Statistics and Probability Letters 114 (2016) 99–103

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Projection pursuit multi-index (PPMI) models

Michael G. Akritas

Department of Statistics, Penn State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 3 November 2015
Received in revised form 12 March 2016
Accepted 12 March 2016
Available online 29 March 2016

Keywords:
Dimension reduction
First projective direction
Joint projective directions
Multi-index model
Projection pursuit regression

ABSTRACT

The concept of *joint projective directions* and the class of *projection pursuit multi-index (PPMI) models* are introduced. PPMI models are MI models with hierarchically defined directions spanning the central mean subspace, and bridge the gap between PP and MI models.

1. Introduction

For a univariate response Y, and covariate vector $\mathbf{X} \in \mathbb{R}^d$, set $\mu(\mathbf{X}) = E(Y|\mathbf{X})$. Attempts to overcome the "curse of dimensionality" in the nonparametric estimation of $\mu(\mathbf{X})$, led to intensive work on models for high-dimensional data which culminated with the works of Huber (1985) and Stone (1985). Thirty years later, it can be said that among the various dimension reduction models proposed, the multi-index (MI) model is probably the most studied. Special cases of the MI model are the projection pursuit regression (PPR) model, and the single index (SI) model which has received particular attention.

The SI model specifies

$$\mu(\mathbf{X}) = g(\boldsymbol{\beta}^T\mathbf{X}), \tag{1}$$

where the $d \times 1$ vector $\boldsymbol{\beta}$ and function g are unknown. For identifiability, it is assumed either that $\|\boldsymbol{\beta}\| = 1$, with $\beta_1 > 0$, where β_1 denotes the first coordinate of $\boldsymbol{\beta}$, or $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_{-1}^T)^T$, with $\|\boldsymbol{\beta}_{-1}\| < 1$ and $\beta_1 = (1 - \|\boldsymbol{\beta}_{-1}\|^2)^{1/2}$, or $\boldsymbol{\beta} = (1, \boldsymbol{\beta}_{-1}^T)^T$; see (7) for the parametrization adopted here. The term single index model was coined by Stoker (1986), though the model was first introduced by Brillinger (1983) who also introduced the least squares estimator as first estimator for $\boldsymbol{\beta}$, under a "linearity" condition.

The PPR model, proposed by Friedman and Stuetzle (1981), specifies

$$\mu(\mathbf{X}) = \sum_{k=1}^{K} g_k(\boldsymbol{\beta}_k^T\mathbf{X}), \quad \text{with } g_1, \ldots, g_k \text{ unknown.} \tag{2}$$

See Huber (1985) for a comprehensive discussion. The parameters $\boldsymbol{\beta}_k$ are called *projective directions*, and are defined recursively as follows. Assuming the first $k-1$ projective directions, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}$, and corresponding functions $g_1, \ldots, g_{k-1}$,

E-mail address: mga@stat.psu.edu.

http://dx.doi.org/10.1016/j.spl.2016.03.008
0167-7152/

have been determined, $\boldsymbol{\beta}_k$ and g_k are determined by

$$\boldsymbol{\beta}_k = \arg\inf_{\mathbf{b}} E\left[(R - E(R|\mathbf{b}^T\mathbf{X}))^2\right], \tag{3}$$

where $R = Y - \sum_{i=1}^{k-1} g_i(\boldsymbol{\beta}_i^T\mathbf{X})$, and $g_k(t) = E(R|\boldsymbol{\beta}_k^T\mathbf{X} = t)$. Identifiability of the projective directions is similar to that of the SIM. The hierarchy among the projective directions is conceptually appealing, but the forced additivity restricts generality. When $K = 1$, the PPR model reduces to the SIM, so that the parameter $\boldsymbol{\beta}$ in (1) equals the first projective direction.

The multi-index model specifies

$$\mu(\mathbf{X}) = g(\mathbf{B}^T\mathbf{X}), \tag{4}$$

for some unknown function g, and unknown full rank matrix $\mathbf{B}$. The MI model (4) is of order k if $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$, with $\boldsymbol{\beta}_i \in \mathbb{R}^d$. Identifiability of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ is discussed by several authors; cf. Xia (2008).

A conceptual disconnect between the single and multi-index models has to do with the fact that none of the parametrizations involves the first projective direction. In fact, it is not clear from the literature whether or not the first projective direction lies in the central mean subspace, which is defined as the subspace spanned by the columns of $\mathbf{B}$.

In this paper we propose a new class of multi-index models, called PPMI because of the use of projection pursuit ideas in their definition and parameter identification. The most general member of this class is (under conditions) a multi-index model with a different type of parametrization. This new parametrization starts with the first projective direction and the additional directions are chosen hierarchically according to their relative usefulness. From the modeling standpoint, the desirable feature of the new parametrization is that a PPMI model of order k will account for at least as much variability as any MI model of order k. Consequently, it is likely that in a model building context, PPMI modeling will result in a more parsimonious model.

2. The PPMI model

For any given d-dimensional vectors $\mathbf{b}_1, \ldots, \mathbf{b}_r$, $r \leq d$, we define the *corresponding r-dimensional projective approximation* of $\mu(\mathbf{X})$ by

$$g_r(u_1, \ldots, u_r|\mathbf{b}_1, \ldots, \mathbf{b}_r) = E(Y|\mathbf{b}_1^T\mathbf{X} = u_1, \ldots, \mathbf{b}_r^T\mathbf{X} = u_r). \tag{5}$$

Short hand notations for $g_r(u_1, \ldots, u_r|\mathbf{b}_1, \ldots, \mathbf{b}_r)$, such as $g_r(\mathbf{u}|\mathbf{B})$, where $\mathbf{u} = (u_1, \ldots, u_r)^T$ and $\mathbf{B}$ is the $d \times r$ matrix $(\mathbf{b}_1, \ldots, \mathbf{b}_r)$, will also be used.

Define the first projective direction, $\boldsymbol{\vartheta}_1$, by

$$\boldsymbol{\vartheta}_1 = \arg\inf_{\mathbf{b}} E\left[(Y - g_1(\mathbf{b}^T\mathbf{X}|\mathbf{b}))^2\right], \tag{6}$$

where, for identifiability purposes, $\boldsymbol{\vartheta}_1$ is taken to be of the form

$$\boldsymbol{\vartheta}_1 = (1, \boldsymbol{\theta}_1^T)^T, \quad \boldsymbol{\theta}_1 \in \mathbb{R}^{d-1}. \tag{7}$$

Thus, the minimization in (6) is specialized to $\mathbf{b}$ of the form $\mathbf{b}_{\mathbf{t}} = (1, \mathbf{t}^T)^T$, $\mathbf{t} \in \mathbb{R}^{d-1}$.

For $r = 2, \ldots, d$, define recursively the *rth joint projective direction*, $\boldsymbol{\vartheta}_r$, by

$$\boldsymbol{\vartheta}_r = \arg\inf_{\mathbf{b}} E\left[(Y - g_r(\boldsymbol{\Theta}_{r-1}^T\mathbf{X}, \mathbf{b}^T\mathbf{X}|\boldsymbol{\Theta}_{r-1}, \mathbf{b}))^2\right], \tag{8}$$

subject to the parametrization specified below, where $\boldsymbol{\Theta}_{r-1} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_{r-1})$. For example, the 2nd joint projective direction is defined as

$$\boldsymbol{\vartheta}_2 = \arg\inf_{\mathbf{b}} E\left[(Y - g_2(\boldsymbol{\vartheta}_1^T\mathbf{X}, \mathbf{b}^T\mathbf{X}|\boldsymbol{\vartheta}_1, \mathbf{b}))^2\right], \tag{9}$$

i.e., as the direction which, when taken jointly with $\boldsymbol{\vartheta}_1$, accounts for the most variability in Y. If

$$E\left[(Y - g_r(\boldsymbol{\Theta}_{r-1}^T\mathbf{X}, \mathbf{b}^T\mathbf{X}|\boldsymbol{\Theta}_{r-1}, \mathbf{b}))^2\right] = E\left[(Y - g_{r-1}(\boldsymbol{\Theta}_{r-1}^T\mathbf{X}|\boldsymbol{\Theta}_{r-1}))^2\right]$$

holds for all $\mathbf{b}$ then the rth joint projective direction is undefined and we can arbitrarily set $\boldsymbol{\vartheta}_r = \mathbf{0}$.

Note that, when non-zero, $\boldsymbol{\vartheta}_r$ is well defined up to the space generated by it and the columns of $\boldsymbol{\Theta}_{r-1}$. That is, if $\mathcal{M}(\mathbf{B})$ denotes the subspace spanned by the columns of $\mathbf{B}$, any $\boldsymbol{\vartheta}_r^*$ such that

$$\mathcal{M}(\boldsymbol{\Theta}_{r-1}, \boldsymbol{\vartheta}_r) = \mathcal{M}(\boldsymbol{\Theta}_{r-1}, \boldsymbol{\vartheta}_r^*)$$

also minimizes $E\left[(Y - g_r(\boldsymbol{\Theta}_{r-1}^T\mathbf{X}, \mathbf{b}^T\mathbf{X}|\boldsymbol{\Theta}_{r-1}, \mathbf{b}))^2\right]$. Thus, the search for the minimum can be confined on $\mathcal{M}(\boldsymbol{\Theta}_{r-1})^c$, the complement of $\mathcal{M}(\boldsymbol{\Theta}_{r-1})$. We adopt a parametrization which facilitates the search while also ensuring identifiability. Consider first $\boldsymbol{\vartheta}_2$. Given the parametrization (7) for $\boldsymbol{\vartheta}_1$, using $\boldsymbol{\vartheta}_2$ of the form $\boldsymbol{\vartheta}_2 = (0, \overline{\boldsymbol{\theta}}_2^T)^T$, with $\overline{\boldsymbol{\theta}}_2 \in \mathbb{R}^{d-1}$, ensures that $\boldsymbol{\vartheta}_2 \in \mathcal{M}(\boldsymbol{\vartheta}_1)^c$ and any other vector in $\mathcal{M}(\boldsymbol{\vartheta}_1)^c$ can be expressed as a linear combination of $\boldsymbol{\vartheta}_1$ and some $\boldsymbol{\vartheta}_2$. Finally, to ensure