Mathematics > Statistics Theory

Tight conditions for consistency of variable selection in the context of high dimensionality

Laëtitia Comminges (LIGM), Arnak Dalalyan (LIGM)

(Submitted on 21 Jun 2011 (v1), last revised 29 Jun 2011 (this version, v2))

We address the issue of variable selection in the regression model with very high ambient dimension, i.e., when the number of variables is very large. The main focus is on the situation where the number of relevant variables, called intrinsic dimension and denoted by d^*, is much smaller than the ambient dimension d. Without assuming any parametric form of the underlying regression function, we get tight conditions making it possible to consistently estimate the set of relevant variables. These conditions relate the intrinsic dimension to the ambient dimension and to the sample size. The procedures that are provably consistent under these tight conditions are simple: they are based on comparing the empirical Fourier coefficients with an appropriately chosen threshold value. The asymptotic analysis reveals the presence of two quite different regimes. The first regime is when d^* is fixed. In this case the situation in nonparametric regression is the same as in linear regression, i.e., consistent variable selection is possible if and only if $\log d$ is small compared to the sample size n. The picture is completely different in the second regime, $d^*\to\infty$ as $n\to\infty$, where we prove that consistent variable selection in nonparametric set-up is possible only if $d^*+\log\log d$ is small compared to $\log n$.

Subjects: **Statistics Theory (math.ST)**
Cite as: **arXiv:1106.4293v2 [math.ST]**

Submission history

From: Arnak Dalalyan [view email]
[v1] Tue, 21 Jun 2011 19:07:25 GMT (546kb,D)
[v2] Wed, 29 Jun 2011 10:06:55 GMT (546kb,D)

Which authors of this paper are endorsers?