

M2 Internship on Statistical methods for robust prediction under covariate shifts

MIA-PS, AgroParisTech/INRAE, Palaiseau
Supervisors: Gloria Buriticá, Tabea Rebafka

2027

Keywords: Extreme value statistics, linear model, variable selection, machine learning.

Context: This internship is part of the Agricoolttools research project aiming at improving predictive tools for agricultural systems and financed by the French national research agency ANR.

Application deadline: February 2027.

1 Context

A major goal of agroecological research is to support the transition toward more sustainable agricultural systems and practices. This transition is driven by urgent challenges, including climate change, biodiversity loss, and the growing scarcity of essential resources such as arable land. As the availability and diversity of data continue to increase, statistical approaches have become increasingly important for addressing these challenges and supporting evidence-based decision-making in agroecological research.

In France, maize production is an important issue, as maize is a key component of animal feed and therefore plays an essential role in the human food supply chain. At the same time, growing attention to animal health and well-being means that the nutritional quality of animal feed can no longer be overlooked. One of the key challenges in maize production is therefore to select the most appropriate variety to grow, one that can achieve high yields under increasingly variable meteorological conditions while providing good nutritional quality for livestock, in particular high digestibility. In maize, digestibility is closely related to lignin concentration. Lignin is a natural biopolymer found in plant cell walls that is poorly digestible by animals and can therefore reduce the nutritional value of maize.

One recent approach to variety selection is phenomic selection, which uses low-cost phenotypic measurements as predictors, such as near-infrared spectroscopy (NIRS) data (Robert et al., 2022). However, classical statistical prediction methods based on NIRS data often struggle to predict extreme lignin concentrations, which are typically observed under stressful environmental conditions, such as water or heat stress. In the context of climate change, improving prediction strategies for variety selection has therefore become increasingly important, as growing environmental variability may lead to conditions that differ substantially from those observed in historical data. The prediction challenge is further compounded by the high-dimensional nature of NIRS data, where the number of predictors often greatly exceeds the number of observations.

From a statistical perspective, the problem can be formulated as a supervised learning problem, where a generic scalar response variable Y is predicted from a set of predictors $X \in \mathbb{R}^p$ in a high-dimensional setting with $p > n$, where n is the number of observed replicates. The aim is to obtain reliable predictions of the response for future observations of covariates, including under changing environmental conditions of predictors that may differ substantially from those represented in the training data. A classical approach consists in modelling independent observations $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ according to the linear model

$$Y_i = \theta^T X_i + \epsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

where the errors ϵ_i are uncorrelated, satisfy $\mathbb{E}[\epsilon_i] = 0$, and have $\text{Var}(\epsilon_i) = \sigma^2$. The objective is then to infer the vector of effects $\theta \in \mathbb{R}^p$. One of the main advantages of this model is its interpretability, as the components of θ quantify the contribution of the different predictors to the response. In a high-dimensional setting, θ can be estimated using either penalisation strategies, such as the lasso, or dimension reduction approaches, such as partial least squares (PLS) (Hastie et al., 2005). However,

although these inference methods can provide good predictive performance under conditions similar to those observed in the training data, they do not in general guarantee accurate predictions when the distribution of the covariates changes. In particular, prediction performance can deteriorate when future observations correspond to environmental conditions that are poorly represented or absent in the training data, such as extreme heat or water stress. This motivates the study of covariate shifts (Li et al., 2020; Shimodaira, 2000), a framework that addresses prediction when the distribution of the covariates in future data differs from that of the training data. In the present context, this framework provides a natural statistical perspective for studying robustness of variety selection under changing environmental conditions.

The objective of this internship is to investigate distributional covariate shifts and their impact on classical inference methods for high-dimensional data as lasso or PLS and to develop strategies to enhance prediction performance for future observations.

2 Scientific description of the internship

The problem of predicting under changing environmental conditions received important attention under the name of covariate shifts. In this case we distinguish the distribution of the training sample of covariates admitting a density p_0 from the distribution of covariates under the j -th environmental condition admitting a density p_j . In a covariate shift context, one main goal is then to understand the loss function

$$\text{loss}_j(\theta) = \mathbb{E}_{\substack{X \sim p_j, \\ Y|X \sim q(y|\theta^*, x)}} \left[(Y - \theta^T X)^2 \right], \quad (2.2)$$

where $q(y|\theta^*, x)$ is the conditional density of the response given the covariate vector under the true parameter model $\theta^* \in \mathbb{R}^p$. In the case where p_j and p_0 coincide, the loss function is the classical mean squared error. We are particularly interested in understanding how the loss function behaves as we replace θ by an estimate $\hat{\theta}_j^n$ depending on the covariate shift p_j . Studying this loss function will be one of the objectives of the internship.

Concerning prediction, we aim at exploring the strategy that consists in adapting the least squares criterion by introducing weights w_i in the sum of squared residuals such as

$$\hat{\theta}_j^n = \arg \min_{\theta \in \Theta} \sum_{i=1}^n w_i (y_i - \theta^T x_i)^2. \quad (2.3)$$

The weight w_i may account for the distribution shift in x_i and as such depends on both $p_j(x_i)$ and $p_0(x_i)$. Moreover, to address the high dimension of the covariates, a ℓ_1 -penalty term may be added. Such inference strategies are deployed in the `glmnet` package in R.

One interesting example is the case, where the distributional shifts point to an overdispersion of the data in future scenarios. Two cases can then be considered, an increase of variance, or a change in the tail-index parameter quantifying the frequency at which extreme records are reached (Resnick, 2007). In the first case, the weights should behave almost as a constant, while in the second case they are expected to be polynomial, that is

$$w_i(x_i) \propto x_i^\lambda, \quad i = 1, \dots, n.$$

for a coefficient $\lambda > 0$ depending on the tail differences. In this case we are interested in inference strategies for learning the weights, while estimating the parameter of effects. Studying lasso estimation, when the coefficient λ is unknown, is the second objective of the internship.

Another interesting observation is that the lasso method is also derived by assuming a centered Laplace prior on the parameter of effects within a Bayesian framework (Hans, 2009). We could combine the covariate shift strategy with a second strategy consisting in setting a more disperse prior on the parameter of effects.

3 Scientific goals of the internship

The internship aims at improving prediction in a high-dimensional setting under changing environmental conditions. During the internship the student is expected to perform the following tasks:

1. Conduct a literature review on covariate shifts and extreme value statistics.

2. Investigate the loss function under different distributional assumptions on the covariates. Run numerical experiments and theoretical analysis.
3. Study the weighted maximum likelihood estimates, theoretical properties of the weights under different distributional assumptions and numerical experiments on different weight strategies.
4. Define a validation protocol to evaluate robustness properties of the estimation methods.
5. Compare the performance of the proposed approach with the state of the art.

The outcome of this internship is a review on classical methods, and a mathematical analysis of covariates shifts for lasso methods under distributional shifts with heavy tails. A second outcome of the internship is an implementation of the methods and numerical experiments comparing methods in terms of robustness. A validation protocol for assessing robustness should also be defines. Application of the inference strategy to the maize dataset will be the final outcome of the project.

4 Environment

- Research unit: UMR MIA Paris-Saclay, AgroParisTech/INRAE, Palaiseau
- Location: AgroParisTech Campus, 22 place de l'agronomie, 91120 Palaiseau
- Supervision: Gloria Buritica, a researcher in extreme value theory with an expertise in climate applications and extrapolation regression algorithms, and Tabea Rebafka, a researcher in statistical machine learning and computational statistics.
- Starting date: flexible, starting in February 2027 or later.
- Duration: 6 months.
- Salary: as an intern, you'll receive a gratification which is capped around 650 euros/month net.

The candidate will have an office, and benefit from the work environment of the MIA Paris-Saclay laboratory, with many PhD students and postdocs working on statistical modeling and machine learning for the life and environmental sciences.

5 Expected profile

We are looking for a motivated and curious student with a background in statistics, mathematics, or machine learning and an interest in applications to the life sciences. The candidate should be in the second year of a Master's degree in Mathematics or Statistics, or in the final year of an engineering school. We do not expect candidates to have mastered all the skills listed below before starting the internship. The most important qualities are a willingness to learn, curiosity, and the ability to invest in the project and develop new skills throughout the internship. In particular, the candidate should have, or be willing to develop, some experience in:

- bibliographic research, analysis, and synthesis,
- statistical modelling,
- R programming and, more generally, an interest in scientific programming,
- applications of statistics and machine learning to the life sciences.

6 Application

Send an email with your CV, a cover letter and an academic transcript to gloria.buritica@agroparistech.fr and tabea.rebafka@agroparistech.fr with the subject line "Application for extreme impact studies internship". Applications are considered as they are received. Latest application deadline is February 2027.

References

- Chris Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.
- Fengpei Li, Henry Lam, and Siddharth Prusty. Robust importance weighting for covariate shift. In *International conference on artificial intelligence and statistics*, pages 352–362. PMLR, 2020.
- Sidney Ira Resnick. *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York, 2007.
- Pauline Robert, Charlotte Brault, Renaud Rincent, and Vincent Segura. Phenomic selection: A new and efficient alternative to genomic selection. *Genomic prediction of complex traits: Methods and protocols*, pages 397–420, 2022.
- Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.