Inria Saclay, MIND team http://team.inria.fr/mind

NeuroSpin,
bat 145, CEA Saclay

# Developer position: Variable Importance analysis and visualization in Python

**Main topic:** open-source software for variable importance analysis
**Keywords:** machine learning, variable importance, inference
**Research team:** MIND (Inria Saclay and CEA)
**Contact:** Bertrand Thirion, bertrand.thirion@inria.fr
**Start and duration of contract:** 01/12/2024, for 2 years
**Salary:** Depending on experience: 28 to 48 k€/ year --free from charges
**Application:** Interested candidate should send CV and motivation letter

## Context

**Variable importance** The use of Artificial Intelligence (AI) techniques has become pervasive in many fields of sciences, owing to their ability to perform prediction of an outcome of interest given some complex input data. However, for the sake of scientific knowledge, AI cannot be used as a ready-made recipe, but should lead to scientific explanations, requiring the use of Explainable AI, in particular variable importance (VI) analysis. It consists in assessing how much each variable matters in the prediction of an outcome.
VI can come in many different flavors: local (i.e. related to a given instance or a group of instance) or global (across all instances), it becomes difficult for high-dimensional data. High-dimensional cases arise in many applications, such as neuroscience, genetics / genomics, population analysis. We aim to share a software that yields valid confidence intervals on VI, so that it can be used for settings where strict guarantees are needed: medicine, and more generally, scientific discovery. Most techniques based on sensitivity analysis do not offer such guarantees. In recent works, researchers of the MIND team have uncovered some approaches that provide reliable solutions for the linear case (meaning that the outcome is a linear function of some of the covariates) in high dimension(1), that we have assessed empirically in different applications (2; 3; 4). We have then successfully addressed the case of more general learners (5; 6; 4). These results are based on two key insights:

- The core approach for this type of inference is conditional association tests (2; 5). Such tests can then be interpreted as such or combined into Shapley values (7). An efficient and generic way to perform such a test is to perform conditional permutations (5).
- When the number of features is very large, grouping features becomes essential to make the problem well-posed. This can rely on prior knowledge or machine learning approaches, such as clustering (1; 6).

**We propose to share these technological advances by building a generic library, hidimstat for variable importance analysis.**

**Need of a Variable importance analysis software** Such a library is needed in the Python ecosystem, because current frameworks do not provide good enough solutions:

- Scikit learn only provided limited support for VI, in particular, it mostly relied on permutation importance, which is known to be inaccurate (5).
- Several frameworks are available in R caret, varimp, iml but their usability is questionable, and the link with Python is too brittle.
- the iml library shows many features and particularly nice visualization, but it is obviously confined to small data and does not use state-of-the-art techniques
- Many of the theoretical and algorithmic developments are still recent, hence the community has not addressed the question yet.

**Importance for the MIND team** Within MIND, several works have been performed to address the question:

- For high-dimensional linear models, the EncluDL framework have been built (1; 2; 3)
- For generalized linear models, the knockoff approach has been shown to be versatile (8). MIND has developed several improvements upon it (9)
- For the non linear case, *Conditional Permutation Inference* and the block version that includes grouping, have been proposed (5; 6).

The corresponding code has been pushed to the hidimstat package, but this still stands as publication-related code rather than a well-designed library. Our aim is to enhance it to align its quality with community standards (tests, documentation, features, API homogeneity, code quality).
While a VI library is naturally generic in terms of application, it is good to retain a few application scenarii to ensure that the library is indeed addressing them properly. The MIND team is particularly interested in the following cases:

- Population imaging, that consists is assessing what image-related features predict relevant mental health or disease information
- Identification of cognitive function, that consists in assessing regions whose activity predicts a certain cognitive state or cognitive task of an individual, based on many brain images labeled with cognitive terms. In a slightly different setting, this question can be addressed based on stroke patients databases, where one assesses which lesion patterns predict performance drops for a certain cognitive task.

We also want to outline that VI is key to many collaborative projects of the Mind team: the VITE ANR project, the EBRAIN-Health European project, a joint PhD thesis between Roche and Inria.

**Objectives of the development project** In a nutshell, the aims of the planed developments are two-fold

- Improve the quality of hidimstat (testing, CI, documentation, API homogeneity, code quality)
- Draw the attention of scientists on the technology on VI bu providing meaningful illustrations and a rich example gallery

# Mission

The main mission of the developed will be to improve the quality of hidimstat:

**testing** ensure that units tests cover all the code, that the tests are reliable, implemented on several representative environments, informative about code quality and that running them takes a reasonable amount of time.

**documentation** ensure that the docstrings are written properly and lead to a meaningful documentation using sphinx-doc, that the library features are properly showcased in the examples, and that there is a narrative documentation that clarifies the main concepts for users

**code quality** Ensure that the code conforms to the standards of modern Python development, that the API is homogeneous, variable and function naming is consistent and informative. The most expensive tools of the library will be evaluated with profiling.

**illustration** VI concepts will have to be rendered graphically in a meaningful way.

Additionally, the developers, students (PhD students), post-docs and researchers working on the project will deal with user interactions and requests.

## Job Offer description

More in detail, the following actions will be undertaken by the:

- Improve the installation on standard platforms, setup the continuous integration
- Measure and Improve the tests coverage
- Make some dependencies optional to minimize the burden of installing the library
- Building the documentation using Sphinx-doc
- Improve the API homogeneity and quality
- Write a sufficient set of examples, without redundancies, and ensure that the documentation is fast enough to generate automatically
- Improve the examples (clarity of explanations, information conveyed)
- Profiling the ode to identify bottlenecks and speed-up the code whenever this is possible
- Improve the style of the code
- Introduce nice visualization to attract more users

## Skills and profile

- Love high-quality code and open source
- Worry about users and like to communicate
- Be curious about data (ie like looking at data and understanding it)
- Have an affinity for problem-solving tradeoffs
- Good scientific Python coders
- Enjoy interacting with a community of developers
- Interest in brain imaging and its applications.
- Experience in optimization is a plus.

## Working at Inria

Established in 1967, Inria is the only public research body fully dedicated to computational sciences.
Combining computer sciences with mathematics, Inria's 3,500 researchers strive to invent the digital technologies of the future. Educated at leading international universities, they creatively integrate basic research with applied research and dedicate themselves to solving real problems, collaborating with the main players in public and private research in France and abroad and transferring the fruits of their work to innovative companies.
The researchers at Inria published over 4,500 articles in 2022. They are behind over 300 active patents and 120 start-ups. The 220 project teams are distributed in eight research centers located throughout France.

### Working with Mind team

Besides permanent researchers, the developer will be in contact with PhD students that do software development as part of their PhD contract, and with the developer team that contributes to many tools of the scientific Python ecosystem (sklearn, joblib, Benchopt). He/she will also be in contact with cognitive and clinical neuroscientists at NeuroSpin.
Mind researchers use English as a common language for their activities (daily interactions, weekly meetings, yearly retreats).

## Benefits

- Canteen and cafeteria;
- Sports equipment;
- Partial transport reimbursement

# References


[1] J.-A. Chevalier, T.-B. Nguyen, B. Thirion, and J. Salmon, "Spatially relaxed inference on high-dimensional linear models," *Statistics and Computing*, vol. 32, no. 5, p. 83, 2022.

[2] J.-A. Chevalier, T.-B. Nguyen, J. Salmon, G. Varoquaux, and B. Thirion, "Decoding with confidence: Statistical control on decoder maps," *NeuroImage*, vol. 234, p. 117921, 2021.

[3] J.-A. Chevalier, J. Salmon, A. Gramfort, and B. Thirion, "Statistical control for spatio-temporal meg/eeg source imaging with desparsified mutli-task lasso," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1759–1770, 2020.

[4] Y. Schwartz, B. Thirion, and G. Varoquaux, "Mapping cognitive ontologies to and from the brain," in *NIPS (Neural Information Processing Systems)*, (United States), Dec. 2013.

[5] A. CHAMMA, D. A. Engemann, and B. Thirion, "Statistically valid variable importance assessment through conditiona l permutations," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, and M. H. and S. Levine, eds.), vol. 36, pp. 67662–67685, Curran Associates, Inc., 2023.

[6] A. Chamma, B. Thirion, and D. Engemann, "Variable importance in high-dimensional settings requires grouping," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 11195–11203, 2024.

[7] I. Covert and S.-I. Lee, "Improving kernelshap: Practical shapley value estimation via linear regression," 2021.

[8] E. Candès, Y. Fan, L. Janson, and J. Lv, "Panning for gold: model-x knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 3, pp. 551–577.

[9] A. Blain, B. Thirion, O. Grisel, and P. Neuvial, "False discovery proportion control for aggregated knockoffs," *Advances in Neural Information Processing Systems*, vol. 36, 2024.