# On diagnostics in generalized partially linear single-index models

Danilo V. Silva[1], Gilberto A. Paula[1], Francimário A. de Lima[2]

[1] Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil
[2] Procurement Data Analysis Department, Petróleo Brasileiro S.A., Brazil

E-mail for correspondence: `danilo.silva@ime.usp.br`

**Abstract:** In this paper some diagnostic procedures, such as residual analysis and sensitivity studies, are developed for generalized partially linear single-index models. A P-GAM type iterative process that includes the model matrix from single-index coefficients is derived and an application to real data set is presented.

**Keywords:** additive models; diagnostic procedures; dimension reduction.

## 1 Introduction

Generalized partially linear single-index models (GPLSIMs) have been a great approach to lead with simultaneous nonlinear continuous covariates accommodated into the same linear predictor term modeled by a unique additive function (see, for instance, Yu et al., 2017). Thus, one may gain efficacy with the parametric dimension reduction and the single-index term coefficients may be interpreted similarly to those in linear regression. Studies have been published on single-index models, but few have been developed on diagnostic procedures. This paper presents residual analysis and sensitivity studies in generalized partially linear single-index models with P-spline smoothing. The paper is organized as follows: GPLSIMs are defined in Section 2, a P-GAM type iterative process is derived in Section 3, diagnostic procedures are discussed in Section 4 and an application with a short discussion for ozone concentration data is presented in Section 5.

## 2 The model

The generalized partially linear single-index models are defined as

$$\text{(i) } y_i|(\mathbf{x}_i, \mathbf{z}_i) \overset{\text{ind}}{\sim} \text{EF}(\mu_i, \phi) \quad \text{and} \quad \text{(ii) } g(\mu_i) = \eta_i + f(u_i),$$

where $\text{EF}(\mu, \phi)$ denotes exponential family distribution of mean $\mu$ and dispersion parameter $\phi^{-1}$, $g(\mu)$ and $f(u)$ are, respectively, the link and additive functions,

$\eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is the linear predictor and $u_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}$ is the inside part of the single-index term with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{is})^{\mathrm{T}}$ containing values of covariates, whereas $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s)^{\mathrm{T}}$ are the coefficients of interest. Parameter identifiability requires the constraints $||\boldsymbol{\alpha}|| = 1$ with $\alpha_1 > 0$, and a usual reparameterization is

$$\boldsymbol{\alpha} = (1, \boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}/\sqrt{1 + ||\boldsymbol{\xi}||^2},$$

in which $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{s-1})^{\mathrm{T}}$ does not have constraints. We consider a P-spline smoothing (Eilers and Marx, 1996) such that $f(u) = \sum_{j=1}^{q} \mathrm{N}_j^k(u)\gamma_j$ with $\mathrm{N}_j^k(u)$ denoting the B-spline basis of degree $k$ with $m = q+k+1$ internal knots (equality spaced), namely $a < t_1^0 < \cdots < t_m^0 < b$. Note that in this notation $k = 3$ corresponds to a cubic B-spline and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^{\mathrm{T}}$ are coefficients to be estimated. The penalized log-likelihood function may be expressed as

$$\mathrm{L}_p(\boldsymbol{\theta}, \lambda) = \mathrm{L}(\boldsymbol{\theta}) - \frac{\lambda}{2}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{P}_d\boldsymbol{\gamma}, \tag{1}$$

where $\mathrm{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \phi\{y_i\tau_i - b(\tau_i)\} + c(y_i, \phi)$ is the regular log-likelihood function, with $\boldsymbol{\theta} = \boldsymbol{\theta}_\xi = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{T}}, \phi)^{\mathrm{T}}$, $\tau = \tau(\mu)$ denotes the canonical parameter, $b(\tau)$ and $c(y, \phi)$ are differentiable functions, whereas $\lambda > 0$ is the smoothing parameter and $\mathbf{P}_d = \mathbf{D}_d{}^{\mathrm{T}}\mathbf{D}_d$ with $\mathbf{P}_d$ being the penalty difference matrix of order $d$. The maximum penalized likelihood estimate (MPLE) is obtained by maximizing (1) for $\lambda$ fixed.

## 3 P-GAM type iterative process

Let the model matrices $\mathbf{X}$ and $\mathbf{Z}$ with rows $\mathbf{x}_i^{\mathrm{T}}$ and $\mathbf{z}_i^{\mathrm{T}}$, respectively, and the base matrix $\mathbf{N}$ with rows $\mathbf{N}_i = [\mathrm{N}_1^k(u_i), \ldots, \mathrm{N}_q^k(u_i)]^{\mathrm{T}}$, $i = 1, \ldots, n$. Fixing $\lambda$ and taking $\alpha = \lambda\phi^{-1}$ as the smoothing parameter, the Fisher scoring procedure for obtaining the MPLE leads to the following P-GAM type iterative process:

$$\boldsymbol{\psi}^{(m+1)} = [\{\mathbf{M}^{(m)}\}^{\mathrm{T}}\mathbf{W}^{(m)}\mathbf{M}^{(m)} + \alpha\mathbf{P}]^{-1}\{\mathbf{M}^{(m)}\}^{\mathrm{T}}\mathbf{W}^{(m)}\tilde{\mathbf{y}}^{(m)}, \tag{2}$$

for $m = 0, 1, 2, \ldots$, with $\tilde{\mathbf{y}} = \mathbf{M}\boldsymbol{\psi} + \mathbf{W}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu})$ denoting the dependent modified variable, where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, $\mathbf{W} = \mathrm{diag}\{\omega_1, \ldots, \omega_n\}$ with $\omega_i = \{g'(\mu_i)\}^{-2}\mathrm{V}_i^{-1}$, $\mathbf{V} = \mathrm{diag}\{\mathrm{V}_1, \ldots, \mathrm{V}_n\}$ with $\mathrm{V}_i = d\mu_i/d\tau_i$, $\mathbf{T} = \mathbf{F}^{(1)}\mathbf{Z}\mathbf{J}_{\alpha\xi}$ in which $\mathbf{J}_{\alpha\xi}$ is the Jacobian matrix of $\boldsymbol{\alpha}$ with respect to $\boldsymbol{\xi}$ and $\mathbf{F}^{(1)} = \mathrm{diag}\{f'(u_1), \ldots, f'(u_n)\}$ in which $f'(u)$ is another B-spline of degree $k-1$ with coefficients given from the original curve $f(u)$. Also, $\mathbf{P} = \mathrm{blockdiag}\{\mathbf{0}_p, \mathbf{P}_d, \mathbf{0}_{s-1}\}$ and $\mathbf{M} = (\mathbf{X}, \mathbf{N}, \mathbf{T})$ is a matrix of $p+q+s-1$ columns. Defining the nodes $t_1^{(0)}, \ldots, t_m^{(0)}$, the degree $k$ of the B-spline and keeping $\alpha$ fixed, we propose the following algorithm:

(i) Give starting values $\boldsymbol{\psi}^{(0)}$, obtaining $\hat{\mu}_i = g^{-1}\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(0)} + \mathbf{N}_i^{\mathrm{T}}\boldsymbol{\gamma}^{(0)}\}$, $\hat{t}_{ij} = f'\{\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\alpha}^{(0)}\}\mathbf{z}_i^{\mathrm{T}}(\partial\boldsymbol{\alpha}/\partial\xi_j)|_{\xi^{(0)}}$ and $\hat{\nu}_i = \hat{\mathbf{t}}_i^{\mathrm{T}}\boldsymbol{\xi}^{(0)}$.

(ii) Compute $\mathbf{M} = (\mathbf{X}, \mathbf{N}, \widehat{\mathbf{T}})$, the weights $\omega_i = \{g'(\hat{\mu}_i)^2 V_i(\hat{\mu}_i)\}^{-1}$ and the dependent modified variable $\tilde{y}_i = g(\hat{\mu}_i) + \hat{\nu}_i + g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)$.

(iii) Perform the iterative process (2) getting $\widehat{\boldsymbol{\psi}}$, then update $\hat{\mu}_i$, $\hat{t}_{ij}$, $\hat{\nu}_i$.

(iv) Alternate algorithm steps (ii) and (iii) until the joint convergence.

(v) Solve the equation $\mathrm{U}_p^{\phi}|_{\psi=\hat{\psi}} = 0$ to the precision parameter estimate.

## 4 Diagnostic procedures

For sensitivity analysis we will consider the normal conformal curvature in the unitary direction $\boldsymbol{\ell}$ defined by

$$\mathrm{B}_{\ell}(\boldsymbol{\theta}, \lambda) = \boldsymbol{\ell}^{\mathrm{T}}\mathbf{B}\boldsymbol{\ell}/\sqrt{\mathrm{tr}(\mathbf{B}^2)}, \text{ with } \mathbf{B} = \Delta^{\mathrm{T}}\{-\ddot{\mathrm{L}}_p(\boldsymbol{\theta}, \lambda)\}^{-1}\Delta,$$

$0 \leq \mathrm{B}_{\ell} \leq 1$, is a symmetric non-negative definite matrix with $\Delta$ being the perturbation matrix and $-\ddot{\mathrm{L}}_p$ denoting the penalized observed information matrix (Poon and Poon, 1999). We consider the aggregate measure $\mathrm{B}_i$ corresponding to $\mathrm{B}_{\ell}$ evaluated in the direction of the $i$th observation. Under the usual case-weight perturbation scheme, an efficient approximation is

$$\mathbf{B}_{\xi} = \hat{\phi}\mathrm{diag}\{r_i^p\}\widehat{\mathbf{Q}}_{\xi}\widehat{\mathbf{Q}}_{\xi}^{\mathrm{T}}\mathrm{diag}\{r_i^p\},$$

in which $\mathbf{Q}_{\xi}$ is orthogonal matrix of decomposition $\mathbf{W}^{\frac{1}{2}}\mathbf{T} = \mathbf{Q}_{\xi}\mathbf{R}_{\xi}$ with the same dimension of $\mathbf{T}$ and $r_i^p = \hat{\phi}^{\frac{1}{2}}(y_i - \hat{\mu}_i)V(\hat{\mu}_i)^{-\frac{1}{2}}$ is the Pearson residual. That way the $\mathrm{B}_i$'s may be computed directly with cost $O[n(s-1)^2]$.

## 5 Ozone data set

The `ozone` data set from `faraway` package (Faraway, 2025) on the relationship between atmospheric ozone concentration and meteorological covariates in the Los Angeles Basis in 1976 was considered. We proposed the following GPLSIM: (i) $\texttt{O3}_i|(\mathbf{x}_i, \mathbf{z}_i) \overset{\mathrm{ind}}{\sim} \texttt{PO}(\mu_i)$ and (ii) $\log(\mu_i) = \eta_i + f(u_i)$, with $\eta_i = \beta_1 + \beta_2\texttt{humidity}_i + \beta_3\texttt{temp}_i$ and $u_i = \alpha_1\texttt{dpg}_i + \alpha_2\texttt{wind}_i + \alpha_3\texttt{ibt}_i + \alpha_4\texttt{vis}_i + \alpha_5\texttt{vh}_i + \alpha_6\texttt{ibh}_i$, for all covariates are standardized. We adopted a cubic P-spline smoothing with $q = 9$ of order $d = 2$, which is suitable from residual analyses. Sensitivity analysis under case-weight perturbation scheme for single-index term presented in Figure 1 (left) indicates 8 highlight observations such that $\mathrm{B}_i > \bar{\mathrm{B}} + 3 \times \mathrm{sd}(\mathrm{B})$. Using complete and partial (removing highlighted observations) ozone data, we may notice from Figure 1 (right) that the coefficient interval estimates for single-index term are influenced by these observations, whereas for the parametric part are not. In the derived P-GAM type iterative process, the direction of the maximization always exists, resulting in a stable algorithm even with poor starting values. The cost of computing the conformal curvature has been an obstacle in sensitivity methods of this type and the way presented in this paper to obtain the vector of aggregated measures in the direction of each observation deals with this issue. The highlighted observations derived by the developed sensitivity procedure only impact the form of the derivative of the additive term and are not extremes in other residual graphs. Finally, one has reduced the parametric dimension from 18.64 degrees of freedom, under GAPLM with six additive functions, to 8.04 under GPLSIM.

**Acknowledgments:** This work was partially supported by CNPq, Brazil.

### References

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-spline and penalties. *Statistical Science*, **11**, 89 – 121.

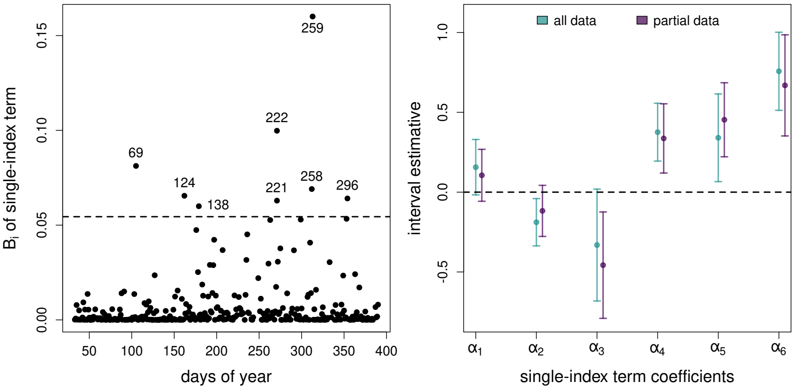

FIGURE 1. Index plot of $\mathrm{B}_i$ under case-weight perturbation scheme for $\widehat{\boldsymbol{\alpha}}$ (left) and 95% interval estimates (right) of the single-index coefficients from the fitted GPLSIM with complete and partial (removing highlighted observations) data.

Faraway, J. (2025). *faraway: Datasets and Functions for Books by Julian Faraway.* R package version 1.0.9.

Poon, W. and Poon, Y.S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **61**, 51 – 61.

Yu, Y., Wu, C. and Zhang, Y. (2017). Penalised spline estimation for generalised partially linear single-index models. *Statistics and Computing*, **27**, 571 – 582.