UPS Delivers Optimal Phase Diagram in High Dimensional Variable Selection

Pengsheng Ji

Cornell University

Consider a linear model $Y = X\beta + z$, $z \sim N(0, I_n)$. Here, $X = X_{n,p}$, where both p and n are large, but $p > n$. We model the rows of X as i.i.d. samples from $N(0, \frac{1}{n}\Omega)$, where Ω is a $p \times p$ correlation matrix, which is unknown to us but is presumably sparse. The vector β is also unknown but has relatively few nonzero coordinates, and we are interested in identifying these nonzeros.

We propose the **U**nivariate **P**enalization **S**creeing (UPS) for variable selection. This is a screen and clean method where we screen with univariate thresholding, and clean with penalized MLE. It has two important properties: sure screening and separable after screening. These properties enable us to reduce the original regression problem to many small-size regression problems that can be fitted separately. The UPS is effective both in theory and in computation.

We measure the performance of a procedure by the Hamming distance, and use an asymptotic framework where $p \to \infty$ and other quantities (e.g., n, sparsity level and strength of signals) are linked to p by fixed parameters. We find that in many cases, the UPS achieves the optimal rate of convergence. Also, for many different Ω, there is a common three-phase diagram in the two-dimensional phase space quantifying the signal sparsity and signal strength. In the first phase, it is possible to recover all signals. In the second phase, it is possible to recover most of the signals, but not all of them. In the third phase, successful variable selection is impossible. UPS partitions the phase space in the same way that the optimal procedures do, and recovers most of the signals as long as successful variable selection is possible.

The lasso and the subset selection are well-known approaches to variable selection. However, somewhat surprisingly, there are regions in the phase space where neither of them is rate optimal, even in very simple settings, such as Ω is tridiagonal, and when the tuning parameter is ideally set.