0.1 Preliminary tests

In order to answer our research question, we first need to determine an effective set up for the group sampling algorithm. For this, we try to answer two questions in this section.

Q_1 Which variant of group sampling is more effective?

In **??**, two variants to create groupings with the help of a SAT-Solver are described. While the creation of the influence model is the same, the actual result is dependent on the groupings created. To determine the superior strategy to create groupings, both variants are tested on our real world examples.

Q_2 How can we determine the group size?

The results of the group sampling algorithm are dependent on the group size used during sampling. With bigger groups, more information for a single feature can be extracted, but it comes at the cost of more actual samples required to create the group. We try to find a group size, which minimizes the samples required while still giving us a decent model. As with Q_1, this is done by testing the group sampling on our real world datasets with different group sizes.

To give an answer to those question, we sample our four real world examples with both variants described in **??** and train an influence model as described in **??**. This is done with group sizes of 2,3,4 and 5 and a sample size of one to 20. The models are evaluated on a test dataset of 100 samples and the MAPE value for each is calculated. The results can be found in Figure 1. Worth mentioning here is, that the sample size reflects the actual number of measurements needed, not the amount of groupings done.

Q_1 Which variant of group sampling is more effective?

When comparing the models created by the first and second variant, "Hamming-Distance" and "Independent Feature", a fundamental flaw in the first variant is visible. If we look at the Apache dataset, for the group sizes of 3,4 and 5, the model stagnates at a MAPE value of 0.75, making it practicably unusable. This is caused by the fact, that the number of features for each group, which the SAT-Solver tries to enable is smaller than the number of required features. If we look at **??** the number of features per group is 8, 6, 4 for the group sizes of 3, 4, 5 respectively on the Apache dataset. With 10 required features on the Apache dataset, the SAT-Solver can not create a group with less than the required amount of features. This results in groups, where only the required features are enabled and thus completely leaving out all other features in the dataset.

Even if the inadequate group sizes are ignored, the grouping variant based on maximizing the Hamming distance is producing worse results than the grouping of independent features.

Q_2 How can we determine the group size?

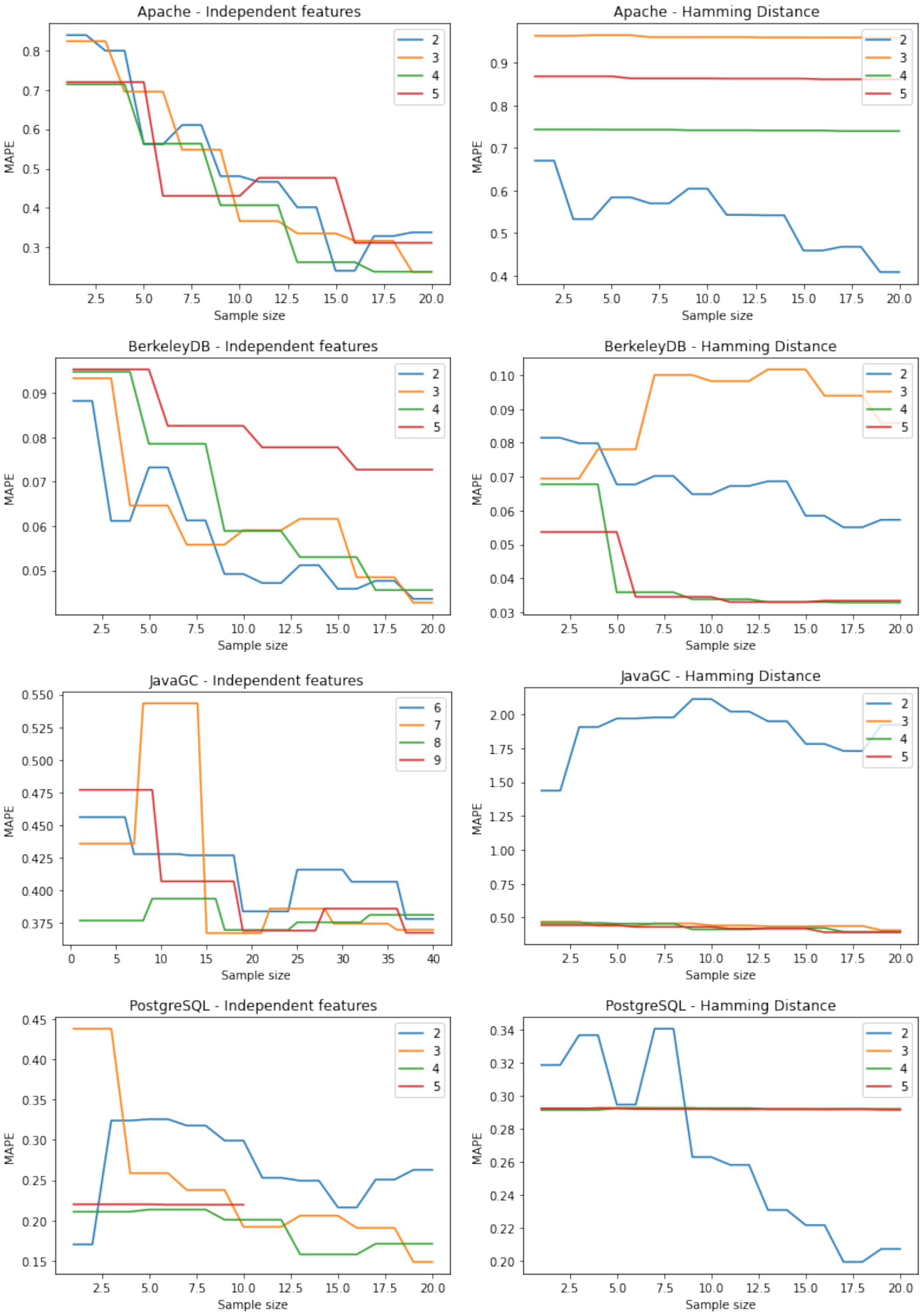

Figure 1: Different group sizes on the real world datasets.