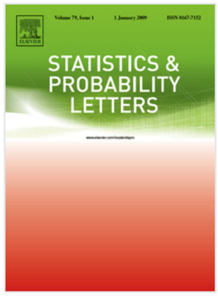
Accepted Manuscript



A robust penalized estimation for identification in semiparametric additive models

Jing Yang, Hu Yang

PII: S0167-7152(15)00345-4
DOI: http://dx.doi.org/10.1016/j.spl.2015.10.002
Reference: STAPRO 7437

To appear in: *Statistics and Probability Letters*

Received date: 10 April 2015
Accepted date: 5 October 2015

Please cite this article as: Yang, J., Yang, H., A robust penalized estimation for identification in semiparametric additive models. *Statistics and Probability Letters* (2015), http://dx.doi.org/10.1016/j.spl.2015.10.002

A robust penalized estimation for identification in semiparametric additive models

Jing Yang[a,*], Hu Yang[a]

[a]*College of Mathematics and Statistics, Chongqing University, Chongqing, 401331, China*

Abstract

Based on the modal regression estimation (Yao et al., 2012) and spline based estimation method, we provide a novel and robust approach to identify nonzero components as well as linear components in semiparametric additive models by applying a two-fold smoothly clipped absolute deviation (SCAD, Fan and Li, 2001) penalty. The performance of the new method is demonstrated in terms of theoretical and simulation results.

Keywords: Semiparametric additive model; Modal regression; B-spline; SCAD penalty; Robustness
MSC: 62G05, 62G08

1. Introduction

Consider the additive regression model

$$Y = u + \sum_{j=1}^{p} f_{0j}(X_j) + \varepsilon, \tag{1}$$

where ε is the model error with conditional mean $\mathrm{E}(\varepsilon \mid X) = 0$, $X = (X_1, X_2, \ldots, X_p)^T$ is p-dimensional covariate and $\{f_{0j}(\cdot), j = 1, 2, \ldots, p\}$ are unknown smooth functions satisfying $\mathrm{E}\{f_{0j}(X_j)\} = 0$ for the sake of model identifiability. This additive combination of univariate functions-one for each covariate X_j-is less general than joint multivariate nonparametric models but can be more interpretable and easier to fit. More importantly, the so-called "curse of dimensionality" that besets multivariate nonparametric regression is largely circumvented. But Opsomer and Ruppert (1999) noticed that in practice some covariates may have linear or even no effects on the response variable while other covariates enter nonlinearly, and proposed the so-called semiparametric additive model (SPAM) with both parametric and nonparametric components. Statistically, the parametric components typically have a faster convergence than that of nonparametric components, treating linear coefficients as nonlinear degrades estimation efficiency. Therefore, the SPAM could be more parsimonious than general additive model in some cases and have attracted considerable attention, related literature refer to Härdle et al. (2004), Jiang et al. (2007), Deng and Liang (2010), Liu et al. (2011), Wei and Liu (2012), among others. Nevertheless, all these works for SPAM are based on the assumption that the linear and nonlinear part are known in advance, which is usually unreasonable in practice. So it is of great interest to develop some efficient methods to distinguish nonzero components as well as linear components from nonlinear ones. Although this goal could be achieved by performing some hypothesis testing, it might be cumbersome to perform in practice whether there are more than just a few predictors to test. In addition, the theoretical properties of such identifications based on hypothesis testing can be somewhat hard to analyze.

Recently, Huang et al. (2010) presented a new type of usage for SCAD penalty as well as related methods, and successfully applied it to nonparametric additive models for the purpose of selecting zero components and parametric components. Following the similar idea, Zhang et al. (2011) studied the model selection using two penalties to simultaneously identify the zero and linear components in partially linear models. Lian (2012a) provided a way to determine linear components in additive models. Lian (2012b) successfully identified nonzero and linear components of model (1) by applying a two-fold SCAD penalty in the conditional quantile regression. Wang and Song (2013) applied the SCAD penalty to identify the model structure in semiparametric varying coefficient partially linear models. However, all these papers were built on either least square (LS) regression which is very sensitive and low-efficiency to many commonly used non-normal errors, or quantile regression that its efficiency is proportional to the density at the median. Hence, it is highly desirable to develop an efficient and robust method that can simultaneously conduct model identification and estimation.

As demonstrated in Yao et al. (2012), the modal regression estimation has a great efficiency gain across a wide spectrum of non-normal error distributions and almost not lose any efficiency for the normal error compared with least square regression. Similar

*Corresponding author at: College of Mathematics and Statistics, Chongqing University, Chongqing, 401331, China.
Email address: `yang2009jing@163.com` (Jing Yang)