G True to the Model or True to the Data?

The "Explanation Shift Detector" proposed in this work relies on the explanation distributions that satisfy efficiency and uninformative theoretical properties. We have used the Shapley values as an explainable AI method that satisfies these properties. A variety of (current) papers discusses the application of Shapley values for feature attribution in machine learning models [76, 77, 78, 79]. However, the correct way to connect a model to a coalitional game, which is the central concept of Shapley values, is a source of controversy, with two main approaches (i) an interventional [47, 45, 43] or (ii) an observational formulation of the conditional expectation[80, 81, 82].

In the following experiment, we compare what are the differences between estimating the Shapley values using one or the other approach. We benchmark this experiment on the four prediction tasks based on the US census data [83] and using the "Explanation Shift Detector", where both the model $f_\theta(X)$ and $g_\psi(\mathcal{S}(f_\theta, X))$ are linear models. We will calculate the Shapley values using the SHAP linear explainer. [4]

The comparison depends on a feature perturbation hyperparameter: whether the approach to compute the SHAP values is either *interventional* or *correlation dependent*. The interventional SHAP values break the dependence structure between features in the model to uncover how the model would behave if the inputs were changed (as it was an intervention). This option is said to stay "true to the model", meaning it will only give allocation credit to the features that the model actually uses.

On the other hand, the full conditional approximation of the SHAP values respects the correlations of the input features. If the model depends on one input that is correlated with another input, then both get some credit for the model's behaviour. This option is said to say "true to the data", meaning that it only considers how the model would behave when respecting the correlations in the input data [84].

In our case, we will measure the difference between the two approaches by looking at the linear coefficients of the model g_ψ and comparing the performance using the geo-political and temporal experiment of the previous section 5, for this case between California 2014 and Puerto Rico 2018.

Table 7: AUC comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for the four prediction tasks based on the US census dataset [83]. The % character represents the relative difference. The performance differences are negligible.

	Interventional	Correlation	%
Income	0.736438	0.736439	1.1e-06
Employment	0.747923	0.747923	4.44e-07
Mobility	0.690734	0.690735	8.2e-07
Travel Time	0.790512	0.790512	3.0e-07

Table 8: Linear regression coefficients comparison of the "Explanation Shift Detector" between estimating the Shapley values between the interventional and the correlation-dependent approaches for one of the US census-based prediction tasks (ACS Income). The % character represents the relative difference. The coefficients show negligible differences between the calculation methods

	Interventional	Correlation	%
Marital	0.348170	0.348190	2.0e-05
Worked Hours	0.103258	-0.103254	3.5e-06
Class of worker	0.579126	0.579119	6.6e-06
Sex	0.003494	0.003497	3.4e-06
Occupation	0.195736	0.195744	8.2e-06
Age	-0.018958	-0.018954	4.2e-06
Education	-0.006840	-0.006840	5.9e-07
Relationship	0.034209	0.034212	2.5e-06

In Table 7 and Table 8, we can see the comparison of the effects of using the aforementioned approaches to learn our proposed method, the "Explanation Shift Detector". Even though the two approaches differ theoretically, the differences become negligible when explaining the protected characteristic, i.e. when providing the linear regression coefficients.

[4] https://shap.readthedocs.io/en/latest/generated/shap.explainers.Linear.html