Accepted Manuscript

An RKHS-based approach to double-penalized regression in high-dimensional partially linear models

Wenquan Cui, Haoyang Cheng, Jiajing Sun

PII: S0047-259X(17)30151-3
DOI: https://doi.org/10.1016/j.jmva.2018.07.013
Reference: YJMVA 4393

To appear in: *Journal of Multivariate Analysis*

Received date : 12 March 2017

Please cite this article as: W. Cui, H. Cheng, J. Sun, An RKHS-based approach to double-penalized regression in high-dimensional partially linear models, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.07.013

An RKHS-based approach to double-penalized regression in high-dimensional partially linear models

Wenquan Cui[a,*], Haoyang Cheng[a], Jiajing Sun[b]

[a]Department of Statistics and Finance, School of Management, University of Science and Technology of China
[b]School of Economics and Management, University of the Chinese Academy of Sciences

Abstract

We study simultaneous variable selection and estimation in high-dimensional partially linear models under the assumption that the nonparametric component is from a reproducing kernel Hilbert space (RKHS) and that the vector of regression coefficients for the parametric component is sparse. A double penalty is used to deal with the problem. The estimate of the nonparametric component is subject to a roughness penalty based on the squared semi-norm on the RKHS, and a penalty with oracle properties is used to achieve sparsity in the parametric component. Under regularity conditions, we establish the consistency and rate of convergence of the parametric estimation together with the consistency of variable selection. The proposed estimators of the non-zero coefficients are also shown to have the asymptotic oracle property. Simulations and empirical studies illustrate the performance of the method.

Keywords: Eigen-analysis, High-dimensional data, Oracle property, Partially linear model, Representer theorem, Reproducing kernel Hilbert space, Sacks–Ylvisaker conditions, SCAD (smoothly clipped absolute deviation) penalty

1. Introduction

High-dimensional data are the object of intense current interest due to the rapid development of information technologies and their applications in scientific experiments. Important areas with an abundance of high-dimensional data include bioinformatics, signal processing, neural imaging and communications networks. However, high-dimensional data modeling is intrinsically challenging. Not only do we need to consider a large number of covariates, but also the possibility that some of them may contribute in a non-linear fashion. The high dimension calls for an adequate method of identifying the most relevant covariates, while the model needs enough flexibility to account for possible non-linearity. There are also associated issues related to estimation, prediction and inference.

In this paper, we consider the problem of simultaneous variable selection and estimation in the partially linear model (PLM) in a high-dimensional setting. The PLM is defined by

$$Y = \mathbf{X}^\top \beta + g(\mathbf{T}) + \varepsilon,$$

where β is a $p \times 1$ vector of regression coefficients associated with the linear covariates $\mathbf{X}$, g is an unspecified function of $\mathbf{T}$, where $\mathbf{T}$ denotes the non-linear covariate(s), and $g(\mathbf{T})$ is unspecified up to a finite number of parameters. The PLM may be viewed as a a general additive model [20]. It is more flexible than a parametric model, and yet more efficient than a nonparametric model. It also helps to overcome the curse of dimensionality in multivariate smoothing.

*Corresponding author
Email address: wqcui@ustc.edu.cn (Wenquan Cui)