Nonregular and Minimax Estimation of Individualized Thresholds in High dimension with Binary Responses

Yang Ning

Cornell University
E-mail: yn265@cornell.edu

Abstract: Given a large number of covariates Z, we consider the estimation of a high-dimensional parameter θ in an individualized linear threshold θTZ for a continuous variable X, which minimizes the disagreement between sign(X−θTZ) and a binary response Y. While the problem can be formulated into the M-estimation framework, minimizing the corresponding empirical risk function is computationally intractable due to discontinuity of the sign function. Moreover, estimating θ even in the fixed-dimensional setting is known as a nonregular problem leading to nonstandard asymptotic theory. To tackle the computational and theoretical challenges in the estimation of the high-dimensional parameter θ, we propose an empirical risk minimization approach based on a regularized smoothed loss function. The statistical and computational trade-off of the algorithm is investigated. Statistically, we show that the finite sample error bound for estimating θ in ℓ2 norm is (slogd/n)β/(2β+1), where d is the dimension of θ, s is the sparsity level, n is the sample size and β is the smoothness of the conditional density of X given the response Y and the covariates Z. The convergence rate is nonstandard and slower than that in the classical Lasso problems. Furthermore, we prove that the resulting estimator is minimax rate optimal up to a logarithmic factor. The Lepski's method is developed to achieve the adaption to the unknown sparsity s and smoothness β. Computationally, an efficient path-following algorithm is proposed to compute the solution path. We show that this algorithm achieves geometric rate of convergence for computing the whole path. Finally, we evaluate the finite sample performance of the proposed estimator in simulation studies and a real data analysis.