Phase Diagram for Variable Selection and Non-optimal Regions for L^1 and L^0 Penalization Methods

*Jiashun Jin and Pengsheng Ji**
Carnegie Mellon University and Cornell University
pj54@cornell.edu

Abstract

Consider a linear model $Y = X\beta + z, z \sim N(0, I_n)$, where the rows of X are iid samples from $N(0, \Sigma)$, with Σ being a p by p matrix. It is believed that only a small fraction of the coordinates of β is nonzero, and we are interested in identifying such coordinates.

We adopt an asymptotic framework where both p are n are large. In certain ranges, we find that the above problem reduces to a normal means problem: $\tilde{Y} = \Sigma^{-1}\beta + \tilde{z}, \tilde{z} \sim N(0, \Sigma^{-1})$, which is relatively easier to analyze. We introduce the notion of it phase space, the two-dimensional domain calibrated by the number of nonzero coordinates of β and the magnitude of them. With careful calibrations, we identify three regions in the phase space: it Region of Exact Recovery, it Region of Almost Full Recovery, and it Region of No Recovery. In the first region, exact recovery of all signals (i.e. nonzero coordinates of β) is possible. In the second region, exact recovery is impossible, but it is possible to recover most of the signals. In the last region, it is impossible to identify any significant portion of the signals.

The L^1-penalization methods are well-known approaches to variable selection. Surprisingly, we find that the regions where such methods achieve the optimal rate of convergence are substantially smaller than that of the optimal procedures. The phenomenon persists even when we replace the L^1-penalization by the L^0-penalization (the latter may yield more efficient approaches in signal recovery, but are also computationally more difficult).

We explain why such approaches yield non-optimal rates. We also introduce a new approach whose partition of the phase space coincides with that of the optimal procedures.