Robust-CPI: A Double Robust Approach to improve Variable Selection

Angel REYERO LOBO

Workshop on statistical inference for complex data
IMT & Inria Paris-Saclay
Ongoing work with:
Pierre NEUVIAL & Bertrand THIRION.

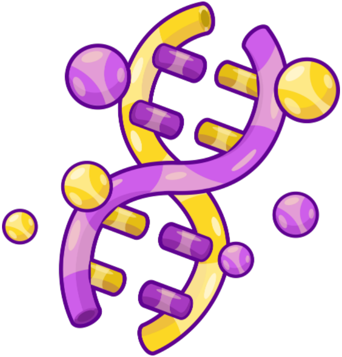
December 9, 2024

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

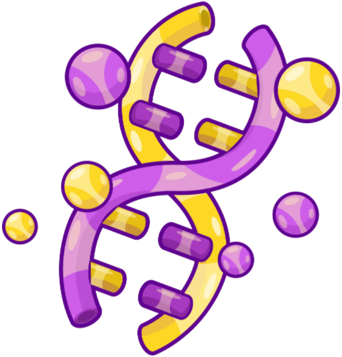
y

Try to study their relationship using a ML model:

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

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:

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

Problematic: Model misspecification

Goals for a VI measure:

- statistically valid
- model-agnostic
- computationally feasible

Main challenges:

- non-linearity
- high-dimensionality
- correlation

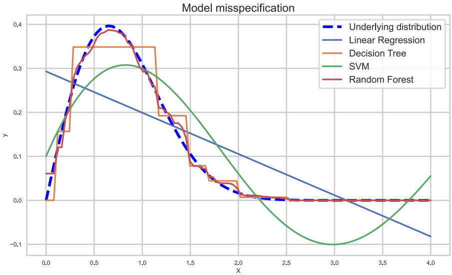


Figure 1: Interpreting the underlying distribution with simple models may be misleading. We need a **model-agnostic** measure.

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:** Refit a model $\widehat{m}_{-j}$ to regress y given X^{-j}.
  - **Permutation-based:** Modify X^j and predict with $\widehat{m}$.

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:** Refit a model $\widehat{m}_{-j}$ to regress y given X^{-j}.
  - **Permutation-based:** Modify X^j and predict with $\widehat{m}$.

Goal: the Generalized Total Sobol Index

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

(Williamson et al. (2021) Biometrics, Williamson et al. (2021) JASA, Bénard et al. (2022) Biometrika, Verdinelli et al. (2023) Statistical Science).

Literature review: General model-agnostic framework

- **Leave One Covariate Out(LOCO)**:

$$\widehat{\psi}_{\mathrm{LOCO}}(j, P_0) = \frac{1}{n_{\mathrm{test}}} \sum_i \mathscr{L}(y_i, \widehat{m}_{-j}(x_i^{-j})) - \mathscr{L}(y_i, \widehat{m}(x_i)). \quad (2)$$

✓ It estimates an interpretable quantity ($\psi_{\mathrm{TSI}}(j, P_0)$).
✓ Type-I error control (Williamson et al. (2021) JASA).
× In practice: unstable and computational intensive.

Literature review: General model-agnostic framework

- **Leave One Covariate Out(LOCO)**:

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

- **Permutation Feature Importance(PFI)**:

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

where the j-th covariate is permuted.

- ✓ Fast (no need to retrain $\widehat{m}$).
- × Extrapolation bias (Chamma et al. (2023) NeurIPS).
- × Not interesting theoretically (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!

Literature review: General model-agnostic framework

- **Leave One Covariate Out(LOCO)**:

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

- **Permutation Feature Importance(PFI)**:

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

  where the j-th covariate is permuted.

- **Conditional Permutation Importance(CPI)**:

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

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

  ✓ Fast and stable with type-I error(Chamma et al. (2023) NeurIPS)

  × Not an interesting theoretical quantity.

Robust-CPI: a new Total Sobol Index estimate

- **Goal:** $\psi_{\mathrm{TSI}}(j, P_0) = \mathbb{E}\left[\mathscr{L}(y, m_{-j}(X^{-j})) - \mathscr{L}(y, m(X))\right]$.
- We can use the Tower's property:

$$m_{-j}(X^{-j}) = \mathbb{E}\left[y|X^{-j}\right] \quad (5)$$

$$= \mathbb{E}\left[\mathbb{E}\left[y|X\right]|X^{-j}\right] \quad (6)$$

$$= \mathbb{E}\left[m(X)|X^{-j}\right] \quad (7)$$

$$= \mathbb{E}\left[m(\widetilde{X}^{(j)})|X^{-j}\right]. \quad (8)$$

Robust-CPI: a new Total Sobol Index estimate

- **Goal:** $\psi_{\mathrm{TSI}}(j, P_0) = \mathbb{E}\left[\mathscr{L}(y, m_{-j}(X^{-j})) - \mathscr{L}(y, m(X))\right]$.
- $m_{-j}(X^{-j}) = \mathbb{E}\left[m(\widetilde{X}^{(j)}) | X^{-j}\right]$.
- Generate n_{cal} conditionally independent samples/ observation.
- **Robust-CPI**:

$$\widehat{\psi}_{\mathrm{Robust-CPI}}(j, P_0) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathscr{L}\left(y_i, \frac{1}{n_{\mathrm{cal}}} \sum_{l=1}^{n_{\mathrm{cal}}} \widehat{m}(\widetilde{x}_{i,l}^{\prime(j)})\right) - \mathscr{L}\left(y_i, \widehat{m}(x_i)\right). \quad (5)$$

✓ It is **consistent**, **fast**, and **stable**.

✓ It links removal with permutation-based approaches.

Robust-CPI: a new Total Sobol Index estimate

- **Goal:** $\psi_{\mathrm{TSI}}(j,P_0) = \mathbb{E}\left[\mathscr{L}(y, m_{-j}(X^{-j})) - \mathscr{L}(y, m(X))\right]$.
- $m_{-j}(X^{-j}) = \mathbb{E}\left[m(\widetilde{X}^{(j)}) | X^{-j}\right]$.
- Generate n_{cal} conditionally independent samples/ observation.
- **Robust-CPI**:

$$\widehat{\psi}_{\mathrm{Robust-CPI}}(j,P_0) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathscr{L}\left(y_i, \frac{1}{n_{\mathrm{cal}}} \sum_{l=1}^{n_{\mathrm{cal}}} \widehat{m}(\widetilde{x}_{i,l}^{\prime(j)})\right) - \mathscr{L}\left(y_i, \widehat{m}(x_i)\right). \tag{5}$$

- ✓ It is **consistent**, **fast**, and **stable**.
- ✓ It links removal with permutation-based approaches.
- If $\mathscr{L} = \ell^2$, fixing n_{cal}, then
  $\widehat{\psi}_{\mathrm{Robust-CPI}}(j,P_0) \xrightarrow{n_{\mathrm{train}}, n_{\mathrm{test}} \to \infty} (1 + 1/n_{\mathrm{cal}})\psi_{\mathrm{TSI}}(j,P_0)$.

Improving variable selection: double robustness

- To detect a null covariate $j \in \mathscr{H}_0$, it is sufficient to either have a good estimate of $\widehat{m}$ or a good conditional sampler:
  - If $\widehat{m} \approx m \in \mathscr{F}_{-j} := \{f, f(u) = f(v) \text{ for } u_{-j} = v_{-j}\}$, then $m(\widetilde{X}'^{(j)}) = m(X)$ and

$$\mathbb{E}\left[\mathscr{L}(y, m(\widetilde{X}'^{(j)}))|\mathscr{D}_{\text{train}}\right] - \mathbb{E}\left[\mathscr{L}(y, m(X))|\mathscr{D}_{\text{train}}\right] \approx 0.$$

  - If $\widetilde{X}'^{(j)} \approx \widetilde{X}^{(j)}$, then $\widetilde{X}^{(j)} \overset{\text{i.i.d.}}{\sim} X$ and $\widetilde{X}^{(j)j} \perp\!\!\!\perp y|X^{-j}$ so

$$\mathbb{E}\left[\mathscr{L}(y, \widehat{m}(\widetilde{X}^{(j)}))|\mathscr{D}_{\text{train}}\right] - \mathbb{E}\left[\mathscr{L}(y, \widehat{m}(X))|\mathscr{D}_{\text{train}}\right] \approx 0.$$

Improving variable selection: double robustness

- To detect a null covariate $j \in \mathscr{H}_0$, it is sufficient to either have a good estimate of $\widehat{m}$ or a good conditional sampler:
  - If $\widehat{m} \approx m \in \mathscr{F}_{-j} := \{f, f(u) = f(v) \text{ for } u_{-j} = v_{-j}\}$, then $m(\widetilde{X}'^{(j)}) = m(X)$ and

$$\mathbb{E}\left[\mathscr{L}(y, m(\widetilde{X}'^{(j)}))|\mathscr{D}_{\text{train}}\right] - \mathbb{E}[\mathscr{L}(y, m(X))|\mathscr{D}_{\text{train}}] \approx 0.$$

  - If $\widetilde{X}'^{(j)} \approx \widetilde{X}^{(j)}$, then $\widetilde{X}^{(j)} \overset{\text{i.i.d.}}{\sim} X$ and $\widetilde{X}^{(j)j} \perp\!\!\!\perp y|X^{-j}$ so

$$\mathbb{E}\left[\mathscr{L}(y, \hat{m}(\widetilde{X}^{(j)}))|\mathscr{D}_{\text{train}}\right] - \mathbb{E}\left[\mathscr{L}(y, \hat{m}(X))|\mathscr{D}_{\text{train}}\right] \approx 0.$$

Proposition 1

Assuming $y = X\beta + \varepsilon$ with X Gaussian, then

$$\mathbb{E}[\widehat{\psi}_{\text{LOCO}}(j, P_0)] = \psi_{\text{TSI}}(j, P_0) + O(1/\boldsymbol{n}_{\text{train}}),$$
$$\mathbb{E}[\widehat{\psi}_{\text{Robust-CPI}}(j, P_0)] = \psi_{\text{TSI}}(j, P_0) + O(1/\boldsymbol{n}_{\text{train}}^2).$$

Double robustness in practice

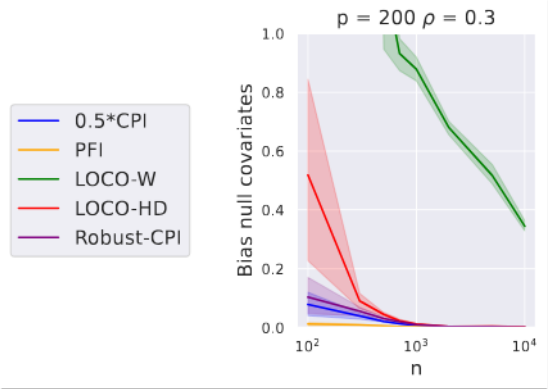


Figure 2: ψ_{TSI} estimation bias in linear setting with random $0.2 * p$ signal and $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = \rho^{|i-j|}$.

Double robustness in practice

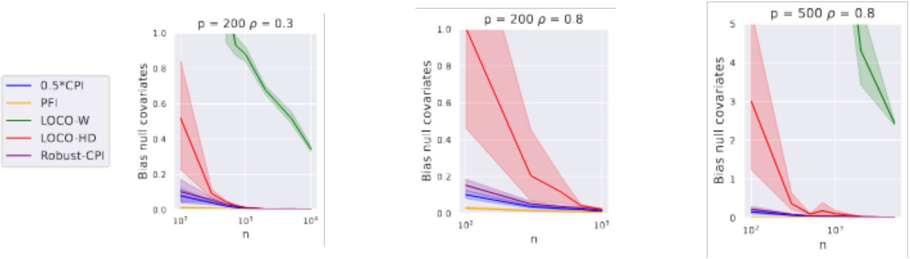


Figure 2: ψ_{TSI} estimation bias in linear setting with random $0.2 * p$ signal and $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = \rho^{|i-j|}$.

Double robustness in practice

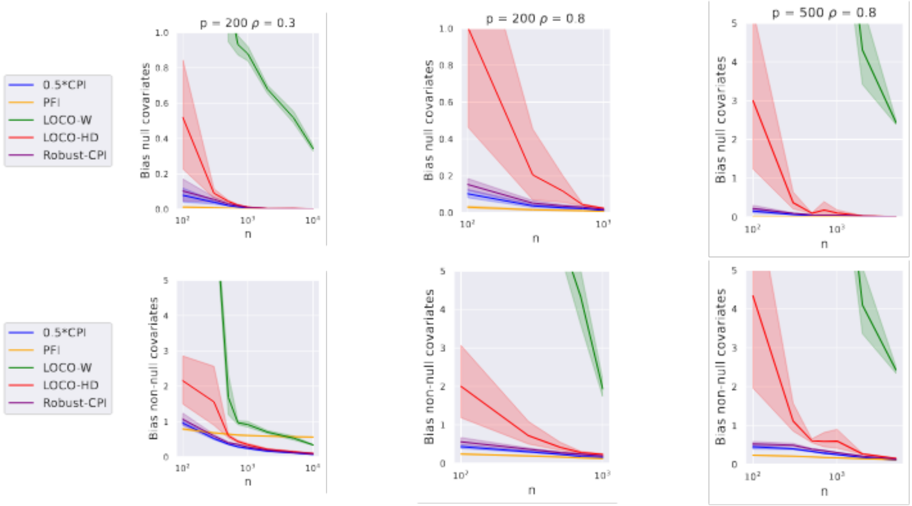


Figure 2: ψ_{TSI} estimation bias in linear setting with random $0.2 * p$ signal and $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = \rho^{|i-j|}$.

Index

Conclusion

- ✓ Robust-CPI provides a **general** and **consistent** VIM.
- ✓ A simple, valid, and fast conditional sampler exists.
- ✓ It is **fast**: it remains a permutation-based approach!
- ✓ If $\mathscr{L} = \ell^2$, it **corrects the CPI bias** to estimate ψ_{TSI}.
- ✓ It leverages CPI's **double robustness**:
  To detect j null, a good $\widehat{m}$ or a good conditional sampler suffices.

Perspectives

Methods	**LOCO**	**PFI**	**CPI**	**Robust-CPI**
Fast	✗	✓	✓	✓
No extrapolation	✓	✗	✓	✓
Interpretable	✓	✗	✗	✓
Double Robustness	✗	✗	✓	✓
Type-I error control	✓	✗	✓	**?**

Thank You, Questions?

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.

H. Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52: 345–370, 1987. doi: 10.1007/BF02294361. URL https://doi.org/10.1007/BF02294361.

References

L. Breiman. Manual on setting up, using, and understanding random forests v3.1. Technical Report 1:58, Statistics Department, University of California, Berkeley, CA, USA, 2002.

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.

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.

References

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.

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/chizat20a.html.

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. A primer on linear classification with missing data, 2024a. URL https://arxiv.org/abs/2405.09196.

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

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.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

References

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.

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.

References

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.

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.

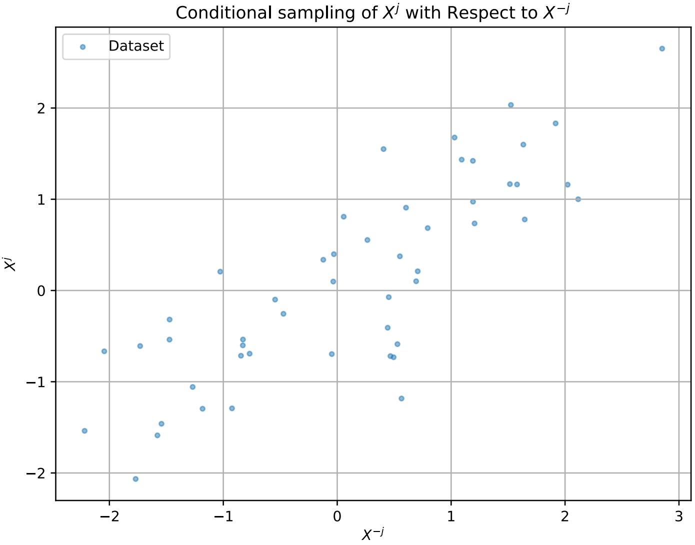
References

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`.

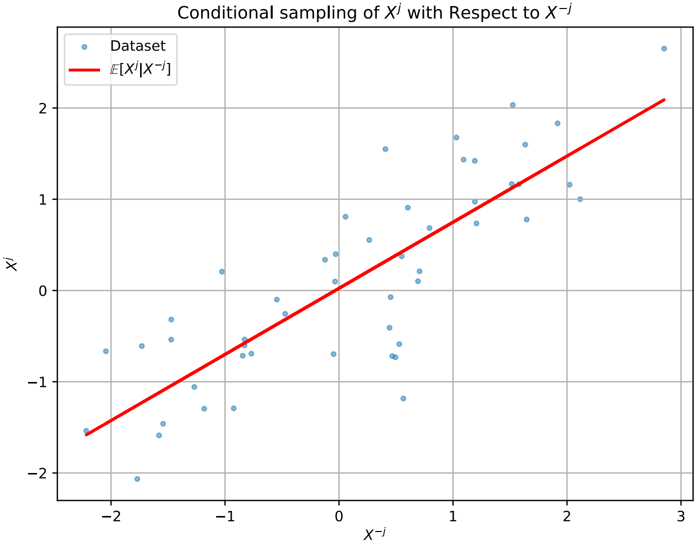
Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22, 2021a. doi: https://doi.org/10.1111/biom.13392. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13392`.

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

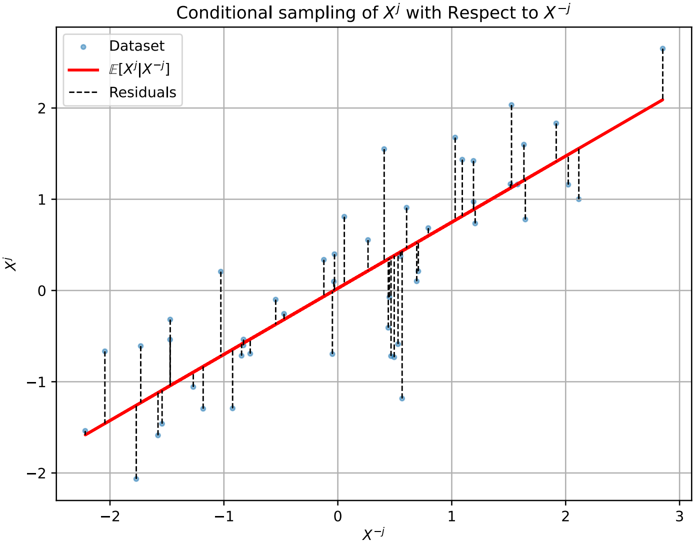
An idea to permute conditionally on X^{-j}



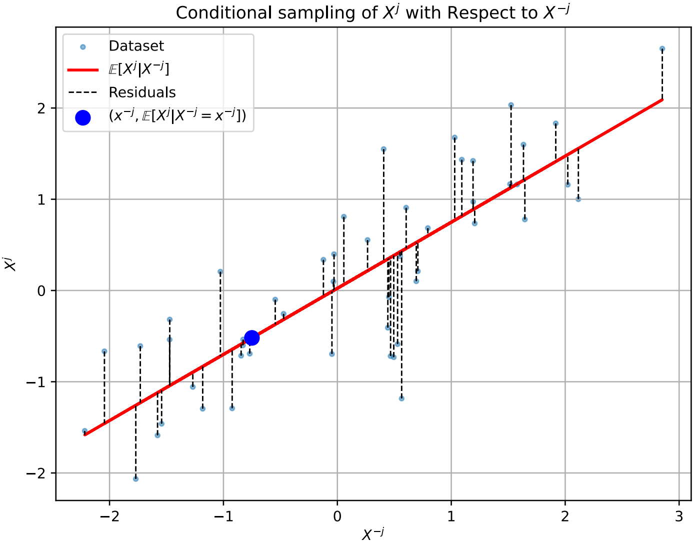
An idea to permute conditionally on X^{-j}



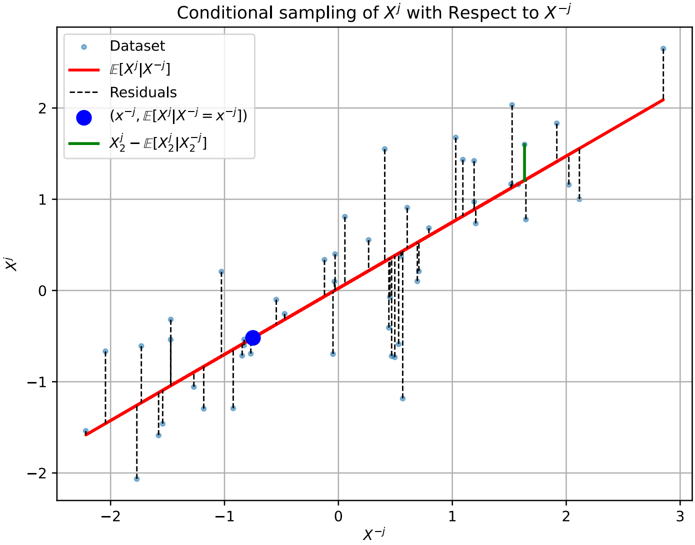
An idea to permute conditionally on X^{-j}



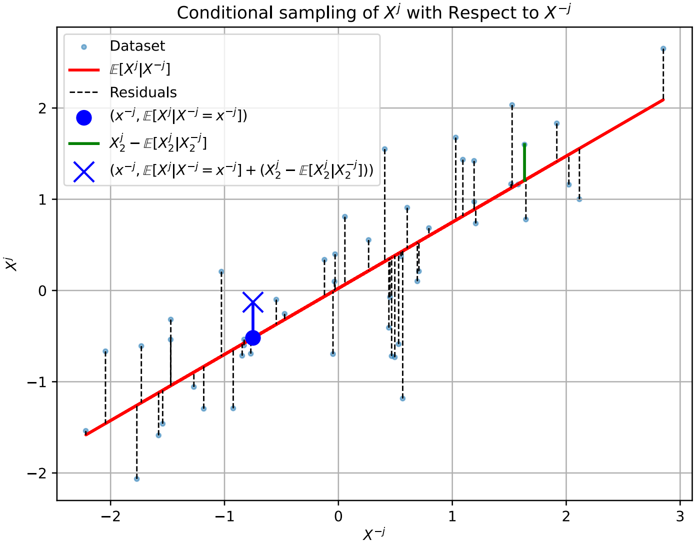
An idea to permute conditionally on X^{-j}



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.

Double robustness using complex models

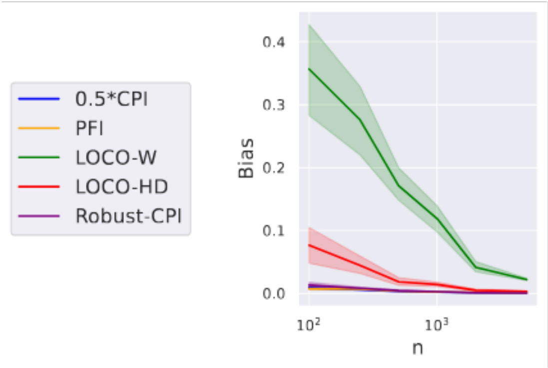


Figure 3: Bias on estimating ψ_{TSI} for null covariates with $y = X_1 X_2 \mathbb{1}_{X_3>0} + 2X_4 X_5 \mathbb{1}_{X_3<0}$, $p = 50$ and $\rho = 0.6$ using super-learner.

Effect of n-cal

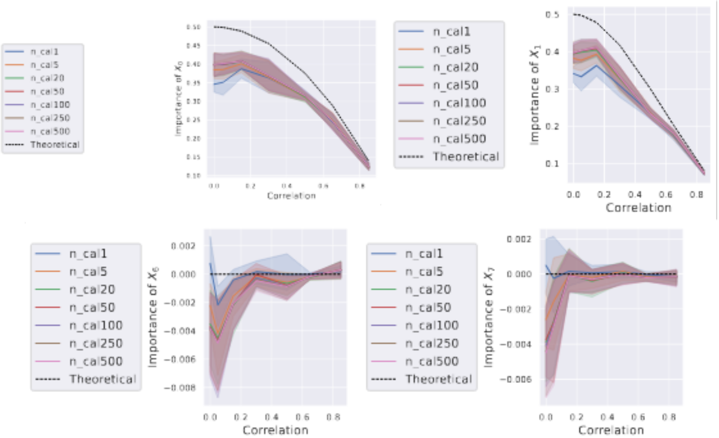


Figure 4: Bias of different Robust-CPI(n_{cal}) on estimating ψ_{TSI} for X_1, X_2 on the top and X_6, X_7 on the bottom, with $y = X_1 X_2 \mathbb{1}_{X_3>0} + 2X_4 X_5 \mathbb{1}_{X_3<0}$, $p = 50$ and $n = 5000$ using super-learner.