Robust model selection using likelihood as data

Jeffrey Miller

Model selection is a central task in statistics, but standard methods are not robust in misspecified settings where the true data-generating process (DGP) is not in the set of candidate models. The key limitation is that existing methods – including information criteria and Bayesian posteriors – do not quantify uncertainty about how well each candidate model approximates the true DGP. In this paper, we introduce a novel approach to model selection based on modeling the likelihood values themselves. Specifically, given K candidate models and n observations, we view the n x K matrix of negative log-likelihood values as a random data matrix and observe that the expectation of each row is equal to the vector of Kullback-Leibler divergences between the K models and the true DGP, up to an additive constant. We use a multivariate normal model to estimate and quantify uncertainty in this expectation, providing calibrated inferences for robust model selection under misspecification. The procedure is easy to compute, interpretable, and comes with theoretical guarantees, including consistency.