Statistica Sinica: Preprint
doi:10.5705/ss.2011.074

VARIABLE SELECTION AND ESTIMATION WITH THE SEAMLESS-L_0 PENALTY

Lee Dicker, Baosheng Huang, and Xihong Lin

Rutgers University, Beijing Institute of Technology, and Harvard School of Public Health

Abstract: Penalized least squares procedures which directly penalize the number of variables in a regression model (L_0 penalized least squares procedures) enjoy nice theoretical properties and are intuitively appealing. On the other hand, L_0 penalized least squares methods also have significant drawbacks. For instance, implementing these procedures is NP-hard and computationally unfeasible when the number of variables is even moderately large. One of the challenges in implementing L_0 penalized least squares procedures is discontinuity of the L_0 penalty. We propose the seamless-L_0 (SELO) penalty, a smooth function on $[0, \infty)$ which very closely resembles the L_0 penalty. The SELO penalized least squares procedure is shown to consistently select the correct model and is asymptotically normal, provided the number of variables grows slower than the number of observations. SELO is efficiently implemented using a coordinate descent algorithm. Tuning parameter selection is crucial to the performance of the SELO procedure. We propose a BIC-like tuning parameter selection method for SELO and show that it consistently identifies the correct model, while allowing the number of variables to diverge. Simulation results show that the SELO procedure with BIC tuning parameter selection performs very well in a variety of settings – outperforming other popular penalized least squares procedures by a substantial margin. Using SELO, we analyze a publicly available HIV drug resistance and mutation dataset and obtain interpretable results.

Key words and phrases: Penalized least squares, oracle property, coordinate descent, tuning parameter selection, BIC.