IA pour l'identification d'interactions entre variables, application à l'effet individuel d'un traitement

AI for identifying features interaction, application to individual treatment effect

" This PhD offer is provided by the ENACT AI Cluster and its partners. Find all ENACT PhD offers and actions on https://cluster-ia-enact.ai/*."*
"This work is supported by the LIFE TRAVEL Interdisciplinary Program, as part of Initiative d'Excellence Lorraine, an Excellence Initiative funded by France 2030."

Université de Lorraine, Ecole doctorale IAEM (ED 77), Laboratoire IECL
Directrice de thèse Anne Gégout-Petit, PR, IEC, Université de Lorraine
Co-directeur de thèse : Nicolas GIRED, PU-PH Professeur des Universités – CICP, CHRU de Nancy, INSERM, Université de Lorraine

Context

The Institut Élie Cartan de Lorraine is one of the largest French mathematics laboratories. The laboratory conducts research in fundamental mathematics [geometry, analysis, number theory] and in applied mathematics [probability, statistics, partial differential equations]. The institute offers a complete scientific training provided by dedicated teacher-researchers. A joint research unit (UMR 7502) of the University of Lorraine, and the CNRS Mathematics, IECL works in close collaboration with Centre Inria de l'Université de Lorraine with the three project-teams it hosts.
The IECL's Probability and Statistics team has now established collaborations with the Nancy and Strasbourg CHRUs on biomarker research for the relapse of cardiac events, and also in oncology on issues of tumor heterogeneity modeling and gene networks. Anne Gégout-Petit has strong skills in biostatistics particularly in the field of variables selection in complex data (A. Gégout-Petit *et al.* 2022, B. Bastien *et al.* 2022)

The CIC-P in Nancy, part of the CHRU, is a research team highly involved in translational research. It integrated AI after the RHU FIGHT-HF program, hiring two data scientists in addition to the existing statisticians. Prof. Girerd, a cardiologist with a PhD in biomathematics, coordinator of the CIC-P, is actively involved in clinical trials and physiopathological research, as well as advanced data analysis. His contributions to machine learning in Medicine are reflected in several publications (Kobayashi, …Girerd JACC CVI 2022; Monzo,… Girerd EJHF 2024). He also co-authored a methodological study on clustering approaches, combining simulations and real data to compare performance (Preud'homme,… Girerd, Sci Rep 2021). Additionally, he holds one of the AI & Medicine chairs within the ENACT initiative, further highlighting his leadership in the field.

The PhD project is also part of the interdisciplinary program Life Travel (Life Trajectories, multimorbidities, functional ability, quality of life and longevity) of the Lorraine Initiative of Excellence. We are applying to Life Travel for the second co-funding. We also have an ANR application with thesis co-funding in progress.

Keywords: features, interactions, statistical learning, CART tree, Random Forest, Causal Random Forest, personalized medicine,

Sujet de thèse

Variable selection is a highly active research field in statistics and machine learning since the 2000s, with the development of numerous methods (LASSO, knockoff, variable importance measures, etc.). It is of paramount importance for the interpretation of models and decision-making algorithms. However, variable selection alone is not sufficient for interpretation; it is also essential to identify and understand interactions between variables.

Standard statistical models, such as generalized linear models (linear regression, logistic regression, analysis of variance, etc.) or mixed-effects models, allow for the modeling of interactions between covariates or latent phenomena by introducing specific terms and testing their statistical significance. However, the number of interactions to test quickly becomes unwieldy as the number of covariates increases. Moreover, with these models, the search for interactions can only be guided by a priori hypotheses or prior knowledge about plausible interactions, which are then introduced and tested. Additionally, the types of interactions are often limited to products of quantitative variables or to cases involving binary variables (e.g., crossings between or with binary variables).

A further limitation of classical interaction modeling is its low intrinsic statistical power, particularly for higher-order or moderate-size interactions. In practice, only first-order interactions (between two variables) are considered, and only those with large effect sizes can be detected. Taken together, these limitations result in an unacceptably high false-negative rate, severely restricting discovery in complex, high-dimensional settings.

In this project, we propose a new framework based on machine learning models to discover and quantify interactions between covariates. Random forests are particularly well suited for this purpose. The hierarchical structure of decision trees naturally captures heterogeneous effects of variables across subpopulations defined by early splits, thereby encoding interaction pathways without explicit specification.

The literature on discovering interactions without prior assumptions using random forests is primarily limited to methods for detecting interactions between single nucleotide polymorphisms (SNPs) (Chen et al., 2011; Jiang et al., 2009; Li et al., 2016). Another body of literature focuses on more targeted interactions, particularly when one of the variables is binary (e.g., treatment vs. placebo in personalized medicine, to identify differential treatment effects). In this latter case, interactions are implicitly embedded in individualized outcome variables, such as individual treatment effects (ITE), which are then predicted using covariates within a counterfactual framework (Lu, Min, et al., 2018; Kobayashi, Masatake, et al., 2025). Here, the interaction is not modeled as an explicit cross-product between variables. Instead, it is embedded within a single individualized variable, defined at the patient level, which captures the differential effect of a factor (e.g., treatment or comorbidity) for each individual.

In this thesis, we will address both scenarios.

For the case without prior assumptions, an initial step will involve studying **CART trees**. Since the trees in a random forest are bootstrap replicates of CART trees, the most frequent patterns of similar hierarchical splits can be identified, and their frequency can be used to quantify the importance of the interaction. The objective of this part is to clarify these concepts, develop algorithms to detect and quantify interactions, and propose statistical tests.

For interactions involving a targeted variable, we will build on our expertise in identifying **individual treatment effects (ITE)**, for which we have estimated a forest for each group of the targeted variable (Kobayashi, Masatake, et al., 2025), within the larger framework of counterfactual reasoning. More broadly, we will leverage **causal machine learning methods** (Lu, Min, et al., 2018) or digital twins. This framework can be generalized beyond treatment effects, for example to quantify individualized differential effects of comorbidities such as diabetes. In this setting, interactions are individualized at the patient level and subsequently modeled and predicted.

In parallel, extensive simulation studies will be conducted. Scenarios involving first- and higher-order interactions will be generated to compare the performance of classical statistical models with interaction terms (e.g., logistic regression) against machine learning approaches, including random forests and causal machine learning methods, in terms of detection power and false-negative rates.

This thesis is methodological, aiming to develop general methods applicable to various fields (hence the submission under the theme "AI for Scientific Discovery"), although our approach is guided by biological and clinical examples, such as diagnostic and personalized medicine questions in cardiology. We will work with large cardiovascular databases, both observational (e.g., the PARADISE database with over 19,000 observations) and randomized (e.g., the RALES, EPHESUS, EMPHASIS and HOMAGE trial, testing the effect of spironolactone in patients with or at risk of heart failure).

The outputs of this PhD will be broadly applicable across the analytic community, beyond health and biomedical research.

In addition to publications in statistical learning journals and applications in clinical studies, we will develop **software packages for R or Python**.

Resumé anglais

Variable selection is a highly active research field in statistics and machine learning since the 2000s, with the development of numerous methods. It is of paramount importance for the interpretation of models and decision-making algorithms. However, variable selection alone is not sufficient for interpretation; it is also essential to identify and understand interactions between variables.

In this project, we propose a new framework based on machine learning models to discover and quantify interactions between covariates. Random forests (RF) are particularly well suited for this purpose. A body of literature focuses on more targeted interactions, particularly when one of the variables is binary (e.g., treatment vs. placebo in personalized medicine, to identify differential treatment effects (ITE)). In this thesis, we will address both scenarios.

This thesis is methodological, aiming to develop general methods applicable to various fields (hence the submission under the theme "AI for Scientific Discovery"), although our approach is guided by biological and clinical examples, such as diagnostic and personalized medicine questions in cardiology. We will work with large cardiovascular databases, both observational or and randomized.

The outputs of this PhD will be broadly applicable across the analytic community, beyond health and biomedical research.

In addition to publications in statistical learning journals and applications in clinical studies, we will develop **software packages for R or Python**.

Résumé en Français

La sélection de variables est un domaine de recherche très actif en statistique et en apprentissage automatique depuis les années 2000, avec le développement de nombreuses méthodes. Elle revêt une importance capitale pour l'interprétation des modèles et des algorithmes décisionnels. Cependant, la sélection de variables seule ne suffit pas pour l'interprétation ; il est également essentiel d'identifier et de comprendre les interactions entre les variables.
Dans ce projet, nous proposons un nouveau cadre basé sur des modèles d'apprentissage automatique pour découvrir et quantifier les interactions entre les covariables. Les forêts aléatoires (RF) sont particulièrement bien adaptées à cet effet. Une partie de la littérature se concentre sur des interactions plus ciblées, en particulier lorsqu'une des variables est binaire (par exemple, traitement vs placebo en médecine personnalisée, pour identifier les effets différentiels du traitement (ITE)). Dans cette thèse, nous aborderons les deux scénarios.
Cette thèse est méthodologique et vise à développer des méthodes générales applicables à divers domaines (d'où sa soumission sous le thème « IA pour la découverte scientifique »), bien que notre approche s'appuie sur des exemples biologiques et cliniques, tels que des questions de diagnostic et de médecine personnalisée en cardiologie. Nous travaillerons avec de grandes bases de données cardiovasculaires, tant observationnelles que randomisées.
Les résultats de cette thèse seront largement applicables à l'ensemble de la communauté analytique, au-delà de la recherche en santé et biomédicale.
Outre des publications dans des revues spécialisées en apprentissage statistique ou d'études cliniques, nous développerons des paquets pour les logiciel R ou Python.

References

Bastien, B., Gégout-Petit et al (2022). *A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles in oncology*. **Journal of Applied Statistics**, *49*(3), 764-781.

Chen, C. C. M., et al. (2011). *Methods for identifying SNP interactions: A review on variations of Logic Regression, Random Forest and Bayesian logistic regression.*

IEEE/ACM Transactions on Computational Biology and Bioinformatics, *8*(6), 1580–1591.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). *Subgroup identification from randomized clinical trial data.* **Statistics in Medicine**, *30*(24), 2867–2880.

Gégout-Petit, A., Gueudin-Muller, A., & Karmann, C. (2022). *The revisited knockoffs method for variable selection in L_1-penalized regressions.* **Communications in Statistics—Simulation and Computation**, *51*(10), 5582–5595.

Jiang, R., et al. (2009). *A random forest approach to the detection of epistatic interactions in case-control studies.* **BMC Bioinformatics**, *10*(Suppl 1), S65.

Kobayashi, M., Girerd et al. (2025). *An individual treatment effect approach to predict response to mineralocorticoid receptor antagonists in patients with heart failure and reduced ejection fraction.* **European Journal of Heart Failure**. 2025 Dec;27(12):2800-2809. doi: 10.1002/ejhf.70047

Li, J., et al. (2016). *Detecting gene-gene interactions using a permutation-based random forest method.* **BioData Mining**, *9*(1), 14. https://doi.org/10.1186/s13040-016-0097-9

Lu, M., et al. (2018). *Estimating individual treatment effect in observational data using random forest methods.* **Journal of Computational and Graphical Statistics**, *27*(1), 209–219.

Monzo, L., et al. (2025). *Machine learning approach to identify phenotypes in patients with ischaemic heart failure with reduced ejection fraction.* **European Journal of Heart Failure**. *Dec;27(12):3378-3391*. doi: 10.1002/ejhf.3547. Epub 2024 Dec 10.PMID: 39654426

Sauve, M., & Tuleau-Malot, C. (2014). *Variable selection through CART.* **ESAIM: Probability and Statistics**, *18*, 770–798. https://doi.org/10.1051/ps/2014015

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). *Variable selection using random forests.* **Pattern Recognition Letters**, *31*(14), 2225–2236.

Profil

The student must hold a master's degree in applied mathematics or data science and be motivated by applications in the healthcare field.