Shapley Performance Attribution for Least Squares

Logan Bell Nikhil Devanathan Stephen Boyd

Stanford University

October 27, 2023

Disclaimer

Shapley performance attribution (SPA) is very different from Shapley additive explanations (SHAP).

- SPA attributes the performance of a model across its features.
- SHAP attributes a specific output of a model across its features for model explainability.
- Both methods share several keywords, but they are only superficially related.
- We will only discuss SPA, and we will not discuss SHAP in this talk.

Toy example

- We have a data set $X \in \mathbf{R}^{100 \times 3}$ and labels $y \in \mathbf{R}^{100}$.
- We split the dataset into 50 training and 50 testing points.
- We fit a linear, least-squares model to the data.
  - It attains an out-of-sample R^2 of 0.92.
  - We now know how good our model is, but how useful are our features?
  - Which one is most important?

Model performance

The out-of-sample R^2 is computed as

$$R^2 = \frac{\|y^{\text{tst}}\|_2^2 - \|X^{\text{tst}}\theta^\star - y^{\text{tst}}\|_2^2}{\|y^{\text{tst}}\|_2^2}.$$

where X^{tst} and y^{tst} are our training dataset, and $\theta^\star$ is the vector of feature coefficients that we fit using the training data.

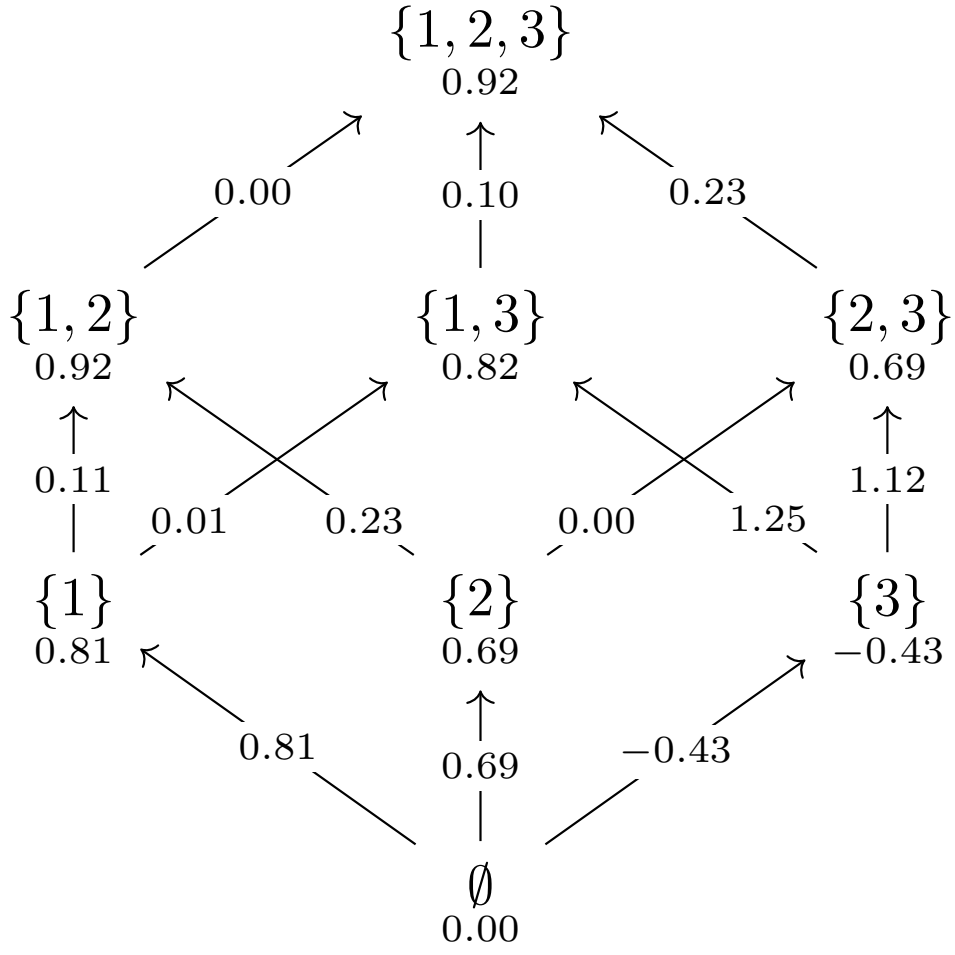
- R^2 is not separable!

Shapley attribution

If we view the R^2 as a reward granted to a coalition of features, then we can use Shapley attribution to determine the importance of our features.

- In 1952, Lloyd Shapley introduced **Shapley attribution**, a method for splitting a reward granted to a coalition of cooperative players.
- The value that Shapley attribution assigns to each player is the player's **Shapley value**.

Shapley attribution lattice representation



Shapley attribution scaling

Suppose we have p features.

- The lattice has 2^p vertices and $p!$ paths from $\emptyset$ to $\{1, \ldots, p\}$.

While the Shapley attribution can be directly computed in our case with $p = 3$, it becomes computationally infeasible after $p \approx 15$.

Least-squares Shapley performance attribution

Least-squares Shapley performance attribution (LS-SPA) is a tool we developed for efficiently and accurately estimating Shapley attributions for least-squares problems.

Using LS-SPA

Install from

`https://github.com/cvxgrp/ls-spa`.

Computing a Shapley attribution is as simple as

```
from ls_spa import ls_spa

attrs = ls_spa(X_train, X_test, y_train, y_test)
```

LS-SPA Performance

- We sample feature chains using a quasi-Monte Carlo method.
  - We get better coverage of the space of feature chains in fewer samples than Monte Carlo.
- We exploit the QR factorization of our observation matrices to reduce the dimension of the least squares problem.
- We also use the QR factorization to efficiently compute the R^2 along a feature chain.

LS-SPA Performance

- With $p = 100$ features and 10^5 training and test observations, we estimate a Shapley attribution in < 3.2 seconds with $< 1\%$ error (2-norm difference to the "true" attribution).

We compute the "true" Shapley attribution by running LS-SPA for 2^{28} sampled permutations.

LS-SPA Performance

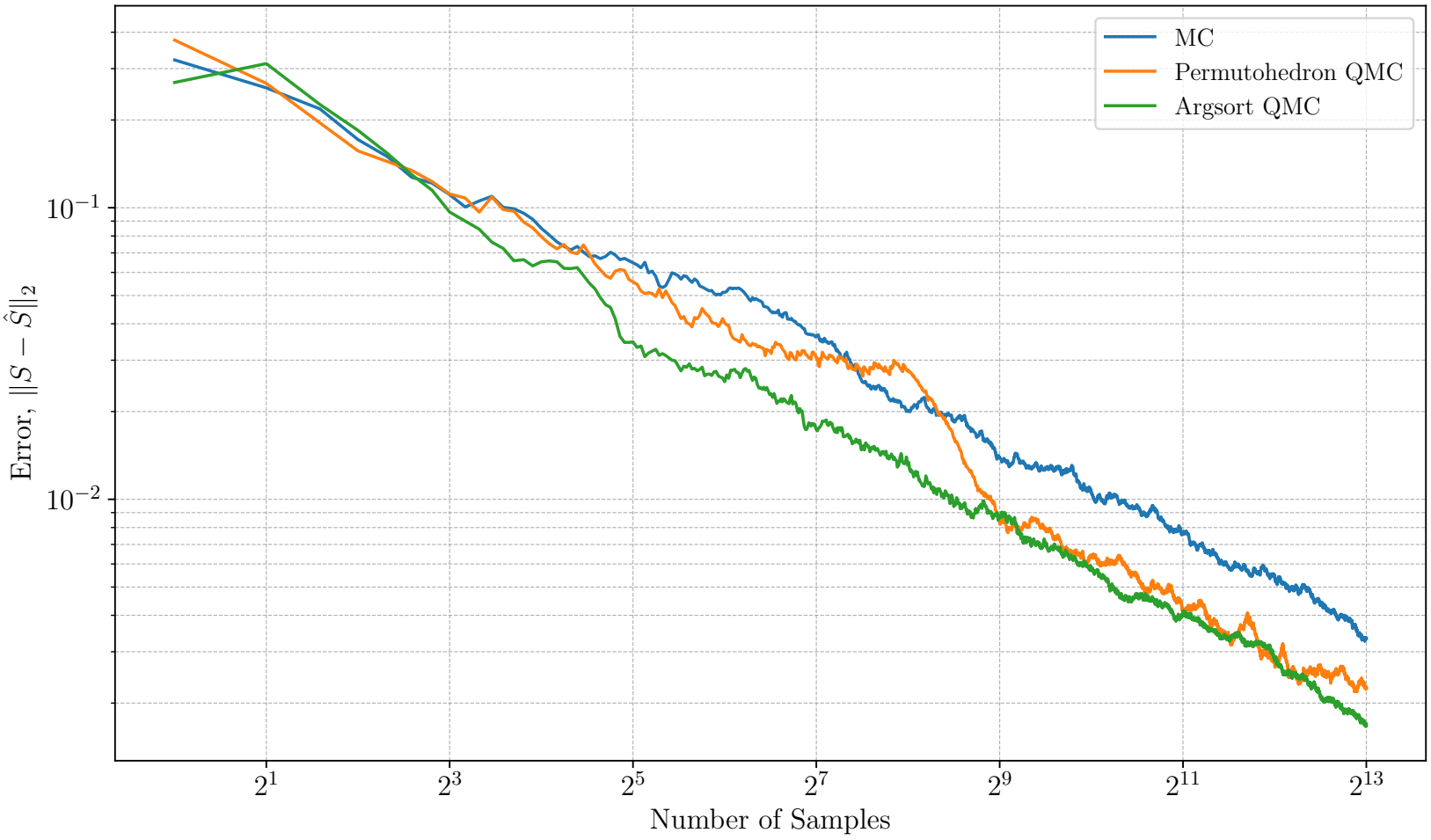


Figure: Number of sampled feature chains vs. error for three different sampling schemes.

Acknowledgements

We thank Ron Kahn for suggesting the topic, Kunal Menda for recommending the use of quasi-Monte Carlo, and Trevor Hastie and Thomas Schmelzer for helpful feedback on an early draft.