Statistica Sinica **14**(2004), 1147-1164

BAYESIAN ESTIMATION IN SINGLE-INDEX MODELS

Anestis Antoniadis[1], Gérard Grégoire[1] and Ian W. McKeague[2]

[1] *Université Joseph Fourier and* [2] *Columbia University*

Abstract: Single-index models offer a flexible semiparametric regression framework for high-dimensional predictors. Bayesian methods have never been proposed for such models. We develop a Bayesian approach incorporating some frequentist methods: B-splines approximate the link function, the prior on the index vector is Fisher-von Mises, and regularization with generalized cross validation is adopted to avoid over-fitting the link function. A random walk Metropolis algorithm is used to sample from the posterior. Simulation results indicate that our procedure provides some improvement over the best frequentist method available. Two data examples are included.

Key words and phrases: B-splines, Fisher-von Mises, projection pursuit regression, random walk Metropolis.

1. Introduction

Single-index models offer a flexible alternative to standard linear regression, with the conditional expectation of the response Y_i allowed to be an arbitrary link function of a finite linear combination of predictors: $E(Y_i|X_i) = f(X_i'\boldsymbol{\theta})$, $i = 1, \ldots, n$. The vector of regression coefficients $\boldsymbol{\theta}$, or index vector, is identifiable up to a constant of proportionality. The link function f is considered an infinite-dimensional nuisance parameter. Such models arise in Friedman and Stuetzle's (1981) projection pursuit regression, and they have extensive applications in econometrics. Single-index models with Gaussian predictors were formulated by Brillinger (1977, 1982). In this paper we propose a Bayesian approach incorporating some frequentist methods, as well as useful prior information, into the inference machinery.

Existing approaches to estimating the index vector are based on two methods: (1) average derivative estimation and (2) semiparametric M-estimation. The average derivative estimation approach, introduced by Stoker (1986) and Powell, Stock and Stoker (1989), exploits the fact that the index vector is proportional to the expected value of the (weighted) gradient of the regression function $E(Y_i|X_i = \cdot)$, which can be estimated nonparametrically. The method provides a $\sqrt{n}$-consistent estimator of the index vector, but may fail in practice with high-dimensional predictors because of the curse-of-dimensionality. Various

improvements and extensions have been studied by Härdle and Stoker (1989), Härdle and Tsybakov (1993), Härdle, Marron and Tsybakov (1992) and Samarov (1993). Asymptotic efficiency of the estimator has been studied by Newey and Stoker (1993). A fast implementation using binning was developed by Turlach (1994). Recently, a sophisticated method of iterative improvement was introduced by Hristache, Juditsky and Spokoiny (2001a); this method extends to multiple-index models, see Hristache, Juditsky, Polzehl and Spokoiny (2001b). A goodness-of-fit test for the single index model has recently been developed by Delecroix, Hall and Vial-Roget (2002).

The semiparametric efficiency approach derives an M-estimator for $\boldsymbol{\theta}$ based on a nonparametric estimator of f with $X_i'\boldsymbol{\theta}$, $i = 1, \ldots, n$ as the design points. Various semiparametric estimators have been studied by Klein and Spady (1993), Ichimura (1993), Horowitz and Härdle (1996) and Delecroix and Hristache (1991), among others. Härdle, Hall and Ichimura (1993) considered the problem of bandwidth selection. In some cases the estimators are shown to be $\sqrt{n}$-consistent and asymptotically efficient. Chiou and Müller (1998) developed a three-stage iterative procedure to estimate the index vector, the link function and the variance (modeled as a function of the mean response) in a quasi-likelihood version of the single-index model, but a strong condition is imposed on the design and there is no guarantee that the iterative procedure converges. Despite good theoretical properties, the semiparametric approach often leads to computational difficulties when attempting to evaluate the estimate of the index vector (which requires the solution to a high-dimensional maximization problem).

The numerical instability of the existing approaches persists even with sophisticated bandwidth selection or iterative improvement techniques; see, e.g., Yu and Ruppert (2002) and Carroll, Fan, Gijbels and Wand (1997). A Bayesian approach offers the hope of more stable estimates, especially for small or moderate data sets with low signal-to-noise ratio. The price to pay is that the implementation needs computationally intensive Markov chain Monte Carlo (MCMC) techniques. Nevertheless, the basic simplicity of the single-index model suggests that Bayesian methods should be relatively easy to develop and implement. Our aim in this paper is to implement the Bayesian approach in Matlab so that it runs within a few minutes on standard desktop computers for moderately large data sets.

Bayesian approaches have been successful with many other nonlinear regression models in recent years, e.g., additive semiparametric regression (Biller (2000)), semiparametric hazard function regression (McKeague and Tighiouart (2000)), nonparametric regression with measurement error (Berry, Carroll and Ruppert (2002)), and generalized linear models with parametric link function determination (Ntzoufras, Dellaportas and Forster (2003)). To the best of our

knowledge, however, a Bayesian approach has not previously been developed for single-index models.

The prior distribution of $\boldsymbol{\theta}$ is taken as Fisher–von Mises, which is a traditional model in the analysis of directional data (Mardia and Jupp (2002)). The link function f is treated as another unknown and approximated by a linear combination of B-spline basis functions with regularly spaced knots. The prior distribution for the B-spline coefficients $\boldsymbol{\beta}$, conditional on $\boldsymbol{\theta}$, is specified as normal with the same dispersion matrix as the least squares estimator. This popular conjugate prior is known as Zellner's (1986) g-prior. We assign an inverse gamma prior for the observation error variance σ^2. This framework allows $\boldsymbol{\beta}$ and σ^2 to be analytically integrated out of the posterior, greatly simplifying the computations.

We add a regularization feature to avoid over-fitting the link function f. Regularization could be achieved by adapting Smith and Kohn's (1996) (see also Kohn, Smith and Chan (2001)) variable selection procedure for the B-spline basis coefficients. In our setting, however, this would slow down the MCMC by at least an order of magnitude. Instead, we develop two simple ridge-regression type modifications, either of which can be used to reduce the bias in the estimates, leading to a much improved regression fit without additional computational cost. The regularization is controlled by a smoothing parameter which is selected by an empirical Bayes procedure.

We use a random walk Metropolis algorithm to sample from the regularized posterior. The proposal distribution is taken as Fisher–von Mises, which is sampled using an algorithm of Ulrich (1984). The resulting Markov chain is easily seen to be geometrically ergodic.

We carry out a small simulation study to compare the performance of our estimator with that of Hristache, Juditsky and Spokoiny (2001a) (HJS for short), which seems to be the best of the existing frequentist estimators. The results indicate that our procedure provides some improvement over HJS.

The paper is organized as follows. The Bayesian single-index model is formulated in Section 2. The regularization procedure is described in Section 3, and the random walk Metropolis algorithm in Section 4. Section 5 contains the results of the simulation study, and two data examples are given in Section 6.

2. Model Formulation

We consider the single-index model

$$Y_i = f(X_i'\boldsymbol{\theta}) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where the Y_i are scalar response variables, the X_i are d-dimensional predictors, and the ϵ_i are independent $N(0, \sigma^2)$ random errors. The unknown parts of the model are the d-dimensional index vector $\boldsymbol{\theta}$, which is normalized to have unit euclidean norm, the link function f, and the error variance σ^2.

The prior is defined in a hierarchical fashion, starting with $\boldsymbol{\theta}$ and σ^2, and then specifying the conditional distribution of f given $\boldsymbol{\theta}$ and σ^2 in terms of a B-spline approximation. Some guidance on the choice of various hyperparameters that are fixed (and thus treated as tuning constants) is given later. The prior is specified as follows.

1. The prior of $\boldsymbol{\theta}$ is Fisher–von Mises with concentration parameter $\lambda_{\text{prior}} \geq 0$ and modal direction $\boldsymbol{\theta}_{\text{prior}}$. In other words, the prior density for $\boldsymbol{\theta}$ is proportional to $\exp(\lambda_{\text{prior}}\boldsymbol{\theta}'\boldsymbol{\theta}_{\text{prior}})$, for $\boldsymbol{\theta}$ in the unit sphere in $\mathbb{R}^d$.
2. The prior of σ^2 is inverse-gamma with density proportional to $\sigma^{-2(A+1)}\exp(-B^{-1}\sigma^{-2})$, $\sigma^2 > 0$. Here $A > 0$ and $B > 0$ are tuning constants.
3. Given $\boldsymbol{\theta}$ and σ^2, the link function f is represented by a linear combination of B-spline basis functions:

$$f(t) = \sum_{j=1}^{K} \beta_j B_j(t), \tag{2.2}$$

where B_j is the jth B-spline basis function of degree q based on $m+1$ equispaced knots over the interval $[a_\theta, b_\theta]$, and $K = m+q$. Here $a_\theta = \min\{X_i'\boldsymbol{\theta}, i = 1,\ldots,n\} - \delta$ and $b_\theta = \max\{X_i'\boldsymbol{\theta}, i = 1,\ldots,n\} + \delta$ for some $\delta > 0$. For given $\boldsymbol{\theta}$ and σ^2, the basis coefficients have prior

$$\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)' \sim N\left(\hat{\boldsymbol{\beta}}_\theta, \sigma^2(\tilde{X}_\theta'\tilde{X}_\theta)^{-1}\right), \tag{2.3}$$

where $\tilde{X}_\theta$ is the design matrix in the B-spline representation of f evaluated at the 'transformed design points' $X_i'\boldsymbol{\theta}$, $i = 1,\ldots,n$, and $\hat{\boldsymbol{\beta}}_\theta = (\tilde{X}_\theta'\tilde{X}_\theta + \rho I)^{-1}\tilde{X}_\theta'\mathbf{Y}$ is the corresponding penalized least squares (or ridge regression) estimator of $\boldsymbol{\beta}$. The positive integers q and m, and $\delta > 0$, $\rho \geq 0$, are tuning constants. The full definitions of B_j and $\tilde{X}_\theta$ are given below.

B-splines provide an attractive system of basis functions. Their simplicity allows $\tilde{X}_\theta$ to be computed quickly compared with other systems, which is an important consideration, because this matrix is recalculated at each step of the sampler.

A B-spline basis function of degree q consists of $q+1$ polynomial pieces of degree q such that the derivative of order $q-1$ is continuous. Each basis function is positive on its support, which is spanned by $q+2$ knots. To approximate a function f on a finite interval $[a,b]$, divide the interval into m subintervals using $m+1$ equispaced knots $x_i = a+(i-1)h_m$, $i = 1,\ldots,m+1$, where $h_m = (b-a)/m$. Each subinterval is covered by $q+1$ non-zero B-splines of degree q. The total number of knots for the construction is $m+1+2q$, with the extra $2q$ knots entering because we need to add q knots to the left of a and q knots to the right of b. The number of basis functions needed to approximate f on $[a,b]$ is

$K = m + q$. The basis functions are written as B_j, $j = 1, \ldots, K$ with the index j connecting the basis function to a unique knot.

The B-spline representation of f in (2.2) allows the single-index model to be written in the form

$$\mathbf{Y} = B(X\boldsymbol{\theta})\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.4}$$

where B is the matrix-valued function defined by $B(\mathbf{t})_{ij} = B_j(t_i)$, $\mathbf{t} = (t_1, \ldots, t_n)' \in \mathbb{R}^n$, $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the vector of responses, $X = (X_1, \ldots, X_n)'$ is the original design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is the vector of observation errors. The design matrix in this form of the model is denoted $\tilde{X}_\theta = B(X\boldsymbol{\theta})$. Note that the basis functions B_j used here implicitly depend on $\boldsymbol{\theta}$ through the interval $[a_\theta, b_\theta]$ that was arranged to cover the transformed design points $X_i'\boldsymbol{\theta}$.

We suggest choosing λ_{prior} in the range 100−700, depending on the degree of confidence in the prior modal direction. The modal direction $\boldsymbol{\theta}_{\text{prior}}$ can be taken as one of the existing frequentist estimates of $\boldsymbol{\theta}$; we have obtained good results using the HJS estimator and when we tried an arbitrary $\boldsymbol{\theta}_{\text{prior}}$, the results were poor. When nothing is known about the error variance, A should be small (say 0.001) and B large (say 100). For typical applications, $q = 2$, m =15−20, and $\delta = 0.001$ work well, cf. the number of knots used in Yu and Ruppert (2002). These are only general guidelines, however, and for any specific application it is important to experiment with various choices of λ_{prior}, A, B and m by examining the plots of the estimated fit, as we illustrate in Sections 5 and 6. Automatic data-driven selection of these tuning parameters could be handled by an empirical Bayes procedure (we discuss this further in the next section), or, alternatively, by assigning higher level priors, as in hierarchical Bayesian modeling. However, we found in our numerical studies that the results are relatively insensitive to variations in these parameters, so the extra effort required to implement such an extension was not considered worthwhile.

3. Regularization

The posterior density of $\boldsymbol{\theta}$, derived in the Appendix, is proportional to

$$D(\boldsymbol{\theta}) = (S(\boldsymbol{\theta}) + 2/B)^{-(A+n/2)} \exp(\lambda_{\text{prior}}\boldsymbol{\theta}'\boldsymbol{\theta}_{\text{prior}}), \tag{3.5}$$

where

$$S(\boldsymbol{\theta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\tilde{X}_\theta \left(\Sigma_0 + \frac{1}{2}\Sigma_\rho(I - \Sigma_0^{-1}\Sigma_\rho)\right)\tilde{X}_\theta'\mathbf{Y}, \tag{3.6}$$

$\Sigma_\rho = (\tilde{X}_\theta'\tilde{X}_\theta + \rho I)^{-1}$ and Σ_0 is Σ_ρ with ρ set to zero. For a given $\boldsymbol{\theta}$, use of the ordinary least squares estimator (i.e., $\hat{\boldsymbol{\beta}}_\theta$ with $\rho = 0$) in place of $\hat{\boldsymbol{\beta}}_\theta$, and with $S(\boldsymbol{\theta})$ in (3.6) reducing to the error sum of squares, cf. Kohn, Smith and Chan (2001), results in severe over-fitting and high bias in the predictions; this is hardly

surprising because then there would be no variable selection or regularization feature built into the prior for f.

We refer to our method of specifying the prior mean of $\boldsymbol{\beta}$ to be the penalized least squares estimator $\hat{\boldsymbol{\beta}}_\theta = \operatorname{argmin}_\beta(\|\mathbf{Y} - \tilde{X}_\theta\boldsymbol{\beta}\|^2 + \rho\|\boldsymbol{\beta}\|^2)$ as *internal* regularization. Here $\|\cdot\|$ is the euclidean norm, $\rho \geq 0$ controls the penalty and I is the $K \times K$ identity matrix. More generally, the P-spline methodology of Eilers and Marx (1996) could be applied: I is replaced by a $K \times K$ symmetric positive semidefinite matrix Q, and the penalty term $\rho\|\boldsymbol{\beta}\|^2$ is replaced by $\rho\boldsymbol{\beta}'Q\boldsymbol{\beta}$.

This contrasts with a simpler though not-strictly-Bayesian method we call *external* regularization in which the ordinary least squares estimator is used as the prior mean for $\boldsymbol{\beta}$ and the posterior is directly regularized by replacing $S(\boldsymbol{\theta})$ in (3.6) by

$$S_E(\boldsymbol{\theta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\tilde{X}_\theta\Sigma_\rho\tilde{X}_\theta'\mathbf{Y}. \tag{3.7}$$

Both methods of regularization are computationally much less demanding than the Bernoulli prior of Smith and Kohn (1996) for variable selection of the B-spline basis coefficients.

The smoothing parameter ρ is an important hyperparameter that needs to be selected automatically from the data. For this purpose we consider an empirical Bayes procedure adapted to the MCMC technique of the next section. An optimal ρ is selected at each MCMC update step via a type of generalized cross validation (GCV) used by Yu and Ruppert (2001). Given $\boldsymbol{\theta}$, we minimize

$$\mathrm{GCV}(\rho) = \frac{n^{-1}\|\mathbf{Y} - \tilde{X}_\theta\hat{\boldsymbol{\beta}}_\theta\|^2}{\{1 - n^{-1}\mathrm{tr}(L(\rho))\}^2}, \tag{3.8}$$

where $L(\rho)$ is the smoothing spline operator corresponding to ρ with

$$L(\rho) = L_s(\rho) + \{I - L_s(\rho)\}\tilde{X}_\theta\{\tilde{X}_\theta'(I - L_s(\rho))\tilde{X}_\theta\}^{-1}\tilde{X}_\theta'\{I - L_s(\rho)\} \tag{3.9}$$

and $L_s(\rho)) = \tilde{X}_\theta\Sigma_\rho\tilde{X}_\theta'$. Note that the trace of $L(\rho)$ can be calculated in such a way that only the traces of matrices of dimension $\dim(\boldsymbol{\beta}) \times \dim(\boldsymbol{\beta})$ need to be evaluated. The minimization of $\mathrm{GCV}(\rho)$ is carried out with $\boldsymbol{\theta}$ replaced by its current estimate.

When this procedure is repeated at each step of the MCMC, the sampler converges to the empirical Bayes posterior distribution of $\boldsymbol{\theta}$, see Casella (2001). However, the knots and therefore the B-spline basis are modified at each step of the MCMC, so this would be computationally demanding. We have already noted that ρ mainly controls the smoothness of the estimated link function f but has relatively little effect on the estimation of the single index $\boldsymbol{\theta}$. Thus it

is reasonable to minimize GCV(ρ) just at the start of the MCMC run with the HJS estimate in place of the unknown $\boldsymbol{\theta}$; this is what we do in the sequel.

4. MCMC Algorithm

We use a random walk Metropolis algorithm to sample from the regularized posterior. The proposal distribution is taken as Fisher-von Mises with concentration parameter $\lambda_{\text{prop}} \geq 0$ and modal vector given by the current value of $\boldsymbol{\theta} = \boldsymbol{\xi}$. The algorithm of Ulrich (1984) is used to generate the proposals from the Fisher-von Mises distribution. The concentration parameter λ_{prop} needs to be calibrated to ensure that the sampler explores the state space adequately. We have found by experiment that $\lambda_{\text{prop}} = 1,000$ produces good results for small and moderate datasets; this value is larger than that recommended for the prior concentration parameter λ_{prior}. The acceptance probability for the proposal is $\alpha(\boldsymbol{\xi}, \boldsymbol{\theta}) = \min\{1, D(\boldsymbol{\theta})/D(\boldsymbol{\xi})\}$.

The following result shows that the sampler is geometrically ergodic, so we can expect good performance in practice. In particular, ergodic averages converge rapidly, the Central Limit Theorem holds, and the standard technique of assessing Monte Carlo error by the method of batch means is applicable. Even though the proof is routine, we include it here for completeness.

Lemma 1. *The random walk Metropolis sampler for the regularized posterior density of $\boldsymbol{\theta}$ is geometrically ergodic.*

Proof. Note that the state space (the unit-sphere in $\mathbb{R}^d$) is compact and the target density is bounded away from 0 and ∞, as is the proposal density. The random walk Metropolis Markov chain is therefore μ-irreducible, aperiodic and its state space is a small set, cf. Theorem 2.2 of Roberts and Tweedie (1996). Here μ denotes Haar measure on the unit-sphere in $\mathbb{R}^d$. It suffices now to check the Foster-Lyapunov drift conditions in Theorem 14.0.1 and Theorem 15.0.1 of Meyn and Tweedie (1993). In their notation, we can take the function $V = 1$, C as the whole state space, $b = 1$ and $\lambda = 1/2$. This completes the proof.

The sampler generates a sequence of index vector values with unit norm. We estimate $\boldsymbol{\theta}$ by the mean or the component-wise median of those values, renormalizing to obtain unit norm. Given such an estimate $\hat{\boldsymbol{\theta}}$, the estimated fit for the response vector at the tranformed design points is taken as $\tilde{X}_\theta \hat{\boldsymbol{\beta}}_\theta$ with $\boldsymbol{\theta}$ set to $\hat{\boldsymbol{\theta}}$. An alternative way of estimating the fit (as pointed out by a referee) would be to average $\tilde{X}_\theta \hat{\boldsymbol{\beta}}_\theta$ over the sampler output for $\boldsymbol{\theta}$, but we found this produced less satisfactory results.

5. Simulation Study

The main goal of this section is to assess the performance of our estimator of the index vector and the resulting fit through a simulation study. We compare

our estimator with that of HJS, and with the results from a generalized additive model. We also examine the behavior of the internal and external regularization methods. The programs are written in Matlab and are available upon request from the authors.

We first consider a single index model with $d = 4$ index coefficients described by equation (2.1) and link function given by $f(u) = u^2 e^u$. We use a moderate sample size, $n = 100$, and two noise levels $\sigma = 0.2$ and $\sigma = 0.5$, which, given the link function, correspond to a moderate and a low signal-to-noise ratio, respectively (by signal-to-noise ratio we mean the integral of f^2 over the interval $[a_\theta, b_\theta]$ divided by σ^2). Each predictor X_i, $i = 1, \ldots, n$ is generated independently from the uniform distribution in the cube $[-1, 1]^4$. The index vector is taken to be $\boldsymbol{\theta} = (2, 1, 1, 1)'/\sqrt{7}$. A typical simulated sample from (2.1) with such parameters and a noise level $\sigma = 0.5$ is displayed on the left panel of Figure 1.

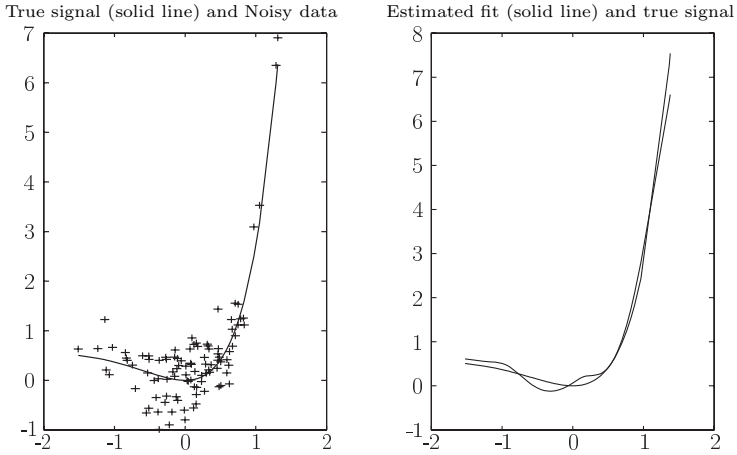

Figure 1. A simulated example. Left display: the true link function (solid line) and the response data at the true transformed design points; right display: the estimated fit (solid line) using external regularization and the true link function (dashed line).

For each value of the noise level, and each type of regularization, 100 simulation runs were performed holding the design and all other model parameters constant, except the $\{\epsilon_i\}$ which were regenerated for each run.

To measure the quality of our estimator (and that of HJS), we use two different criteria: $\text{angle}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \cos^{-1}(\hat{\boldsymbol{\theta}}'\boldsymbol{\theta})$, and sup norm $\sup_{1 \le j \le d} |\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j|$. Throughout this section, the prior concentration parameter is taken as $\lambda_{\text{prior}} = 150$ and the modal direction $\boldsymbol{\theta}_{\text{prior}}$ as the estimate of HJS. For the parameters of

the inverse gamma distribution specifying the prior distribution of σ^2, we chose $A = 0.001$ and $B = 100$. The B-splines had 20 knots ($m = 19$) and were of degree $q = 2$, reflecting assumed knowledge that the link function is smooth. Edge effects were controlled using $\delta = 0.001$. The minimization of GCV(ρ) is carried out on a grid of 20 ρ-values ranging from 0.1 to 4.

The random walk Metropolis algorithm was implemented as follows. We chose $\lambda_{\text{prop}} = 1,000$ for the concentration parameter of the proposal distribution, reflecting our confidence in HJS's estimate as a good starting point. This gave an acceptance rate for the proposals of 52%. Roberts and Rosenthal (2001) have shown that the optimal acceptance rate in terms of minimal Monte Carlo error for this type of Metropolis sampler is 23.4%; although we have a slightly suboptimal acceptance rate, the issue of optimality is not very important in our case as the MCMC output is obtained so quickly. After a burn-in of 500 iterations, we used 4,000 iterations to obtain $\hat{\boldsymbol{\theta}}$; the Monte Carlo error based on the 4,000 sampled values of $\boldsymbol{\theta}$ was negligible.

The left panel of Figure 1 displays the true link function and the responses at each of the transformed design points. The right panel displays the true and estimated fit by our method; 'fit' here is defined as the link function interpolated from the estimated transform of the design points. It can be seen that the estimated fit displays large variability in the lower part of the design interval, probably due to the low signal-to-noise ratio on this part of the design.

That regularization is necessary for estimating the link function is illustrated in Figure 2, which is similar to Figure 1, but with $\rho = 0$ (no regularization). Clearly, the absence of any regularization has a disastrous effect on the estimates.

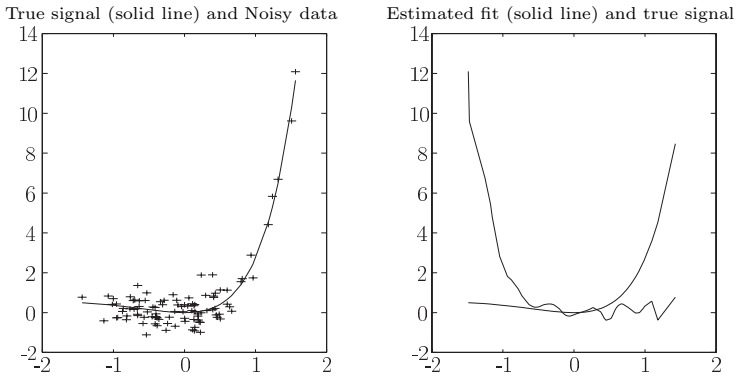

Figure 2. Same as Figure 1 but with $\rho = 0$.

Figures 3 and 4 compare the performance of our estimator with the internal and external regularizations to that of the HJS estimator, for two noise levels.

The posterior mean is the best estimator in each case, and outperforms the estimator of HJS. There are no significant differences between the two types of regularization.

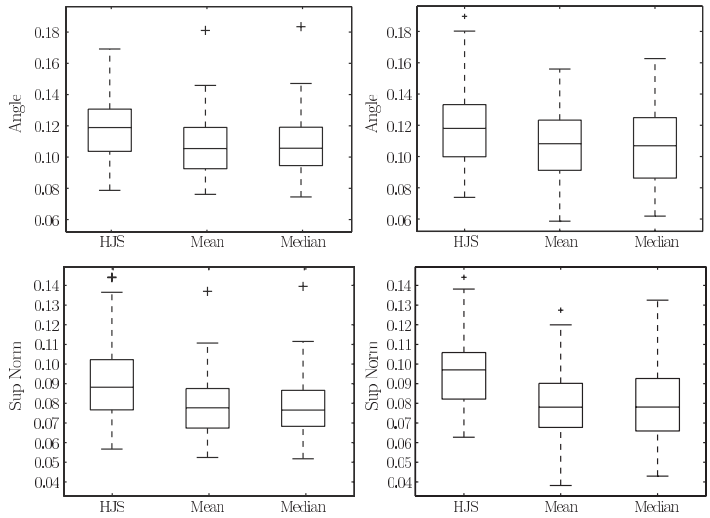

Figure 3. Box plots of the error criteria over 100 simulation runs with noise level $\sigma = 0.2$, $d = 4$; left panels: external regularization; right panels: internal regularization.

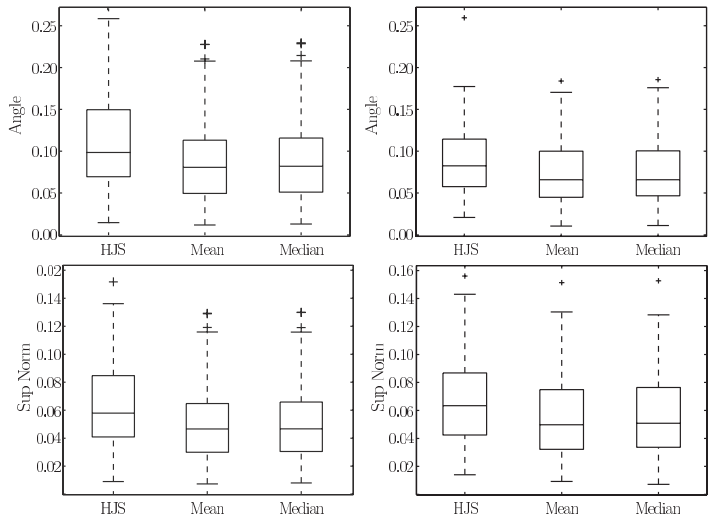

Figure 4. Same as Figure 3 but with $\sigma = 0.5$.

Next we consider higher dimensional predictors. Using the same single index model as before, but with $d = 10$ and $\boldsymbol{\theta} = (1, 1, 1, 2, 0, 0, 0, 0, 0, 0)'/\sqrt{7}$, we find that the Bayesian estimators continue to offer some improvement over HJS, see Figure 5. For the same model, Figure 6 shows boxplots of the average squared error of the fit from nonlinear least squares compared with our Bayesian approach; the nonlinear least squares fit is obtained under the assumption that an oracle provides the link function. Notice that the average squared errors have greater dispersion for the Bayesian fit, but the means are almost identical, suggesting that our approach provides an efficient adaption to the unknown link function.

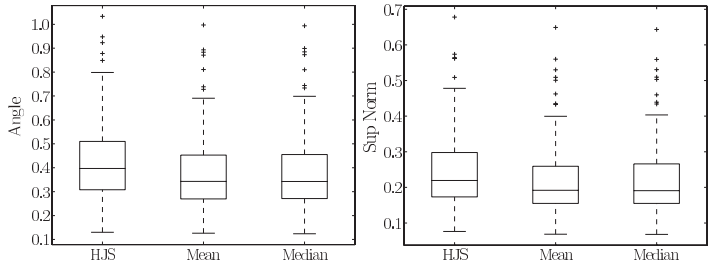

Figure 5. Box plots of the error criteria over 100 simulation runs with noise level $\sigma = 0.5$ and $d = 10$. Internal regularization.

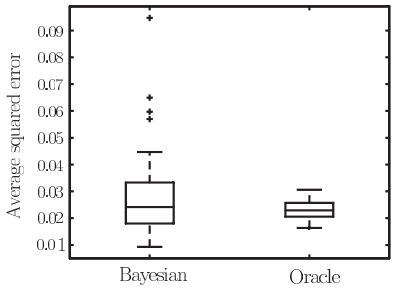

Figure 6. Box plots of averaged squared errors over 100 simulation runs with noise level $\sigma = 0.5$ and $d = 10$; left: Bayesian fit based on the posterior mean with internal regularization; right: fit based on nonlinear least squares with an oracle providing the link function.

We next examine the behavior of the Bayesian single index model fit to a departure from the model through nonlinear interactions between the predictors. We simulated data from the model $E(Y|U, V) = f(g(U) + h(V))$, where f is the standard normal cdf, $g(u) = 0.5(\exp(0.35u) - 1)$, $h(v) = 0.5\sin(\pi v/3)$, and the predictors U, V are independent $N(0, 4)$. The sample size is $n = 100$ and $\sigma = 0.1$. Note that g has a much larger squared-$L^2[-10, 10]$-norm (29.2) than

h (3.8), so the single index parameter estimate corresponding to U is expected to dominate. Indeed, the posterior mean $\hat{\boldsymbol{\theta}} = (0.9995, 0.0325)'$, indicating that only the variation due to the first predictor (U) is captured by the single index model. Figure 7 (left panel) displays the single index model fit and the function $u \mapsto f(g(u))$ that it tends to approximate. In the right panel we plot the first generalized additive model component obtained using the Splus script `gam` of Hastie and Tibshirani, which is seen to be in quite close agreement with the single index model fit.

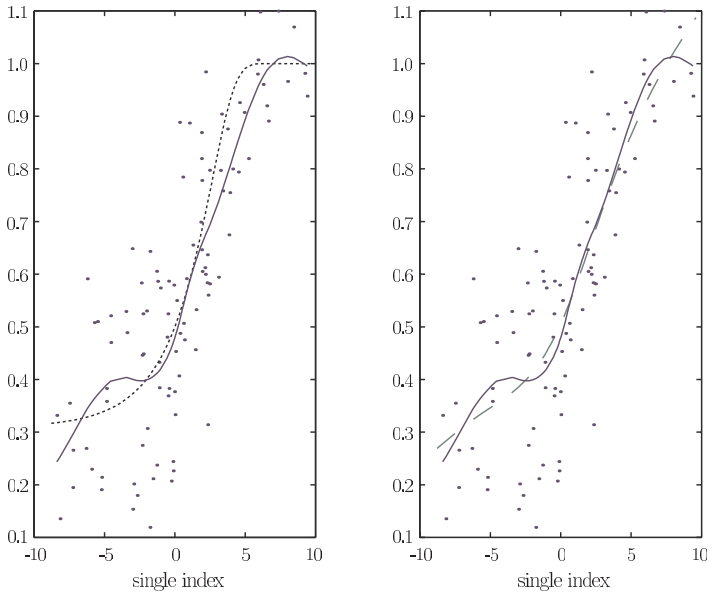

Figure 7. A simulated generalized additive model with two components and a nonlinear link function; left: the single index fit (solid line), and the approximated link function (dotted line); right: the single index model fit (solid line), and the first component of the `gam` fit (dashed line).

6. Examples

We now test our methodology on two data examples. The first dataset originated from a petroleum reservoir study conducted by BP Research, and contains measurements on four cross-sections from each of twelve core samples. Each core sample was measured for permeability, and the cross-sections analysed for total area of pores, total perimeter of pores, and shape. The data are available in the MASS Splus library of Venables and Ripley (1999). The aim of the study is to predict permeability Y (a property of fluid flow) measured in milli-Darcies from

the cross-section measurements, which are the end-product of a complex image processing procedure performed at the University of Oxford, providing the total area in pixels (`area`), total perimeter in pixels (`peri`), and $\texttt{shape} = \texttt{peri}/\sqrt{\texttt{area}}$, a measure of roundness of the pores in the cross-section. As permeabilities vary greatly (6.3–1300), we have used a log scale.

To fit our Bayesian single-index model to these data, and in the second example below, we used the following values of the tuning parameters: $\lambda_{\text{prior}} = 700$, $\boldsymbol{\theta}_{\text{prior}}$ is again the HJS estimate, $A = 0.001$, $B = 100$, $q = 2$, $\delta = 0.001$, and $m = 14$. Here we used 15 knots rather than 20 due to the small sample size. The concentration parameter of the proposal distribution is $\lambda_{\text{prop}} = 1,000$, and the smoothing parameter, determined by GCV, is $\rho = 1.82$. The plot of the resulting fit is shown in Figure 8.

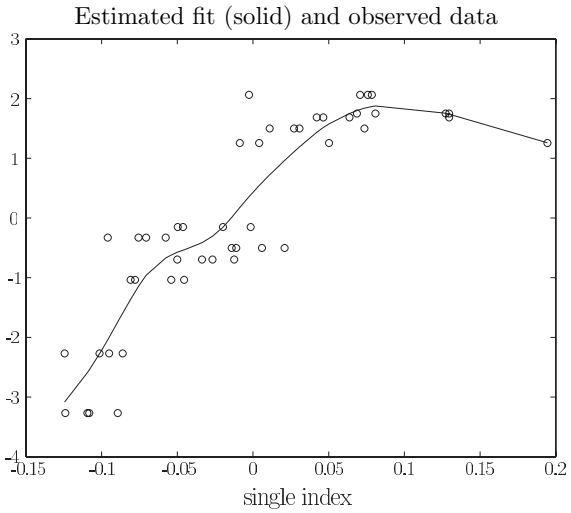

Figure 8. Permeability data example: estimated link function over the transformed design points.

After a burn-in of 500 iterations, 4,000 iterations of the sampler gave posterior mean $\hat{\boldsymbol{\theta}} = (0.3418, -0.9343, -0.0463)'$ with corresponding componentwise posterior standard deviations $(0.0379, 0.0150, 0.0808)'$. Note that the variables `area` and `peri` dominate. Note also that the resulting index vector matches closely with the first projection pursuit direction given in Venables and Ripley (1999) for the same data. It also matches well with the HJS estimate $(0.3714, -0.9267, -0.0580)'$ with corresponding standard deviations $(0.0425, 0.0183, 0.0896)'$.

For our second example, the dataset originated from an environnemental study relating ozone concentration Y to three meteorological variables, namely

wind speed, temperature and radiation. The data are daily measurements of the four variables for $n = 111$ days (Chambers and Hastie (1992)). As the ozone concentrations vary greatly, we have used a log scale. The smoothing parameter, determined by GCV, is $\rho = 1.86$. The plot of the resulting fit is shown in Figure 9.

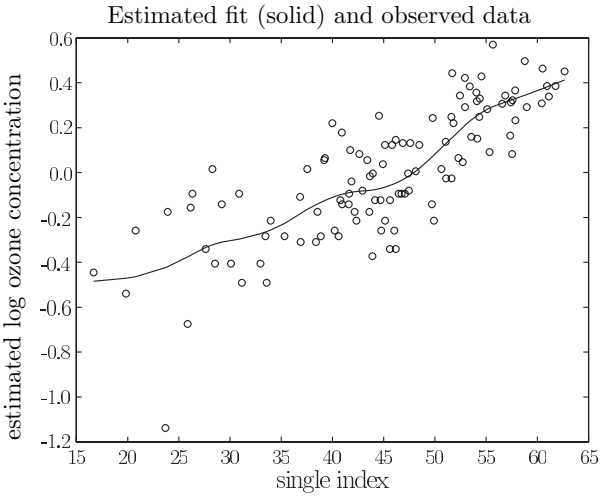

Figure 9. Ozone data example.

The posterior mean estimate of the index vector is $\hat{\boldsymbol{\theta}} = (0.0817, 0.5565, -0.8103)'$ with corresponding componentwise posterior standard deviations $(0.0677, 0.1248, 0.0832)'$. This is close to the HJS estimate $(0.0407, 0.5263, -0.8493)'$, with corresponding standard deviations $(0.0821, 0.1469, 0.0886)'$. The resulting fit also matches closely with the generalized additive model fit given in Chambers and Hastie (1992) for the same data. Note also that the `temperature` and `wind` variables in the second and third components dominate, which is consistent with the conclusion of Chambers and Hastie.

In future work it would be of interest to extend our Bayesian single-index model to allow σ^2 to depend on the index, or even on the mean response, cf. the quasi-likelihood model of Chiou and Müller (1998). This would not be a straightforward extension however, because it would probably no longer be possible to integrate out σ^2 from the posterior, and a much more sophisticated MCMC strategy would be needed.

Acknowledgements

The authors thank Anatoli Juditsky for providing Matlab code implementing the HJS estimator. Ian McKeague acknowledges support from Université Joseph

Fourier and NSF Grant 0204688.

Appendix. Proof of (3.5)

The derivation of the posterior marginal distribution of $\boldsymbol{\theta}$ under internal regularization requires some care, so we have included it here for completeness. First recall the notation $\hat{\boldsymbol{\beta}}_\theta = \Sigma_\rho \tilde{X}_\theta' \mathbf{Y}$ for the prior mean of $\boldsymbol{\beta}$ given $\boldsymbol{\theta}$, where $\Sigma_\rho = (\tilde{X}_\theta' \tilde{X}_\theta + \rho I)^{-1}$.

From Bayes formula, the joint posterior density of the parameters in the model is

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{Y}) &\propto p(\mathbf{Y} \,|\, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{\beta} |\, \boldsymbol{\theta}, \sigma^2)\, p(\boldsymbol{\theta}) p(\sigma^2) \\ &\propto \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{Y} - \tilde{X}_\theta \boldsymbol{\beta})' (\mathbf{Y} - \tilde{X}_\theta \boldsymbol{\beta}) \right\} \\ &\quad \times \frac{1}{|\det \Sigma_0|^{1/2} (2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\theta)' \Sigma_0^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\theta) \right\} \\ &\quad \times \exp(\lambda_{\text{prior}} \boldsymbol{\theta}' \boldsymbol{\theta}_{\text{prior}}) \frac{1}{\sigma^{2(A+1)}} \exp\left\{ \frac{-1}{B\sigma^2} \right\}. \end{aligned}$$

By straightforward calculations, the first two exponential terms can be written as

$$\exp\left\{ -\frac{1}{2\sigma^2} 2 \left(\boldsymbol{\beta}' \Sigma_0^{-1} \boldsymbol{\beta} - 2\boldsymbol{\beta}' (I + \Sigma_0^{-1} \Sigma_\rho) \tilde{X}_\theta' \mathbf{Y} + \mathbf{Y}' \mathbf{Y} + \mathbf{Y}' \tilde{X}_\theta \Sigma_\rho \Sigma_0^{-1} \Sigma_\rho \tilde{X}_\theta' \mathbf{Y} \right) \right\}.$$

Thus, integrating out $\boldsymbol{\beta}$ yields

$$\begin{aligned} &\int \exp\left\{ -\frac{1}{2\sigma^2} \left((\mathbf{Y} - \tilde{X}_\theta \boldsymbol{\beta})' (\mathbf{Y} - \tilde{X}_\theta \boldsymbol{\beta}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\theta)' \Sigma_0^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_\theta) \right) \right\} \\ &\quad \times \frac{1}{|\det \Sigma_0|^{1/2} (2\pi\sigma^2)^{n/2}} d\boldsymbol{\beta} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} S_1(\boldsymbol{\theta}) \right\} \int \exp\left\{ -\frac{1}{\sigma^2} \boldsymbol{\beta}' (I + \Sigma_0^{-1} \Sigma_\rho) \tilde{X}_\theta' \mathbf{Y} \right\} \\ &\quad \times \frac{1}{|\det \Sigma_0|^{1/2} (2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{\sigma^2} \boldsymbol{\beta}' \Sigma_0^{-1} \boldsymbol{\beta} \right\} d\boldsymbol{\beta} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} S_1(\boldsymbol{\theta}) \right\} \phi\left(\frac{1}{\sigma^2} (I + \Sigma_0^{-1} \Sigma_\rho) \tilde{X}_\theta' \mathbf{Y} \right), \end{aligned}$$

where $S_1(\boldsymbol{\theta}) = \mathbf{Y}'\mathbf{Y} + \mathbf{Y}' \tilde{X}_\theta \Sigma_\rho \Sigma_0^{-1} \Sigma_\rho \tilde{X}_\theta' \mathbf{Y}$ and ϕ is the Laplace transform of the $N(0, (\sigma^2/2)\Sigma_0)$ distribution. Replacing ϕ by its familiar expression allows us to write the last line above in the form

$$\exp\left\{ -\frac{1}{2\sigma^2} S_1(\boldsymbol{\theta}) \right\} \exp\left\{ \frac{1}{4\sigma^2} \mathbf{Y}' \tilde{X}_\theta \Lambda \tilde{X}_\theta' \mathbf{Y} \right\},$$

where we denote $\Lambda = (I + \Sigma_\rho \Sigma_0^{-1})\Sigma_0(I + \Sigma_\rho \Sigma_0^{-1})$. Note that $\Lambda = 4\Sigma_0$ when $\rho = 0$. The result of the above calculus can be restated as

$$p(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{Y}) \propto \exp(\lambda_{\text{prior}} \boldsymbol{\theta}' \boldsymbol{\theta}_{\text{prior}}) \frac{1}{\sigma^{2(A+1+n/2)}} \times \exp\left\{ -\frac{1}{2\sigma^2} \left(\frac{1}{B/2} + S_1(\boldsymbol{\theta}) - \frac{1}{2} \mathbf{Y}' \tilde{X}_\theta \Lambda \tilde{X}_\theta' \right) \right\}.$$

The final step is to integrate out σ^2. This gives

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{Y}) \propto D(\boldsymbol{\theta}) &= \exp(\lambda_{\text{prior}} \boldsymbol{\theta}' \boldsymbol{\theta}_{\text{prior}}) \left(S(\boldsymbol{\theta}) + 2/B \right)^{-(A+n/2)}, \\ S(\boldsymbol{\theta}) &= S_1(\boldsymbol{\theta}) - \frac{1}{2} \mathbf{Y}' \tilde{X}_\theta \Lambda \tilde{X}_\theta' \mathbf{Y} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}' \tilde{X}_\theta \left(\Sigma_0 + \frac{1}{2} \Sigma_\rho (I - \Sigma_0^{-1} \Sigma_\rho) \right) \tilde{X}_\theta' \mathbf{Y}, \end{aligned}$$

completing the derivation.

References

Berry, S., Carroll, R. and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.* **97**, 160-169.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *J. Comput. Graph. Statist.* **9**, 122-140.

Brillinger, D. R. (1997). The identification of a particular nonlinear time series system. *Biometrika* **64**, 509-515.

Brillinger, D. R. (1982). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehman* (Edited by P. J. Bickel, K. A. Doksum and J. L. Hodges), 97-114. Wadsworth, New York.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Geberalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics* **2**, 485-500.

Chambers, J. M. and Hastie, T. (1992). *Statistical Models in S*. Wadsworth, New York.

Chiou, J.-M. and Müller, H.-G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.* **93**, 1376-1387.

Delecroix, M. and Hristache, M. (1999). *M*-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bull. Belgian Math. Soc.* **6**, 161-185.

Delecroix, M., Hall, P. and Vial-Roget, C. (2002). Test des modèles à direction révélatrice unique. *Abstracts of the 34th meeting of the French Statistical Society*, 364-365.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.* **11**, 89-121.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817-823.

Härdle, W. and Tsybakov, A. B. (1993). How sensitive are average derivatives? *J. Econometrics* **58**, 31-48.

Härdle, W. and Stoker, T. M. (1989). Invertigating smoothing multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Härdle, W., Marron, J. S. and Tsybakov, A. B. (1992). Bandwidth choice for average derivative estimation. *J. Amer. Statist. Assoc.* **87**, 218-226.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91**, 1632-1640.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001a). Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29**, 595-623.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001b). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29**, 1537-1566.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58**, 71-120.

Klein, R. L. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387-421.

Kohn, R., Smith, M. and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statist. Comput.* **11**, 313-322.

Mardia, K. V. and Jupp, P. V. (2002). *Statistics and Computing*, Wiley, New York.

McKeague, I. W. and Tighiouart, M. (2000). Bayesian estimators for conditional hazard functions. *Biometrics* **56**, 1007-1015.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer-Verlag, London.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica* **61**, 11911-1223.

Ntzoufras, I., Dellaportas, P. and Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *J. Statist. Plann. Inference* **111**, 165-180.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-1430.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.* **16**, 351-367.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multivariate Hastings and Metropolis algorithms. *Biometrika* **83**, 95-110.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88**, 836-847.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317-344.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54**, 1461-1481.

Turlach, B. A. (1994). Fast implemetation of density-weighted average derivative estimation. In *Computing Science and Statistics* **26** (Edited by J. Sall and A. Lehman), 28-33. Computationally Intensive Statistical Methods.

Ulrich, G. (1984). Computer generation of distributions on the m-sphere. *J. Roy. Statist. Soc. Ser. C* **33**, 158-163.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus.* Springer, New York.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimations for partially linear single index models. *J. Amer. Statist. Assoc.* **97**, 1042-1054.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233-243. North-Holland.

Laboratoire de Modelisation et Calcul, Université Joseph Fourier, Tour IRMA, B.P.53, 38041 Grenoble CEDEX 9, France.

E-mail: anestis.antoniadis@imag.fr

Laboratoire de Modelisation et Calcul, Université Joseph Fourier, Tour IRMA, B.P.53, 38041 Grenoble CEDEX 9, France.

E-mail: gerard.gregoire@imag.fr

Department of Biostatistics, Columbia University, 722 West 168th Street, 6th Floor, New York, NY 10032, U.S.A.

E-mail: im@columbia.edu

(Received February 2003; accepted November 2003)