

Learning the Smoothness of Weakly Dependent Functional Times Series

Hassan Maissoro[1,2*] Valentin Patilea[1] Myriam Vimond[3]

[1] *CREST, Ensai, and Datastorm; hassan.maissoro@datastorm.fr*
[2] *CREST, Ensai; valentin.patilea@ensai.fr*
[3] *CREST, Ensai; myriam.vimond@ensai.fr*

Abstract

We consider stationary functional time series where each observation is a trajectory, measured with error at discretely, possibly randomly, sampled domain points. We consider the estimator for the local regularity of the trajectories introduced by Golovkine et al. (2022) in the context of dependent observations. A non-asymptotic bound for the concentration of the local regularity estimator is derived for functional time series which are L^p-m-approximable in the sense of Hörmann and Kokoszka (2010). We also derive a non-asymptotic concentration bound for the Hölder constant estimator. Given the estimates of the local regularity and the Hölder constant, one can diagnose changes in regularity along the trajectory, build optimal recovery of the trajectories, *etc.* Our estimates perform well in simulations. Real data sets illustrate the finite sample performance.

Keywords. Concentration bounds, Kernel smoothing, Nagaev inequality, Stochatic processes

References

1. Golovkine, S., Klutchnikoff, N., and Patilea, V. (2022). Learning the smoothness of noisy curves with application to online curve estimation. *Electronic Journal of Statistics*, 16(1):1485–1560.

2. Hörmann, S., Horváth, L., and Reeder, R. (2013). A functional version of the arch model. *Econometric Theory*, 29(2):267–288.

3. Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884.

*Presenting author

Testing a parametric circular regression function

Andrea Meilán-Vila[1*], Mario Francisco-Fernández[2] & Rosa M. Crujeiras[3]

[1] *Universidad Carlos III de Madrid*
[2] *Universidade da Coruña*
[3] *Universidade de Santiago de Compostela*

Abstract

In this work, new approaches for testing a parametric linear-circular regression model (circular response and Euclidean covariates) are proposed and analyzed [3]. The test statistics employed in these procedures are based on a comparison between a (non-smoothed or smoothed) parametric fit under the null hypothesis and a nonparametric estimator of the circular regression function. Notice that, in this framework, a suitable measure of circular distance must be employed. The null hypothesis that the regression function belongs to a certain parametric family is rejected if the distance between both fits exceeds a certain threshold. To perform the parametric estimation, procedures based on least squares or maximum likelihood are used [1, 2]. For the nonparametric alternative, a local linear-type estimator [4] is considered. For the application in practice, different bootstrap methods are designed, and their performance is analyzed and compared in empirical experiments. The testing proposals are also illustrated with a real dataset.

Keywords. Model checking; circular data; local linear regression; bootstrap.

References

[1] Fisher, N. I. and Lee, A. J. (1992). Regression models for an angular response. *Biometrics*, 48(3):665–677.

[2] Lund, U. (1999). Least circular distance regression for directional data. *Journal of Applied Statistics*, 26(6):723–733.

[3] Meilán-Vila, A., Francisco-Fernández, M., and Crujeiras, R. M. (2022). Goodness-of-fit tests for multiple regression with circular response. *Journal of Statistical Computation and Simulation*, 92(9):1941–1963.

[4] Meilán-Vila, A., Francisco-Fernández, M., Crujeiras, R. M., and Panzera, A. (2021). Nonparametric multiple regression estimation for circular response. *TEST*, 30(3):650–672.

*Presenting author. Email: ameilan@est-econ.uc3m.es

Independence tests for randomly censored data: novel proposal and the review of recent developments

Marija Cuparić[1], Bojana Milošević[2,*]

[1] *University of Belgrade, Faculty of Mathematics, marijar@matf.bg.ac.rs*
[2] *University of Belgrade, Faculty of Mathematics, bojana@matf.bg.ac.rs*

Abstract

Here we consider the problem of testing independence between two random variables in the presence of random censoring. In particular, we consider three different censoring scenarios: one of the targeted variables is censored, both targeted variables are censored with the same censoring variable and both targeted variables are censored with different censoring variables. In all three cases, some well-known, as well as novel adaptations of several famous test statistics for testing independence (designed for complete data) are presented. Their limiting null distributions are shown, accompanied by the proposal of resampling procedures that might be used for their approximation. Finally, the results of wide empirical power are summarized and directions for further research are presented.

Keywords. IPCW method, survival analysis, Kendall coefficient, bootstrap

[*]Presenting author

A uniform kernel trick for high-dimensional two-sample problems

Javier Cárcamo[1], Antonio Cuevas[2] & Luis-Alberto Rodríguez[3,*]

[1] *University of the Basque Country, javier.carcamo@ehu.eus*
[2] *Universidad Autónoma de Madrid, antonio.cuevas@uam.es*
[3] *Universidad Autónoma de Madrid, luisalberto.rodriguez@uam.es*

Abstract

We use a suitable version of so-called "kernel trick" to devise a family of two-sample (homogeneity) tests, especially focussed on high-dimensional and functional data. Our proposal provides some simplification in the important practical problem of selecting an appropriate kernel function. Specifically, we apply a "uniform" version of the kernel trick which involves the supremum within a class of kernel-based distances.

We obtain the asymptotic distribution (under both, the null and alternative hypotheses) of the test statistic. The proofs rely on empirical processes theory, combined with Hadamard directional differentiability techniques and functional Karhunen-Loève expansions of the underlying processes. This methodology has some advantages over other standard approaches in the literature. We also give insight into the performance of our proposal compared to the current kernel-based approach (Gretton *et al.*, 2007) and other high-dimensional techniques, such as those based on energy distances (Szekely and Rizzo, 2017).

Keywords. Goodness of fit, mean embedding, plug-in estimation, probability metric, RKHS.

References

1. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 513–520.

2. Szekely, G. J., & Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, 4, 447–479.

*Presenting author

KSD Aggregated Goodness-of-fit Tests

Antonin Schrab[1,2,3,*], Ilmun Kim[4], Benjamin Guedj[1,3] & Arthur Gretton[2]

[1] *Centre for Artificial Intelligence, University College London*
[2] *Gatsby Computational Neuroscience Unit, University College London*
[3] *Inria London*
[4] *Department of Statistics & Data Science, Yonsei University*

Abstract

We investigate properties of goodness-of-fit tests based on the Kernel Stein Discrepancy (KSD). We introduce a strategy to construct a test, called KSDAgg, which aggregates multiple tests with different kernels. KSDAgg avoids splitting the data to perform kernel selection (which leads to a loss in test power), and rather maximises the test power over a collection of kernels. We provide theoretical guarantees on the power of KSDAgg: we show it achieves the smallest uniform separation rate of the collection, up to a logarithmic term. KSDAgg can be computed exactly in practice as it relies either on a parametric bootstrap or on a wild bootstrap to estimate the quantiles and the level corrections. In particular, for the crucial choice of bandwidth of a fixed kernel, it avoids resorting to arbitrary heuristics (such as median or standard deviation) or to data splitting. We find on both synthetic and real-world data that KSDAgg outperforms other state-of-the-art adaptive KSD-based goodness-of-fit testing procedures. KSDAgg is a quadratic-time test; we also propose a linear-time variant which uses an incomplete U-statistic, which we refer to as KSDAggInc. We quantify exactly the cost incurred in the minimax rate for this improvement in computational efficiency. We support our claims with numerical experiments on the trade-off between computational efficiency and test power.

Keywords. Goodness-of-fit testing, minimax adaptivity, kernel methods.

References

1. Schrab, A., Guedj, B., and Gretton, A. (2022a). KSD Aggregated Goodness-of-fit Test. *arXiv preprint 2202.00824*.

2. Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022b). Efficient Aggregated Kernel Tests using Incomplete U-statistics. *arXiv preprint 2206.09194*.

*Presenting author

The Maximally Selected Likelihood Ratio Test In Random Coefficient Models

Lajos Horváth[1], Lorenzo Trapani[2] & Jeremy VanderDoes[3,*]

[1] *Department of Mathematics, University of Utah, Salt Lake City, UT 84112-0090 USA*
horvath@math.utah.edu
[2] *School of Economics, Granger Centre for Time Series Econometrics, The University of Nottingham, University Park Nottingham NG7 2RD U.K.*
lorenzo.trapani1@nottingham.ac.uk
[3] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, N2L 3G1, Canada*
jeremy.vanderdoes@uwaterloo.ca

Abstract

We investigate changepoint detection in the context of a Random Coefficient Autoregressive model of order 1. In order to ensure power versus breaks occurring very closely to sample endpoints, we study the (maximally selected) Likelihood Ratio statistic, showing that this has power versus breaks occurring even as close as $O(\log \log N)$ periods from the beginning/end of sample. Our test statistic has the same distribution - of the Darling-Erdos type - irrespective of whether the data are stationary or not, and can therefore be applied with no prior knowledge on this. Further, no nuisance parameters need to be estimated in order to use the test statistic. Our simulations show that our test has very good power and, when applying a suitable correction to the asymptotic critical values, the correct size. We illustrate the usefulness and generality of our approach through three applications to epidemiological, economic, and medical time series.

Keywords. RCA(1) model, change point detection, stationary and non stationary models, limit results, estimable parameters

---

*Presenting author

On the effect of the Kaplan-Meier estimator's assumed tail behaviour on goodness-of-fit testing

IJH Visagie[1,*], JS Allison[1] & E Bothma[1]

[1] *Subject Group Statistics*
North-West University
South Africa
jaco.visagie@nwu.ac.za

Abstract

When analysing lifetime data in the presence of censoring one is often required to estimate the distribution function of the lifetimes non-parametrically. The most popular estimator used for this purpose is the Kaplan-Meier estimator. For values larger than the sample maximum two different assumptions are commonly used for this estimator in the statistical literature. The first is to set the value of the estimate to one while the second is to use the value of the estimate at the sample maximum when estimating the tail of the distribution function. We illustrate the profound effect of these assumptions on the sizes and powers of goodness-of-fit tests for three classes of distributions often used in survival analysis. The considered classes of distributions are the exponential, Weibull and gamma.

Keywords. Exponential distribution, Gamma distribution, Goodness-of-fit testing, Random right censoring, Weibull distribution.

[*] Presenting author

Adaptive Functional Principal Component Analysis

Valentin Patilea[1] Sunny Wang[2,*]

[1] *CREST, Ensai; valentin.patilea@ensai.fr*
[2] *CREST, Ensai; sunny.wang@ensai.fr*

Abstract

We build kernel estimators for the mean and the covariance functions of functional data, and use them for functional PCA. The random trajectories are, not necessarily differentiable, have unknown, possibly non constant regularity, and are measured with possibly heteroscedastic error, at discrete design points. We propose specific bandwidth rules for the eigenvalues and the eigenfunctions, respectively. The bandwidth adapts to the local regularity of the trajectories, and minimizes the mean squared error between our eigenelements estimates and the ideal ones, which would be obtained if the curves were observed in continuous time, without noise. They can be applied with both sparsely or densely sampled curves, are easy to calculate and to update, and perform well in simulations. Simulations illustrate the effectiveness of the new approach.

Keywords. Covariance function, Kernel smoothing, Local regularity

References

1. Cai, T. and Yuan, M. (2010). Nonparametric covariance function estimation for functional and longitudinal data. Working paper, University of Pennsylvania and Georgia Institute of Technology.

2. Golovkine, S., Klutchnikoff, N., and Patilea, V. (2022). Learning the smoothness of noisy curves with application to online curve estimation. *Electronic Journal of Statistics*, 16(1):1485–1560.

3. Golovkine, S., Klutchnikoff, N., and Patilea, V. (2021). Adaptive optimal estimation of irregular mean and covariance functions. arXiv preprint arXiv:2108.06507.

4. Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Annals of Statistics*, 44(5):2281–2321.

*Presenting author

Support from the following partners is gratefully acknowledged

- ENSAI (Ecole Nationale de la Statistique et de l'Analyse de l'Information)
- CREST (Center for Research in Economics and Statistics; UMR CNRS 9194), France
- North-West University, South Africa
- Joint Research Initiative *'Models and mathematical processing of very large data'* under the aegis of Risk Foundation and Institut Louis Bachelier, in partnership with MEDIAMETRIE and GENES, France
- Laboratoire de Mathématiques et de leurs Applications de Pau (LMAP, UMR CNRS 5142), France