

English version below

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

Stage niveau M2 début 2026

Université Paris Saclay - CentraleSupélec - INRAE

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.lemmler@centralesupelec.fr.

Contexte

On considère une population d'individus pour laquelle une caractéristique est mesurée plusieurs fois dans des conditions différentes, par exemple la hauteur d'une plante au cours du temps. Les modèles à effets mixtes sont bien adaptés pour ce type de données, appelées mesures longitudinales, car ils permettent de modéliser à la fois la variabilité présente au sein des mesures 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. On cherche à estimer des paramètres du modèle, à sélectionner les covariables influentes et à prédire la caractéristique.

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 pour la caractéristique considérée. L'objectif est alors de sélectionner ces covariables, par exemple via un estimateur pénalisé de type LASSO. 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.

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 sur données simulées,
- * comparer les performances de la méthode développée à d'autres méthodes existantes
- * étendre les développements au cas de modèles non linéaires à effets mixtes

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. L'estimateur proposé pourra être calculé via des algorithmes de type gradient stochastique adaptés aux modèles à variables latentes [5].

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 AgriCoolTools. Il se déroulera au laboratoire MICS à l'École CentraleSupélec sur le plateau de Saclay. La durée du stage sera de cinq ou six mois, entre février et septembre 2026. La gratification mensuelle est d'environ 650 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] 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.
- [2] 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.
- [3] Van De Geer, S. (2019). On the asymptotic variance of the debiased Lasso, *Electronic Journal of Statistics* (13), 2970-3008.
- [4] Xia, L., Nan, B., & Li, Y. (2023). Debiased lasso for generalized linear models with a diverging number of covariates. *Biometrics*, 79(1), 344-357.
- [5] 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 2026

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.lemmler@centralesupelec.fr.

Context

Consider a population of individuals for which a characteristic is measured several times under different conditions, for example the height of a plant over time. Mixed-effects models are well suited to this type of data, known as longitudinal measurements, because they allow us to model both the variability within an individual's measurements and the variability between individuals in the population. From a statistical modeling perspective, these models are latent variable models, which include fixed effects and unobserved random effects. They can also incorporate covariates that characterize individuals. The aim is to estimate the model parameters, select the influential covariates, and predict the characteristic.

When considering a large number of covariates, such as genetic markers, we often work under the assumption of parsimony, according to which there are only a small number of covariates that influence the characteristic under consideration. The objective is then to select these covariates, for example using a penalized estimator such as LASSO. A reduced model obtained by retaining only the selected covariates can then be considered. However, several classical asymptotic statistical results are no longer valid in this model, because it is itself random, having been selected from the data [2]. In particular, it is no longer possible to construct valid confidence intervals with expected coverage rates for estimators and predictions based on standard results.

To address this issue, an approach based on a debiasing step for the LASSO estimator has been proposed in the case of linear and generalized linear models [3,4]. Theoretical results have been established for the unbiased estimator in these contexts and also provide predictive guarantees.

The objective of the internship is to extend these methods to the case of mixed-effects models.

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.

Mathematical aspects The approach considered for constructing an unbiased estimator based on the LASSO estimator is that proposed by [3]. The objective is to define a new estimator with lower bias via an analytical form obtained from the Karush-Kuhn-Tucker conditions. The proposed estimator can be calculated using stochastic gradient algorithms adapted to latent variable models [5].

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 AgriCoolTools project. It will take place at the MICS laboratory at the École CentraleSupélec on the Saclay plateau. The duration of the internship will be five to six months, between February and September 2026. The monthly stipend is approximately 650 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] 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.
- [2] 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.
- [3] Van De Geer, S. (2019). On the asymptotic variance of the debiased Lasso, *Electronic Journal of Statistics* (13), 2970-3008.
- [4] Xia, L., Nan, B., & Li, Y. (2023). Debiased lasso for generalized linear models with a diverging number of covariates. *Biometrics*, 79(1), 344-357.
- [5] 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.