English version below

Inférence post sélection de modèles en grande dimension dans des modèles à effets mixtes

Stage niveau M2 début 2025

Université Paris Saclay INRAE
Unité MaIAGE Mathématiques et informatique appliquées du génome à l'environnement, Jouy-en-Josas

Pour postuler, merci d'envoyer un **dossier complet** contenant **votre CV, votre lettre de motivation et vos derniers relevés de notes de Master 1 ou équivalent** à estelle.kuhn@inrae.fr et à sarah.lemler@centralesupelec.fr.

Contexte

Dans le cadre des mesures répétées, chaque individu d'une population est mesuré plusieurs fois dans des conditions différentes, par exemple au cours du temps ou dans plusieurs conditions environnementales. Les modèles à effets mixtes sont bien adaptés pour ce type de données car ils permettent de modéliser à la fois la variabilité présente au sein des mesures répétées d'un individu et la variabilité entre les individus de la population. Du point de vue de la modélisation statistique, ces modèles sont des modèles à variables latentes, qui comportent des effets fixes et des effets aléatoires non observés. Ils peuvent également intégrer des covariables caractérisant les individus. S'agissant de l'inférence, on peut s'intéresser à l'estimation des paramètres du modèle, à sélectionner les covariables influentes et à la prédiction de la sortie du modèle.

Lorsqu'on souhaite considérer un grand nombre de covariables, par exemple des marqueurs génétiques, on travaille souvent sous l'hypothèse de parcimonie selon laquelle il existe un petit nombre de covariables influentes. L'objectif est alors d'identifier ces covariables. Cette sélection peut s'effectuer par exemple via un estimateur pénalisé de type LASSO [1]. Un modèle réduit obtenu en ne conservant que les covariables sélectionnées peut alors être considéré. Cependant, plusieurs résultats classiques de statistique asymptotique ne sont plus valides dans ce modèle, car il est lui même aléatoire, ayant été sélectionné à partir des données [2]. En particulier, il n'est plus possible de construire des intervalles de confiance valides ayant des taux de couverture attendus pour les estimateurs et les prédictions à partir des résultats standards.

Pour faire face à cette problématique, une approche basée sur une étape de débiaisage de l'estimateur LASSO a été proposée dans le cas de modèles linéaires et linéaires généralisés [3,4]. Des résultats théoriques ont été établis pour l'estimateur débiaisé dans ces contextes et permettent de fournir également des garanties en prédiction.

L'objectif du stage est d'étendre ces méthodes au cas des modèles à effets mixtes et d'appliquer les méthodes développées Ã l'analyse de la variabilité du processus de sénescence pour une population de génotypes de blé.

Objectifs du stage

* proposer un estimateur débiaisé obtenu à partir de l'estimateur régularisé de type LASSO dans un modèle linéaire à effets mixtes,

* étudier ses propriétés théoriques de convergence et de normalité asymptotique

* développer et implémenter un algorithme pour calculer l'estimateur débiaisé et le valider sur données simulées,
* proposer un intervalle de confiance en estimation et en prédiction et le valider en simulation.
* comparer les performances de la méthode développée à d'autres méthodes existantes
* étendre les développements au cas de modèles non linéaires à effets mixtes
* analyser les données réelles de blé

Aspects mathématiques L'approche envisagée pour construire un estimateur débiaisé à partir de l'estimateur de type LASSO est celle proposée par [3]. L'objectif est de définir un nouvel estimateur ayant un biais plus faible via une forme analytique obtenue à partir des conditions de Karush-Kuhn-Tucker et d'un estimateur de l'information de Fisher du modèle. Ce dernier pourra être obtenu à partir des développements proposés par [6]. L'estimateur proposé pourra être calculé via des algorithmes de type gradient stochastique adaptés aux modèles à variables latentes [7].

Profil recherché

Formation niveau BAC+5 (Master 2 ou école d'ingénieurs), connaissance en statistiques théoriques et computationnelles, maîtrise d'un langage de programmation indispensable; rigueur scientifique, curiosité intellectuelle, facilité de communication.

Modalités pratiques

Le stage s'inscrit dans le cadre du projet ANR Stat4Plant. Il se déroulera au centre INRAE de Jouy-en-Josas dans l'unité MaIAGE. La durée du stage sera de cinq ou six mois, entre février et septembre 2025. La gratification mensuelle est d'environ 550 euro (taux légal). L'encadrement sera réalisé par Estelle Kuhn (INRAE, MaIAGE) et Sarah Lemler (CentraleSupélec, MICS). Le stage pourra possiblement déboucher sur un sujet de thèse combinant des statistiques mathématiques, computationnelles et des applications en sciences du vivant.

Références bibliographiques

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.
[2] Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the LASSO. Annals of Statistics, 44(3), 907-927.
[3] Van de Geer, S., Bühlmann, P., Ritov, Y. A., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics, 42(3),1166-1202.
[4] Van De Geer, S. (2019). On the asymptotic variance of the debiased Lasso, Electronic Journal of Statistics (13),2970-3008.
[5] Xia, L., Nan, B., & Li, Y. (2023). Debiased lasso for generalized linear models with a diverging number of covariates. Biometrics, 79(1), 344-357.
[6] Delattre, M., & Kuhn, E. (2023). Computing an empirical Fisher information matrix estimate in latent variable models through stochastic approximation. Computo
[7] Baey, C., Delattre, M., Kuhn, E., Leger, J. B., & Lemler, S. (2023). Efficient preconditioned stochastic gradient descent for estimation in latent variable models. In International Conference on Machine Learning (pp. 1430-1453). PMLR.

Post-selection inference for high-dimensional mixed-effects models

Master's Level Internship Starting in Early 2025

Université Paris-Saclay, INRAE
Unit MaIAGE Mathematics and Computer Science Applied to Genome and Environment, Jouy-en-Josas

To apply, please send a **complete application package** containing **your CV, cover letter, and your most recent transcripts from Master 1 or equivalent studies** to estelle.kuhn@inrae.fr and sarah.lemler@centralesupelec.fr.

Context

The framework of repeated measures involves situations where each individual in a population is measured multiple times under different conditions, for instance, over time or across various environmental conditions. Mixed-effects models are well-suited for such data as they allow for modeling both the variability within repeated measurements for an individual and the variability between individuals in the population. From a statistical modeling perspective, these models are latent variable models that include both fixed effects and unobserved random effects. They can also incorporate covariates characterizing the individuals. In terms of inference, one may focus on estimating model parameters, selecting relevant covariates, and predicting model outcomes.

When dealing with a large number of covariates, such as genetic markers, it is common to work under the sparsity assumption, where only a small subset of covariates are relevant. The goal is then to identify these covariates. This selection can be performed, for example, using a penalized estimator like the LASSO [1]. A reduced model, keeping only the selected covariates, may then be considered. However, several classical results of asymptotic statistics are no longer valid for this model because it is itself random, having been selected based on the data [2]. In particular, it is no longer possible to construct based on standard results valid confidence intervals with expected coverage rates for the estimators and predictions.

To address this issue, a debiasing approach for the LASSO estimator has been proposed in the context of linear and generalized linear models [3,4]. Theoretical results have been established for the debiased estimator in these contexts, which also provide guarantees in prediction.

The aim of this internship is to extend these methods to the context of mixed-effects models and to apply the methods developed to the analysis of the variability of the senescence process for a population of wheat genotypes.

Internship objectives

* Propose a debiased estimator derived from the LASSO-type regularized estimator in a linear mixed-effects model,
* Study its theoretical properties of convergence and asymptotic normality,
* Develop and implement an algorithm to compute the debiased estimator and validate it on simulated data,
* Propose confidence intervals for estimation and prediction and validate them on simulated data,

* Compare the performance of the developed method with existing methods,
* Extend the developments to the case of nonlinear mixed-effects models.
* Analyze real wheat data.

Mathematical aspects The approach considered to construct a debiased estimator from the LASSO-type estimator is that proposed in [3]. The goal is to define a new estimator with reduced bias using an analytical form derived from the Karush-Kuhn-Tucker conditions and an estimator of the Fisher information for the model. The latter can be obtained based on the developments proposed in [6]. The proposed estimator can be computed using stochastic gradient algorithms adapted to latent variable models [7].

Desired Profile

Candidates should have a BAC+5-level education (Master's degree or engineering school), skills in theoretical and computational statistics, proficiency in a programming language is essential, along with scientific rigor, intellectual curiosity, and strong communication skills.

Practical Details

The internship is part of the ANR Stat4Plant project. It will take place at the INRAE center in Jouy-en-Josas within the MaIAGE unit. The duration of the internship will be five to six months, between February and September 2025. The monthly stipend is approximately 550 euros (legal rate). The internship will be supervised by Estelle Kuhn (INRAE, MaIAGE) and Sarah Lemler (CentraleSupélec, MICS). This internship may potentially lead to a PhD topic combining mathematical and computational statistics with applications in life sciences.

References

[1] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.
[2] Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the LASSO. Annals of Statistics, 44(3), 907-927.
[3] Van de Geer, S., Bühlmann, P., Ritov, Y. A., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics, 42(3),1166-1202.
[4] Van De Geer, S. (2019). On the asymptotic variance of the debiased Lasso, Electronic Journal of Statistics (13),2970-3008.
[5] Xia, L., Nan, B., & Li, Y. (2023). Debiased lasso for generalized linear models with a diverging number of covariates. Biometrics, 79(1), 344-357.
[6] Delattre, M., & Kuhn, E. (2023). Computing an empirical Fisher information matrix estimate in latent variable models through stochastic approximation. Computo
[7] Baey, C., Delattre, M., Kuhn, E., Leger, J. B., & Lemler, S. (2023). Efficient preconditioned stochastic gradient descent for estimation in latent variable models. In International Conference on Machine Learning (pp. 1430-1453). PMLR.