Journées de Biostatistique 2023

Rapport sur les résumés

https://indico.math.cnrs.fr/e/Biostat2023

ID de résumé : **3**

Robustesse aux données manquantes : comparaison entre modèle mixte et analyse en composantes principales fonctionnelles

Contenu

En recherche médicale, la collecte de données longitudinales est très fréquente. L'analyse de ces données permet de décrire l'évolution temporelle d'un processus biologique tout en prenant en compte l'importante variabilité individuelle. Par exemple, les cohortes qui étudient l'évolution cognitive du sujet âgé intègrent les résultats des patients à différents tests psychométriques collectés sur des intervalles de temps réguliers. L'étude rétrospective de ces trajectoires permet de mieux décrire l'évolution cognitive du sujet âgé et constitue un outil précieux dans la compréhension de l'histoire naturelle de la démence.

Le modèle statistique le plus couramment utilisé dans l'étude de données longitudinales dans ce contexte est le modèle mixte. Il permet de modéliser avec une grande flexibilité des trajectoires variées, hétérogènes d'un sujet à l'autre avec souvent peu de mesures par sujet et la présence de données manquantes. Ce modèle est connu pour être, dans certains cas, robuste à la présence de données manquantes, fréquentes dans les données longitudinales. Néanmoins, il s'agit d'un modèle paramétrique et des hypothèses souvent fortes sont nécessaires pour estimer le modèle.

Si l'on considère ces données longitudinales comme des données fonctionnelles, c'est-à-dire comme des réalisations aléatoires d'une fonction inconnue sous-jacente, il devient envisageable d'appliquer d'autres outils statistiques. En particulier, l'analyse en composantes principales fonctionnelles (ACPF) est un outil qui permet de décrire des trajectoires temporelles, sans hypothèses paramétriques. Néanmoins, si cette méthode a été éprouvée sur des données denses et régulières, son utilisation sur des données longitudinales avec, en particulier, la présence de données manquantes reste à explorer.

Nous avons réalisé une étude empirique qui montre que l'ACPF se comporte bien malgré la présence de données manquantes et peut se comparer avec l'approche par modèle mixte. Une application de l'ACPF sur des données issues de la cohorte 3C qui étudie l'évolution du sujet âge permet d'illustrer son intérêt.

Auteurs principaux: GENUER, Robin (Université de Bordeaux / Bordeaux Population Health); PROUST-LIMA, Cécile (Université de Bordeaux / Bordeaux Population Health); SÉGALAS, Corentin (Université de Bordeaux / Bordeaux Population Health)

Orateur: SÉGALAS, Corentin (Université de Bordeaux / Bordeaux Population Health)

Commentaires:

Le résume est fourni en français mais une version anglaise peut être fournie si besoin.

Statut: DÉPOSÉ

Déposé par **SÉGALAS, Corentin** le **lundi 18 septembre 2023**

ID de résumé : **4**

Analyse des trajectoires de soin, clustering et survie auprès des patients souffrant d'insuffisance cardiaque

Contenu

L'insuffisance cardiaque (IC) entraîne environ 200 000 hospitalisations par an en France et est associée à une surmortalité élevée. Alors que les patients IC sont de plus en plus âgés, la compréhension des causes de décès en milieu hospitalier est une problématique de santé publique majeure. Dans ce travail, nous proposons une méthodologie innovante visant à identifier les trajectoires de soin fréquentes et à étudier leur impact sur la survie globale.
Les données utilisées sont extraites de l'Echantillon Généraliste des Bénéficiaires sur les patients IC avec une première hospitalisation entre 2010 et 2016. Dans un premier temps, nous procédons à l'exploration de motifs séquentiels pour caractériser les séquences de soin à partir des diagnostics d'hospitalisation (Groupes Homogènes de Malades, GHM). La deuxième étape consiste à utiliser un algorithme de clustering afin d'apporter un caractère interprétable. Pour ce faire, une nouvelle métrique a été développée pour évaluer la distance entre deux GHM. Enfin, une analyse de survie est réalisée en utilisant des approches semi-paramétriques (modèle de Cox à hasards proportionnels, pénalisé et splines) et des méthodes d'ensemble.
10 051 patients ont été inclus, représentant 85 594 hospitalisations. Un total de 6 618 séquences a été identifié et cinq clusters ont été construits. Les multiples hospitalisations pour IC étaient les motifs les plus fréquents, souvent combinées au décès, et en particulier pour deux clusters incluant des patients plus âgés et avec moins d'hospitalisations. Les seconds motifs fréquents identifiés étaient les complications pulmonaires et cardiaques. Les principaux facteurs associés au décès étaient l'âge, le sexe et la durée de séjour. Les associations trajectoires et décès n'étaient en revanche pas significatives.
Dans l'ensemble, l'approche proposée présente un intérêt méthodologique pour l'analyse des trajectoires de soin, clustering et survie chez les patients souffrant d'insuffisance cardiaque et peut être aisément transposée à d'autres problématiques cliniques.

Auteurs principaux: AMADEI, Tristan (ENSAE Paris –IP Paris); KIRSCHER, Tristan (ENSAE Paris –IP Paris); KLEIN, Antoine (ENSAE Paris –IP Paris)

Co-auteurs: MURRIS, Juliette (HeKA, Inserm, Inria Paris); Dr TROPEANO, Anne-Isabelle (Hôpital Européen Georges Pompidou, AP-HP, Inserm); Prof. KATSAHIAN, Sandrine (Hôpital Européen Georges Pompidou, AP-HP, HeKA, Inserm, Inria Paris)

Orateurs: AMADEI, Tristan (ENSAE Paris –IP Paris); KIRSCHER, Tristan (ENSAE Paris –IP Paris); KLEIN, Antoine (ENSAE Paris –IP Paris)

Commentaires:

Le travail proposé a été réalisé en collaboration avec les étudiants de l'ENSAE Paris dans le cadre de la supervision du projet 'Stat app'.

Statut: DÉPOSÉ

Déposé par **MURRIS, Juliette** le **jeudi 21 septembre 2023**

ID de résumé : **5**

Données RNA-seq semi-synthétiques : pourquoi normaliser alors qu'on peut générer des faux positifs ?

Contenu

En revisitant les analyses de données RNA-seq semi-synthétiques simulées par Li et al. (*Genome Biology* 2022), nous soulignons l'importance de la prise en compte de la taille de librairie et de l'utilisation d'une normalisation adéquate avant d'envisager toute manipulation des observations supposant leur échangeabilité. Après avoir proposé une nouvelle stratégie de simulation corrigée, nous montrons que seules certaines méthodes d'analyse différentielle souffrent réellement d'une inflation exagérée du taux de faux positifs dans les études de grandes tailles avec une forte variabilité biologique (telles que les études chez l'humain) —en particulier edgeR et DESeq2 . De plus, nous montrons que le comportement du test des rangs de Wilcoxon n'est en rien supérieur à d'autres tests spécialisés pour l'analyse différentielle de données RNA-seq tels que dearseq. Ces derniers sont de surcroît souvent plus versatiles et peuvent prendre en compte des designs expérimentaux relativement complexes, tandis que le test des rangs ne peut s'appliquer que pour la comparaison non-ajustée entre deux groupes.

Auteurs principaux: HEJBLUM, Boris (Inserm U1219); BA, Kalidou; THIÉBAUT, Rodolphe; AGNIEL, Denis

Orateur: HEJBLUM, Boris (Inserm U1219)

Statut: DÉPOSÉ

Déposé par **HEJBLUM, Boris** le **mercredi 27 septembre 2023**

ID de résumé : **6**

ShiBa logiciel d'inférence bayésienne sans codage

Contenu

La diffusion de l'inférence bayésienne dans la communauté scientifique, y compris parmi les spécialistes de biométrie reste limitée. Cette timidité s'explique en partie par l'absence d'outil logiciel omnibus et ne nécessitant pas de codage. Notre logiciel ShiBa, sous licence CC-by-nc 4.0, vient combler ces manques.

ShiBa est un développement utilisant Shiny pour R (Chang W et al (2023). shiny: Web Application Framework for R. https://shiny.posit.co/, https://github.com/rstudio/shiny) permettant une interface homme-machine conviviale et une prise en main facile. Le logiciel est un outil d'analyse statistique n'utilisant que l'inférence bayésienne (d'où le nom ShiBa associant Shiny et Bayes), et repose essentiellement sur STAN (https://mc-stan.org/support/). L'utilisation des Hamiltonian Monte Carlo permet une convergence plus rapide, appréciée par l'utilisateur, tout en gardant de bonnes propriétés de convergence. Outre la possibilité de réaliser des inférences simples (estimation de moyennes, de la fréquence des modalités d'une variable catégorielle), ShiBa permet l'estimation des paramètres de modèles linéaires généralisés : régression linéaire, logistique, Poisson et Beta. Les distributions a priori sont par défaut celles de STAN mais l'utilisateur a la possibilité de changer les hyperparamètres. Par exemple dans la régression linéaire de y sur x, STAN propose un prior normal sur le paramètre de moyenne nulle et d'écart-type 2,5 fois le rapport des écarts-type entre y et x. Mais à l'aide d'une fenêtre pop-up, l'utilisateur peut les changer et voit graphiquement la représentation du prior.

Les résultats de l'inférence sont présentés graphiquement et sous forme tabulaire (médiane et intervalle empirique, par défaut à 95%) avec un indicateur de bonne convergence basé sur l'absence d'autocorrélation résiduelle et sur un Rhat inférieur à 1,10 (pour chaque paramètre). A des fins de sélection de modèle, le WAIC est présenté.

L'aide à l'utilisation se fait par info-bulles pour chaque paramètre ou option ainsi que par un Wiki.

Il reste à tester!

Auteur principal: Dr FABACHER, Thibaut (Université de Strasbourg, Laboratoire ICube UMR7357)

Co-auteurs: Prof. GODET, Julien (Université de Strasbourg); Dr LEFEVBRE, François (Hôpitaux Universitaires de Strasbourg); Prof. MEYER, Nicolas (Hôpitaux Universitaires de Strasbourg); Prof. SAULEAU, Erik-A. (Université de Strasbourg, Laboratoire ICube UMR7357); Dr SEVERAC, François (Hôpitaux Universitaires de Strasbourg)

Orateur: Prof. SAULEAU, Erik-A. (Université de Strasbourg, Laboratoire ICube UMR7357)

Commentaires:

L'orateur sera ou le Dr Fabacher (recherche de financement pour le déplacement au congrès) ou moi.

Statut: DÉPOSÉ

Déposé par **SAULEAU, Erik-A.** le **jeudi 28 septembre 2023**

ID de résumé : **7**

Analyse des pseudo-observations par la méthode des moments généralisée comme alternative à l'analyse de survie Bayésienne

Contenu

En Bayésien, la formulation d'un modèle de survie à risques proportionnels nécessite généralement la modélisation de la fonction de risque de base. Elle peut être paramétrique et suppose alors des hypothèses fortes ou non paramétrique conduisant à une implémentation complexe. En fréquentiste, les pseudo-observations définies par Andersen sont devenues une alternative à l'analyse de survie par le modèle de Cox mais elles sont surtout avantageuses pour des modélisations plus complexes telles que les modèles multi-états ou les événements récurrents. L'avantage des pseudo-observations est de s'affranchir de la complexité des données censurées en les transformant en données longitudinales, ensuite analysées par les équations d'estimations généralisées (GEE). L'objectif est de proposer une nouvelle alternative à l'analyse de survie Bayésienne reposant sur l'analyse des pseudo-observations. Nous proposons d'utiliser la méthode des moments généralisée (GMM) qui repose sur la définition d'une fonction quadratique de moments. Dans le cadre fréquentiste, certains auteurs ont montré que l'approche GMM donne des estimateurs plus efficients que l'approche GEE lorsque la matrice de travail est mal spécifiée. Contrairement au GEE, une version Bayésienne a également été proposée, basée sur une pseudo-vraisemblance. Nous avons donc étendu les approches GMM (fréquentiste et Bayésien) aux spécificités de l'analyse des pseudo-observations et comparé leurs performances, par une étude de simulation d'essais randomisés, à celles des modèles de Cox, GEE et Bayésien exponentiel par morceaux. La version fréquentiste donne des performances similaires au GEE. Le GMM Bayésien surestime légèrement l'effet traitement pour des petits échantillons. Pour illustration, trois analyses post-hoc ont été réalisées sur des essais cliniques, de différentes tailles, incluant des patients atteint du Sarcome d'Ewing. Les modèles GMM ont donné des estimations proches du modèle de Cox. L'analyse Bayésienne des pseudo-observations ouvre de nouvelles perspectives pour l'analyse de survie Bayésienne ne nécessitant pas la spécification de la fonction de risque de base.

Auteur principal: ORSINI, Léa (CESP, INSERM U1018, Université Paris-Saclay, Villejuif, France)

Co-auteurs: Mme BRARD, Caroline (Clinical Development Operations, Ipsen Innovationn, Les Ulis, France); M. DEJARDIN, David (Product Development, Data Sciences, F. Hoffmann-La Roche AG, Basel, Switzerland); M. LE TEUFF, Gwénaël (CESP, INSERM U1018, Université Paris-Saclay, Villejuif, France); Prof. LESAFFRE, Emmanuel (I-Biostat, KU-Leuven, Leuven, Belgium)

Orateur: ORSINI, Léa (CESP, INSERM U1018, Université Paris-Saclay, Villejuif, France)

Statut: DÉPOSÉ

Déposé par **ORSINI, Léa** le **jeudi 28 septembre 2023**

ID de résumé : **8**

Instrumental Variables: state of the art and practical recommandations

Contenu

In this work, we provide a comprehensive theoretical and empirical exploration of the integration of instrumental variables (IV) in causal analysis. Specifically, we focus on the estimation of the Average Treatment Effect (ATE) when confronted with the challenge of unmeasured confounding variables.

We begin by introducing the conceptual foundations and methodological underpinnings of the IV estimator, highlighting the critical assumptions, potential violations, and strategies for mitigating such violations. Comparative simulations involving well-known ATE estimators, including Inverse Propensity Score Weighting, the G-Formula, and IV estimation, are presented, demonstrating their performance across a diverse range of scenarios.

Then, acknowledging that practitioners often rely on the sometimes unrealistic linearity of outcome assumption in ATE estimation, we detail a more flexible nonparametric approach that facilitates the computation of the Local Average Treatment Effect (LATE). This method requires an additional assumption, monotonicity, ensuring a monotonous relationship between treatment and the instrumental variable, and integrates it within the framework of Principal Stratification. Empirical and analytical results are showcased, emphasizing the efficacy of this methodology while advocating for the need for caution in LATE estimation.

Our findings reveal challenges in ATE estimation using IV in scenarios with limited sample sizes and the inherent complexity of interpreting results in nonparametric approaches, where the target population for LATE estimation may remain unidentified solely based on available data.

In conclusion, this presentation aims at gaining a comprehensive understanding of when and how to judiciously incorporate instrumental variables into causal analysis, leading to more accurate and insightful conclusions.

Auteur principal: KHELLAF, Remi (INRIA)

Orateur: KHELLAF, Remi (INRIA)

Statut: DÉPOSÉ

Déposé par **KHELLAF, Remi** le **vendredi 29 septembre 2023**

ID de résumé : **9**

Comparison between different randomization methods for treatment allocation with a continuous factor as stratification criterion

Contenu

Randomization is a key step in clinical trials to ensure a valid estimation of treatment effect. Most popular randomization method is stratification with blocks. This method can cause serious imbalances which makes this method unworkable in case of small sample size trials or incorporation of several prognostic factors. Minimization can overcome these problems, by accounting for many factors at the design stage, even when the number of patients is low.

The origin of this work is a future study with Ceva Santé Animale, about testing a new medicine on dogs. Stratification had to be applied to allocate 150 individuals to three treatment arms, depending on four qualitative factors and a continuous variable. Continuous criteria are often imbalanced at baseline without any clinically relevant classes to categorize the factor. A new randomization method accounting for both continuous and categorical factors with many strata was developed. This method is derived from classical Pocock and Simon's minimization and accounting for continuous factors directly without transformation into categorical variables while maintaining randomness.

The main objective of this work was to compare different randomization methods, including stratification with blocks, classical minimization and the new randomization method. Simulations were run on randomly generated sets of individuals, to estimate the impact of the randomization method, the number of factors used and the number of individuals on the imbalance between treatment arms.

The method proposed, i.e. non-deterministic neighborhood-based minimization, allows to consider continuous covariate in a way which is at least as efficient as categorized-based usual methods. Minimization is preferable in terms of balance when the number of patients decreases. In the case of this study, the final decision was to use the new randomization method, which creates similar imbalance as stratification with blocks and classical minimization for qualitative factors, but a smaller one for the quantitative variable.

Auteurs principaux: M. BLONDEL, Thomas (Ceva Santé Animale); MOLINARD, Aymeric (INSA Rennes); M. MONTESTRUC, François (eXYSTAT)

Orateur: MOLINARD, Aymeric (INSA Rennes)

Statut: DÉPOSÉ

Déposé par **MOLINARD, Aymeric** le **vendredi 29 septembre 2023**

ID de résumé : **10**

Statistical properties of minimal sufficient balance randomization approach in a stratification context, a simulation study

Contenu

Introduction – Randomization is a crucial step in clinical trials and ensures balance across treatment groups. Several approaches exist (e.g. stratified permuted blocks or covariate adaptive minimization). Some of them were introduced recently such as Zhao et al Minimum Sufficient Balance (MSB) in 2015. The aim of this work is to assess the performance of MSB to grasp a better understanding of its strengths and limitations.
Methods – (i) Stratified permuted 4-blocks, (ii) Pocock & Simon's minimization (with 2 and 4 classes) and (iii) MSB were applied and compared one to another in simulated stratified and unstratified settings. The simulated scheme was based on Sepsicool-1 trial data. Data augmentation was performed using NORTA method in order to generate correlated datasets with any specified distribution. Scenarios of 5,000 datasets each were considered with variations in terms of sample sizes and correlation structures. Evaluation criteria included imbalance tests, adjusted and unadjusted statistical power, and RMSE of observed and estimated treatment effect.
Results – MSB outperformed at minimizing covariates imbalance with stratified Students-t-test and stratified Wilcoxon signed-rank test, as well as non-stratified Students-t-test (for both augmented and non-augmented scenarios). Minimization was better than all the others for non-stratified Wilcoxon signed-rank test. Since, in practice, covariate adaptive randomization procedures are rarely stratified, as the important covariates are included in the randomization, MSB remains a relevant candidate. However, no randomization procedure was clearly better than the other at reaching the true treatment effect.
Conclusion – This study showed that MSB is a valuable randomization approach in adaptive design to control for group imbalance, and to maintain a high probability in showing treatment effect.
Keywords - Randomized clinical trial, adaptive design, data simulation

Auteur principal: M. CANNAFARINA, Hugo (ENSAI)

Co-auteurs: GUENEGOU ARNOUX, Armelle (Assistance Publique-Hôpitaux de Paris URC APHP.Centre - INSERM CIC1418-EC - INSERM-INRIA HeKA); Prof. KATSAHIAN, Sandrine (Assistance Publique-Hôpitaux de Paris URC APHP.Centre - INSERM CIC1418-EC - INSERM-INRIA HeKA)

Orateur: M. CANNAFARINA, Hugo (ENSAI)

Statut: DÉPOSÉ

Déposé par **GUENEGOU ARNOUX, Armelle** le **dimanche 1 octobre 2023**

ID de résumé : **11**

Puissance conditionnelle lors de l'analyse intermédiaire en essai clinique

Contenu

Introduction – L'analyse intermédiaire (AI) lors d'un essai clinique peut permettre d'évaluer le critère de jugement principal avant le recrutement ou la fin de suivi de l'ensemble des patients. Cela mène à l'arrêt précoce ou à la poursuite de l'essai. Cette étape inclut notamment l'évaluation de la puissance conditionnelle (PC, probabilité d'obtenir un résultat significatif à la fin de l'étude, conditionnellement aux hypothèses initiales et aux données collectées à l'AI). Notre objectif est de fournir les éléments nécessaires et suffisants au calcul de cette PC, ainsi que ses propriétés.
Méthodes – Trois zones sont définies pour étudier la PC : 1) une zone défavorable à la poursuite de l'essai, 2) une zone prometteuse à la poursuite après réévaluation du nombre de patients, et 3) une zone favorable à la poursuite dans les conditions initiales. A partir des données d'un essai clinique randomisé contrôlé en réanimation, l'évolution de la PC est explorée selon le nombre total de patients inclus et la différence observée de critère de jugement (binaire) entre les deux bras de traitement au moment de l'AI.
Résultats – La PC est positivement associée au nombre de patients total inclus à l'AI. La marge de différence entre randomisation et critère de jugement principal au moment de l'AI puis finale est plus importante. Une PC calculée lors de l'AI considérée dans la zone prometteuse a de plus grande chance de conduire à une puissance théorique finale élevée. Néanmoins, nous n'avons pas pu mettre en évidence d'association directe entre PC et significativité de la différence de critère de jugement en fin d'étude. Enfin, la PC peut être calculée au global ou par sous-groupe d'intérêt.
Conclusion – L'exploration des propriétés de la PC a mené à l'élaboration de fiches de bonnes pratiques pour des statisticiens.
Mots-clés - Essai clinique, analyse intermédiaire, puissance conditionnelle

Auteur principal: MURRIS, Juliette (HeKA, Inserm, Inria Paris)

Co-auteurs: Prof. KATSAHIAN, Sandrine (Assistance Publique-Hôpitaux de Paris URC APHP.Centre - INSERM CIC1418-EC - INSERM-INRIA HeKA); Mlle MEGRET, Maud (ENSAI)

Orateur: Mlle MEGRET, Maud (ENSAI)

Statut: DÉPOSÉ

Déposé par **GUENEGOU ARNOUX, Armelle** le **dimanche 1 octobre 2023**

ID de résumé : **12**

High-dimension Mechanistic Model Building using LASSO Approaches : Application to Ebola Vaccination

Contenu

Constructing non-linear mixed-effects models (NLMEM) deepens our comprehension of biological processes. Specifically, NLMEM facilitates the incorporation of inter-individual variability by parameterizing models at the individual level within a population framework. To do so, parameters combine fixed effects, capturing population-level relationships with covariates, and random effects, accounting for individual deviations. Estimation in NLMEM is achievable through maximum likelihood methods, such as the Stochastic Approximation Expectation-Maximization (SAEM-Dempster, 1977; Kuhn & Lavielle, 2005) algorithm. However, this approach is computationally intensive, and selecting covariates that define individual-level parameters cannot be done by comparing all possible models.

For the optimized construction of NLMEM, traditional methodologies rely on modified stepwise approaches (SCM-Jonsson, 1998; COSSAC-Ayral, 2021). Alternatively, the Stochastic Approximation for Model Building Algorithm (SAMBA-Prague & Lavielle, 2022) builds the covariate model on the posterior realization of the parameters. Within a low-dimensional context, SAMBA efficiently and more rapidly constructs models by minimizing an information criterion. However, we aim to extend it for high-dimensional settings—such as those involving transcriptomic data. Initially, SAMBA employs a stepwise AIC algorithm for covariate selection. Our proposal integrates a multivariate LASSO approach, offering a more nuanced treatment of parameter correlations. This methodology incorporates a whitening step (Perrot-Dockès, 2018) and a stability selection process (Meinshausen & Bühlmann, 2010).

We validated our approach through simulations imitating the dynamics of the humoral immune response to an Ebola vaccine (Pasin, 2019). These simulations were replicated 100 times, each involving 100 individuals and 200 covariates. Remarkably, the False Discovery Rate for the proposed method was reduced by a factor of 10, while maintaining a similar False Negative Rate. This indicates enhanced control over the False Positive Rate. We applied our method using data from the Prevac/Prevac-UP trial (Prevac-UP Team, 2022), which compares two licensed vaccines for Ebola in Africa.

Auteurs principaux: GABAUT, Auriane (Université de Bordeaux, Inria, Inserm, Bordeaux Population Health Research Center, SISTM Team ; Vaccine Research Institute, Créteil, France); Mme PRAGUE, Mélanie (Université de Bordeaux, Inria, Inserm, Bordeaux Population Health Research Center, SISTM Team ; Vaccine Research Institute, Créteil, France)

Orateur: GABAUT, Auriane (Université de Bordeaux, Inria, Inserm, Bordeaux Population Health Research Center, SISTM Team ; Vaccine Research Institute, Créteil, France)

Statut: DÉPOSÉ

Déposé par **Mlle GABAUT, Auriane** le **lundi 2 octobre 2023**