LARGE-SCALE MODEL SELECTION WITH MISSPECIFICATION

EMRE DEMIRKAYA
University of Southern California, Los Angeles, USA

YANG FENG
Columbia University, New York, USA
e-mail: yang.feng@columbia.edu

PALLAVI BASU
Tel Aviv University, Tel Aviv, Israel

JINCHI LV
University of Southern California, Los Angeles, USA

Model selection is crucial to high-dimensional learning and inference for contemporary big data applications in pinpointing the best set of covariates among a sequence of candidate interpretable models. Most existing work assumes implicitly that the models are correctly specified or have fixed dimensionality. Yet both features of model misspecification and high dimensionality are prevalent in practice. In this paper, we exploit the framework of model selection principles in misspecified models originated in Lv and Liu (2014) [1] and investigate the asymptotic expansion of Bayesian principle of model selection in the setting of high-dimensional misspecified models. With a natural choice of prior probabilities that encourages interpretability and incorporates Kullback–Leibler divergence, we suggest the high-dimensional generalized Bayesian information criterion (HGBIC) for large-scale model selection with misspecification. Our new information criterion characterizes the impacts of both model misspecification and high dimensionality on model selection. We further establish the consistency of covariance contrast matrix estimation and the model selection consistency of $HGBIC_p$ in ultra-high dimensions under some mild regularity conditions. The advantages of our new method are supported by numerical studies.

References