Statistical Inference for Variable Importance in High-Dimensional and Highly Correlated Settings

Internship presentation

Angel REYERO

MSc Mathematics & Artificial Intelligence
Institut de Mathématiques d'Orsay (IMO)
Université Paris-Saclay
Under the supervision of:
Pierre NEUVIAL & Bertrand THIRION.

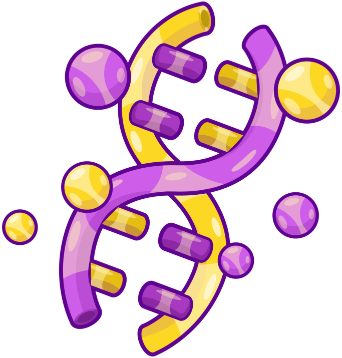
30th of September

Contents

Index

Motivation: Intrinsic Variable Importance

How can we define / learn the importance of each covariate X^j with respect to an outcome y?



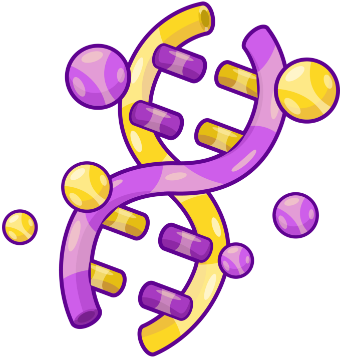
X

y

Try to study their relationship using a ML model:

$$\widehat{m} \in \underset{f \in \mathscr{F}}{\operatorname{argmin}} \widehat{\mathbb{E}} \left[\mathscr{L}(f(X), y) \right]. \tag{1}$$

How can we define / learn the importance of each covariate X^j with respect to an outcome y?



X

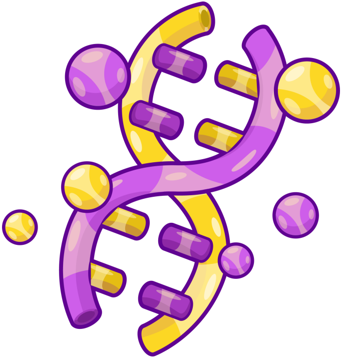
y

- Try to study their relationship using a ML model:

$$\hat{m} \in \underset{f \in \mathscr{F}}{\operatorname{argmin}} \widehat{\mathbb{E}}\left[\mathscr{L}(f(X), y)\right]. \tag{1}$$

Motivation: Intrinsic Variable Importance

How can we define / learn the importance of each covariate X^j with respect to an outcome y?



X

$y = m(X) + \varepsilon \in \mathbb{R}$

- Try to study their relationship using a ML model:

$$\widehat{m} \in \underset{f \in \mathscr{F}}{\operatorname{argmin}} \widehat{\mathbb{E}} \left[(f(X) - y)^2 \right]. \tag{2}$$

Problematic

Goals for a VI measure:

- statistically valid
- model-agnostic
- computationally feasible
- conditional approach

Main challenges:

- non-linearity
- high-dimensionality
- correlation

☹ Current approaches do not offer sufficient statistical guarantees and do not work in these complex settings.

Index

Standard approaches

- The importance of j, $\psi(j, P_0)$, is usually obtained by:

Predictability **using** the covariate j **VS** Predictability **without** the covariate j

- Approaches to measure the predictability without j (Covert et al. (2021) JMLR):
  - **Removal-based:** They refit a model $\widehat{m}_{-j}$ to regress y given X^{-j} (for example LOCO and Shapley values)
  - **Permutation-based:** They break the relationship between y and X^{-j} reusing $\widehat{m}$ (for example PFI and CPI)

Leave One Covariate Out (LOCO)

- It is defined as the plug-in estimate of $(m_{-j}(X^{-j}) := \mathbb{E}\left[y|X^{-j}\right])$

$$\psi_{\mathrm{LOCO}}(j, P_0) = \mathbb{E}\left[(y - m_{-j}(X^{-j}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right].$$

✓ It estimates the unnormalized Total Sobol Index $(\mathbb{E}\left[\mathbb{V}(y|X^{-j})\right])$.

✓ Type-I error control (Williamson et al. (2021) JASA).

Leave One Covariate Out (LOCO)

- It is defined as the plug-in estimate of $(m_{-j}(X^{-j}) := \mathbb{E}\left[y|X^{-j}\right])$

$$\psi_{\mathrm{LOCO}}(j, P_0) = \mathbb{E}\left[(y - m_{-j}(X^{-j}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right].$$

✓ It estimates the unnormalized Total Sobol Index $(\mathbb{E}\left[\mathbb{V}(y|X^{-j})\right])$.

✓ Type-I error control (Williamson et al. (2021) JASA).

× In practice: instability and invalid null hypothesis testing.

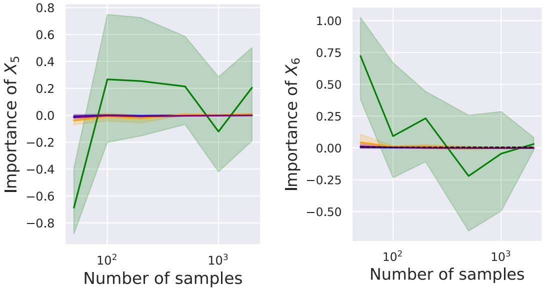


Figure 1: In green, $\widehat{\psi}_{\mathrm{LOCO}}$ for two null covariates ($p = 50$, $y = X_0X_1\mathbb{1}_{X_2>0} + 2X_3X_4\mathbb{1}_{X_2<0}$ and $X \sim \mathcal{N}(1, \Sigma)$ with $\Sigma_{i,j} = 0.6^{|i-j|}$).

Permutation Feature Importance(PFI)

- It is given by:

$$\widehat{\psi}_{\mathrm{PFI}}(j, P_0) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left((y_i - \widehat{m}(x_i^{(j)}))^2 - (y_i - \widehat{m}(x_i))^2 \right). \qquad (3)$$

where the j-th covariate is permuted.

- It tries to estimate

$$\psi_{PFI}(j, P_0) := \mathbb{E}\left[(y - m(X^{(j)}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right],$$

where $X^{(j)-j} = X^{-j}, X^{(j)j} \perp\!\!\!\perp X^{-j}, y$ and $X^{(j)j} \sim X^j$.

✓ Fast (no need to retrain $\widehat{m}$).

× Extrapolation bias (Chamma et al. (2023) NeurIPS).

× Not an interesting theoretical quantity (Bénard et al (2022) Biometrika).

Instead of breaking the relationship of X^j with X^{-j} and y, we only need to break it with y!

Permutation Feature Importance(PFI)

- It is given by:

$$\widehat{\psi}_{\mathrm{PFI}}(j, P_0) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left((y_i - \widehat{m}(x_i^{(j)}))^2 - (y_i - \widehat{m}(x_i))^2 \right). \quad (3)$$

where the j-th covariate is permuted.

- It tries to estimate

$$\psi_{PFI}(j, P_0) := \mathbb{E}\left[(y - m(X^{(j)}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right],$$

where $X^{(j)-j} = X^{-j}, X^{(j)j} \perp\!\!\!\perp X^{-j}, y$ and $X^{(j)j} \sim X^j$.

✓ Fast (no need to retrain $\widehat{m}$).

× Extrapolation bias (Chamma et al. (2023) NeurIPS).

× Not an interesting theoretical quantity (Bénard et al (2022) Biometrika).

Instead of breaking the relationship of X^j with X^{-j} and y, we only need to break it with y!

Permutation Feature Importance(PFI)

- It is given by:

$$\widehat{\psi}_{\mathrm{PFI}}(j, P_0) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left((y_i - \widehat{m}(x_i^{(j)}))^2 - (y_i - \widehat{m}(x_i))^2 \right). \tag{3}$$

where the j-th covariate is permuted.

- It tries to estimate

$$\psi_{PFI}(j, P_0) := \mathbb{E}\left[(y - m(X^{(j)}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right],$$

where $X^{(j)-j} = X^{-j}, X^{(j)j} \perp\!\!\!\perp X^{-j}, y$ and $X^{(j)j} \sim X^j$.

✓ Fast (no need to retrain $\widehat{m}$).

× Extrapolation bias (Chamma et al. (2023) NeurIPS).

× Not an interesting theoretical quantity (Bénard et al (2022) Biometrika).

Instead of breaking the relationship of X^j with X^{-j} and y, we only need to break it with y!

Conditional Permutation Importance

- (Chamma et al.(2023) NeurIPS) It is given by:

$$\widehat{\psi}_{\mathrm{CPI}}(j, P_0) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left((y_i - \widehat{m}(\widetilde{x}_i^{(j)}))^2 - (y_i - \widehat{m}(x_i))^2 \right), \qquad (4)$$

where the j-th covariate is *conditionally* permuted.

- It tries to estimate

$$\psi_{CPI}(j, P_0) := \mathbb{E}\left[(y - m(\widetilde{X}^{(j)}))^2 \right] - \mathbb{E}\left[(y - m(X))^2 \right],$$

where $\widetilde{X}^{(j)-j} = X^{-j}, \widetilde{X}^{(j)j} \perp\!\!\!\perp y | X^{-j}$ and $\widetilde{X}^{(j)j} \sim X^j | X^{-j}$.

✓ Fast and stable in practice with type-I error control.

× Not an interesting theoretical quantity.

× No theoretical foundation on the conditional permutation.

Conditional Permutation Importance

- (Chamma et al.(2023) NeurIPS) It is given by:

$$\widehat{\psi}_{\text{CPI}}(j,P_0) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left((y_i - \widehat{m}(\widetilde{x}_i^{(j)}))^2 - (y_i - \widehat{m}(x_i))^2 \right), \qquad (4)$$

where the j-th covariate is *conditionally* permuted.

- It tries to estimate

$$\psi_{CPI}(j,P_0) := \mathbb{E}\left[(y - m(\widetilde{X}^{(j)}))^2\right] - \mathbb{E}\left[(y - m(X))^2\right],$$

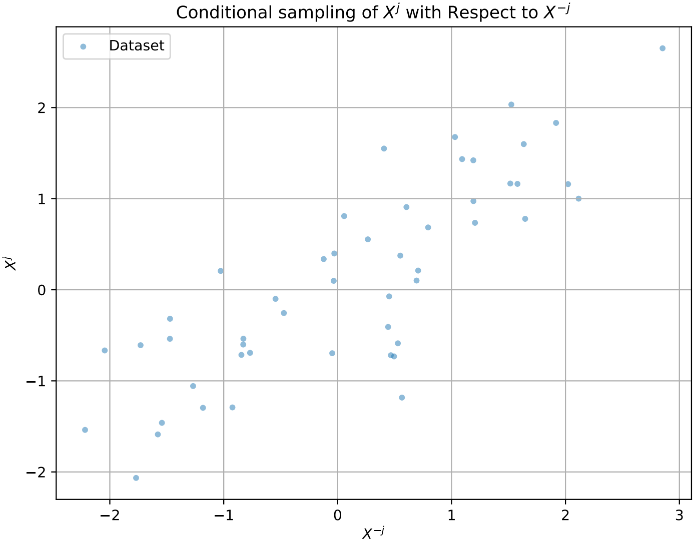
where $\widetilde{X}^{(j)-j} = X^{-j}, \widetilde{X}^{(j)j} \perp\!\!\!\perp y|X^{-j}$ and $\widetilde{X}^{(j)j} \sim X^j|X^{-j}$.

✓ Fast and stable in practice with type-I error control.

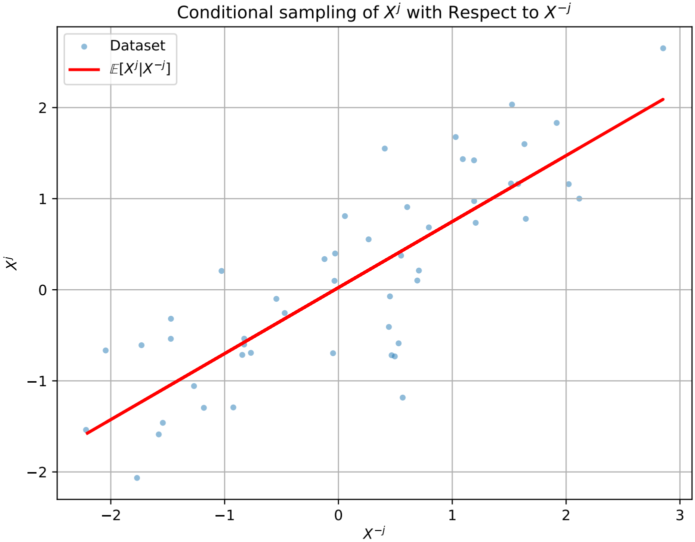
Lemma 1 (Internship contribution)

$\psi_{\text{CPI}}(j,P_0) = 2\psi_{\text{LOCO}}(j,P_0)$.

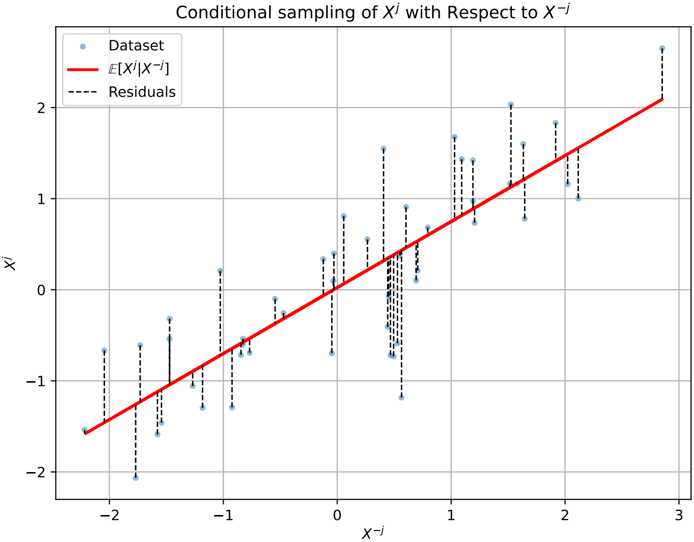
An idea to permute conditionally on X^{-j}



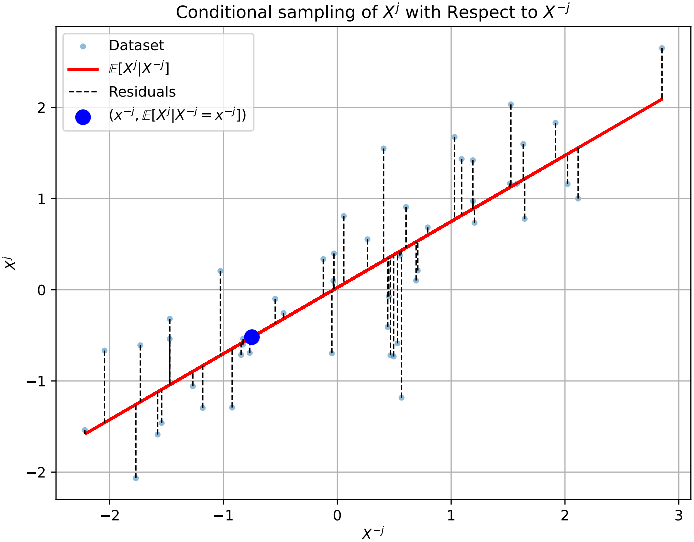
An idea to permute conditionally on X^{-j}



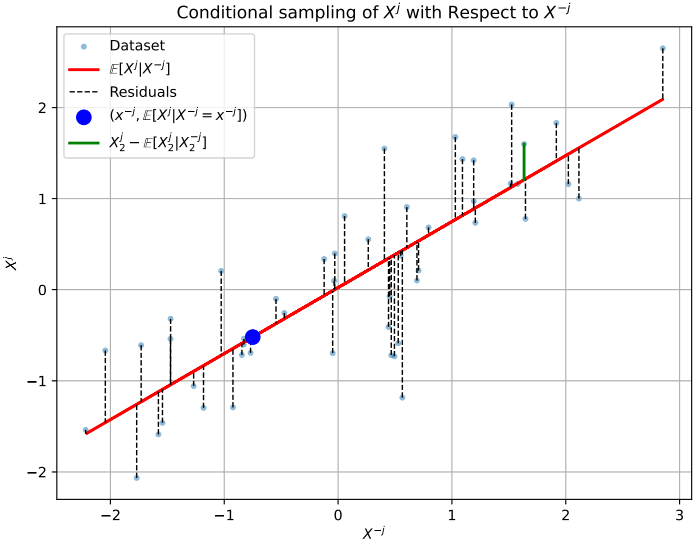
An idea to permute conditionally on X^{-j}



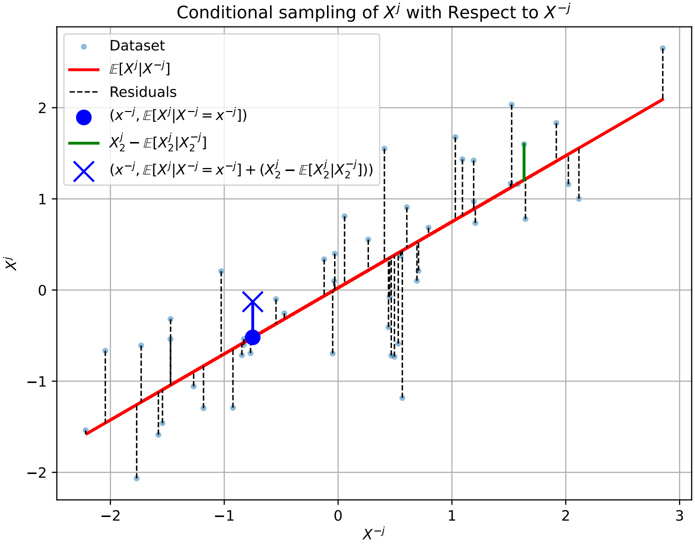
An idea to permute conditionally on X^{-j}



An idea to permute conditionally on X^{-j}



An idea to permute conditionally on X^{-j}



Validity of the conditional sampling

- In practice, we need to train a regressor $\widehat{\nu}_{-j}$ of X^j on X^{-j}. Then, for an x, we predict $\widehat{\nu}_{-j}(x^{-j})$ and add a permuted residual $(x'^j - \widehat{\nu}_{-j}(x'^j))$.

Assumption 1

$X^j = \nu_{-j}(X^{-j}) + \varepsilon$ *with* $\varepsilon \perp\!\!\!\perp X^{-j}$.

Lemma 2 (Internship contribution)

Under Assumption 1 and assuming the consistency of $\widehat{\nu}_{-j}$, the conditional step of the CPI, presented in Chamma et al.(2023) NeurIPS, is valid.

Key intermediary takeaways

- Removal-based approaches suffer from high variability.
- Permutation-based approaches are faster and more stable.
- It is possible to estimate LOCO using permutation approaches!
- LOCO is heuristically used for variable selection.
- They provide type-I error control and in practice it does not work.

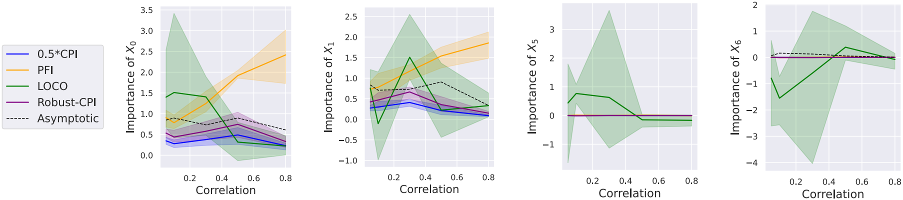


Figure 2: **Setting:** $y = X_0 X_1 \mathbb{1}_{X_2>0} + 2X_3 X_4 \mathbb{1}_{X_2<0}$ and $X \sim \mathcal{N}(1, \Sigma)$ with $\Sigma_{i,j} = \rho^{|i-j|}$ and $p = 50, n = 300$. The black dotted line represents LOCO with $n = 100000$. On the left two important covariates. On the right two null covariates. On the x-axis, we vary the correlation ρ.

Index

Knockoffs framework(Candès et al. (2018) JRSS)

Knockoffs provide a framework for controlled variable selection (find $\mathscr{H}_1 := \mathscr{H}_0^c, \mathscr{H}_0 := \{j : X^j \perp\!\!\!\perp y | X^{-j}\}$) combining three ingredients:

1. **Knockoff variables ($\widetilde{X}$):** *imitations* of X that do not preserve the relationship with y.

Knockoffs framework(Candès et al. (2018) JRSS)

Knockoffs provide a framework for controlled variable selection (find $\mathscr{H}_1 := \mathscr{H}_0^c, \mathscr{H}_0 := \{j : X^j \perp\!\!\!\perp y | X^{-j}\}$) combining three ingredients:

1. **Knockoff variables ($\widetilde{X}$):** *imitations* of X that do not preserve the relationship with y.
2. **Knockoffs statistics($\mathbf{W} \in \mathbb{R}^p$):** each coordinate W_j measures the importance of each covariate j comparing the *predictability* of original covariate with the knockoff covariate.

Example 3 (Lasso Coefficients Difference)

Regress $y \in \mathbb{R}$ on $[X, \widetilde{X}] \in \mathbb{R}^{2p}$ using LASSO and compute the estimated coefficient difference: $W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}|$.

Knockoffs framework(Candès et al. (2018) JRSS)

Knockoffs provide a framework for controlled variable selection (find $\mathscr{H}_1 := \mathscr{H}_0^c, \mathscr{H}_0 := \{j : X^j \perp\!\!\!\perp y | X^{-j}\}$) combining three ingredients:

1. **Knockoff variables ($\widetilde{X}$):** *imitations* of X that do not preserve the relationship with y.
2. **Knockoffs statistics(W $\in \mathbb{R}^p$):** each coordinate W_j measures the importance of each covariate j comparing the *predictability* of original covariate with the knockoff covariate.
3. **Threshold:** it is data-dependent and given by

$$T_q^\star = \min\left\{ t \in \mathscr{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\}. \tag{5}$$

   and $+\infty$ if empty.

This procedure provides and FDR control!

$$\mathrm{FDP}(\widehat{S}) := \frac{\left|\widehat{S} \cap \mathscr{H}_0\right|}{\left|\widehat{S}\right| \vee 1} \qquad \mathrm{FDR}(\widehat{S}) := \mathbb{E}\left[\mathrm{FDP}(\widehat{S})\right]. \tag{6}$$

Some pitfalls of the procedure

- Not easy to construct $\widetilde{X}$ in practice. Candès et al. (2018) JRSS assumed Gaussianity and they estimated the covariance matrix, which does not work in high dimension (Blain et al.(2024)).

✓ [Internship contribution] We proposed a sequential algorithm based on the CPI conditional sampling.

- The most performing statistic is the LCD, which may not work in highly non-linear settings.

✓ [Internship contribution] We proposed another statistic: the Shapley-knockoffs.

Some pitfalls of the procedure

- Not easy to construct $\widetilde{X}$ in practice. Candès et al. (2018) JRSS assumed Gaussianity and they estimated the covariance matrix, which does not work in high dimension (Blain et al.(2024)).

✓ [Internship contribution] We proposed a sequential algorithm based on the CPI conditional sampling.

- The most performing statistic is the LCD, which may not work in highly non-linear settings.

✓ [Internship contribution] We proposed another statistic: the Shapley-knockoffs.

CPI-Knockoffs

We propose another procedure to control the FDR:

1. We construct $\widetilde{X}$ in which each coordinate is conditionally sampled:

$$\widetilde{X}^j = \nu_{-j}(X^{-j}) + (X'^j - \nu_{-j}(X'^{-j})).$$

2. We construct $\mathbf{W}_{\mathrm{CPI}}$ in which each coordinate is given by:

$$\mathbf{W}_{\mathrm{CPI}}(X, \widetilde{X}, y)^j = \left(y - \widehat{m}(X^1, \ldots, \widetilde{X}^j, \ldots, X^p)\right)^2 - (y - \widehat{m}(X))^2.$$

3. We apply the threshold $T_q^\star$.

⚠ Neither $\widetilde{X}$ provides a knockoff variable, nor is $\mathbf{W}_{\mathrm{CPI}}$ a knockoff statistic.

Theorem 3 (Internship contribution)

Under some mild assumptions, CPI-Knockoffs provides an FDR control.

On the assumptions

We need Assumption (1) on the covariates, consistency of $\widehat{\nu}_{-j}$ and:

Assumption 2 (sign-flip property)

$$\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)}) \sim \varepsilon^j \left(\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)})\right), \textit{ with } \varepsilon \sim \mathscr{U}(\{1, -1\}), j \in \mathscr{H}_0.$$

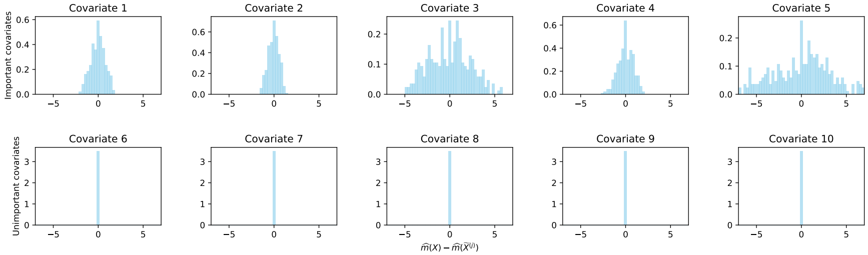


Figure 3: **Setting:** $y = X_1 - X_2 + 2X_3 + X_4 - 3X_5$, $n_{\text{train}} = 700, n_{\text{test}} = 300$, $d = 5000$ and $X \sim \mathscr{N}(1, \Sigma)$ with $\Sigma_{i,j} = 0.6^{|i-j|}$. Each plot corresponds to a specific coordinate.

Index

Take-home message

The main contributions of this internship:

- We established the **validity of the conditional sampling step**.
- We linked removal-based approaches with permutation-based ones, providing more stable LOCO estimates: **0.5CPI** and **RobustCPI**.
- We proposed a algorithm for **constructing knockoff variables**.
- We introduced a new knockoff statistic: **Shapley-knockoffs**.
- We proposed an efficient and parallelizable procedure that controls FDR without the strong computational issues of the knockoffs: **CPI-knockoffs**.

Perspectives

Next steps:

- Instead of controlling the **FDR**, it would be interesting to control the **FDP**.
- Perform numerical experiments to **compare** the performance of the proposed methods with the state-of-the art methods.
- Introduce **grouping** of the covariates to handle highly correlated settings.

Index

References

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Minimax rate of consistency for linear models with missing values, 2022.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Random features models: a way to study the success of naive imputation, 2024.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055 – 2085, 2015. doi: 10.1214/15-AOS1337. URL https://doi.org/10.1214/15-AOS1337.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological*, 57:289–300, 1995. URL https://api.semanticscholar.org/CorpusID:45174121.

References

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165 – 1188, 2001. doi: 10.1214/aos/1013699998. URL https://doi.org/10.1214/aos/1013699998.

Alexandre Blain, Bertrand Thirion, Olivier Grisel, and Pierre Neuvial. False discovery proportion control for aggregated knockoffs, 2023. URL https://arxiv.org/abs/2310.10373.

Alexandre Blain, Bertrand Thirion, Julia Linhart, and Pierre Neuvial. When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs, 2024. URL https://arxiv.org/abs/2407.06892.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

References

Clément Bénard, Gérard Biau, Sébastien da Veiga, and Erwan Scornet. Shaff: Fast and consistent shapley effect estimates via random forests, 2022a. URL https://arxiv.org/abs/2105.11724.

Clément Bénard, Sébastien da Veiga, and Erwan Scornet. MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA, 2022b.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection, 2017. URL https://arxiv.org/abs/1610.02351.

Ahmad Chamma, Denis A. Engemann, and Bertrand Thirion. Statistically valid variable importance assessment through conditional permutations, 2023.

References

Ian Covert, Scott M. Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *CoRR*, abs/2011.14878, 2020. URL https://arxiv.org/abs/2011.14878.

Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 2nd edition, 2021. doi: 10.1201/9781003158745. URL https://doi.org/10.1201/9781003158745.

Derek Hansen, Brian Manzo, and Jeffrey Regier. Normalizing flows for knockoff-free controlled feature selection, 2022. URL https://arxiv.org/abs/2106.01528.

References

Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, 1996. ISSN 0951-8320. doi: https://doi.org/10.1016/0951-8320(96)00002-6. URL https://www.sciencedirect.com/science/article/pii/0951832096000026.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance, 2021.

Wei Jiang, Julie Josse, and Marc Lavielle. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2019.106907. URL

References

https://www.sciencedirect.com/science/article/pii/S0167947319302622.

I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures, 2020. URL https://arxiv.org/abs/2002.11097.

Angel D Reyero Lobo, Alexis Ayme, Claire Boyer, and Erwan Scornet. Harnessing pattern-by-pattern linear classifiers for prediction with missing data, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. URL http://arxiv.org/abs/1705.07874.

References

Norm Matloff and Pete Mohanty. *A Method for Handling Missing Values in Prediction Applications*, 2023. URL `https://github.com/matloff/toweranNA`. R package version 0.1.0.

Xinlei Mi, Baiming Zou, Fei Zou, and Jianhua Hu. Permutation-based identification of important biomarkers for complex diseases via machine learning models. *Nature Communications*, 12(1):3008, 2021. doi: 10.1038/s41467-021-22756-2. URL `https://doi.org/10.1038/s41467-021-22756-2`. Published: 21 May 2021.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models, 2021.

References

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values, 2020.

Tuan-Binh Nguyen, Jérôme-Alexis Chevalier, Bertrand Thirion, and Sylvain Arlot. Aggregation of Multiple Knockoffs. In *ICML 2020 - 37th International Conference on Machine Learning*, number 119 in Proceedings of the ICML 37th International Conference on Machine Learning,, Vienne / Virtual, Austria, July 2020. URL `https://hal.science/hal-02888693`.

Art B. Owen. Sobol' indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. doi: 10.1137/130936233. URL `https://doi.org/10.1137/130936233`.

References

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2021. URL https://arxiv.org/abs/1912.02762.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335739.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization, 2016.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4), August 2015. ISSN 0090-5364. doi: 10.1214/15-aos1321. URL http://dx.doi.org/10.1214/15-AOS1321.

References

Matteo Sesia, Eugene Katsevich, Stephen Bates, Emmanuel Candès, and Chiara Sabatti. Multi-resolution localization of causal variants across the genome. *Nature Communications*, 11(1):1093, feb 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14791-2. URL `https://doi.org/10.1038/s41467-020-14791-2`.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), June 2020. ISSN 0090-5364. doi: 10.1214/19-aos1857. URL `http://dx.doi.org/10.1214/19-AOS1857`.

Eunhye Song, Barry L. Nelson, and Jeremy Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016. doi: 10.1137/15M1048070. URL `https://doi.org/10.1137/15M1048070`.

References

Wesley Tansey, Victor Veitch, Haoran Zhang, Raul Rabadan, and David M. Blei. The holdout randomization test for feature selection in black box models, 2021. URL https://arxiv.org/abs/1811.00645.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Isabella Verdinelli and Larry Wasserman. Feature importance: A closer look at shapley values and loco, 2023.

References

Brian Williamson and Jean Feng. Efficient nonparametric statistical inference on population feature importance using shapley values. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10282–10291. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/williamson20a.html`.

Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021a. doi: https://doi.org/10.1111/biom.13392. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13392`.

Brian D. Williamson, Peter B. Gilbert, Noah R. Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance, 2021b.

Thank You, Questions?

Index

On the assumption of symmetry of the difference $\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)}) \sim \widehat{m}(\widetilde{X}^{(j)}) - \widehat{m}(X)$ for $j \in \mathscr{H}_0$:

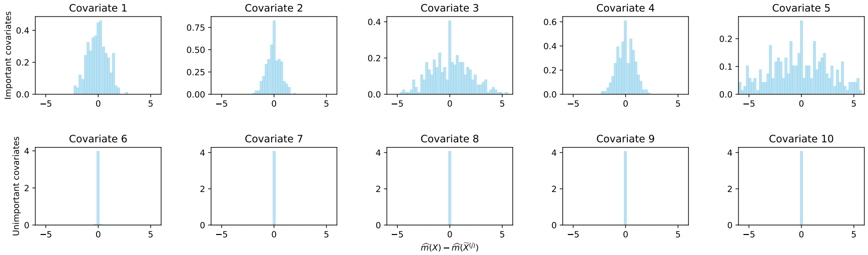


Figure 4: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. Each plot corresponds to a specific coordinate.

Linear setting

Power across the individuals:

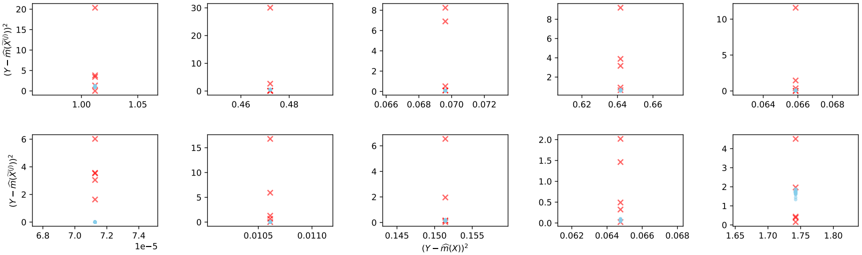


Figure 5: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. Each plot represents an individual. On the x-axis, we have the prediction error made on the individual, and on the y-axis, the error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

Linear setting

Power by aggregating the individuals using the mean:

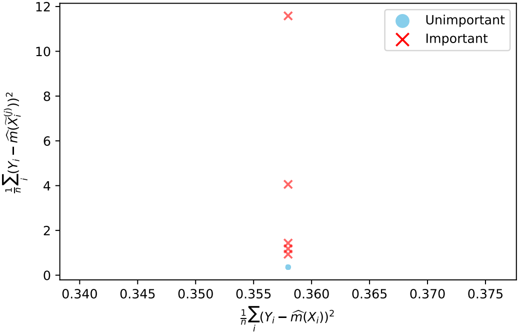


Figure 6: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. It represents the mean of the errors made across the individuals. On the x-axis, we have the mean prediction error, and on the y-axis, the mean error made by changing a coordinate using a conditionally independent sample.

Linear setting

The distribution of the statistic on the null covariates:

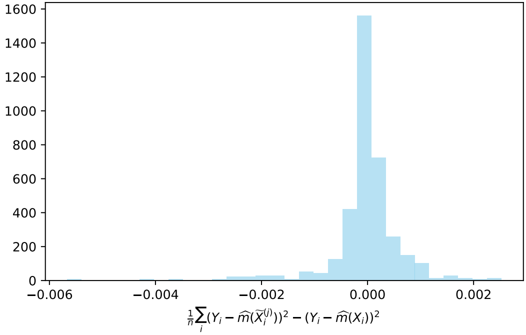


Figure 7: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 500$. It represents the histogram of the mean of difference between the errors made across the individuals by using the original and a conditionally independent sample on the null covariates.

High dimensional linear setting

On the assumption of symmetry of the difference
$\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)}) \sim \widehat{m}(\widetilde{X}^{(j)}) - \widehat{m}(X)$ for $j \in \mathscr{H}_0$:

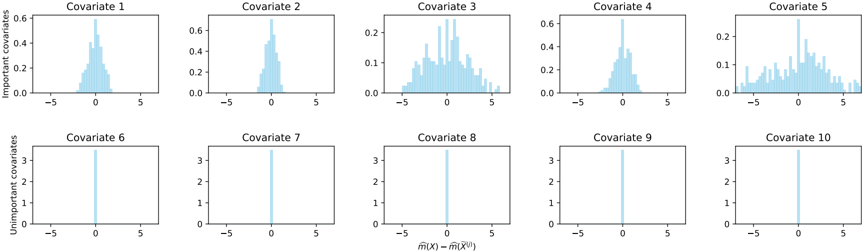


Figure 8: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. Each plot corresponds to a specific coordinate.

Power across the individuals:

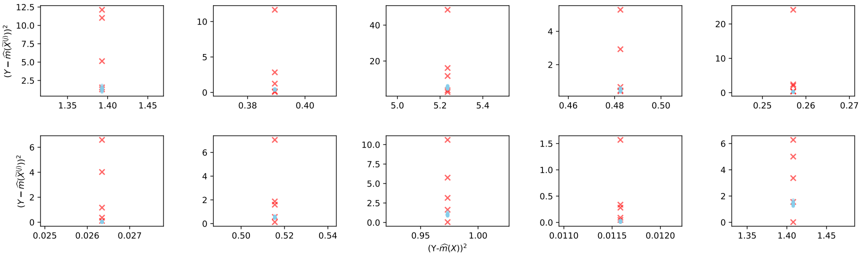


Figure 9: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. Each plot represents an individual. On the x-axis, we have the prediction error made on the individual, and on the y-axis, the error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

High dimensional linear setting

Power by aggregating the individuals using the mean:

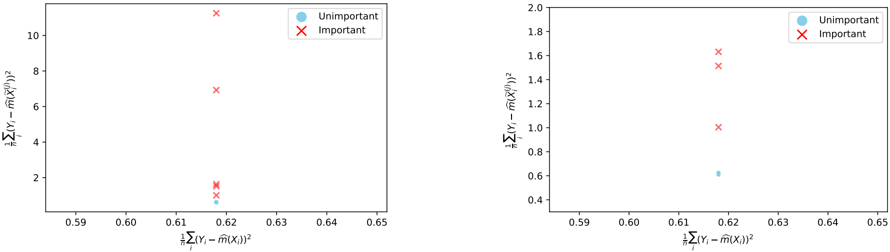


Figure 10: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. The figure represents the mean of the errors made across the individuals. On the x-axis, we have the mean prediction error, and on the y-axis, the mean error made by changing a coordinate using a conditionally independent sample. On the right, the augmented figure.

High dimensional linear setting

The distribution of the statistic on the null covariates:

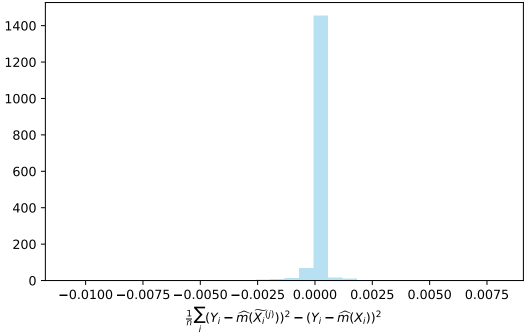


Figure 11: $y = X_0 - X_1 + 2X_2 + X_3 - 3X_4$, $n_{\text{train}} = 700, n_{\text{test}} = 300$ and $d = 5000$. It represents the histogram of the mean of difference between the errors made across the individuals by using the original and a conditionally independent sample on the null covariates.

Non-linear setting

On the assumption of symmetry of the difference $\widehat{m}(X) - \widehat{m}(\widetilde{X}^{(j)}) \sim \widehat{m}(\widetilde{X}^{(j)}) - \widehat{m}(X)$ for $j \in \mathscr{H}_0$:

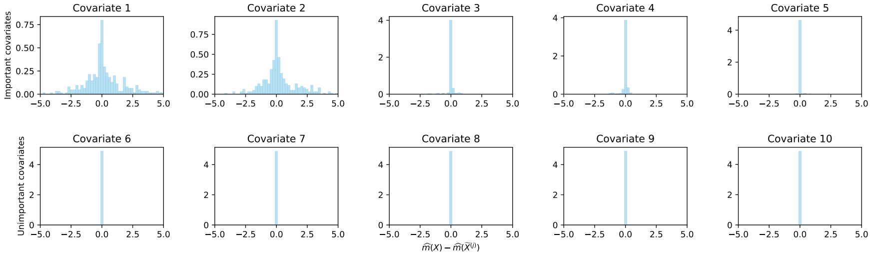


Figure 12: $y = X_1 X_2 \mathbb{1}_{X_3>0} + 2X_4 X_5 \mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300$, $d = 500$ and $X \sim \mathscr{N}(1, \Sigma)$ with $\Sigma_{i,j} = 0.6^{|i-j|}$. Each plot corresponds to a specific coordinate.

Non-linear setting

Power across the individuals:

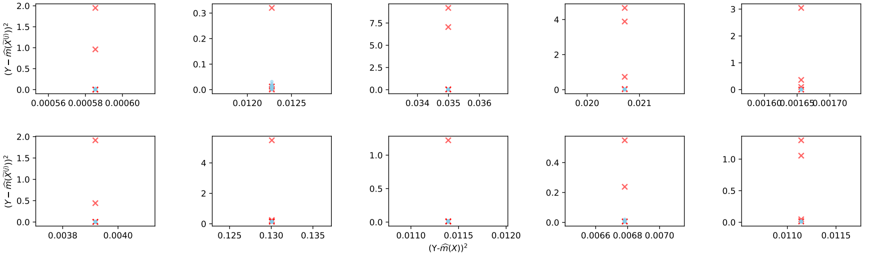
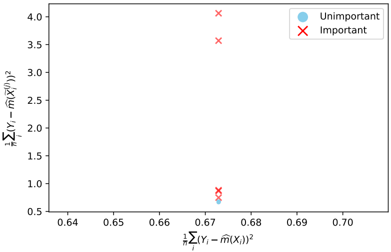


Figure 13: $y = X_1 X_2 \mathbb{1}_{X_3>0} + 2X_4 X_5 \mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300, d = 500$ and $X \sim \mathcal{N}(1, \Sigma)$ with $\Sigma_{i,j} = 0.6^{|i-j|}$. Each plot represents an individual. On the x-axis, we have the prediction error made on the individual, and on the y-axis, the error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

Non-linear setting

Power by aggregating the individuals using the mean:



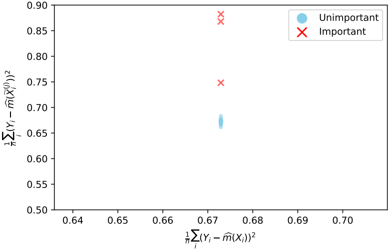


Figure 14: $y = X_1 X_2 \mathbb{1}_{X_3>0} + 2X_4 X_5 \mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300$, $d = 500$ and $X \sim \mathcal{N}(1, \Sigma)$ with $\Sigma_{i,j} = 0.6^{|i-j|}$. It represents the mean of the errors made across the individuals. On the x-axis, we have the mean prediction error, and on the y-axis, the mean error made by changing a coordinate using a conditionally independent sample. The red crosses stand for the relevant covariates and the blue dots for the null covariates.

Non-linear setting

The distribution of the statistic on the null covariates:

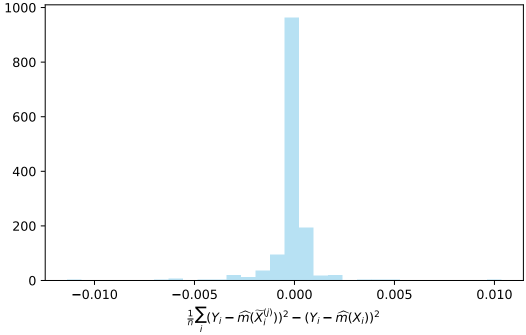


Figure 15: $y = X_1 X_2 \mathbb{1}_{X_3>0} + 2X_4 X_5 \mathbb{1}_{X_3<0}$, $n_{\text{train}} = 700, n_{\text{test}} = 300$, $d = 500$ and $X \sim \mathcal{N}(1, \Sigma)$ with $\Sigma_{i,j} = 0.6^{|i-j|}$. It represents the histogram of the mean of difference between the errors made across the individuals by using the original and a conditionally independent sample on the null covariates.

Index

LOCO estimates: 0.5CPI

Definition 4 (0.5CPI)

Given a covariate j, a training sample of size n_{train} and a test sample of size n_{test}, we train the regressors $\widehat{m}$ and $\widehat{\nu}_{-j}$ over the train set, we compute the residuals over the test set, and the new LOCO estimate is given by

$$\widehat{\psi}_{0.5\text{CPI}}(j, P_0) = \frac{1}{2n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left((y_i - \widehat{m}(\widetilde{x}_i^{(j)}))^2 - (y_i - \widehat{m}(x_i))^2 \right).$$

LOCO estimates: Robust-CPI

Definition 5 (Robust-CPI)

Given a train, test and calibration set, we define the Robust-CPI estimate $\widehat{\psi}_{\mathrm{Robust-CPI}}(j, P_0)$ as

$$\frac{n_{\mathrm{cal}}}{n_{\mathrm{cal}}+1}\left(\frac{1}{n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}}\left(y_i - \frac{1}{n_{\mathrm{cal}}}\sum_{k=1}^{n_{\mathrm{cal}}}\widehat{m}(\widetilde{x}_{i,k}^{(j)})\right)^2 - \left(y_i - \widehat{m}(x_i)\right)^2\right), \tag{7}$$

where $\widetilde{x}_{i,k}^{(j)}$ is computed as:

$$\widetilde{x}_{i,k}^{(j),l} = \begin{cases} x_i^l & \text{if } l \neq j \\ \widehat{\nu}_{-j}(x_i^{-j}) + \left[x_k^j - \widehat{\nu}_{-j}(x_k^{-j})\right] & \text{if } l = j. \end{cases} \tag{8}$$