MCP Conference 2022

12th International Conference on Multiple Comparison Procedures

Book of Abstracts

August 30 - September 02, 2022
University of Bremen, Germany

Panel Discussions

Beyond conventional error rate control: decision-theoretic, conditional, Bayesian approaches

organized by: Bretz, Frank; Hsu, Jason

Abstract

Control of (Type I) error rate should translate to control of incorrect decision rate, taking into account the characteristics specific to pharmaceutical drug development. Using aggregate numbers has its limitations. For example, standard familywise error rate (FWER) control assumes all rejections of true null hypotheses are equally serious, which may be inadequate when the consequences of a larger number of rejections are worse than those of one incorrect rejection or consequences of some incorrect rejections are more serious than those of other rejections. The objective of this panel session is to discuss error rate concepts that may complement FWER or are alternatives to it. In setting such as ordered endpoints, basket trials, and platform trials, we will discuss whether conditional, decision-theoretic, and Bayesian error rates may more directly take consequences of incorrect decisions into account.

Chair: Xinping Cui
Speakers: Werner Brannath, Frank Bretz, Jason Hsu
Panelists: Hsien Ming (James) Hung, Martin Posch, Susanne Urach plus all speakers

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Beyond conventional error rate control: decision-theoretic, conditional, Bayesian approaches
Start: 1:45 p.m. (90 minutes)

Correct and Logical Causal Inference

organized by: Bretz, Frank; Hsu, Jason

Abstract

With odds ratio (OR) and hazard ratio (HR) subject to marginalization paradoxes, this session will cover different approaches to dealing with the marginal and conditional inference issue. Specifically, based on Liu et al (2021) and associated Discussions of that paper in Biometrical Journal, the session will share perspectives from regulatory, industry, and academic experts on simultaneous inference that can respect logical relationships among efficacy in subgroups and their combinations. In the context of targeted therapies which naturally have subgroups, this session will discuss the following specific issues in randomized controlled trials (RCTs):

1. Using efficacy measures such as Odds ratio (OR) and hazard ratio (HR) can make a prognostic biomarker appear predictive, leading to potentially wrong targeting of patient. The predictive illusion occurs because mixing OR and HR tend to dilute them, in the presence of a prognostic factor of the biomarker, even with ignorable treatment assignment in an RCT.

2.Subgroup Mixable Estimation (SME) will give causal inference in an RCT (without needing explicit standardization formulas), by automatically accounting for the prognostic effect, when logic-respecting efficacy measure such as Relative response (RR) and ratio of median survival times (RoM) are used.

3. A third insight is that, under the condition that (within each biomarker subgroup) the two possible outcomes of each patient (one observable the other potential) are independent, efficacy measures equivalent to OR and HR can become logic respecting.

Moderator: Xinping Cui
Framer of the discussions: Jason Hsu & Frank Bretz
Panelists: Norbert Benda, Yi Liu, Gene Pennello, Sue-Jane Wang, Dong Xi

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Correct and Logical Causal Inference
Start: 3:45 p.m. (90 minutes)

Oral Presentations

Grammar of Group Sequential Design

Anderson, Keaven M; Zhao, Yujie; Xiao, Nan; Zhang, Yilong

Abstract

We discuss a general grammar intended for design and evaluation of group sequential designs. The objective is to allow writing and review of simple, readable code for a wide range of designs as well as simulations. This begins with simpler designs such as two-arm trials with binomial, normal or time-to-event designs under a proportional hazards assumption. Next, we consider asymptotic methods for non-proportional hazards with logrank and other test statistics. We then extend to more complex designs with multiple endpoints in a trial. This last topic can involve many endpoints, some of which have tests with known correlations. Both asymptotic methods and simulation are incorporated. All methods have been used in multiple trials over the last several years.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Software developments
Start 11:00 a.m. (30 minutes)

Inference on Winners

Andrews, Isaiah

Abstract

Many empirical questions concern target parameters selected through optimization. For example, researchers may be interested in the effectiveness of the best policy found in a randomized trial, or the best-performing investment strategy based on historical data. Such settings give rise to a winner's curse, where conventional estimates are biased and conventional confidence intervals are unreliable. This paper develops optimal confidence intervals and median-unbiased estimators that are valid conditional on the target selected and so overcome this winner's curse. If one requires validity only on average over targets that might have been selected, we develop hybrid procedures that combine conditional and projection confidence intervals to offer further performance gains relative to existing alternatives.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Selective inference
Start: 9:00 a.m. (30 minutes)

Replicability issues in medical research: Science and Politics

Benjamini, Yoav; Jaljuli, Iman Jaljuli; Heller, Ruth; Panagiotou, Orestis

Abstract

Selective inference and irrelevant variability are two statistical issues hindering replicability across science. I will review the first in the context of secondary endpoint analysis in clinical and epidemiological research. This leads us to discuss the debate about p-values and statistical significance and the politics involved. I will present practical approaches that seem to accommodate the concerns of NEJM editors, as reflected in their guidelines.

I shall discuss more briefly the issue of addressing the relevant variability, in the context of in preclinical animal experiments, and the implication of this work about assessing replicability in meta-analysis.

Date and time

Day 1 (Wednesday, August 31, 2022)
Sesssion: Replicability and Reproducibility (I)
Start: 11:20 a.m. (30 minutes)

Notip: Non-parametric True Discovery Proportion estimation for brain imaging

Blain, Alexandre; Thirion, Bertrand; Neuvial, Pierre

Abstract

Cluster-level inference procedures are widely used for brain mapping. These methods compare the size of clusters obtained by thresholding brain maps to an upper bound under the global null hypothesis, computed using Random Field Theory or permutations. However, the guarantees obtained by this type of inference - i.e. at least one voxel is truly activated in the cluster - are not informative with regards to the strength of the signal therein. There is thus a need for methods to assess the amount of signal within clusters; yet such methods have to take into account that clusters are defined based on the data, which creates circularity in the inference scheme. This has motivated the use of post hoc estimates that allow statistically valid estimation of the proportion of activated voxels in clusters. In the context of fMRI data, the All-Resolutions Inference framework introduced in [Rosenblatt et al., NeuroImage 2018] provides post hoc estimates of the proportion of activated voxels. However, this method relies on parametric threshold families, which results in conservative inference.

In this paper, we leverage randomization methods to adapt to data characteristics and obtain tighter false discovery control. We obtain Notip: a powerful, non-parametric method that yields statistically valid estimation of the proportion of activated voxels in data-derived clusters. Numerical experiments demonstrate substantial power gains compared with state-of-the-art methods on 36 fMRI datasets. The conditions under which the proposed method brings benefits are also discussed.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Post-hoc FDP control (I)
Start: 12:20 p.m. (20 minutes)

Methods for utilising non-concurrent controls in platform trials

Bofill Roig, Marta; Posch, Martin

Abstract

Platform trials aim at evaluating the efficacy of several experimental treatments within a single trial. The number of experimental arms is not prefixed, as arms may be added or removed as the trial progresses. Platform trials offer the possibility of comparing the efficacy of experimental arms using a shared control group. Compared to separate trials with their own controls, this increases the statistical power and requires fewer patients. Shared controls in platform trials include concurrent and non-concurrent control data. For a given experimental arm, non-concurrent controls refer to data from patients allocated to the control arm before the arm enters the trial. Using non-concurrent controls is appealing because it may improve the trial's efficiency while decreasing the sample size. However, since arms are added sequentially, randomization occurs at different times. This lack of true randomization over time might introduce bias due to time trends. The challenge is to discern when and how to use non-concurrent controls to increase the trial's efficiency without introducing bias.

In this talk, we review methods to incorporate non-concurrent control data in treatment-control comparisons allowing for time trends. We focus mainly on frequentist approaches that model the time trend and Bayesian strategies that limit the borrowing level depending on the heterogeneity between concurrent and non-concurrent controls. We examine the impact of time trends on the operating characteristics of treatment effect estimators for each method under different patterns for the time trends. We outline under which conditions the methods lead to unbiased estimators and discuss the gain in power compared to trials only using concurrent controls.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Multiple testing methodology for group-sequential and platform trials
Start: 4:45 p.m. (30 minutes)

Testing partial conjunction hypotheses under dependency

Bogomolov, Marina

Abstract

In many statistical problems the hypotheses are naturally divided into groups, and the investigators are interested to perform group-level inference, possibly along with inference on individual hypotheses. We consider the goal of discovering groups containing u or more signals with group-level false discovery rate (FDR) control. This goal can be addressed by multiple testing of partial conjunction hypotheses with a parameter u, which reduce to global null hypotheses for $u = 1$. We consider the case where the partial conjunction p-values are combinations of within-group p-values, and obtain sufficient conditions on (1) the dependencies among the p-values within and across the groups, (2) the combining method for obtaining partial conjunction p-values, and (3) the multiple testing procedure, for obtaining FDR control on partial conjunction discoveries. We consider separately the dependencies encountered in the meta-analysis setting, where multiple features are tested in several independent studies, and the p-values within each study may be dependent. Based on the results for this setting, we generalize the procedure of Benjamini, Heller, and Yekutieli (2009) for assessing replicability of signals across studies, and extend their theoretical results regarding FDR control with respect to replicability claims.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Replicability and Reproducibility (I)
Start: 11:50 a.m. (30 minutes)

Generalized Pairwise Comparisons with Prioritized Outcomes

Buyse, Marc

Abstract

Generalized pairwise comparisons (GPCs) have been proposed to simultaneously analyse several outcomes of any type (discrete, continuous, possibly censored). GPCs are a generalization of the Mann-Whitney formulation of the Wilcoxon test which is especially useful when (1) the outcomes of interest can be prioritized (from clinically most important to least important), and (2) clinical thresholds are deemed relevant for some of these outcomes (for instance, survival gains should exceed 6 months to be considered clinically worthwhile). In randomized clinical trials comparing Treatment to Control, GPCs consist of comparing all possible pairs of patients formed by taking one patient from the Treatment group and one patient from the Control group. Each pair is classified as a win, a loss or a tie for the outcome of highest priority. Ties are then classified using the next outcome of lower priority, and the process is repeated until all outcomes have been analysed. Possible measures of treatment effect include the Win Ratio ([number of wins] divided by [number of losses]), the Win Odds ([number of wins plus half the number of ties] divided by [number of losses plus half the number of ties]), or the Net Treatment Benefit ([number of wins minus number of losses] divided by [number of pairs]). The Net Treatment Benefit is an absolute measure that directly addresses patient-centric questions about the probabilities of benefits and harms from treatment. As such, GPC can potentially be used to individualize treatment choices. The general properties of GPC will discussed for both a single outcome and multiple outcomes. An attractive feature of the Net Treatment Benefit is that the contribution of each outcome is additive (conditional on the outcome priorities). A natural testing procedure to account for the multiple outcomes is to test the Net Treatment Benefit for all outcomes first, and then proceed sequentially by eliminating the outcomes in reverse order of clinical importance. All of these concepts will be illustrated in different clinical trial settings.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Recent Advances in Generalized Pairwise Comparisons of Multiple Prioritized Endpoints
Start: 8:30 a.m. (30 minutes)

Multiple comparison procedures for discrete uniform and homogeneous tests

Cousido-Rocha, Marta; de Uña-Álvarez, Jacobo; Döhler, Sebastian

Abstract

Homogeneous discrete uniform (hdu) p-values often arise in applications with multiple testing. For example, this occurs in genome wide association studies whenever a nonparametric one-sample (or two-sample) test is applied throughout the gene loci. Even though discrete p-values arise in many applications, few research explicitly deal with this aspect of multiple testing. Furthermore, the proposed discrete corrections of multiple comparison procedures are irrelevant for hdu p-values.

Then, we consider multiple comparison procedures for such setting based on several existing estimators for the proportion of true null hypotheses, π_0, which take the discreteness of the p-values into account. The theoretical guarantees of the several approaches with respect to the estimation of π_0 and the false discovery rate control are reviewed. The performance of the discrete procedures is investigated through intensive Monte Carlo simulations considering both independent and dependent p-values. The methods are applied to three real data sets for illustration purposes too. Since the particular estimator of π_0 used on the multiple comparison procedure may influence its performance, relative advantages and disadvantages of the reviewed procedures are discussed. Practical recommendations are given.

The methods have been implemented inthe user-friendly DiscreteQvalue package of the free software R.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Miscellaneous Topics (I)
Start: 2:35 p.m. (20 minutes)

Adaptive group sequential designs for multiple time-to-event outcomes

Danzer, Moritz Fabian; Faldum, Andreas; Schmidt, Rene

Abstract

Adaptive designs for the assessment of a single time-to-event outcome are well established. However, care has to be taken when interim data from further endpoints is used for data-dependent design changes (e.g. sample size recalculation). It is particularly problematic to base design changes on interim data from an additional endpoint that may serve as a surrogate for the chosen primary endpoint.

Similar problems arise if several time-to-event endpoints are assessed simultaneously as one of these variables may be used to make predictions about another variable for patients who enter the trial before an interim analysis and remain event-free beyond it. Existing group sequential designs for multivariate survival trials cannot be extended to adaptive designs as this additional information is not accounted for.

We provide adaptive group sequential designs for testing hypotheses on the joint distribution of multiple time-to-event endpoints. Our approach enables data-dependent design changes based on the information from all involved time-to-event endpoints. To make this possible, a few distributional assumptions have to be made. More precisely, we assume the underlying multi-state model to be Markovian or Semi-Markovian. Large sample distributions of the testing procedure are derived using counting process approaches. Small sample properties and the behaviour under deviation from the above-mentioned conditions are studied by simulation.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Multiple testing methodology for group-sequential and platform trials
Start: 4:15 p.m. (30 minutes)

FDP control in multivariate linear models using the bootstrap

Davenport, Samuel; Thirion, Bertrand; Neuvial, Pierre

Abstract

In this article we develop a method for performing post-hoc inference of the False Discovery Proportion (FDP) over multiple contrasts of interest in the multivariate linear model. To do so we use the bootstrap to simulate from the null distribution of the null contrasts. We combine the bootstrap with the post-hoc inference bounds of Blanchard et al (2020) and prove that doing so provides simultaneous asymptotic control of the FDP over all subsets of hypotheses. This requires us to demonstrate consistency of the multivariate bootstrap in the linear model which we do via the Lindeberg CLT, providing a simpler proof of this result than that of Eck (2018). We demonstrate, via simulations, that our approach controls the Joint Error Rate and is typically more powerful than existing, state of the art, parametric methods. We illustrate our methods on fMRI data from the Human Connectome project and on a transcriptomic dataset.

Gilles Blanchard, Pierre Neuvial, Etienne Roquain, et al. Post hoc confidence bounds on false positives using reference families. Annals of Statistics, 48(3):1281{1303, 2020}.

Daniel J Eck. Bootstrapping for multivariate linear regression models. Statistics & Probability Letters, 134:141{149, 2018}.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Post-hoc FDP control (I)
Start: 11:50 a.m. (20 minutes)

Inference in generalized linear models with robustness to misspecified variances

De Santis, Riccardo; Goeman, Jelle J.; Hemerik, Jesse; Finos, Livio

Abstract

One of the crucial assumptions of generalized linear models is related to variance specification. The general assumption of a common dispersion parameter can be too strict in practice, which can be seen in analogy with the assumption of homoscedasticity in the linear model. If we focus on hypothesis testing, variance misspecification can easily provoke the failure of type I error control of the standard parametric approach. Conditional tests are generally used to implement a more robust approach with respect to possible model misspecifications. We will present a novel method which requires only the unbiased estimation of the mean, which requests the correct specification of the link function, by means of an appropriate modification of score contributions, whilst being robust against any general biased estimation of the Fisher information. This approach is quite general and flexible and allows the extension to the multivariate framework, and therefore to any resampling-based multiple testing procedure (Familywise Error Rate, False Discovery Proportion, etc.).

Date and time

Day 3 (Friday, September 2, 2022)
Session: Miscellaneous Topics (II)
Start: 10:10 a.m. (20 minutes)

Operational Characteristics of Hierarchical Generalized Pairwise Comparisons Test

Deltuvaite-Thomas, Vaiva

Abstract

The Generalized Pairwise Comparisons (GPC) method is a multivariate extension of the non-parametric Wilcoxon-Mann-Whitney test, allowing comparisons of two groups of observations based on multiple hierarchically ordered outcomes of any type (e.g., discrete, continuous, time to event). The summary measure of the difference between the groups called the net treatment benefit quantifies the difference between the probabilities that a random observation from one group is doing better than an observation from the other group. Due to its hierarchical formulation the GPC can take into account the correlations between the endpoints. We will show through theoretical considerations how the expected value and the variance of the net treatment benefit statistic depend, in a complicated manner, on the entire variance-covariance structure of the set of the outcomes. We will discuss the patterns in which all these parameters influence the operational characteristics (power and type-I error) of the test and use concrete clinical examples to demonstrate these influences in real life.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Recent Advances in Generalized Pairwise Comparisons of Multiple Prioritized Endpoints
Start: 9:20 a.m. (20 minutes)

Multiple testing of partial conjunction null hypotheses with application to replicability analysis of high-dimensional studies

Dickhaus, Thorsten

Abstract

The partial conjunction null hypothesis is tested in order to discover a signal that is present in multiple studies. The standard approach of carrying out a multiple test procedure on the partial conjunction (PC) p-values can be extremely conservative. We suggest alleviating this conservativeness, by eliminating many of the conservative PC p-values prior to the application of a multiple test procedure. This leads to the following two-step procedure: First, select the set with PC p-values below a selection threshold; second, within the selected set only, apply a family-wise error rate or false discovery rate controlling procedure on the conditional PC p-values. We prove that the conditional PC p-values are valid for certain classes of one-parametric statistical models (including one-parameter natural exponential families), and provide conditions for (asymptotic) FDR control for several multiple test procedures operating on conditional PC p-values. We also compare the proposed methodology with other recent approaches by means of computer simulations.

This is joint work with Ruth Heller and Anh-Tuan Hoang.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Replicability and Reproducibility (II)
Start: 8:30 a.m. (30 minutes)

Analysis of ordered composite endpoints

Follmann, Dean; Fay, Michael P; Hamasaki, Toshimitsu; Evans, Scott

Abstract

Composite endpoints are frequently used in clinical trials, but simple approaches, such as the time to first event, do not reflect any ordering among the endpoints. However, some endpoints, such as mortality, are worse than others. A variety of procedures have been proposed to reflect the severity of the individual endpoints such as pairwise ranking approaches, the win ratio, and the desirability of outcome ranking. When patients have different lengths of follow-up, however, ranking can be difficult and proposed methods do not naturally lead to regression approaches and require specialized software. This paper defines an ordering score O to operationalize the patient ranking implied by hierarchical endpoints. We show how differential right censoring of follow-up corresponds to multiple interval censoring of the ordering score allowing standard software for survival models to be used to calculate the nonparametric maximum likelihood estimators (NPMLEs) of different measures. Additionally, if one assumes that the ordering score is transformable to an exponential random variable, a semiparametric regression is obtained, which is equivalent to the proportional hazards model subject to multiple interval censoring. Standard software can be used for estimation. We show that the NPMLE can be poorly behaved compared to the simple estimators in staggered entry trials. We also show that the semiparametric estimator can be more efficient than simple estimators and explore how standard Cox regression maneuvers can be used to assess model fit, allow for flexible generalizations, and assess interactions of covariates with treat-ment. This talk is based on Follmann, Fay, Hamasaki and Evans (2020, Statistics in Medicine).

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Recent Advances in Generalized Pairwise Comparisons of Multiple Prioritized Endpoints
Start: 9:40 a.m. (20 minutes)

Sequential algorithmic modification with test data reuse

Feng, Jean; Pennello, Gene; Petrick, Nicholas; Sahiner, Berkman; Pirracchio, Romain; Gossmann, Alexej

Abstract

After initial release of a machine learning algorithm, the model can be fine-tuned by retraining on subsequently gathered data, adding newly discovered features, or more. Each modification introduces a risk of deteriorating performance and must be validated on a test dataset. It may not always be practical to assemble a new dataset for testing each modification, especially when most modifications are minor or are implemented in rapid succession. Recent works have shown how one can repeatedly test modifications on the same dataset and protect against overfitting by (i) discretizing test results along a grid and (ii) applying a Bonferroni correction to adjust for the total number of modifications considered by an adaptive developer. However, the standard Bonferroni correction is overly conservative when most modifications are beneficial and/or highly correlated. This work investigates more powerful approaches using alpha-recycling and sequentially-rejective graphical procedures (SRGPs). We introduce novel extensions that account for correlation between adaptively chosen algorithmic modifications. In empirical analyses, the SRGPs control the error rate of approving unacceptable modifications and approve a substantially higher number of beneficial modifications than previous approaches.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Adaptive inference
Start: 10:10 a.m. (20 minutes)

On positive association of absolute-valued and squared multivariate Gaussians beyond MTP_2

Finner, Helmut; Roters, Markus

Abstract

Concepts of positive (negative) dependence associated with probability inequalities are often essential for proving conservativeness of multiple decision procedures. For example, multivariate totally positive of order 2 (MTP_2) as well as the weaker notion of positive association (PA) of random variables yield various probability inequalities useful in multiple testing. In this talk we are concerned with the question which absolute-valued p-dimensional multivariate normally distributed random vectors are positively associated (PA). Around 1980 various authors (cf. e.g. Bolviken (1982), Karlin, Rinott (1983), Rüschendorf (1981)) proved that absolute normals $|X|$ are MTP_2 if the inverse of the covariance matrix of DX is an M-matrix for some signature matrix D. In this talk we show that this so-called signed MTP_2 condition is not necessary for PA of absolute-valued normals for $p \geq 3$. Hence, there is at least some free space beyond the celebrated but tiny MTP_2 world for absolute-valued normals to be PA. Our main findings are based on the the fact that conditionally increasing in sequence (CIS) implies PA. For $p = 3$ we show that there exist absolute-valued multivariate normals which are CIS (and hence PA) but not MTP_2 iff the underlying covariance matrix satisfies a certain condition. However, for $p \geq 4$, we also show that the existence of a CIS sequence is not necessary for absolute-valued normals to be PA. Finally, our results disprove Theorem 1 in Eisenbaum (2014) and the conjecture that MTP_2, infinite divisibility and PA of squared multivariate normals are equivalent.

References:

Bølviken, E. (1982). Probability inequalities for the multivariate normal with non-negative partial correlations. Scand. J. Statist. 9, 49-58.

Eisenbaum, N. (2014). Characterization of positively correlated squared Gaussian processes. Ann. Probab. 42, 559-575.

Finner, H., Roters, M. (2022). On positive association of absolute-valued and squared multivariate Gaussians beyond MTP_2. Preprint.

Karlin, S., Rinott, Y. (1981). Total positivity properties of absolute value multinormal variables with applications to confidence interval estimates and related probabilistic inequalities. Ann. Stat. 9, 1035-1049.

Rüschendorf, L. (1981). Characterization of dependence concepts in normal distributions. Ann. Inst. Stat. Math. 33(3), 347-359.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Miscellaneous Topics (II)
Start: 9:30 a.m. (20 minutes)

Ensemble Inference to Enhance Replicability

Finos, Livio; Vesely, Anna; Altè, Gianmarco; Pastore, Massimiliano; Calcagì, Antonio; Girardi, Paolo

Abstract

Data processing of non-trivial datasets often involves choices among several reasonable options for excluding, transforming, coding data, and modeling them. This multiplicity of steps gives rise to a "multiverse" of reasonable models and, therefore, statistical results. Unfortunately, it is a common practice to report only one privileged single analysis, therefore depriving the reader of the taste of this multiplicity and making the interpretation too optimistic.

Together with other questionable research practices, this is one of the main reason of the dramatic Reproducibility Crisis (Yong, 2012) and lack of confidence in many fields from psychology to economics, from sociology to medicine and neuroscience.

Steegen et al (2016) firstly proposed the Multiverse Analysis, which is, roughly speaking, nothing but the idea of frankly reporting all the analyses performed, then allowing the readers to evaluate the "stability" of the results. The proposal certainly represents an evaluable step toward the honest science. Since then, the method has been largely developed and has grown in popularity.

Despite this, it remains relegated to a descriptive role if as a formal inferential approach is not adopted. Simonsohn et al. 2020 firstly proposed a valuable method to derive a permutation-based test in this framework. However, this methodology is restricted to the linear model and does not cover all possible pre-processing steps. Furthermore – in our opinion – a more formal approach to the problem will cast the problem in the right theoretical context. In this contribution we exploit the flip-score test (Hemerik et al, 2020) to develop a very general and flexible approach that account for these issues.

References:

Yong, E. (2012). Replication studies: Bad copy. Nature News, 485(7398), 298.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. Nature Human Behaviour, 4(11), 1208-1214.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702-712.

Hemerik, J., Goeman, J. J., & Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82* (3), 841-864.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Replicability and Reproducibility (II)
Start: 9:00 a.m. (30 minutes)

Online multiple testing with FWER control

Fischer, Lasse

Abstract

While online FDR control is studied extensively, there is less work on FWER control in the online setting. In 2021, Tian & Ramdas introduced the Adaptive-Discard-Spending (ADDIS-Spending) as an online procedure with FWER control. In this talk we apply the concepts of adaptivity and discarding to the graphical approach by Bretz et al. (2009), resulting in what we call the ADDIS-Graph. Due to its graphical representation the ADDIS-Graph is easy to interpret. Moreover, it leads to power improvements compared to the ADDIS-Spending in case of local dependent p-values. In addition, we exhaust the significance level under independence of the p-values to obtain uniformly superior ADDIS procedures with theoretical results that are supported by simulations. Furthermore, we formulate a new closure principle for online multiple testing and present a condition under which a closed procedure is indeed an online procedure.

This is joint work with Werner Brannath and Marta Bofill Roig.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: On-line multiple testing
Start: 11:40 a.m. (20 minutes)

Online False Discovery Rate Control for LORD & SAFFRON Under Positive, Local Dependence

Fisher, Aaron

Abstract

Online testing procedures assume that hypotheses are observed in sequence, and allow the significance thresholds for upcoming tests to depend on the test statistics observed so far. Some of the most popular online methods include alpha investing, LORD++ (hereafter, LORD), and SAFFRON. These three methods have been shown to provide online control of the "modified" false discovery rate (mFDR) under a condition known as conditional superuniformity. However, to our knowledge, LORD & SAFFRON have only been shown to control the traditional false discovery rate (FDR) under an independence condition on the test statistics. Our work bolsters these results by showing that SAFFRON and LORD additionally ensure online control of the FDR under a "local" form of nonnegative dependence. Further, FDR control is maintained under certain types of adaptive stopping rules, such as stopping after a certain number of rejections have been observed. Because alpha investing can be recovered as a special case of the SAFFRON framework, our results immediately apply to alpha investing as well. In the process of deriving these results, we also formally characterize how the conditional superuniformity assumption implicitly limits the allowed p-value dependencies. This implicit limitation is important not only to our proposed FDR result, but also to many existing mFDR results.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: On-line multiple testing
Start: 11:20 a.m. (20 minutes)

A false discovery rate estimator for model selection

Fithian, Will

Abstract

Tunable variable selection algorithms like the lasso or forward stepwise regression are commonly assessed by cross-validation or Stein's unbiased risk estimator, which respectively estimate the algorithm's prediction error and mean squared estimation error at each value of the tuning parameter. As a counterpart to these methods, we propose a new estimator for a generic model selection algorithm's false discovery rate (FDR), defined as the expected fraction of null variables in the selected model. Our method first decomposes the FDR as a sum of the contributions from each null variable, then estimates each term using Rao-Blackwellization. Our estimator is conservative in the sense that its bias is non-negative, and it can be used to tune the algorithm so that the tuned method controls FDR at a prespecified level. I will present the method and illustrate its application in several simple examples.

This is joint work with Yixiang Luo and Lihua Lei.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Multiple testing in the presence of structures
Start: 8:30 a.m. (30 minutes)

Confidence Intervals for Selected Parameters

Frostig, Tzviel; Benjamini, Yoav

Abstract

In many applications of estimation and inference a few parameters are selected to be reported. Usually, the same data is used for selection and inference. This usually lead to an over optimistic confidence intervals which fail to cover their respective parameter at the expected rate.

Methods suggested to tackle the issue often increase the length of the confidence interval beyond what is required. Hechtinger et al. suggested a more exact correction for confidence intervals of the k out of m selected parameters, ensuring the simultaneous over the selected coverage. We extend the result and show the result hold for a positive dependency type and for absolute valued based selection rules. Furthermore, we suggest an improvement for false coverage rate based confidence intervals, proving no correction is required for a specific side of the interval. Ther result yield shorter CIs with the same confidence level.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Post-hoc FDP control (II)
Start: 8:50 a.m. (20 minutes)

Conditional Versus Unconditional Approaches to Selective Inference

Goeman, Jelle; Solari, Aldo

Abstract

We investigate a class of methods for selective inference that condition on a selection event. Such methods operate in a two-stage process. First, a (sub)collection of hypotheses is determined in a data-driven way from some large universe of hypotheses. Second, inference is done within the data-driven collection, conditional on the information that was used for the selection. Examples of such methods include basic data splitting, as well as modern data carving methods and post-selection inference methods based on the polyhedral lemma. In this paper, we adopt a holistic view on such methods, viewing the selection, conditioning and final error control steps together as part of a single method. From this perspective, we show that selective inference methods based on selection and conditioning are always dominated by multiple testing methods defined directly on the full universe of hypotheses. In particular, our result shows that multiple testing by data splitting is inadmissible. Although our main result holds whatever the error rate that is controlled, e.g. false discovery rate or unadjusted, we concentrate in our examples on the situation that familywise error rate control is targeted ("simultaneous on the selected"). For this case we have additional results suggesting that conditioning on a selection can be a way to develop unconditional methods with attractive properties.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Post-hoc FDP control (II)
Start: 8:30 a.m. (20 minutes)

Balancing the Trade-Off between type I and II Errors from a Bayesian Perspective

Grieve, Andrew Peter

Abstract

In the Neyman-Pearson school of inference there are two types of errors that can be made, declaring a positive result when the truth is negative, and declaring a negative result when the truth is positive. Traditionally, control of the probability of the former, the type I error is the most important and is fixed at a low value, following which the sample size is chosen to give an acceptable probability of the latter error, the type II error. When resource is limited, and when is it not, there will always be a trade-off between the probability of type I and II errors, and in this talk based on Walley and Grieve (2021) we paper explore optimising the trade-off for a study with a planned Bayesian analysis, generalising previous work of Grieve (2015). This work provides a scientific basis for a discussion between stakeholders as to the appropriate probabilities of type I and II error.

References:

Grieve, AP (2015). How to test hypotheses if you must. *Pharmaceutical Statistics, 14, 139-150*

Walley RJ and Grieve, AP (2021) Optimising the trade-off between type I and II error rates In the Bayesian context. *Pharmaceutical Statistics,* 20, 710-720.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Lower bounds for type I / II error tradeoffs
Start: 11:40 a.m. (20 minutes)

Optimal allocation of Monte Carlo simulations to multiple hypothesis tests

Hahn, Georg; Lange, Christoph

Abstract

Multiple hypothesis tests are often carried out in practice using p-value estimates obtained with bootstrap or permutation tests since the analytical p-values underlying all hypotheses are usually unknown. In this talk, we consider the allocation of a prespecified integer number of Monte Carlo simulations K (i.e., permutations or draws from a bootstrap distribution) to a given integer number of hypotheses m in order to approximate their p-values p (a vector with m entries in $[0, 1]$) in an optimal way, in the sense that the allocation minimizes the total expected number of misclassified hypotheses. A misclassification occurs if a decision on a single hypothesis, obtained with an approximated p-value, differs from the one obtained if its p-value was known analytically. First, under the assumption that p is known and K is real-valued, and using a normal approximation of the Binomial distribution, the optimal real-valued allocation of K simulations to m hypotheses is derived when correcting for multiplicity with the Bonferroni correction, both when computing the p-value estimates with or without a pseudo-count. Computational subtleties arising in the former case will be discussed. Second, with the help of an algorithm based on simulated annealing, empirical evidence is given that the optimal integer allocation is likely of the same form as the optimal real-valued allocation, and that both seem to coincide asymptotically. Third, an empirical evaluation on the COPDGene (chronic obstructive pulmonary disease) GWAS study demonstrates that a recently proposed sampling algorithm based on Thompson sampling asympotically mimics the optimal (real-valued) allocation when the p-values are unknown and thus estimated at runtime, yielding more accurate classifications than merely approximating all p-values with an equal number of permutations.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Adaptive inference
Start: 9:30 a.m. (20 minutes)

Confident and Logical Selection of the Cut-point of a Biomarker for Patient Targeting

Han, Yang

Abstract

Confidently choosing a cut-point for a continuously valued biomarker to target patients with is challenging because there are two levels of multiplicity: the multiplicity of efficacy in the marker-positive subgroup and in the marker-negative subgroup at each cut-point, and the further multiplicity of searching through infinitely many cut-points. Currently available methods do not strongly control familywise type I error rate (FWER) across both levels of multiplicity. I will present a method that does. Taking a confidence band approach, our method in fact sets forth four principles that we believe every confident biomarker cut-point selection method should strive to adhere to.

For diseases with continuous outcome such as Type II Diabetes and Alzheimer's Disease, our method provides exact simultaneous confidence intervals for efficacy in the marker positive and negative subgroups, simultaneously for all possible cut-point values. I will demonstrate an interactive app for it.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Miscellaneous Topics (I)
Start: 2:15 p.m. (20 minutes)

Designing Optimal Multiple Testing Procedures

Heller, Ruth

Abstract

For a single hypothesis testing problem, the optimal policy maximizes the probability to reject the null (i.e., the power), subject to controlling the type I error probability at a predefined α. Nowadays, conducting a study with a single hypothesis is rare. Even when designing a clinical trial, it is often the case that at least two hypothesis testing problems are simultaneously considered. We formulate the problem of finding the optimal policy for $K > 1$ hypothesis testing problems for various notions of power, subject to controlling an overall measure of false discovery, like family-wise error rate (FWER) or false discovery rate (FDR). We start by describing a complete solution for deriving optimal policies for $K = 2$ hypotheses, which have some desired monotonicity properties, and are computationally simple. We demonstrate the utility of our approach in reanalyzing a clinical trial which aims to infer on the treatment effect of a new drug in two distinct subgroups. Next, we consider the high dimensional setting, where the highly influential two-group model is often assumed to hold. Optimal control of the marginal false discovery rate (mFDR), in the sense that it provides maximal power (expected true discoveries) subject to mFDR control, is known to be achieved by thresholding the local false discovery rate (locFDR, the probability of the hypothesis being null given the set of test statistics), with a fixed threshold. We address the challenge of controlling optimally the popular false discovery rate (FDR) or positive FDR (pFDR) in the general two-group model, which also allows for dependence between the test statistics. We develop an efficient algorithm for finding these policies, and use it for problems with thousands of hypotheses. We illustrate these procedures on gene expression studies.

Joint work with Abba Krieger and Saharon Rosset.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Closing session
Start: 11:00 a.m. (60 minutes)

Flexible control of the median of the false discovery proportion

Hemerik, Jesse; Solari, Aldo; Goeman, Jelle J.

Abstract

We propose a multipe testing method that controls the median of the proportion of false discoveries (mFDP) in a flexible way. Our method only requires a vector of p-values as input and is comparable to the Benjamini-Hochberg (1995) method, which controls the mean of FDP. Benjamini-Hochberg requires choosing α before looking at the data, but our method does not. For example, if using $\alpha = 0.05$ leads to no discoveries, α can be increased to 0.1. We also provide mFDP-adapted p-values, which naturally also have a post hoc interpretation.

Our method is inspired by the popular estimator of the total number of true hypotheses by Storey. We note that this estimator can be adapted to provide an (often) median-unbiased estimator of the FDP, when a fixed rejection threshold is used. Taking this as a starting point, we proceed to construct simultaneously median-unbiased estimators of the FDP. This simultaneity allows for the claimed flexibility. Another good property of our method is that it is naturally adaptive, in the sense that it does not necessarily become conservative (like e.g. Benjamini-Hochberg 1995) if the (unknown) fraction of false hypotheses is high.

We would advise users to not only estimate the FDP, but to also compute a confidence interval for the FDP, using other available methods.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Post-hoc FDP control (II)
Start: 9:30 a.m. (20 minutes)

Simultaneous inference across treatment levels, various generalized linear models, time points, primary endpoints, and effect sizes

Hothorn, Ludwig A.

Abstract

A considerable amount of published papers on multiple testing focuses on a most powerful approach for some to very many p-values derived from mean value comparisons or primary endpoints, mostly independently. Sometimes there is no point to formulate the 'family' in FWER only for one source of multiplicity, but just really experimentwise for several sources, such as treatment levels, correlated endpoints, time points, multiple models, etc., simultaneously. This is achieved with the multiple marginal models approach (Pipper et al., 2012), which allows a maxT-test over multiple glm(m)'s, estimating the correlation matrix for the multivariate t distribution from the included models.

Since part of the functionality is available in the R packages multcomp and tukeytrend (Schaarschmidt, 2021), the tremendous flexibility of this approach is demonstrated using several short case studies: i) modeling dose as qualitative factor and quantitative covariate jointly, ii) analysing correlated, possibly differently scaled multiple primary endpoints in k-sample designs, iii) inference on multiple models, e.g. various Weibull shape parameters in the poly-k trend test, iv) inference over dose and time, and v) analysing sub-groups jointly with the overall population in phase III RCT.

These single-step approaches allow both adjusted p-values and/ or compatible simultaneous confidence intervals, which is important for an adequate interpretation.

References:

Pipper CB, Ritz C, Bisgaard H. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. JRSS-A 2012; 61(2): 315-326.

Schaarschmidt F, Ritz C, Hothorn LA. The Tukey trend test: Multiplicity adjustment using multiple marginal models. Biometrics 2021.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Multiple testing in the presence of structures
Start: 9:40 a.m. (20 minutes)

ETZ: A Clinically Meaningful Model for Assessing Confirmability of Study Results

Hsu, Jason

Abstract

Drug development proceeds in stages, from Phase 1 through Phase 3 and beyond. Among the phase transitions, the Phase 2 to Phase 3 transition has the lowest success rate, around 30%, despite the fact that phase 3 confirmatory studies are designed to achieve 80% or even 90% power. This presentation provides a practical framework towards Confirmability of Randomized Controlled Trials (CRTs). For CRT with repeated measures, we propose an ETZ model which is related to but different from MMRM (Mixed Model Repeated Measures) and the Random Coefficients model. The ETZ model decomposes variability into a pre-treatment (intercept Z) component, a post-treatment (trajectory T) component, and a measurement Error (E) component. We show both numerically and visually how to assess separately the impact of each component on potential failure to confirm, so that each can be targeted for reduction. For example, while Z variability can be decreased by narrowing the patient entry criterion, trajectory T variability can be decreased by enrolling only patients with sufficient biological targets for the therapy to act upon. Measurement error can be reduced by using more highly trained raters. Our methodology can thus be used to direct resources toward reducing the variability most causal for confirmatory failure.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Replicability and Reproducibility (I)
Start: 12:20 p.m. (30 minutes)

Controlled Discovery and Localization of Signals via Bayesian Linear Programming

Spector, Asher; Janson, Lucas

Abstract

In many high-dimensional statistical problems, it is necessary to simultaneously discover signals and localize them as precisely as possible. For instance, genetic fine-mapping aims to discover causal genetic variants, but the strong local dependence structure of the genome makes it hard to identify the exact locations of those variants. So the statistical task is to output as many regions as possible and have those regions be as small as possible, while controlling how many outputted regions contain no signal. The same type of problem arises in any signal discovery application where signals cannot be perfectly localized, such as locating stars in astronomical sky surveys and change-point detection in time series. However, there are two competing objectives: maximizing the number of discoveries and minimizing the size of those discoveries (all while controlling false discoveries), so our first contribution is to propose a single unified measure we call the resolution-adjusted power that formally trades off these two objectives and hence, at least in principle, can be maximized subject to a constraint on false discoveries. We take a Bayesian approach, but the resulting constrained posterior optimization over candidate discovery regions is nonconvex and extremely high-dimensional. Thus our second contribution is Bayesian Linear Programming (BLiP), which uses linear programming to find a feasible solution (i.e., it controls false discoveries) that verifiably nearly maximizes the expected resolution-adjusted power. BLiP is remarkably computationally efficient and can wrap around any Bayesian model and algorithm for approximating the posterior distribution over signal locations. Applying BLiP on top of existing state-of-the-art Bayesian analyses of UK Biobank data (for genetic fine-mapping) and the Sloan Digital Sky Survey (for astronomical point source detection) increased the resolution-adjusted power by 30-120% with just a few minutes of computation. BLiP is implemented in the new packages pyblip (Python) and blipr (R).

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Knockoffs and multiple testing (with biomedical applications)
Start: 1:45 p.m. (30 minutes)

Evaluating treatment effects in group sequential multivariate longitudinal studies with covariate adjustment

Jeffries, Neal

Abstract

Jeffries, Troendle, and Geller (2018) investigated testing for a treatment difference in the setting of a randomized clinical trial with a single outcome measured longitudinally over a series of common follow-up times while adjusting for covariates. That paper examined the null hypothesis of no difference at any follow-up time versus the alternative of a difference for at least one follow-up time. We extend those results here by considering multivariate outcome measurements, where each individual outcome is examined at common follow-up times. We consider the case where there is interest in first testing for a treatment difference in a global function of the outcomes (e.g. weighted average or sum) with subsequent interest in examining the individual outcomes, should the global function show a treatment difference. Testing is conducted for each follow-up time and may be performed in the setting of a group sequential trial. Testing procedures are developed to determine follow-up times for which a global treatment difference exists and which individual combinations of outcome and follow-up time show evidence of a difference while controlling for multiplicity in outcomes, follow-up, and interim analyses. These approaches are examined in a study evaluating the effects of tissue plasminogen activator on longitudinally obtained stroke severity measurements.

This is joint work with James F. Troendle, and Nancy L. Geller.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Multiple testing methodology for group-sequential and platform trials
Start: 3:45 p.m. (30 minutes)

Application of multiple hypotheses testing procedures to market network analysis

Koldanov, Petr

Abstract

Detection of sets of pairwise strongly connected stocks is important problem in market network analysis. To handle the problem one can apply multiple hypotheses testing theory [4]. In the report methods for dividing conclusions on connections in a stock market by two parts: statistically significant (statistically reliable) conclusions and statistically unreliable ones are discussed. Let $\gamma_{i,j}$ be the measure of dependence between random variables X_i, X_j corresponding to the stocks of the market and γ_0 is a chosen threshold. Statistically significant conclusions on strong connections is considered as the set of rejected hypotheses $\{h_{n_{i,j}} : \gamma_{i,j} \leq \gamma_0 \ (\gamma_{i,j} < \gamma_0)\}$ by any multiple hypotheses testing procedure with FWER control in strong sense. Statistically unreliable conclusions on strong connections is considered as the set of accepted hypotheses $\{h_{e_{i,j}} : \gamma_{i,j} \geq \gamma_0 \ (\gamma_{i,j} > \gamma_0)\}$ by any multiple hypotheses testing procedure with FWER control in strong sense. Different problem statements and different multiple hypotheses testing procedures [2], [3], [1],[5] for simultaneous hypotheses testing $h_{n_{i,j}}$, $h_{e_{i,j}}$ are discussed. Example of stock market analysis is presented.

References:

[1] BAUER, P., HACKL, P., HOMMEL, G. and SONNEMANN, E. (1986). Multiple testing of pairs of one-sided hypotheses. Metrika, 33, 121-127.

[2] Gupta, S.S. MULTIPLE DECISION PROCEDURES. Theory and Methodology of Selecting and Ranking Populations/S.S. Gupta, S. Panchapakesan SIAM, 2002.

[3] H. Finner a, G. Giani Closed subset selection procedures for selecting good populations. Journal of Statistical Planning and Inference 38 (1994) 179-200.

[4] Kalyagin V. A., Koldanov A. P., Koldanov P., Pardalos P. M. Statistical Analysis of Graph Structures in Random Variable Networks. Springer, 2020.

[5] Guo W., Romano J.P. On Stepwise Control of Directional Errors under Independence and Some Dependence. Journal of statistical planning and inference, 2015, Vol.163, p.21-33.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Miscellaneous Topics (II)
Start: 9:50 a.m. (20 minutes)

Generalized SLOPE - Variable Selection in Linear Model

Kos, Michał

Abstract

We introduce a new estimator for the vector of coefficients in the linear model $y = Xb^* + e$.

The Generalized SLOPE (GS) estimator is defined in a following way:

$\hat{b}_{GS}(y, X, \lambda, U) = \arg\min_b [0.5||y - Xb||^2 + 0.5||Ub||^2 + l_1|b|_{t(1)} + ... + l_p|b|_{t(p)}]$

where U is a regularization matrix; l_1,...,l_p is a positive nonincreasing sequence; t is a permutation of a set $1, ..., p$ such that $|b|_{t(1)} \geq ... \geq |b|_{t(p)}$.

This procedure is a generalization of SLOPE and ELASTIC NET procedures. During the session we shall present new results illustrating that GS with properly chosen regularization parameters λ and U, controls FDR at level q, when explanatory variables are correlated.

This results generalize theorem 1.1 presented in [1], which proves that SLOPE with properly chosen λ, controls FDR at level q, for orthogonal design matrices ($X'X = I$).

Reference:

[1] Bogdan M. and van den Berg E. and Sabatti C. and Su W. and Candes E. J. (2015). SLOPE – adaptive variable selection via convex optimization. Annals of Applied Statistics, 1103–1140.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Statistical (Machine) Learning
Start: 4:05 p.m. (20 minutes)

What can multiple comparisons offer to machine learning?

Angelopoulos, Anastasios; Bates, Stephen; Candès, Emmanuel; Jordan, Michael; Lei, Lihua

Abstract

We introduce a framework for calibrating machine learning models so that their predictions satisfy explicit, finite-sample statistical guarantees using the graphical approach for familywise error rate (FWER) control. Our calibration algorithm works with any underlying model and (unknown) data-generating distribution and does not require model refitting. The framework addresses, among other examples, false discovery rate control in multi-label classification, intersection-over-union control in instance segmentation, and the simultaneous control of the type-1 error of outlier detection and confidence set coverage in classification or regression. Our main insight is to reframe the risk-control problem as multiple comparisons, enabling techniques and mathematical arguments different from those in the previous ML literature. We use our framework to provide new calibration methods for several core machine learning tasks with detailed worked examples in computer vision.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Statistical (Machine) Learning
Start: 4:45 p.m. (20 minutes)

GGM knockoff filter: False discovery rate control for Gaussian graphical models

Li, Jinzhou; Maathuis, Marloes H.

Abstract

We propose a new method to learn the structure of a Gaussian graphical model with finite sample false discovery rate control. Our method builds on the knockoff framework of Barber and Candès for linear models. We extend their approach to the graphical model setting by using a local (node-based) and a global (graph-based) step: we construct knockoffs and feature statistics for each node locally, and then solve a global optimization problem to determine a threshold for each node. We then estimate the neighbourhood of each node, by comparing its feature statistics to its threshold, resulting in our graph estimate. Our proposed method is very flexible, in the sense that there is freedom in the choice of knockoffs, feature statistics and the way in which the final graph estimate is obtained. For any given data set, it is not clear a priori what choices of these hyperparameters are optimal. We therefore use a sample-splitting-recycling procedure that first uses half of the samples to select the hyperparameters, and then learns the graph using all samples, in such a way that the finite sample FDR control still holds. We compare our method to several competitors in simulations and on a real data set.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Knockoffs and multiple testing (with biomedical applications)
Start: 2:35 p.m. (20 minutes)

Improving knockoffs with conditional calibration

Luo, Yixiang; Fithian, William; Lei, Lihua

Abstract

The knockoff filter of Barber and Candes (2015) is a flexible framework for multiple testing in supervised learning models, based on introducing synthetic predictor variables to control the false discovery rate (FDR). Using the conditional calibration framework of Fithian and Lei (2020), we introduce the *calibrated knockoff procedure*, a method that uniformly improves the power of any knockoff procedure. We implement our method for fixed-X knockoffs and show theoretically and empirically that the improvement is especially notable in two contexts where knockoff methods can be nearly powerless: when the rejection set is small, and when the structure of the design matrix prevents us from constructing good knockoff variables. In these contexts, calibrated knockoffs even outperform competing FDR-controlling methods like the (dependence-adjusted) Benjamini–Hochberg procedure in many scenarios.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Selective inference
Start: 9:50 p.m. (20 minutes)

False clustering rate control in a mixture model

Marandon-Carlhian, Ariane; Rebafka, Tabea; Roquain, Etienne; Sokolovska, Nataliya

Abstract

The clustering task consists in delivering labels to the members of a sample. For most data sets, some individuals are ambiguous and intrinsically difficult to attribute to one or another cluster. However, in practical applications, misclassifying individuals is potentially disastrous. To overcome this difficulty, the idea followed here is to classify only a part of the sample in order to obtain a small misclassification rate. This approach is well known in the supervised setting, and referred to as classification with an abstention option. The purpose of this paper is to revisit this approach in an unsupervised mixture-model framework. The problem is formalized in terms of controlling the false clustering rate (FCR) below a prescribed levelα. Hence, if for instance, α is chosen to be 5% and 100 items are finally chosen to be classified by the method, then the number of misclassified items is expected to be at most 5. This high interpretability is similar to the one of the false discovery rate (FDR) in multiple testing. We introduce new procedures that control the FCR while maximizing the expected number of classified items, which share similarities with the procedure of Sun and Cai (2007). An application to breast cancer data illustrates the benefits of the new approach from a practical viewpoint.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Statistical (Machine) Learning
Start: 3:45 p.m. (20 minutes)

Online multiple testing with super uniformity reward

Meah, Iqraa; Döhler, Sebastian; Roquain, Etienne

Abstract

Online multiple testing refers to the context where a possibly infinite number of hypotheses are tested, and the p-values are available one by one sequentially. This context differs from the usual one where the number of hypotheses to test $m < \infty$ is known beforehand, and the p-values are available together. The online methods proposed so far can suffer from a significant loss of power when the p-values are obtained from discrete tests. To resolve this issue, we introduce the method of super-uniformity reward that incorporates information about the individual null cumulative distribution functions. We prove that the rewarded procedures uniformly improve upon the non-rewarded ones while keeping the type I error control, and illustrate their performance on simulated and real data application of biology.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: On-line multiple testing
Start: 11:00 a.m. (20 minutes)

Controlling false discoveries under partial, nonparametric knowledge of dependence

Nguyen, Drew; Fithian, William

Abstract

We propose a new method, the Graphical Benjamini-Hochberg (GrBH) procedure, for finite-sample false discovery rate (FDR) control in multiple testing problems where we do not assume worst-case dependence of all the p-values on each other. Instead, our method leverages an analyst's knowledge that, for any given p-value, they can identify other p-values that can be safely assumed independent of the given one. This constitutes a constraint on the p-values' dependence graph. Correcting for worst-case dependence when it is unnecessary makes FDR-controlling procedures very conservative, so our procedure calibrates a rejection threshold for each hypothesis' p-value, only using the data through those other p-values which were assumed independent. In our method, we find the conservativeness is lost when most p-values are independent of each other, which is the case in genome-wide association studies (GWAS) for genetic markers located physically far away. In this "short-range" dependence setting, simulation examples show that GrBH performs favorably compared to corrections for worst-case dependence, and that an improved version of GrBH performs just as well as the Benjamini-Hochberg (BH) procedure, which assumes independence.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Multiple testing in the presence of structures
Start: 9:20 a.m. (20 minutes)

Weighted Posterior Odds: A Data Summary for Decision Making

Pennello, Gene Anthony

Abstract

Under the class of linear loss functions in which k is the loss of falsely accepting an hypothesis A (type 1 error) relative to the loss of falsely not accepting it (type 2 error), we show that the weighted posterior odds (WPO) in favor of A can be interpreted as that value of k at which the Bayes rule is indifferent to making either decision. That is, the Bayes rule is to accept A if WPO $>$ k. WPO facilitates decision making because it is on the same scale as the trade-off decision makers must face between the consequences of falsely accepting A and falsely not accepting it. Thus, WPO is an attractive alternative to the p value as a summary measure of evidence. WPO can be used as the basis for constructing a decision-theoretic interval estimate for the parameter of A. We derive formulas for WPO and the interval estimator under linear losses for hypotheses on a mean or a mean difference in one- or two-sample normally distributed data. Under additive linear losses for multiple hypothesis tests, the Bayes rule is a comparisonwise test procedure, that is, the rule is simply to compare the WPO for each hypothesis to the k assigned to it.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Lower bounds for type I / II error tradeoffs
Start: 11:00 a.m. (20 minutes)

Selective inference for false discovery proportion in a Hidden Markov Model

Perrot-Dockès, Marie; Blanchard, Gilles; Neuvial, Pierre; Roquain, Etienne

Abstract

We address the multiple testing problem under the assumption that the true/false hypotheses are driven by a Hidden Markov Model (HMM), which is recognized as a fundamental setting to model multiple testing under dependence since the seminal work of Sun and Cai (2009). While previous work has concentrated on deriving specific procedures with a controlled False Discovery Rate (FDR) under this model, following a recent trend in selective inference, we consider the problem of establishing confidence bounds on the false discovery proportion (FDP), for a user-selected set of hypotheses that can depend on the observed data in an arbitrary way. We develop a methodology to construct such confidence bounds first when the HMM model is known, then when its parameters are unknown and estimated, including the data distribution under the null and the alternative, using a nonparametric approach. In the latter case, we propose a bootstrap-based methodology to take into account the effect of parameter estimation error. We show that taking advantage of the assumed HMM structure allows for a substantial improvement of confidence bound sharpness over existing agnostic (structure-free) methods, as witnessed both via numerical experiments and real data examples.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Miscellaneous Topics (I)
Start: 2:55 p.m. (20 minutes)

Doubly-sequential experimentation

Ramdas, Aaditya

Abstract

If one zooms out appropriately, both scientific inquiry and industry research can be thought of as being a sequence of sequential experiments. We will examine what type of guarantees we may desire both within and across experiments, and design unified frameworks for achieving these. Within experiments, one needs to construct anytime-valid p-values or e-values (via confidence sequences or e-processes) in order to ensure correct inference for a single experiment at data-dependent stopping time. Across experiments, one can use online algorithms for controlling the false discovery rate or false coverage rate. These modular pieces fit together perfectly even if the start and end times for each experiment are not synchronized in any way across experiments.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: On-line multiple testing
Start: 10:30 a.m. (30 minutes)

Subgroup selection with a multivariate isotonic regression function

Reeve, Henry William Joseph; Muller, Manuel; Cannings, Timothy; Samworth, Richard

Abstract

In this work we consider the challenge of subgroup selection with a multivariate isotonic regression function. The goal here is to output a large subset of the feature space which satisfies the following Type 1 error guarantee: On the selected set, the regression function (the conditional expectation of the response given the covariates) is uniformly above some pre-specified threshold, with high probability. In previous work, we have demonstrated the extent of the challenge by providing minimax lower bounds on the regret for classes of distributions with smooth regression functions. In this work we consider a more optimistic perspective by imposing additional structure. In particular, we consider a setting in which the regression function is monotonic with respect to a natural partial ordering on the feature space. In this setting we build upon ideas from the multiple testing literature to construct an adaptive procedure with a near-optimal, high probability regret guarantee. This is joint work with Manuel Muller (Cambridge), Richard Samworth (Cambridge) and Timothy Cannings (Cambridge).

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Multiple testing in the presence of structures
Start: 9:00 a.m. (20 minutes)

Derandomizing Knockoffs

Ren, Zhimei; Wei, Yuting; Candès, Emmanuel

Abstract

Model-X knockoffs is a general procedure that can leverage any feature importance measure to produce a variable selection algorithm, which discovers true effects while rigorously controlling the number or fraction of false positives. Model-X knockoffs is a randomized procedure which relies on the one-time construction of synthetic (random) variables. This paper introduces a derandomization method by aggregating the selection results across multiple runs of the knockoffs algorithm. The derandomization step is designed to be flexible and can be adapted to any variable selection base procedure to yield stable decisions without compromising statistical power. When applied to the base procedure of Janson and Su (2016), we prove that derandomized knockoffs controls both the per family error rate (PFER) and the k family-wise error rate (k-FWER). Further, we carry out extensive numerical studies demonstrating tight type-I error control and markedly enhanced power when compared with alternative variable selection algorithms. Finally, we apply our approach to multi-stage genome-wide association studies of prostate cancer and report locations on the genome that are significantly associated with the disease. When cross-referenced with other studies, we find that the reported associations have been replicated.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Knockoffs and multiple testing (with biomedical applications)
Start: 2:15 p.m. (20 minutes)

Knowing the signs: a sensible formulation of tests, and multiple tests

Rice, Kenneth

Abstract

For real-valued parameters, many controversial issues in statistical testing can be easily resolved if (following e.g. Tukey) we view the test as a decision about just the sign of the parameter - where we either assert it is positive, negative, or say nothing either way. These are not the usual "accept/reject" choices, but the sign-decision approach nevertheless gives simple motivations for many familiar ideas, including two-sided tests at level α, p-values, credible sets, 80% power as a threshold for a not-too-risky study design, using level $\alpha = 0.005$ not 0.05 to improve confidence in results, and post-hoc power calculations being a waste of effort. Extending the framework to multiple sign-decisions, one can recover simple motivations for Bonferroni correction (through both familywise-error and expected number of errors) and the Benjamini-Hochberg algorithm. We briefly review this framework, before providing recent results on its use with multiple tests.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Lower bounds for type I / II error tradeoffs
Start: 10:30 a.m. (30 minutes)

Multiplicity-adjusted confidence intervals for conditional prediction performance measures

Rink, Pascal; Brannath, Werner

Abstract

In machine learning, the selection of a promising model from a potentially large number of competing models and the assessment of its prediction performance are critical tasks that need careful consideration, for instance when trying to establish a medical diagnosis or prognosis rule. Cross-validation and data splitting are both well-established approaches how to come up with a final model and give rise to prediction performance estimates. However, in order to be able to report a valid confidence interval for the conditional prediction performance estimate, i.e., a confidence interval that has coverage at least at the nominal level for the performance estimate of the model fit on the data at hand, the finally selected model usually needs to be evaluated on an unseen and independent test set. In this talk, we will instead propose to perform the model selection and the confidence interval estimation on the very same data. In this way, a larger fraction of the data at hand can be used for training, which can lead to better performing prediction models, especially in machine learning setups where cross-validation is not feasible. We obtain valid confidence intervals for the conditional performance estimate by using the bootstrap, exponential tilting, and a multiplicity correction. As an input to our algorithm we only need the true labels and the out-of-training predictions of all the competing models; no additional training is needed. The predictions may result from varied model hyperparameters or even different learning algorithms. Also, the proposed method yields valid confidence intervals for common prediction performance estimates, such as the classification accuracy or the area under the receiver operating characteristic curve.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Statistical (Machine) Learning
Start: 4:25 p.m. (20 minutes)

onlineFDR: An R package and Shiny app for online error rate control

Robertson, David

Abstract

In this talk, we introduce the R package onlineFDR, which allows users to control the FDR or FWER for online hypothesis testing. In this framework, a null hypothesis is rejected based only on the previous decisions, as the future p-values and the number of hypotheses to be tested are unknown. onlineFDR aims to be a comprehensive software resource for online error rate control, and includes functions for fully sequentially, asynchronous and batched testing. We also introduce Shiny apps which allow user-friendly ways of exploring different online hypothesis testing algorithms.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Software developments
Start: 11:30 a.m. (30 minutes)

Machine learning meets false discovery rate

Lei, Lihua; Marandon, Ariane; Mary, David; Roquain, Etienne

Abstract

Classical false discovery rate (FDR) controlling procedures offer strong and interpretable guarantees, while they often lack of flexibility. On the other hand, recent machine learning classification algorithms, as those based on random forest (RF) or neural networks (NN), have great practical performances but lack of interpretation and of theoretical guarantees. In this paper, we make these two meet by introducing a new adaptive novelty detection procedure with FDR control, called AdaDetect. It extends the scope of recent works of multiple testing literature to the high dimensional setting, notably the one in Yang et al (2021). AdaDetect is shown to both control strongly the FDR and to have a power that mimics the one of the oracle in a specific sense. The interest and validity of our approach is demonstrated with theoretical results, numerical experiments on several benchmark datasets and with an application to astrophysical data. It is in particular shown that, while AdaDetect can be used in combination with any classifier, it is particularly efficient when combined with RF and NN methods. Overall, this work is at the crossroad of multiple testing, conformal prediction and machine learning.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Adaptive inference
Start: 9:50 a.m. (20 minutes)

Confidence bounds for the number of positive effects

Solari, Aldo; Heller, Ruth

Abstract

For n unknown parameters, we are often interested in inference on the number of parameters with positive sign. For example, the number of studies in a meta-analysis with positive effects. In this work we provide confidence bounds for the number of positive effects. Providing a lower bound is related to the replicability goal of establishing that the effect was discovered in at least r of n studies. We show that providing an upper bound in addition to the lower bound comes with no extra cost. If the upper bound is low, it conveys the limit on replicability. The confidence bounds are obtained by sequential testing of partial conjunction hypotheses based on combination tests, e.g. Fisher's method. When the parameters are expected to have mixed signs, we consider combination tests based on conditional p-values. Finally, we derive simultaneous confidence bounds for the number of positive effects in any subset of studies by using the partitioning principle, and we provide a shortcut to allow fast computation.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Replicability and Reproducibility (II)
Start: 9:30 a.m. (30 minutes)

The edge of discovery: Controlling the local false discovery rate at the margin

Soloff, Jake

Abstract

Despite the popularity of the false discovery rate (FDR) as an error control metric for large-scale multiple testing, its close Bayesian counterpart the local false discovery rate (lfdr), defined as the posterior probability that a particular null hypothesis is false, is a more directly relevant standard for justifying and interpreting individual rejections. However, the lfdr is difficult to work with in small samples, as the prior distribution is typically unknown. We propose a simple multiple testing procedure and prove that it controls the expectation of the maximum lfdr across all rejections; equivalently, it controls the probability that the rejection with the largest p-value is a false discovery. Our method operates without knowledge of the prior, assuming only that the p-value density is uniform under the null and decreasing under the alternative. We also show that our method asymptotically implements the oracle Bayes procedure for a weighted classification risk, optimally trading off between false positives and false negatives. We derive the limiting distribution of the attained maximum lfdr over the rejections, and the limiting empirical Bayes regret relative to the oracle procedure.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Empirical Bayes
Start: 11:20 a.m. (25 minutes)

Powerful and efficient knockoffs with knockpy

Spector, Asher Max; Janson, Lucas; Fithian, Will

Abstract

Model-X knockoffs allows analysts to perform feature selection using almost any machine learning algorithm while provably guaranteeing exact FDR control. A wide array of work has shown knockoffs to be a very powerful methodology overall, but some recent work has shown that the "gold standard" methods for applying knockoffs often unnecessarily lose power—indeed, they can be provably powerless in certain settings. Furthermore, naive implementations of knockoffs can be computationally expensive and technically challenging to implement. With this motivation, this talk will review recent developments in the knockoffs literature which can (provably) boost the power of knockoffs, and furthermore, it will introduce knockpy, a unified framework for knockoffs-based inference in python which implements these state-of-the-art procedures. In as little as one line of code, knockpy can call a wide variety of methods from the knockoffs-based literature, and furthermore, it scales efficiently to extremely high-dimensional problems. Overall, knockpy is designed to allow practitioners to use knockoffs in a way that is easy, computationally efficient, and consistently powerful. This talk is based on two joint works, one with Lucas Janson and another with Will Fithian.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Knockoffs and multiple testing (with biomedical applications)
Start: 2:55 p.m. (20 minutes)

Nonparametric methods for clustered data in the several sample case

Sprünken, Erin

Abstract

In many trials and experiments, subjects are not only observed once, but multiple times, resulting in a cluster of possibly correlated observations. For example, mice sharing the same cage or students of the same class are typical examples of clustered data.

Typically, under the assumption of normally distributed data, mixed models are used for analysis.

However, this model assumption is rather strict and hard to justify in most real data analyses. Furthermore, skewed data (e.g. waiting times), discrete data (e.g. count data) or ordered categorical data measured on an ordinal scale are typical endpoints in a variety of trials. This motivates the use of nonparametric methods which do not rely on any specific data distribution.

For the two-sample case, several nonparametric procedures exist. For binary clustered data, a chi-square-test for contingency tables can be used. Furthermore, generalizations of the Wilcoxon-Mann-Whitney-test exist for testing the null hypothesis of equal distributions of clustered data. An extension is provided by a procedure under a less strict null hypothesis formulated in terms of the Wilcoxon-Mann-Whitney effect.

In the present talk, we aim to generalize the procedures for the analysis of several samples. Thus, we propose a general nonparametric framework for comparing multiple groups of clustered data under mild assumptions. We present different inference methods, namely ANOVA-type test statistics and a multiple contrast test procedure and investigate their asymptotic behavior. Extensive simulation studies indicate that the methods control the type-1 error level well, even with small sample sizes. A real data example illustrates the application of the proposed methods.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Miscellaneous Topics (II)
Start: 9:00 a.m. (20 minutes)

Sparse Recovery With Multiple Data Streams: A Sequential Adaptive Testing Approach

Sun, Wenguang

Abstract

Multistage design has been used in a wide range of scientific fields. By allocating sensing resources adaptively, one can effectively eliminate null locations and localize signals with a smaller study budget. We formulate a decision-theoretic framework for simultaneous multi-stage adaptive testing and study how to minimize the total number of measurements while meeting pre-specified constraints on both the false positive rate and missed discovery rate. The new procedure, which effectively pools information across individual tests using a simultaneous multistage adaptive ranking and thresholding approach, achieves precise error rates control and leads to great savings in total study costs. The performance of the method is investigated through simulation and illustrated through large-scale A/B tests and high-throughput screening.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Lower bounds for type I / II error tradeoffs
Start: 11:20 a.m. (20 minutes)

Testing Primary and Secondary Endpoints in Group Sequential Clinical Trials

Tamhane, Ajit

Abstract

In the first half of the talk I will review our work over the last decade on the problem of testing primary and secondary endpoints subject to a gatekeeping constraint in a group sequential clinical trial. This work deals with a single primary and a single secondary endpoint. In the second half of the talk I will discuss our current work on multiple primary and secondary endpoints, focusing on two secondary endpoints.

References:

1. Tamhane, A. C., Mehta, C. R. and Liu, L. (2010), "Testing a primary and a secondary endpoint in a group sequential design," Biometrics, 66, 1174-1184.

2. Tamhane, A. C., Wu, Y. and Mehta, C. R. (2012a), "Adaptive extensions of a two-stage group sequential procedure for testing a primary and a secondary endpoint (I): Unknown correlation between the endpoints," Statistics in Medicine, 31, 2027-2040.

3. Tamhane, A. C., Wu, Y. and Mehta, C. R. (2012b), "Adaptive extensions of a two-stage group sequential procedure for testing a primary and a secondary endpoint (II): Sample size re-estimation, Statistics in Medicine, 31, 2041-2054.

4. Tamhane, A. C., Gou, J., Jennison, C. Mehta, C. R. and Curto, T. (2018), "A gatekeeping procedure for testing a primary and a secondary endpoint in a group sequential design with multiple interim looks," Biometrics, 74. 40-48.

5. Tamhane, A. C., Xi, D. and Gou, J. (2022), "Testing one primary and two secondary endpoints in a two-stage group sequential trial with extensions." In preparation.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Opening session
Start: 9:00 a.m. (60 minutes)

Semiparametric Methods for Covariate Adjustment for Net Benefit Effect Sizes for Multiple Outcomes

Thas, Olivier

Abstract

Generalised pairwise comparisons (GPC) is gaining more and more attention as an effect size in clinical trials. It can take several forms (e.g net benefit, win ratio, win odds, probabilistic index) and can be defined for a single outcome as well as for multiple outcomes. The estimation of this effect size, and its properties, is still an ongoing research area, and correcting the GPC effect size for covariates has not yet attracted much attention. We have developed flexible semiparametric methods for analysing the net benefit with adjustment for baseline covariates. These methods are based on Probabilstic Index Models and they are easy to implement. In this talk, we will outline the construction of the methods and demonstrate them on a case study.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Recent Advances in Generalized Pairwise Comparisons of Multiple Prioritized Endpoints
Start: 9:00 a.m. (20 minutes)

Permutation-based true discovery guarantee by sum tests

Vesely, Anna; Finos, Livio; Goeman, Jelle J.

Abstract

Sum-based global tests are highly popular in multiple hypothesis testing. We propose a general closed testing procedure for sum tests, which provides lower confidence bounds for the proportion of true discoveries (TDP), simultaneously over all subsets of hypotheses; these simultaneous inferences come for free, i.e., without any adjustment of the alpha-level, whenever a global test is used. Our method allows for an exploratory approach, as simultaneity ensures control of the TDP even when the subset of interest is selected post hoc. It adapts to the unknown joint distribution of the data through permutation testing. Any sum test may be employed, depending on the desired power properties. We present an iterative shortcut for the closed testing procedure, based on the branch and bound algorithm, which converges to the full closed testing results, often after few iterations; even if it is stopped early, it controls the TDP. We compare the properties of different choices for the sum test through simulations, then we illustrate the feasibility of the method for high dimensional data on brain imaging and genomics data.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Post-hoc FDP control (II)
Start: 9:10 a.m. (20 minutes)

New tools for post-hoc FDP control

Vovk, Vladimir

Abstract

I will start from a brief review of two alternative tools for statistical hypothesis testing, p-values and e-values. Both have been used in the algorithmic theory of randomness for decades (on the log scale and under other names), but only p-values are widely used in non-Bayesian statistics; e-values are related to Bayes factors, especially in the case of a simple null hypothesis. My plan is to compare and contrast two versions of post-hoc FDP control, one based on p-values and the other on e-values. The advantage of e-values is that they are easy to combine. This makes them a powerful tool for post-hoc FDP control: once we have a testing procedure for each of the elementary hypotheses, we can construct a multiple testing procedure performing efficient post-hoc FDP control. Traditionally, post-hoc FDP control is performed using p-values, or even rejections at a fixed significance level (corresponding to comparing p-values to a fixed threshold). Recent results on the admissibility of procedures for combining p-values also make post-hoc FDP control more efficient. If we assume that the testing procedures for the elementary hypotheses are independent, p-values become easier to combine than e-values, but I will argue that e-values still remain useful for post-hoc FDP control.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Post-hoc FDP control (I)
Start: 11:20 a.m. (30 minutes)

Score Attack: A Lower Bound Technique for Differentially Private Estimation

Wang, Yichen

Abstract

We introduce a general technique, the score attack, for lower bounding the differential-privacy-constrained minimax risk of parameter estimation.

Inspired by the tracing attack idea in differential privacy, the score attack method is applicable to any statistical model with a well-defined score statistic and capable of optimally lower bounding the minimax risk of estimating unknown model parameters when the estimator is required to be differentially private.

The effectiveness of this general method is demonstrated in a variety of examples: the generalized linear model in classical and high-dimensional sparse settings, the Bradley-Terry-Luce model for pairwise comparisons, and non-parametric function estimation in the Sobolev class.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Adaptive inference
Start: 9:00 a.m. (30 minutes)

rpact: an R Package for Confirmatory Adaptive Group Sequential Designs

Wassmer, Gernot; Pahlke, Friedrich

Abstract

Starting in 2018, we have been developing "rpact", an open source R package for the design evaluation and analysis of confirmatory adaptive clinical trials. The idea was to cover with this new R package the methodology as described in the monograph on group sequential and confirmatory adaptive design in clinical trials by Wassmer and Brannath (Springer, 2016). We published the first official release of "rpact" on CRAN in October 2018. The current version 3.2 of rpact covers a wide range of adaptive methodology including sample size reassessment procedues, adaptive multi-arm multi-stage designs, and population enrichment designs. Design characteristics af adaptive strategies can be assessed by extensive simulation functions. rpact also serves as software for planning and analysing classical group sequential designs as well as a sample size calculator for fixed sample sizes. In our talk, we describe the basic features of the current version of rpact and illustrate them by examples.

Date and time

Day 2 (Thursday, September 1, 2022)
Session: Software developments
Start: 10:30 a.m. (30 minutes)

Optimality in statistical inference for permutation invariant problems

Weinstein, Asaf

Abstract

I consider simultaneous inference problems that are invariant under permutations, meaning that all components of the problem are oblivious to the labelling of the multiple instances under consideration. For any such problem I identify the optimal solution which is itself permutation invariant, the most natural condition one could impose on the the set of candidate solutions. Interpreted differently, for any possible value of the parameter I find a tight (non-asymptotic) lower bound on the statistical performance of any procedure that obeys the aforementioned condition. By generalizing the standard decision theoretic notions of permutation invariance, I show that the results apply to a myriad of popular problems in simultaneous inference, so that the ultimate benchmark for each of these problems is established. The connection to the nonparametric empirical Bayes approach of Robbins is discussed in the context of asymptotic attainability of the bound uniformly in the parameter value.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Empirical Bayes
Start: 11:45 a.m. (25 minutes)

Testing at the edge: A frequentist perspective on the local false discovery rate

Xiang, Daniel; Soloff, Jake; Fithian, Will

Abstract

The two-groups model is a Bayesian multiple testing framework that models the truth status of hypotheses as random. Surprisingly, procedures designed to estimate the local false discovery rate (lfdr, Efron et al., 2001) can have desirable frequentist properties, such as FDR control for fixed configurations of hypotheses (Lei and Fithian 2018). We propose a definition of the lfdr for the fixed-effects model, providing a frequentist interpretation for inferences made without the Bayes assumption. Viewing the truth status of each hypothesis as fixed, lfdr is the long-run frequency, in a sequence of repeated multiple testing experiments, that a randomly selected hypothesis is false conditional on the corresponding test statistic. When there are many hypotheses to test, selecting one uniformly at random relates the compound decision problem to a two-groups model in which frequentist and Bayesian local false discovery rates coincide. While FDR (Benjamini and Hochberg 1995) based procedures control the rate of false discoveries on average, some rejections may be recognizably worse than others. We propose instead to control the recognizable false discovery rate (RFDR), defined for a multiple testing method as the long-run frequency of its least significant discovery being false.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Selective Inference
Start: 10:10 a.m. (20 minutes)

Multiple conditional randomization tests

Zhang, Yao; Zhao, Qingyuan

Abstract

We establish a general sufficient condition for constructing multiple "nearly independent" conditional randomization tests, in the sense that the joint distribution of their p-values is almost uniform under the global null. This property implies that the tests are jointly valid and can be combined using standard methods. Our theory generalizes existing techniques in the literature that use independent treatments, sequential treatments, or post-randomization, to construct multiple randomization tests. In particular, it places no condition on the experimental design, allowing for arbitrary treatment variables, assignment mechanisms and unit interference. This framework's flexibility is illustrated by developing conditional randomization tests for lagged treatment effects in stepped-wedge randomized controlled trials. A weighted Z-score test is further proposed to maximize the power when the tests are combined. We compare the efficiency and robustness of the commonly used mixed-effect models and the proposed conditional randomization tests using simulated experiments and real trial data.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Miscellaneous Topics (I)
Start: 1:45 p.m. (30 minutes)

Model-free Multiple Testing using Mirror Statistics (MMM)

Zhao, Zhigen

Abstract

We consider the general regression analysis to study the relation between a univariate response and a p-dimensional covariate. We assume the general multi-index model with unknown link function. It is assumed that the response depends on the covariate via some linear combinations, which is characterized by the central subspace. For each covariate, we want to test the hypothesis whether this covariate plays any role in the central subspace. In this paper, we combine the idea of the sufficient dimension reduction methods and the Gaussian mirror to construct the MMM methods, standing for Model-free Multiple Testing using Mirror Statistics. It is shown that MMM controls the FDR at a desired level asymptotically. Numerically evidence has shown that that MMM is much more powerful than all its alternatives.

Date and time

Day 1 (Wednesday, August 31, 2022)
Session: Empirical Bayes
Start: 12:10 p.m. (25 minutes)

Locally Simultaneous Inference

Zrnic, Tijana; Fithian, Will

Abstract

Selective inference controls type I error for statistical questions, such as hypotheses to test or confidence intervals to construct, that are selected in a data-driven manner. A standard solution to selective inference is *simultaneous inference*, which delivers valid answers to the set of all questions that could possibly have been asked. However, simultaneous inference can be unnecessarily conservative if this set includes many questions that were unlikely to be asked in the first place.

In this talk I will discuss a less conservative approach that we call *locally simultaneous inference*, which only answers those questions that could *plausibly* have been asked, in light of the observed data. For example, if we construct an interval for the "winning" treatment effect in a clinical trial with multiple treatments, and it is obvious in hindsight that only one treatment had a chance to win, then our approach will return an interval that is nearly the same as the uncorrected, "naive" interval. Along with a general framework of locally simultaneous inference, I will discuss several applications including inference on the winning treatment and model selection via the LASSO.

Date and time

Day 3 (Friday, September 2, 2022)
Session: Selective Inference
Start: 9:30 a.m. (20 minutes)

Poster Presentations

Haplotype based testing for a better understanding of the selective architecture

Chen, Haoyu; Pelizzola, Marta; Futschik, Andreas

Abstract

The identification of genomic regions affected by selection is an important goal in population genetics. If temporal data are available, allele frequency changes at SNP positions are often used. When a large number of SNP positions is tested, a multiple testing correction is needed to avoid false positives due to genetic drift. Here we provide a new testing approach that uses haplotype frequencies instead of allele frequencies. With this approach less multiple testing correction is needed, which leads to tests with higher power, especially when the number of candidate haplotypes is small or moderate. For a larger number of haplotypes, we propose haplotype block based methods. The use of haplotypes also permits for a better understanding of selective signatures. For this purpose we propose post-hoc tests for the number and difference in strength of the selected haplotypes.

IIDEA: Interactive Inference for Differential Expression Analyses

Enjalbert Courrech, Nicolas; Neuvial, Pierre

Abstract

Differential gene expression studies aim at identifying genes whose mean expression level differs significantly between two known populations. The state of the art approach to this problem consists in performing one test per gene, followed by a multiple testing correction in order to control the False Discovery Rate (FDR), that is, the expected proportion of errors among selected genes. The obtained gene list is then typically refined by further selecting genes with a large effect size (as in volcano plots) or belonging to a specific gene set or pathway. Unfortunately, such data-driven or multiple selections can invalidate FDR control [1].

Recent statistical developments in post hoc inference, pioneered by [2], make it possible to obtain valid statistical guarantees for such gene lists. We have developed IIDEA, an interactive R/shiny application that implements post hoc inference methods for differential expression studies developed by [3]. Because these methods build on permutation approaches, they are able to adapt to the dependency structure observed in a given data set, as illustrated in [4]. IIDEA makes it possible for users to interactively select genes of interest (either from a volcano plot or corresponding to a particular gene set) and obtain valid statistical guarantees regarding the number of true/false positives among these genes.

Links:

- application: `https://shiny-iidea-sanssouci.apps.math.cnrs.fr/`

- source code: `https://github.com/pneuvial/sanssouci/tree/develop/inst/shiny-examples/volcano-plot`

- web page of the SansSouci package: `https://pneuvial.github.io/sanssouci/`

References:

[1] M. Ebrahimpoor and J. J. Goeman, "Inflated false discovery rate due to volcano plots: problem and solutions", Briefings in Bioinformatics, 2021.

[2] J. J. Goeman and A. Solari, "Multiple testing for exploratory research", Statistical Science, vol. 26, no. 4, pp. 584–597, 2011.

[3] G. Blanchard, P. Neuvial, and E. Roquain, "Post hoc confidence bounds on false positives using reference families", Annals of Statistics, vol. 48, no. 3, pp. 1281–1303, 2020.

[4] N. Enjalbert-Courrech and P. Neuvial, "Powerful and interpretable control of false discoveries in differential expression studies". bioRxiv `https://www.biorxiv.org/content/10.1101/2022.03.08.483449v1`

What implications do analysis choices have on study results?

Krause, Linda; Zapf, Antonia

Abstract

Which variables should we include in the regression analysis as adjusting variables? How do we handle missing data? Using which criteria should we define the binary endpoint? These and more are questions we as statisticians often have to discuss when consulting clinicians who want to plan or analyse clinical trials or observational studies. In clinical trials, those decisions have to be made during writing of the study protocol before collection of any data or at latest during formulating the statistical analysis plan before database lock. In observational studies those choices are often made after data collection.

We aim to systematically investigate the implications of analysis choices on study results in the sense of a robustness analysis using one clinical trial and one cross-sectional study as examples. In the cross-sectional study we examine the influence of covariate sets to include in the regression analyses for adjustment. In the clinical trial we explore the impact of derivation of binary endpoints from numerical values. In both studies we investigate handling of missing data (single or multiple imputation). We inspect the implications on study results by comparing estimators and confidence intervals for all possible combinations of potential analysis choices. Depending on the respective study and hypothesis, we aim to combine the different results to obtain one answer to the initial medical questions or to show that such a combination is not sensible. Ideas on how to incorporate adjusting for multiple testing in this step are proposed and discussed. We visualize all results in a joint manner in an interactive R Shiny app to assist clinicians and collaborators in understanding the potential implications of analysis choices on study results and their influence on study interpretation.

Data fission: splitting a single data point

Leiner, James

Abstract

Suppose we observe a random vector X from some distribution P in a known family with unknown parameters. We ask the following question: when is it possible to split X into two parts $f(X)$ and $g(X)$ such that neither part is sufficient to reconstruct X by itself, but both together can recover X fully, and the joint distribution of $(f(X), g(X))$ is tractable? As one example, if $X = (X_1, \ldots, X_n)$ $X = (X_1, \ldots, X_n)$and P is a product distribution, then for any $m < n$, we can split the sample to define $f(X) = (X_1, \ldots, X_m)$ and $g(X) = (X_{m+1}, \ldots, X_n)$. Rasines and Young (2021) offers an alternative route of accomplishing this task through randomization of X with additive Gaussian noise which enables post-selection inference in finite samples for Gaussian distributed data and asymptotically for non-Gaussian additive models. In this paper, we offer a more general methodology for achieving such a split in finite samples by borrowing ideas from Bayesian inference to yield a (frequentist) solution that can be viewed as a continuous analog of data splitting. We call our method data fission, as an alternative to data splitting, data carving and p-value masking. We exemplify the method on a few prototypical applications, such as post-selection inference for trend filtering and other regression problems.

Controlling the false discovery rate under dependency with the adaptively weighted bh procedure

Lin, Mengqi; Fithian, William

Abstract

We introduce a generic adaptively weighted, covariate-assisted multiple testing methods for finite-sample false discovery rate (FDR) control with dependent test statistics where the dependence could be arbitrary. Our approach uses conditional calibration to tackle dependency between test statistics and make use of the conditional statistics to learn adaptive weights without violating FDR control. We also propose a provably optimal weights and a concrete algorithm to approximate. Together with the conditional calibration, our adaptively optimal weighted procedure controls FDR while manifesting great power. Additionally for fixed weights, our procedure dominates the traditional weighted BH procedures under positive dependence and dominates the general weighted step-up procedures under arbitrary dependence. Simulations illustrate significant power gains over competing approaches to FDR control under dependence.

Hommel BH: an adaptive Benjamini-Hochberg procedure using Hommel's estimator for the number of true hypotheses

Magnani, Chiara Gaia; Goeman, Jelle J.; Solari, Aldo

Abstract

We propose an adaptive Benjamini and Hochberg procedure for control of the false discovery rate (FDR) by using Hommel's estimator for the number of true hypotheses. We show that the proposed procedure (HBH) has FDR control under the assumptions of independence of p-values and positive regression dependence within nulls (PRDN).

Under independence, we derive a closed-form expression for the maximum FDR of our procedure as a function of the significance level and the number of hypotheses. It turns out that the adjustment needed in order to control the FDR at the desired level is negligible.

Under PRDN, we find an upper bound for FDR of our procedure, and the price to pay for using the adaptive procedure is modest.

We illustrate the new method with an application to two well-known examples from the literature.

Combining Multiple Testing with Multivariate Singular Spectrum Analysis

Movahedifar, Maryam; Dickhaus, Thorsten

Abstract

Appropriate preprocessing is a fundamental prerequisite for analyzing a noisy dataset. The purpose of this paper is to apply a nonparametric preprocessing method, called Singular Spectrum Analysis (SSA), to a variety of datasets which are subsequently analyzed by means of multiple statistical hypothesis tests. SSA is a nonparametric preprocessing method which has recently been utilized in the context of many life science problems. In the present work, SSA is compared with three other state-of-the-art preprocessing methods in terms of goodness of denoising and in terms of the statistical power of the subsequent multiple test. These other methods are either parametric or nonparametric. Our findings demonstrate that (multivariate) SSA can be taken into account as a promising method to reduce noise, to extract the main signal from noisy data, and to detect statistically significant signal components.

Multiple testing procedure for crack detection in concrete data

Nguyen, Duc

Abstract

Concrete materials are widely-used in constructions, and maintenance of concrete structures plays an important role. Therefore, crack segmentation for several types of concrete is one of the most important tasks. Our goal is to preindentify regions containing cracks for large 3D input images in linear time using multiple hypotheses procedure with CUSUM test statistic. Based on several geometry attributes in each region, we study the behavior of the tail probability of the test statistics for m-dependent multivariate random fields. To this end, we compare the false discovery proportion and the power of our test with other well-known procedures such as Benjamini-Hochberg or Hochberg-Yekutieli procedures under the independence assumption.

Randomized p-values in Binomial Models and in Group-Testing

Odipo, Daniel Ochieng; Hoang, Anh-Tuan; Dickhaus, Thorsten

Abstract

When screening for rare diseases in large populations, conducting individual tests can be expensive and time-consuming. In group-testing, individuals are pooled and tested together. If a group tests negative, then all the individuals in that group are declared negative. Otherwise, it is concluded that at least one individual in that group is positive. Group-testing can be used for several objectives such as to classify the individuals with respect to their disease status, to estimate the prevalence in the target population, or to conduct a hypothesis test on the unknown prevalence. In this work, we are concerned with hypothesis test problems regarding the unknown prevalence. We consider the case when the population is stratified leading to a multiple testing problem. We define two-stage randomized p-values first under a general binomial model and then for a model pertaining to the proportion of positive individuals in group-testing. Randomized p-values are less conservative compared to non-randomized p-values under the null hypothesis, but they are stochastically not smaller under the alternative. We show that the proposed randomized p-values are valid in the binomial model. Testing individuals in pools for a fixed number of tests improves the power of the tests based on the p-values. The power of the tests based on randomized p-values as a function of the sample size is also investigated. Simulation studies and real data analysis are used to compare and analyze the different variants of the considered p-values.

Machine Learning Techniques for Breast Cancer Risk Prediction: Comparison with the BCRAT

Parcali, Berfu; Mutlu, Fezan

Abstract

Early diagnosis of breast cancer increases the number of possible treatments, the success rate of the treatments and the chance of survival.The aim of this work is compare machine learning methods in breast cancer risk assessment based on the BCRAT (Gail 2 Model). BCRAT is a well accepted cancer risk assessment model which evaluates the main factors in breast cancer. Random Forest(RF), Support Vector Machine(SVM), k-Nearest Neighbor(k-NN), Artificial Neural Network(ANN), Naive Bayes(NB) algorithm with the purpose of breast cancer risk assessment. Firstly, risk factor was determined using BCRAT method on the dataset in RStudio programme, then dataset was divided into training - testing sets using 80 - 20 splits which resulted in seperate training and testing sets. Afterwards, k-NN, ANN, SVM, RF and NB algorithms were applied and results were compared based on the classification performance. According to the comparison results, the classification performance for was NB (AUC=0.8536), k-NN (AUC=0.8585), SVM (AUC=0.9150), ANN (AUC=0.9294) and RF(AUC=0.9599) directly from from lowest to highest.

Time-course analysis of multiple endpoints in real-world medical data, with an example on hospitalized COVID-19 patients

Reiner-Benaim, Anat

Abstract

Real world data, which is retrieved from continuously accumulating electronic databases, poses many statistical challenges, including selection bias, missing data and sparsity. Specifically, time course medical data is often characterised by heterogenous measurement times and measurement frequencies between patients, as well as repeated admissions of typically a small portion of the patients. Furthermore, when many endpoints are of concern, multiple testing of diversely distributed variables is involved. Post hoc analysis, aimed to focus the timing of measurement dynamics, is often required. I will discuss approaches to address these concerns, including a non-parametric mixed effect model, hierarchical multiple testing and a standardized heatmap. I will demonstrate an integrated analysis within a study on disease trajectory among hospitalized COVID-19 patients.

Multiple multi-sample testing under arbitrary covariance dependency

Vutov, Vladimir Krasimirov; Dickhaus, Thorsten

Abstract

Numerous datasets in various scientific disciplines are nowadays high-dimensional. Such datasets are often analysed by means of large-scale multiple testing simultaneously for some phenotype – for example, cancer subtypes – on each feature among thousands of features. This talk describes a new procedure that evaluates the strength of associations between a nominal (categorical) outcome and a large number of features simultaneously. We have proposed an inferential approach for conducting multiple multi-sample tests under arbitrary correlation dependency among test statistics. Specifically, the approach decomposes the (multinomial) target variable into multiple baseline-category pairs and approximates the false discovery proportion (FDP) under arbitrary correlation dependency within each pair. In this way, our methodology offers a sensible trade-off between the expected numbers of true and false rejections. Moreover, we have illustrated that the proposed procedure can be used even in cases where the sample size is considerably smaller than the number of features.

Furthermore, this talk demonstrates a practical application of the proposed workflow on hyperspectral imaging data. This dataset is generated by a matrix-assisted laser desorption/ionization (MALDI) instrument, where the nominal response categories describe cancer subtypes.

Post-selection inference for e-value based confidence intervals

Xu, Ziyu; Wang, Ruodu; Ramdas, Aaditya

Abstract

Suppose that one can construct a valid $(1-\delta)$-CI for each of K parameters of potential interest. If a data analyst uses an arbitrary data-dependent criterion to select some subset S of parameters, then the aforementioned confidence intervals for the selected parameters are no longer valid due to selection bias. We design a new method to adjust the intervals in order to control the false coverage rate (FCR). The main established method is the "BY procedure" by Benjamini and Yekutieli (JASA, 2005). Unfortunately, the BY guarantees require certain restrictions on the the selection criterion and on the dependence between the CIs. We propose a natural and much simpler method—both in implementation, and in proof—which is valid under any dependence structure between the original CIs, and any (unknown) selection criterion, but which only applies to a special, yet broad, class of CIs. Our procedure reports $(\frac{1-\delta|S|}{K})$-CIs for the selected parameters, and we prove that it controls the FCR at delta for confidence intervals that implicitly invert e-values; examples include those constructed via supermartingale methods, or via universal inference, or via Chernoff-style bounds on the moment generating function, among others. The e-BY procedure is proved to be admissible, and it recovers the BY procedure as a special case via calibration. Our work also has implications for multiple testing in sequential settings, since it applies at stopping times, to continuously-monitored confidence sequences with bandit sampling.

Acknowledgments

We are grateful to Boehringer Ingelheim and CRC Press, Taylor & Francis Group for supporting the conference.