Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Structural Adaptive Dimension Reduction

Jörg Polzehl [1], Stefan Sperlich [2] [3]

submitted: 15th May 2007

[1] Weierstrass Institute
for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117
Berlin, Germany
E-Mail: polzehl@wias-berlin.de

[2] Georg-August Universität Göttingen
Institut für Statistik und Ökonometrie
Platz der Göttinger Sieben 5
D - 37073 Göttingen

No. 1227
Berlin 2007

[3]This research was financially supported by the "Dirección General de Investigación del Ministerio de Ciencia y Tecnología" SEJ2004-04583/ECON.

2000 *Mathematics Subject Classification.* 62G05 .

Key words and phrases. Dimension-reduction, multi-index model, index space, structural adaptation, R.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

The paper introduces and discusses different estimation methods for multi index models where the indices are parametric and the link function is nonparametric. More specific, the here introduced methods follow the idea of Hristache et al. (2001), modify and try to improve it. Moreover, they constitute alternatives to the so called MAVE-based methods (Xia et al, 2002). We concentrate on an intuitive presentation of what each procedure is doing to the data and its implementation. All methods considered here we have made freely available in R. We conclude with a comparative simulation study based on the provided package EDR.

1 Introduction

Dimensionality continues to be a challenging problem in nonparametric estimation and testing. Many different methods have been proposed to circumvent the so called curse of dimensionality. In nonparametric estimation one could distinguish basically between two different approaches. One is the data explorative method searching for a structural adaptation. The alternative is exploring structural restrictions motivated from model theory. The second one refers to additional knowledge available in the specific context, e.g. economics, medicine, biology, physics, etc. which might impose separability conditions like additivity or similar knowledge of structure.

In this article we concentrate on the first approach. We suppose to have data (Y_i, X_i), $i = 1, \dots, n$, which are generated by a model of the form

$$Y_i = f(X_i) + \varepsilon_i = g(\theta_1^T X_i, \theta_2^T X_i, \cdots, \theta_M^T X_i) + \varepsilon_i = g(\Theta X_i) + \varepsilon_i \ , \tag{1}$$

where Y_i are scalar response variables, X_i are d-dimensional explanatory variables, ε_i are random errors and $f(\cdot)$, respectively $g(\cdot)$, are unknown functions $f : I\!R^d \to I\!R$, $g : I\!R^M \to I\!R$ with $M \le d$.

In other words, Θ is a linear (orthogonal) mapping from the high-dimensional space $I\!R^d$ onto $I\!R^M$. For identification we impose that $\Theta\,\Theta^\top = I_M$, where I_M is the $M \times M$ identity matrix. Note that in our estimation procedures this restriction is neither necessary nor wanted. Moreover, the length of vector θ_j characterizes the variability of the function f from (1) in that direction. Therefore, g is homogeneous, i.e. has the same smoothness in all M directions, what simplifies the choice of smoothing parameters.

If $M = d$ then we are back in the fully nonparametric case. In practice, however, model (1) explains most of the variation of Y for rather small M (actually, for $M = 1, 2$ sometimes 3). Relations as in (1) are referred to as multi-index regression models. All the information about $f(x)$ is concentrated in a (low-dimensional) projection Θx. The aim is to reach dimension reduction for the regression problem, and to describe the index space $\mathcal{I} = \operatorname{Im}\Theta^\top$ which is also referred to as the effective dimension space, see e.g. [8], [9] and [3].

Many different methods have been proposed to address the problem of adaptive dimension reduction. We do not intend to give a comprehensive overview but refer only to some recent contributions and references therein. An interesting new approach has been introduced in [13]. They first span the mean central subspace by the Fourier transform of the density weighted gradient of f. This way they avoid the difficult estimation of f and its derivative(s). A particularly interesting contribution of their work is that afterwards, they succeed to describe the whole central subspace. This is done by the means of the mean central subspaces of all possible transformations of response Y. This paper also contains a good review of existing methods including recent advances like the contour regression procedure of [7], and inverse regression with a minimum discrepancy approach, see [4].

Our methods come closest to the following contributions. [5], [11] and [12] proposed algorithms for estimating the index space for a given effective dimension M, which allows to bypass this *curse of dimensionality* problem using the structural adaptation approach.

All methods discussed here consist of three main steps. The first is to estimate the $d \times d$ matrix Υ of squared averaged derivatives $(\frac{1}{n}\sum_{i=1}^{n} \nabla f(X_i)\ \nabla f^\top(X_i))$ by an iterative procedure. This procedure does not rely on the unknown effective dimension M. In the second step the M-dimensional index space is estimated for a given M. Finally we estimate the link function to obtain a complete description of the model.

All procedures introduced and compared here are made available as a package (EDR) of the R-Statistical System [10].

The rest of the paper is organized as follows. In the next section we explain the basic ideas underlying all estimation procedures considered here. In Section 3.1 we describe first the original [5] estimation method based on these ideas, including a description of our implementation and a discussion of the choice of (initial) parameters. We then propose a modification that leads to improved numerical results. Finally, we present a procedure that involves an additional penalization for directions outside a presumed lower dimensional space. This again improves results in our numerical study. We try to provide fully automatic procedures in the sense that in case of doubts, reasonable defaults for the parameters are available. In Section 4 we compare the numerical performance of the introduced methods.

2 Basic Ideas

2.1 Estimating the Generalized Principle Components

Recall that we consider the model

$$E[Y|X=x] = f(x) = g\left\{\theta_1^T x, \theta_2^T x, \ldots, \theta_M^T x\right\} = g\left\{\Theta x\right\} \ , \tag{2}$$

where $\Theta = (\theta_1, \theta_2, \dots, \theta_M)^T \in I\!R^{M\times d}$. Note that Θ is a linear orthogonal mapping from the high-dimensional space $I\!R^d$ onto the space $I\!R^M$, $M \le d$, satisfying the identification condition that the maximal eigenvalue of $\Theta\Theta^\top$ is equal to one. For the ease of presentation we will assume $X \in [-1,1]^d$. As it is well known that the optimal design for nonparametric regression is the uniform one, an appropriate data transform is recommended.

The model structure (2) implies that each gradient $\nabla f(X_i)$, belongs to the index space $\mathcal{I}$, which in turn is spanned by the vectors $\nabla f(X_i)$. Therefore, a natural basis in $\mathcal{I}$ can be defined via the single value decomposition of the matrix Υ defined as

$$\Upsilon = \frac{1}{n}\sum_{i=1}^{n} \nabla f(X_i)\nabla f^\top(X_i) \ = O_d \Lambda O_d^\top \tag{3}$$

with an orthonormal $d \times d$-matrix O_d and a $d \times d$-diagonal matrix Λ with decreasing eigenvalues. These matrices deliver information about model (2). Let M be the rank of Υ, then the first M columns of O_d provide an orthonormal basis of the space $\mathcal{I}$. The diagonal elements of Λ show how fast the function g varies in each direction. This suggests to first estimate Υ from the data and then recover the space $\mathcal{I}$ using this estimate. Moreover, this provides a natural ordering of the indices.

Matrix Υ is a quadratic functional of the gradient of the regression function f. [6] propose an estimation procedure based on the expansion of the gradient ∇f with respect to an orthonormal basis. Suppose that we are given a collection $\{\psi_\ell\,,\ \ell = 1,\dots,L\}$ of functions $\psi_\ell :\ I\!R^d \to I\!R$ which satisfy

$$\sum_{i=1}^{n} \psi_\ell(X_i)\psi_{\ell'}(X_i) = \delta_{\ell\ell'},$$

where $\delta_{\ell\ell} = 1$ and $\delta_{\ell\ell'} = 0$ for $\ell \neq \ell'$. Now, let β_ℓ^* with

$$\beta_\ell^* = \sum_{i=1}^{n} \nabla f(X_i)\psi_\ell(X_i) \tag{4}$$

be the ℓ-th coefficient of ∇f with respect to the basis system $\{\psi_\ell\}$. Note that each d-vector β_ℓ^* is a linear functional of the gradient and hence belongs to $\mathcal{I}$. Thus if the dimension of the space spanned by $\beta_1^*,\dots,\beta_L^*$ equals M, this set of vectors completely characterizes the index space $\mathcal{I}$, and one can identify the space $\mathcal{I}$ by seeking for the first M principal components of the set $\beta_1,\dots,\beta_L$.

In order to estimate $\mathcal{B}^*(\mathcal{B}^*)^T = \Upsilon_L$ ($\mathcal{B}^*$ being the matrix having β_k^*, $k = 1,\dots,L$ as columns) one can first construct an estimate $\widehat{\beta}_\ell$ of each coefficient β_ℓ^*, e.g.

$$\widehat{\beta}_\ell = \sum_{i=1}^{n} \widehat{\nabla f}(X_i)\psi_\ell(X_i) \tag{5}$$

on the basis of a pilot estimate $\widehat{\nabla f}$ of the gradient, and then compose the estimate

$$\widehat{\Upsilon}_L = \sum_{\ell=1}^{L} \widehat{\beta}_\ell\widehat{\beta}_\ell^\top$$

of Υ_L. It is sufficient for our purposes to choose L such that $\text{rank}(\Upsilon_L) = M$.

It holds $\Upsilon_L \leq \Upsilon$ and since $\Upsilon\Upsilon_L = \Upsilon_L\Upsilon$, the eigenvectors of Υ are at the same time the eigenvectors of Υ_L. Both matrices are nonnegative and the eigenvalues of Υ_L are uniformly smaller or equal to the eigenvalues of Υ. Finally, it is clear that $\Upsilon_L = \Upsilon$ if $L \geq n$. For the case of the multi-index function $f(x) = g(\theta_1^\top x, \ldots, \theta_M^\top x)$, the matrix Υ is of rank M, so that the rank Υ_L is not greater than M. We suppose that the system $\{\psi_\ell\}$ is selected properly and the rank of Υ_L is also M. Then this matrix can be used for describing the structure of the original model in place of Υ. The reason for using the matrix Υ_L instead of Υ is that the problem of estimating the quadratic functional Υ is much harder than the problem of estimating the family of linear functionals β_ℓ defining the matrix Υ_L provided that the basis functions ψ_ℓ are sufficiently smooth. Therefore, as we always estimate Υ via $\widehat{\Upsilon}_L$, we will skip the index L in the following.

As representation (2) is not unique, it is more convenient for our purposes to work with another one. Each vector β_ℓ^* belongs to $\mathcal{I}$ and hence $\text{rank}(\mathcal{B}^*) \leq M$. If $\mathcal{B}^*$ completely describes the index space $\mathcal{I}$, then we have even $\text{rank}(\mathcal{B}^*) = M$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ be the ordered set of eigenvalues of Υ. Since $\text{rank}(\Upsilon) = M$, only the first M of them are positive and the remaining ones are equal to zero. Lemma 2.1. of [5] offers an explicit representation of the model via the orthogonal decomposition of the symmetric $L{\times}L$-matrix $(\mathcal{B}^*)^\top\mathcal{B}^*$. Due to this lemma, the model (1) can always be rewritten in the form

$$f(x) = g\left((\mathcal{B}^* O_M)^\top x\right) \tag{6}$$

which is used in the sequel. We define also

$$\mathcal{R}^* = \mathcal{B}^* O_M \ . \tag{7}$$

2.2 Estimating the Link

The proposed procedures are designed for getting the unknown structure of the model. The so far published theoretical results indicate that the structure (the unknown index space) can be estimated at the best possible rate $n^{-1/2}$ as long as $M \leq 3$, and so our methods do. We will see that all the procedures also deliver an estimator of the regression function f. However, this estimate is only suboptimal in rate. The optimal choice of the bandwidth depends upon the smoothness of g. Rate optimality for the local linear methods is achieved for the bandwidth of order $n^{-1/(4+M)}$ for the M-index case.

A natural way of improving the quality of estimating the regression function f (or the link function g) is to perform one more estimation step. Denote by $\widehat{\Upsilon}_M$ the best M-rank approximation of $\widehat{\Upsilon}$. I.e. if $\widehat{\Upsilon} = O\,\text{diag}\{\mu_1^2, \ldots, \mu_d^2\}O^\top$ with $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_d$, then define also $\mathcal{R}_M = O_M\,\text{diag}\{\mu_1, \ldots, \mu_M\}$ where O_M is the block of the first M columns of O.

We can now infer on target function f by estimating the g as a function of $t = \mathcal{R}_M^\top x$. A local linear estimator of g and its first (partial) derivatives at $t = \mathcal{R}_M^\top x$ is given by

$$\begin{aligned}\begin{pmatrix}\widehat{g}(t)\\ \widehat{g'})(t)\end{pmatrix} &= \operatorname*{arginf}_{a_i,\theta_i} \sum_{j=1}^{n} \left(Y_j - a_i - \theta_i^\top z_j\right)^2 K\left(|z_j|^2/b^2\right) \\ &= \left(\sum_{j=1}^{n} \begin{pmatrix}1\\ z_j\end{pmatrix}\begin{pmatrix}1\\ z_j\end{pmatrix}^\top K\left(|z_j|^2/b^2\right)\right)^{-1} \sum_{j=1}^{n}\begin{pmatrix}1\\ z_j\end{pmatrix} Y_j K\left(|z_j|^2/b^2\right)\end{aligned}$$

with $z_j = t - t_j$. The bandwidth b can be selected by a data-driven selector like cross validation. Estimation of g for $M \leq 2$ can be performed e.g. using the package *sm* for the R Statistical System [10], see [1] or [2].

3 The Estimation Procedures

We consider three methods, the procedure published in [5] [HJPS], including a slight modification of this method [mod-HJPS] that leads to significant improvements in the performance, and a new method that aims to yield further adaptation to the given dimension M [`Penalized method`].

3.1 The HJPS procedure

We consider the following iterative structural adaptation approach. We start with the estimates $\widehat{\nabla f}$ obtained by a fully nonparametric local linear fit and some bandwidth h_1. We then calculate $\widehat{\beta}_\ell = \sum_{i=1}^{n} \widehat{\nabla f}(X_i)\psi_l(X_i)$, $\ell = 1, \dots, L$. Although this estimate is very rough, it contains some information about the structure of the function f and, in particular, about the mapping Θ. All vectors $\widehat{\beta}_\ell$, up to the estimation error, belong to the index space $\mathcal{I}$. This information can be used for producing another, more careful estimate of the gradient function and hence, of the vectors β_ℓ^*. More precisely, let $\widehat{\mathcal{B}}_1$ be the matrix composed from the vectors $\widehat{\beta}_\ell$, $\ell = 1, \dots, L$. We define the gradient estimate $\widehat{\nabla f}^{(2)}(X_i)$ at X_i by a local linear fit

$$\begin{aligned}\begin{pmatrix}\widehat{f}^{(2)}(X_i)\\ \widehat{\nabla f}^{(2)}(X_i)\end{pmatrix} &= \operatorname*{arginf}_{c\in I\!R,\, b\in I\!R^d} \sum_{j=1}^{n}\left[Y_j - c - b^\top(X_j - X_i)\right]^2 K\left(\frac{|S_2 X_{ij}|^2}{h_2^2}\right) \\ &= \left\{\sum_{j=1}^{n}\begin{pmatrix}1\\ X_{ij}\end{pmatrix}\begin{pmatrix}1\\ X_{ij}\end{pmatrix}^\top K\Big(\frac{|S_2X_{ij}|^2}{h_2^2}\Big)\right\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix}1\\ X_{ij}\end{pmatrix} K\Big(\frac{|S_2X_{ij}|^2}{h_2^2}\Big),\end{aligned}$$

with $X_{ij} = X_i - X_j$. Smoothing is performed restricting positive weights to the ellipsoid $\{x : |S_2(x - X_i)| \leq h_2\}$, with $S_2 = (I + \rho_2^{-2}\widehat{\mathcal{B}}_1\widehat{\mathcal{B}}_1^\top)^{-1/2}$ for some $\rho_2 < 1$ and $h_2 > h_1$. In other words, we shrink the original isotropic support of the kernel in all directions $\widehat{\beta}_\ell$ ($\rho_2 < 1$) and stretch them in all orthogonal directions. This leads to estimates

$$\widehat{\beta}_\ell^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\nabla f}^{(2)}(X_i)\psi_\ell(X_i)$$

of β_ℓ^* producing the matrix $\widehat{\mathcal{B}}_2$. We continue this way each time compressing the averaging windows in the direction of the current estimate $\widehat{\mathcal{B}}_k$ and expanding them in orthogonal directions.

The results presented in [5] show that this procedure allows to estimate the index space $\mathcal{I}$ at the rate $n^{-1/2}$ provided that $M < 4$.

The procedure involves input parameters h_1 and $\rho_{\min} < \rho_1$, such that ρ decreases geometrically from ρ_1 to $\rho_{\min}$ by a factor c_ρ and h increases geometrically by a factor c_h during iterations. The choice of these parameters as well as the set of basis functions $\{\psi_\ell\}$ will be discussed later.

To guarantee convergence of the procedure some local regularity of the design is required. Otherwise the gradient estimates could have a very large standard deviation which may deteriorate the quality of the index estimates. This problem can be avoided by weighting each element of the sum in the expression for $\widehat{\beta}_\ell^{(k)}$ with some coefficients that express the degree of local regularity of the design.

We now provide the algorithm in closed form.

- Step 1. Initialization: specify parameters $\rho_1, \rho_{\min}, c_\rho, h_1, c_h,\ C_w$ and the set of functions $\{\psi_\ell\}$; Define $\overline{w}$ as the square root of the minimal eigenvalue of the matrix $\overline{\mathcal{V}}$ with

$$\overline{\mathcal{V}} = \frac{1}{\boldsymbol{E}K(\zeta^\top\zeta)}\boldsymbol{E}\begin{pmatrix}1\\ \zeta\end{pmatrix}\begin{pmatrix}1\\ \zeta\end{pmatrix}^\top K(\zeta^T\zeta)$$

  where ζ is random and uniformly distributed over the ball $B_1 = \{x \in I\!\!R^d : |x| \le 1\}$: $\overline{w}^2 = \lambda_{\min}\left(\overline{\mathcal{V}}\right)$; set $k = 1$, $\widehat{\mathcal{B}}_0 = 0$;

- Step 2. Compute $\widehat{\Upsilon}^{(k)} = \widehat{\mathcal{B}}_{(k-1)}\widehat{\mathcal{B}}_{(k-1)}^\top$. If $\|\widehat{\Upsilon}^{(k)}\| > 1$, then normalize it by its maximal eigenvalue: $\widehat{\Upsilon}^{(k)} := \widehat{\Upsilon}^{(k)}/\|\widehat{\Upsilon}^{(k)}\|_\infty$; Set $S_k = \left(I + \rho_k^{-2}\widehat{\Upsilon}^{(k)}\right)^{1/2}$;

- Step 3. For every $i = 1, \ldots, n$, compute the matrix $\mathcal{V}_k(X_i)$ with

$$\mathcal{V}_k(X_i) = \sum_{j=1}^n \begin{pmatrix}1\\ W_{ij,k}\end{pmatrix}\begin{pmatrix}1\\ W_{ij,k}\end{pmatrix}^\top K\left(W_{ij,k}^\top W_{ij,k}\right)$$

  where $W_{ij,k} = h_k^{-1}S_k(X_j - X_i)$ and define w_i by $w_i^2 = \lambda_{\min}\left(\mathcal{V}_k(X_i)\right)/\lambda_{\max}\left(\mathcal{V}_k(X_i)\right)$;

- Step 4. If the condition

$$\frac{1}{n}\sum_{i=1}^n w_i \ge C_w\overline{w}$$

  is not fulfilled, then increase h_k by the factor c_h, that is, $h_k := c_h h_k$ and decrease ρ_k (if $k > 1$) by c_h. Repeat from Step 3.

- Step 5. For every $i = 1, \ldots, n$, compute $\widehat{\nabla f}^{(k)}(X_i)$:

$$\begin{pmatrix} \widehat{f}^{(k)}(X_i) \\ \widehat{\nabla f}^{(k)}(X_i) \end{pmatrix} = \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} K\Big(\frac{|S_k X_{ij}|^2}{h_k^2}\Big) \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\Big(\frac{|S_k X_{ij}|^2}{h_k^2}\Big);$$

- Step 6. For every $\ell = 1, \ldots, L$, compute the vector $\widehat{\beta}_\ell^{(k)}$

$$\widehat{\beta}_\ell^{(k)} = \left(\sum_{i=1}^{n} w_i \right)^{-1} \sum_{i=1}^{n} \widehat{\nabla f}^{(k)}(X_i)\, \psi_\ell(X_i)\, w_i$$

  with the previously obtained w_i's. Compose the matrix $\widehat{\mathcal{B}}_k$ with columns $\widehat{\beta}_\ell^{(k)}$, $\ell = 1, \ldots, L$.

- Step 7. Set $\rho_{k+1} = c_\rho \rho_k$, and $h_{k+1} = c_h h_k$. If $\rho_{k+1} \geq \rho_{\min}$, then set $k = k+1$ and continue with Step 2.

In the following, we denote by k_n the number of iterations. We set $\widehat{\mathcal{B}} = \widehat{\mathcal{B}}_{k_n}$, and $\rho_{k_n} h_{k_n} = (\rho h)_{k_n}$.

3.1.1 A Modification: mod-HJPS

A correction in Step 3 of the algorithm seems to significantly improve its numerical behavior. Precisely we replace the definition of $\mathcal{V}_k(X_i)$ by

$$\mathcal{V}_k(X_i) = \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} K\left(W_{ij,k}^{\top} W_{ij,k} \right)$$

which resembles the weighting scheme of the local linear estimate used in Step 5.

3.2 An Alternative: the `Penalized` Algorithm

The algorithm HJPS/mod-HJPS used the idea of structural adaptation to create a sequence of increasingly eccentric ellipsoids that allowed to estimate an M-dimensional effective dimension reduction space by the space spanned by the principal axis corresponding to the M largest eigenvalues of the ellipsoid. Eccentricity of the ellipsoids defined by S_k is only driven by inhomogeneities within the data and is usually small if $M \approx d$ or if the structural information is weak.

If we know the dimension M of the EDR and that $M << d$ we can exploit this information even more. To do this we introduce a basis optimization inside the estimation procedure. Recall that HJPS suggests to take a very large set of basis functions $\{\psi_\ell\}$. As mentioned in [5], the ideal choice of this family is given by orthogonalization of the set of partial derivatives $\nabla f_1, \ldots, \nabla f_d$ of the target function f. Since the gradient $\nabla f(x)$ belongs, for all x, to the M-dimensional index space, we would, in case of full knowledge on the EDR, need only M basis functions. We can again use structural adaptation to utilize the available structural information for approaching this ideal choice of basis functions. We define new basis functions as linear combinations of the original ones. Under this restriction, the optimal choice is given by projecting the gradient ∇f onto the subspace in $I\!R^n$ generated by the ψ_ℓ. This projection is described via the singular value decomposition of the $d \times L$ matrix $\nabla f \cdot \Psi = \mathcal{B}$, or, equivalently, by the eigenvalue decomposition of the $L \times L$ matrix $(\mathcal{B})^\top \mathcal{B}$. This matrix is of rank M and it maps the whole space $I\!R^L$ into a M-dimensional subspace denoted as $\widetilde{\mathcal{I}}$. Let us denote this projector on $\widetilde{\mathcal{I}}$ by Π^*. The product $\widetilde{\Psi} = \Psi\Pi^*$ defines a system of basis functions which effectively contains only M nontrivial elements. If $\widehat{\Pi}^*$ is an estimate of the projector Π^*, then the product $\Psi\widehat{\Pi}^*$ is the data-driven counterpart of the "ideal" $\widetilde{\Psi}$. The use of such a basis system is equivalent to multiplying the matrix $\widehat{\mathcal{B}} = \widehat{\nabla f} \cdot \Psi$ by $\widehat{\Pi}^*$, i.e. using the matrix $\widehat{\nabla f} \cdot \Psi \cdot \widehat{\Pi}$.

A penalization using exactly M nontrivial basis element may be to restrictive, especially if the information about the true EDR is weak. We therefore allow to perform the penalization within a space of spaned by m, $(M \le m << d)$ linear combinations.

To conclude, in our procedure we first define $\widehat{\Pi}^*$ in the kth iteration by

$$\Pi_k^2 \quad = \quad \rho_k^2 I_{L\times L} + \widehat{\mathcal{M}}_{k-1} \tag{8}$$

with the projector $\widehat{\mathcal{M}}_k = U_m U_m^T$, where $U_m \in I\!R^{L\times m}$ consists of the first m Eigenvectors of $(\widehat{\mathcal{B}}_k \Pi_k)^T \widehat{\mathcal{B}}_k \Pi_k$. Then, we replace $\widehat{\mathcal{B}}_k \widehat{\mathcal{B}}_k^T$ in HJPS, mod-HJPS by $\widehat{\mathcal{B}}_k \Pi_k \Pi_k^T \widehat{\mathcal{B}}_k^T$. Note also that $\Pi_k = \Pi_k^T$.

Finally, we normalize the projector giving

$$\widehat{\Upsilon}^{(k)} \quad = \quad \frac{\widehat{\mathcal{B}}_k \Pi_k^2 \widehat{\mathcal{B}}_k^T}{\|\widehat{\mathcal{B}}_k \Pi_k^2 \widehat{\mathcal{B}}_k^T\|_\infty}\,. \tag{9}$$

This yields a numerically stable algorithm.

We now present the description of the estimation method. The main part of the estimation procedure is the iterative structural adaptive algorithm. As a result, some estimates of the vectors $\{\beta_\ell\}$ and matrix Υ are obtained. Afterwards, the index space, the link function g and the regression function f are constructed on the base of these estimates.

3.2.1 Estimating the β_ℓ's for given m

As before, for each iteration k we reduce the parameter h_k, and ρ_k geometrically. The initial values $(k = 1)$ of these parameters correspond to the situation with no structural information about the model (see Step 1), the final values correspond to the situation with almost full information and can be selected in a data driven way.

Define $U_h(x)$ as the number of the design points X_i in the ball of the radius h and the center at x. Then the algorithm reads as follows:

- Step 1. Initialization: Initialize the parameters $\rho_1 = 1.0$, $c_\rho = e^{-1/6}$, $c_h = \sqrt{c_\rho}$, the bandwidth h_1 and define the set of basis functions $\{\psi_\ell\}$ as in (HJPS). Set $k = 1$, $\widehat{\mathcal{M}}^{(0)} = I_{L\times L}$ and $\widehat{\Upsilon}^{(0)} = 0_{d\times d}$.

- Step 2. Compute

$$S_k^2 = \rho_k^2 I + \widehat{\Upsilon}_{k-1}, \qquad \Pi_k^2 = \rho_k^2 I + \widehat{\mathcal{M}}_{k-1},$$

- Step 3. For every $i = 1, \ldots, n$, compute the matrix

$$\mathcal{V}_{(k)}(X_i) \;\; = \;\; \sum_{j=1}^{n} \binom{1}{X_{ij}} \binom{1}{X_{ij}}^{\top} K\Big(\frac{|S_k X_{ij}|^2}{h_k^2}\Big). \tag{10}$$

and define w_i by $w_i^2 = \lambda_{\min}\left(\mathcal{V}_k(X_i)\right)/\lambda_{\max}\left(\mathcal{V}_k(X_i)\right)$

- Step 4. If the condition

$$\frac{1}{n}\sum_{i=1}^{n} w_i \geq C_w \overline{w}$$

is not fulfilled, then increase h_k by the factor c_h, that is, $h_k := c_h h_k$ and decrease ρ_k (if $k > 1$) by c_h. Repeat from Step 3.

- Step 5. For every $i = 1, \ldots, n$ compute:

$$\begin{pmatrix} \widehat{f}^{(k)}(X_i) \\ \widehat{\nabla f}^{(k)}(X_i) \end{pmatrix} = \{\mathcal{V}_k(X_i)\}^{-1} \sum_{j=1}^{n} Y_j \binom{1}{X_{ij}} K\Big(\frac{|S_k X_{ij}|^2}{h_k^2}\Big)$$

- Step 6. Compute the vectors $\widehat{\beta}_\ell^{(k)} = (\sum_{i=1}^{n} w_i^{(k)})^{-1} \sum_{i=1}^{n} w_i^{(k)} \widehat{\nabla f}(X_i)\psi_\ell(X_i)$, $\ell = 1, \ldots, L$ with $(w_i^{(k)})^2 := \lambda_{min}(\mathcal{V}_k(X_i))/\lambda_{max}(\mathcal{V}_k(X_i))$, and compose the $d \times L$ matrix $\widehat{\mathcal{B}}^{(k)}$ with columns $\widehat{\beta}_1^{(k)}, \ldots, \widehat{\beta}_L^{(k)}$;

- Step 7. Perform a singular value decomposition of the matrix

$$\widehat{\mathcal{B}}_k \Pi_k = \sum_{i=1}^{d} \mu_i u_i v_i^T$$

with single values μ_i and corresponding vectors u_i, v_i. Then update

$$\widehat{\Upsilon}_k = \frac{\widehat{\mathcal{B}}_k \Pi_k^2 \widehat{\mathcal{B}}_k^T}{\|\widehat{\mathcal{B}}_k \Pi_k^2 \widehat{\mathcal{B}}_k^T\|_\infty} = \mu_1^{-2} \sum_{i=1}^{d} \mu_i^2 u_i u_i^T \quad \text{and} \quad \widehat{\mathcal{M}}_k = \sum_{i=1}^{m} v_i v_i^T$$

- Step 8. If $\rho_k \leq \rho_1 n^{-1/3}$, stop. Else, set $\rho_{k+1} = c_\rho \rho_k$, $h_{k+1} = c_h h_k$, increase k by one, $k := k + 1$, and continue with Step 2;

Again, for the given m, the estimator $\widehat{\mathcal{I}}_M$ of the index space $\mathcal{I}$ is spanned by the first m principal components of the matrix $\widehat{\mathcal{B}}_{k_n} \widehat{\mathcal{B}}_{k_n}^T$.

3.3 Choice of parameters

It is obvious that the quality of estimation by the proposed methods depends on the rule for changing the parameters h and ρ, and, in particular, on their values at the initial and final iteration. The values ρ_k decrease from ρ_1 to $\rho_{\min}$ while the h_k increase during iteration from h_1 to $h_{\max}$. The value h_1 is to be selected in such a way that for the majority of points X_i, the estimate $\widehat{\nabla f}(X_i)$ is well defined. A necessary (and usually sufficient) condition is that every ball $\{x : |x - X_i| \leq h_1\}$ contains at least $d+1$ design points. The estimate of β_l is restricted to use only such points by the definition of w_i in Step 3 of HJPS and step 4 of the penalized algorithm. Step 4 of the algorithms guarantees that a sufficient number of design points with positive weights exists.

The proposed rule leads to $k_n \approx 6\log(\rho_1/\rho_{\min}) \approx 2\log n$ iterations and provides that $h_{k_n} \approx C_0$. Note also that assuming the structure of the matrix $\widehat{\mathcal{B}}_{(k-1)}\widehat{\mathcal{B}}_{(k-1)}^\top$ to follow the structure of the target matrix Υ^*, neighborhood $E_k(X_i)$ is stretched at each iteration step by factor c_h in all directions and is shrunk by factor c_ρ in directions of the M-dimensional index space $\mathcal{I}$. Therefore, the Lebesgue measure of every such neighborhood is changed each time by the factor $e^{\frac{d}{2(4\vee d)} - \frac{m}{6}}$ which is larger or equal to 1 for all $M \leq 3$ and $d > M$. Under the assumption of a random design with a positive density, this would lead to an increase of the mean number of design points inside each $E_k(X_i)$.

Theoretical results, see e.g. [5] suggest that $k_n \geq \ln(n)$. Our simulation studies suggest that $k_n = 2\ln(n)$ is a good choice. This explains why we set $c_\rho = e^{-1/6}$ in step 1 for given $\rho_{min} = \rho_1 n^{-1/3}$. Certainly, a different combination (c_ρ, ρ_{min}) would also be possible. From a theoretical point of view we need $n^{-1/3} \leq \frac{\rho_{min}}{\rho_1} \leq n^{-2/5}$ and thus $e^{-1/6} = 0.84648 \geq c_\rho \geq 0.81873 = e^{-1/5}$. To our experience, taking e.g. $c_\rho = e^{-1/5}$ or $e^{-1/6}$ does not make a significant difference in practice.

For the case with $M \leq 3$, we propose the following rule of thumb

$$\begin{array}{lll} \rho_1 = 1, & \rho_{\min} = n^{-1/3}, & c_\rho = e^{-1/6}, \\ h_1 = C_0 n^{-\frac{1}{4\vee d}}, & C_w = \frac{1}{6\vee\log(n)} & c_h = e^{\frac{1}{2(4\vee d)}}, \end{array} \tag{11}$$

where C_0 is to be defined depending on the design.

Remark: *We designed all procedures such that they can be used as fully automatic (data adaptive) procedures. However, that requires that all components of X have approximately the same scale. Standardization of the explanatory variables may therefore be necessary. Additional tuning may be possible by modifying the initial values ρ_1 and h_1.*

4 Numerical Performance

We now present the results of a small simulation study. We illustrate and compare the properties of the proposed procedures in two situations characterized by a one and two-dimensional EDR, respectively.

Example 4.1 *We consider a single index model* ($M = 1$)

$$Y_i = X_i^T\theta \ \sin(\sqrt{5}X_i^T\theta) + \epsilon_i \ , \qquad i =, 1, \dots, n \tag{12}$$

with $\theta = (1, 2, 0, \dots, 0)/\sqrt{5}$ *and* X_i *uniformely distributed in* $[-1, 1]^d$*. The errors* ϵ_i *are generated from a Gaussian distribution with standard deviation* 0.3*.*

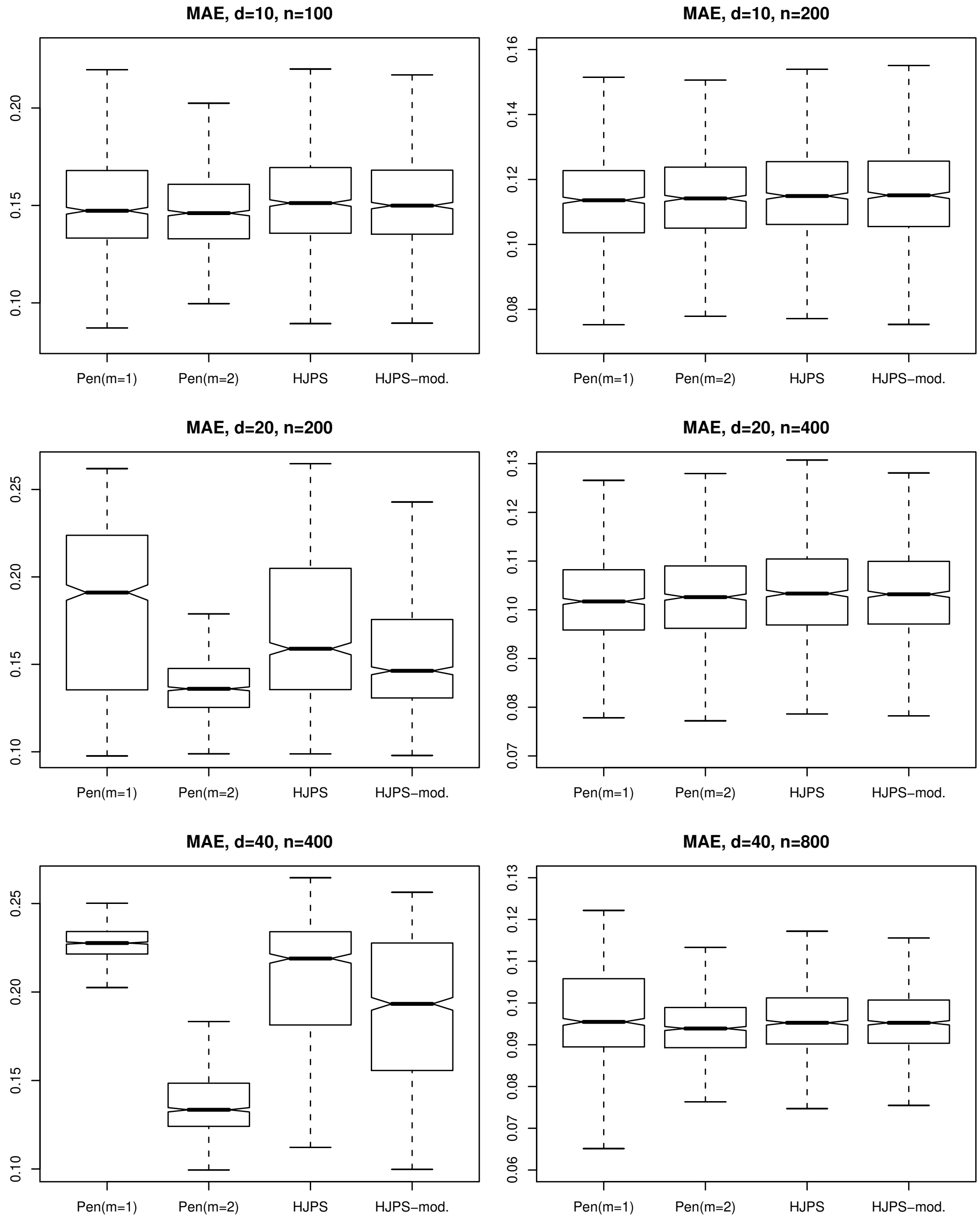

Figure 1: *MAE estimated from 1000 simulations for Example 4.1 with* $M = 1$

Example 4.2 *The second example is a multi-index model with a $M = 2$-dimensional EDR*

$$Y_i = X_i^T\theta_1 \sin(\sqrt{5}X_i^T\theta_2) + X_i^T\theta_2\ \sin(\sqrt(5)X_i^T\theta_1) + \epsilon_i\ , \qquad i = 1, \ldots, n \qquad (13)$$

with $\theta_1 = (1, 2, 0, \ldots, 0)/\sqrt{5}$, $\theta_2 = (-2, 1, 2, 0, \ldots, 0)/3$. Again the X_i are uniformely distributed in $[-1, 1]^d$ and the errors ϵ_i are Gaussian with standard deviation 0.3.

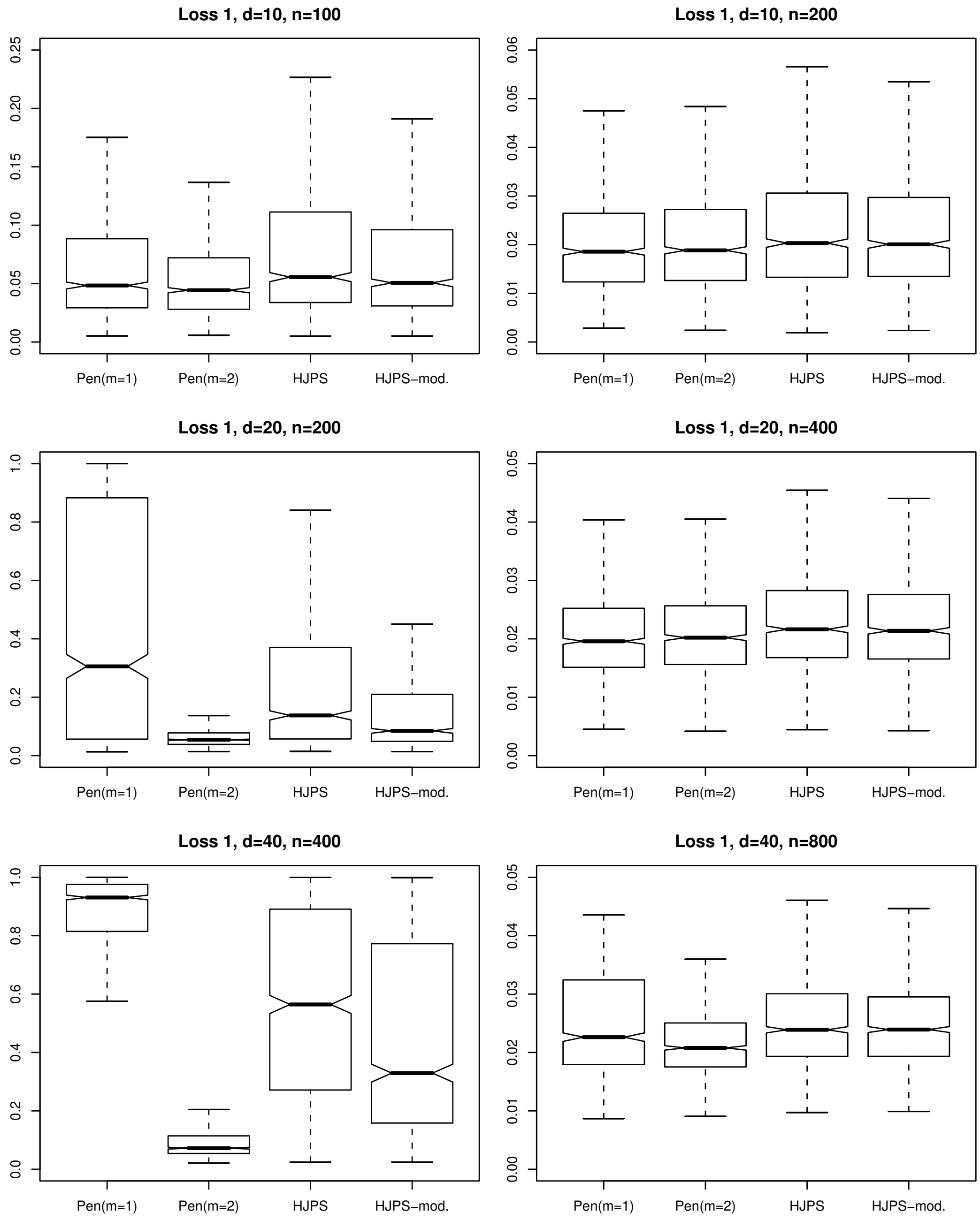

Figure 2: *Loss1 estimated from 1000 simulations for Example 4.1 with $M = 1$*

Simulations of size 1000 were performed specifying different values for dimension d ($d = 10, 20, 40$) and sample size n ($n = 10d$ and $n = 20d$). The initial bandwidth h_1 was specified as $h_1 = .85\sqrt{d}(d/n * \prod_j^d IQR(X_{.j}))^{1/d}$, with $IQR(X_{.j})$ denoting the Inter-Quartile-Range of the jth explanatory variable.

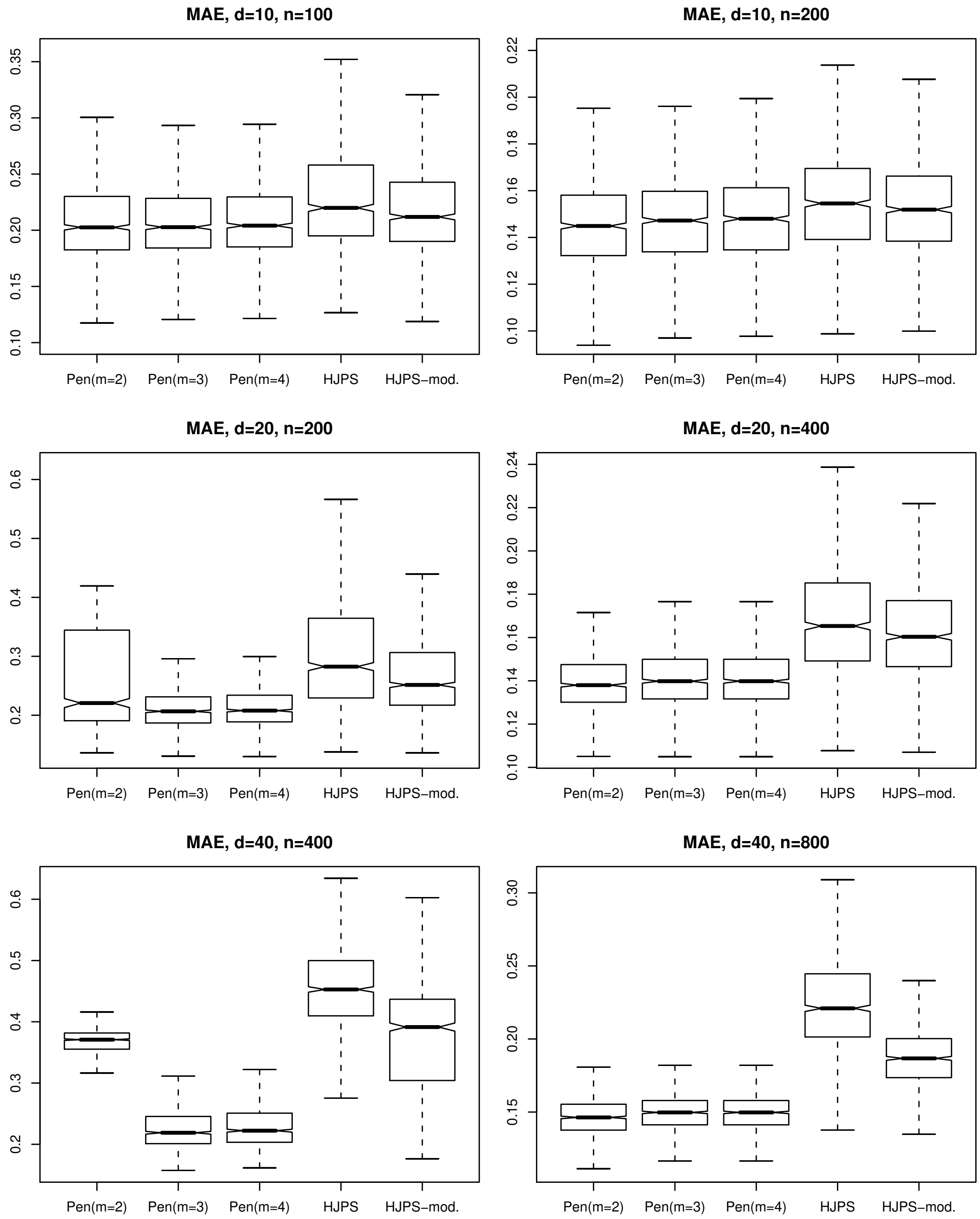

Figure 3: *MAE estimated from 1000 simulations for Example 4.2 with* $M = 2$.

For the `Penalized` algorithm we have, as an additional parameter, to specify the rank m of matrix $\widehat{\mathcal{M}}_k$ that determines the penalization of the basis. Within the simulations we

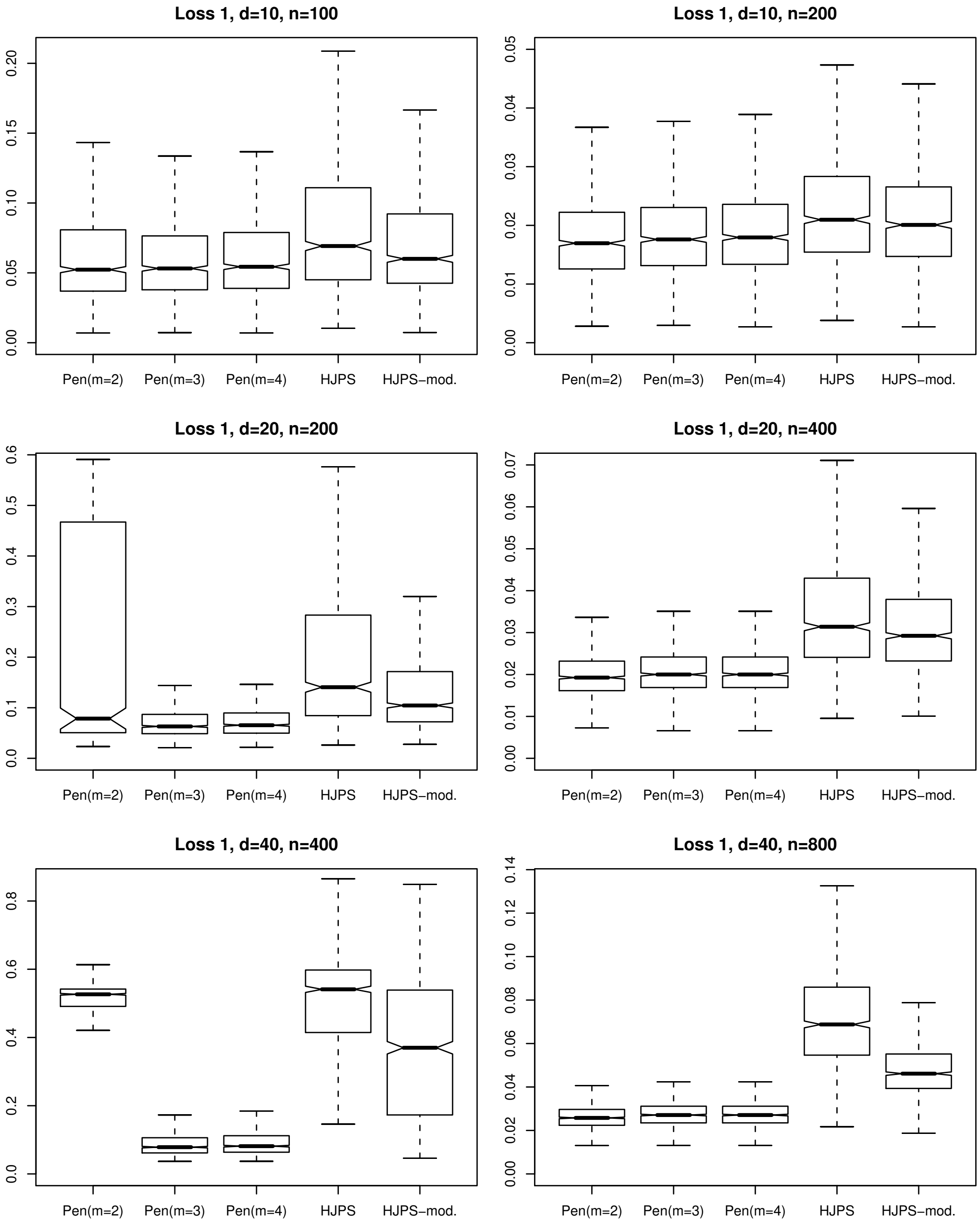

Figure 4: *Loss1 estimated from 1000 simulations for Example 4.2 with* $M = 2$.

compare the HJPS and the modified HJPS algorithms with the penalized procedure using $m = M$, $m = M + 1$ and additionally, in case of Example 4.2, $m = M + 2$. Estimation of the link function is performed using the package *sm* from the R Statistical System [10].

We have calculated two different performance measures for each simulation, the average absolute error MAE

$$\frac{1}{n}\sum_{i=1}^{n} |g(x) - \widehat{g}(x)|$$

and the distance [loss1] between the real projection and the estimated one of x through $\mathcal{R}^*$, respectively $\widehat{\mathcal{R}}_m$

$$\left\|\mathcal{R}^* \left\{I - \mathcal{R}_M^T(\mathcal{R}_M\mathcal{R}_M^T)^{-1}\mathcal{R}_M\right\}\right\|_2 \;/\; \|\mathcal{R}^*\|_2 \qquad \text{with} \quad \|A\|_2 = tr(AA^T), \tag{14}$$

where I is the identity. The results are summarized as box-plots. Figures 1 and 2 provide the MAE and the loss 14 for all considered situations in case of Example 4.1. Figures 3 and 4 contain the corresponding information for Example 4.2.

In general we observe a significant improvement by using the modified version of HJPS in comparison to the original proposal. Further, the penalized algorithm seems to outperform the modified HJPS algorithm if the dimension m is chosen to be slightly larger then the usually unknown true dimension M of the EDR. In situations where d is large or n/d is small choosing $m = M$ leads to over-penalization and loss of information within the adaptation process. In situations where n/d is large the penalized method seems to outperform HJPS for all considered choices of m.

As a conclusion our simulations suggest to use the penalized algorithm with $m = M + 1$ if the true dimension M of the EDR can be guessed and the modified HJPS procedure in other cases.

References

[1] A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: the Kernel Approach With S-Plus Illustrations.* Oxford Univ. Press, 1997.

[2] A. W. Bowman and A. Azzalini. Computational aspects of nonparametric smoothing with illustrations from the sm library. *Comp. Stat. & Data Anal.*, 42:545–560, 2003.

[3] D. Cook. Principal hessian directions revisited. *J. Amer. Statist. Ass.*, 93:84–100, 1998.

[4] D. Cook and L. Ni. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Ass.*, 100:410–428, 2005.

[5] M. Hristache, A Juditski, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Annals of Statistics*, 29(6):1537–1566, 2001.

[6] I. Ibragimov, A. Nemirovskii, and R. Khasminski. Some problems on noparametric estimation in gaussian white noise. *Theory Probab. Appl.*, 3(1):391–406, 1986.

[7] B. Li, H. Zha, and F. Chiaromonte. Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33:1580–1616, 2005.

[8] K.-C. Li. Sliced inverse regression for dimension reduction. (with discussion). *J. Amer. Statist. Ass.*, 86:316–342, 1991.

[9] K.-C. Li. On principal hessian directions for data visualization and dimension reduction: *J. Amer. Statist. Ass.*, 87:1025–1039, 1992.

[10] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[11] Y. Xia and W. K. Li. On single–index coefficient regression models. *J. Amer. Statist. Ass.*, 94:1275–1285, 1999.

[12] Y. Xia, H. Tong, W.K. Li, and L. Zhu. An adaptive estimation of dimension reduction space. (with discussion). *J. R. Statist. Soc. B*, pages 1–28, 2002.

[13] Y. Zhu and P. Zeng. Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Ass.*, 101:1638–1651, 2006.