National Institute for Theory and Mathematics in Biology
Funded by the U.S. National Science Foundation and the Simons Foundation

Stabilizing Model Selection

Melissa Adrian, Jake Soloff, Rebecca Willett

Model selection is the process of choosing a model (or subset of models) from a class of candidate models using a collection of training data. This problem is common to many application domains, including ecological dynamics, bioinformatics, dynamical systems, environmental studies, psychology, and more. A key example in biological settings is estimating dynamical system equations describing interactions of species in an ecosystem from data. Common methods for model selection are highly unstable, meaning if a single training sample is removed from the training set, a different model may be selected.

We present a new approach to stabilizing an arbitrary model selection procedure that uses a combination of bagging and an "inflated" argmax operation. Bagging aggregates estimates computed using different subsets of the training data, and the inflated argmax selects models that are the best or near-best performers in a principled, data-adaptive way. **Our method selects a small collection of models that all fit the data and is guaranteed to be stable -- the removal of any training point will result in a collection of selected models that overlaps with the original collection.** *Example*: we aim to select a symbolic set of equations to identify how competition in an ecosystem influences species' abundances, as illustrated with data from a Lotka-Volterra system in Figure 1. The inflated argmax improves model stability with high accuracies and small numbers of returned models relative to previous approaches, as illustrated in Figure 2.

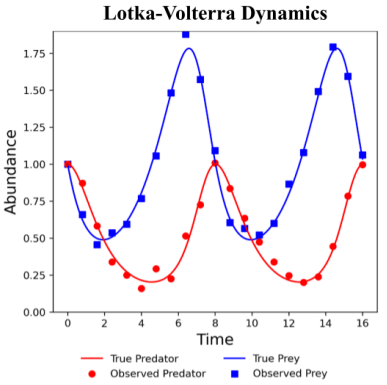

Fig 1. Example training data corresponding to Lotka-Volterra predator-prey trajectories. We seek stable algorithms for estimating dynamical system equations from data.

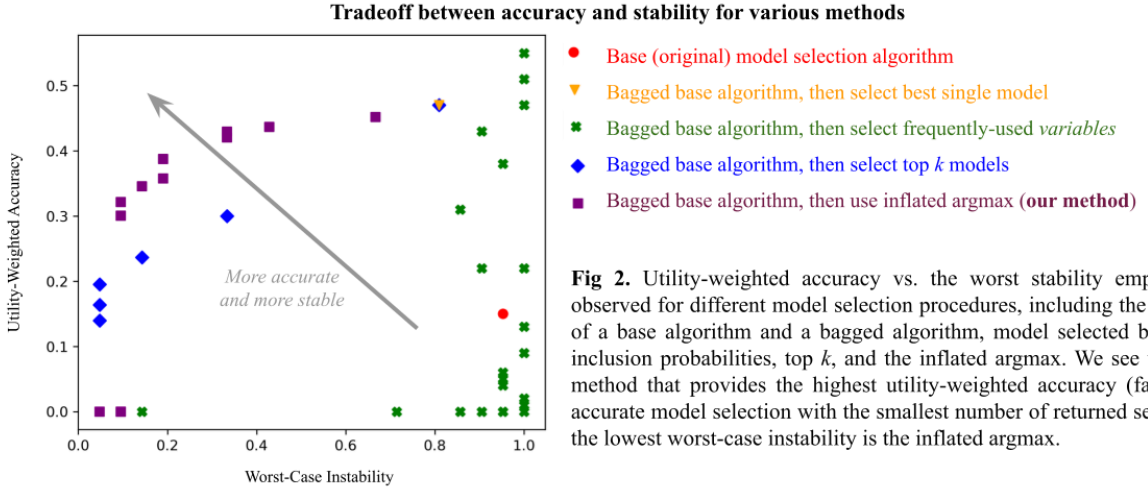

Fig 2. Utility-weighted accuracy vs. the worst stability empirically observed for different model selection procedures, including the argmax of a base algorithm and a bagged algorithm, model selected based on inclusion probabilities, top k, and the inflated argmax. We see that the method that provides the highest utility-weighted accuracy (favors an accurate model selection with the smallest number of returned sets) with the lowest worst-case instability is the inflated argmax.

Reference: https://arxiv.org/abs/2410.18268. This research was supported in part by grants from the NSF (DMS-2235451) and Simons Foundation (MP-TMPS-00005320).