Bayesian Methods for Wavelet Series in Single-Index Models

Chun Gun Park, Marina Vannucci and Jeffrey D. Hart [1]

Last Revision: August 27, 2004

Journal of Computational and Graphical Statistics, to appear

Abstract

Single-index models have found applications in econometrics and biometrics, where multidimensional regression models are often encountered. Here we propose a nonparametric estimation approach that combines wavelet methods for non-equispaced designs with Bayesian models. We consider a wavelet series expansion of the unknown regression function and set prior distributions for the wavelet coefficients and the other model parameters. To ensure model identifiability, the direction parameter is represented via its polar coordinates. We employ ad hoc hierarchical mixture priors that perform shrinkage on wavelet coefficients and use Markov chain Monte Carlo methods for *a posteriori* inference. We investigate an independence-type Metropolis-Hastings algorithm to produce samples for the direction parameter. Our method leads to simultaneous estimates of the link function and of the index parameters. We present results on both simulated and real data, where we look at comparisons with other methods.

[1]Chun Gun Park (cgpark@ncc.re.kr) is Researcher, National Cancer Center, Republic of Korea. Marina Vannucci (mvannucci@stat.tamu.edu) is Associate Professor of Statistics, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, USA. Jeffrey D. Hart (hart@stat.tamu.edu) is Professor of Statistics, Texas A&M University.

Key words: Bayesian methods, MCMC, non-equispaced design, nonparametric regression, single-index models, wavelet series, wavelet shrinkage.

1 Introduction

Let (Y_i, X_i) be generated by the nonparametric regression model

$$Y_i = r(X_i\boldsymbol{\beta}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

with Y_i a scalar response variable, $\mathbf{X}_i$ p-variate explanatory variables, $p > 1$, $\varepsilon_i \sim N(0, \sigma^2)$ *i.i.d.*, $\boldsymbol{\beta}$ a $p \times 1$ vector of unknown parameters such that $\|\boldsymbol{\beta}\| = 1$ and $r : \mathbb{R} \to \mathbb{R}$ an unknown *link* function. Models of type (1) are referred to as *single-index models*, Stoker (1986), Härdle and Stoker (1989) and Ichimura (1993). The direction parameter $\boldsymbol{\beta}$ is often called the *index*.

Single-index models represent a convenient way to handle high-dimensional data, allowing for nonlinear dependency and, at the same time, avoiding the "curse of dimensionality" problem related to a fully nonparametric approach. Typical inferential procedures for such models are two-step, estimating $\boldsymbol{\beta}$ first to form $z = \mathbf{X}\hat{\boldsymbol{\beta}}$ and then r. Härdle *et al.* (1993) considered a kernel estimator of r with bandwidth h and minimised a mean integrated squared error to estimate h and $\boldsymbol{\beta}$. Li and Duan (1989) proposed an alternative method that uses "sliced inverse regression" to estimate the index parameter. Extensions to generalized partially linear single-index models were put forward by Carroll *et al.* (1997), Xia *et al.* (1999) and recently by Yu and Ruppert (2002). See the introduction of Antoniadis *et al.* (2004) for additional references to classical approaches to single-index models.

Many of the existing classical methods often lead to numerical instability when estimating the index vector, especially in high dimensions. Bayesian methods, on the other hand, can provide more stable estimates, especially for small or moderate sample sizes with a low signal-to noise ratio. In Antoniadis *et al.* (2004) the authors propose a Bayesian approach to single-index modelling that incorporates some frequentist methods. They use B-splines to approximate the link function and a prior model with a regularization feature to avoid over-fitting. The index vector is estimated via a random walk Metropolis algorithm.

Here we construct a nonparametric estimator of the link function r that uses wavelet bases expansions. The regression model we deal with is not the typical setup where wavelets have been mostly applied in the literature, in that the design is non-equispaced. We adopt a Bayesian approach and set appropriate prior distributions for the coefficients of the wavelet expansions and for the model parameters. We perform inference via MCMC methods, investigating an independence-type Metropolis-Hastings to produce samples from the posterior distribution of the direction parameter. Our method leads to simultaneous posterior estimates of the link function and of the index parameter, as well as estimates of the other model parameters. To the best of our knowledge this work represents the first attempt to combine wavelet methods for non-equispaced designs with Bayesian methods for single-index models. In what follows we will find it convenient to re-parameterize the direction parameter by its polar coordinates

$$\beta_1 = \cos(\theta_{p-1}) \dots \cos(\theta_2) \cos(\theta_1)$$
$$\beta_2 = \cos(\theta_{p-1}) \dots \cos(\theta_2) \sin(\theta_1)$$

$$\ldots$$

$$\beta_{p-1} = \cos(\theta_{p-1})\sin(\theta_{p-2})$$

$$\beta_p = \sin(\theta_{p-1}),$$

where $0 < \theta_1 < 2\pi$ and $-\frac{\pi}{2} < \theta_i < \frac{\pi}{2}$ for $i = 2, \ldots, p-1$. This is a convenient way of imposing the constraint $\beta_1^2 + \ldots + \beta_p^2 = 1$, which is necessary to make the model identifiable. We can write this transformation as $\boldsymbol{\beta} = T(\boldsymbol{\theta}) = (t_1(\boldsymbol{\theta}), \ldots, t_p(\boldsymbol{\theta}))^T$ with

$$t_d(\boldsymbol{\theta}) = \beta_d = \sin(\theta_{d-1}) \prod_{j=0}^{p-d} \cos(\theta_{p-j}), \quad d = 1, \ldots, p, \tag{2}$$

where $\sin(\theta_0) = \cos(\theta_p) = 1$.

The remainder of the paper is organized as follows: Section 2 is a brief introduction to wavelet series and wavelet estimation methods. Section 3 describes the model and the prior distributions. Section 4 addresses the posterior inference and related computational issues. Section 5 illustrates the method using simulated and real data. For the latter we also look at comparisons with kernel local linear smooths and a splines-based Bayesian method. Section 6 is a concluding discussion.

2 Preliminaries

2.1 Wavelet series

A wavelet basis in $L_2(\mathbb{R})$ is a collection of functions obtained as translations and dilations of a scaling function ϕ and a wavelet function ψ, Daubechies (1992). The function ϕ is constructed as the solution of the dilation equation

$\phi(x) = \sqrt{2}\sum_l h_l \phi(2x-l)$ for a given set of filter coefficients h_l that satisfy suitable conditions. The function ψ is obtained from ϕ as $\psi(x) = \sqrt{2}\sum_l g_l \phi(2x-l)$, with filter coefficients g_l often defined as $g_l = (-1)^l h_{1-l}$. The wavelet collection is constructed by translations and dilations as $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. Wavelet and scaling functions are particularly useful to approximate other functions. In general terms, scaling functions give a good approximation of smooth functions while wavelets are particularly useful to represent local fluctuations, such as discontinuities or cusps.

Orthonormal wavelet bases have been formally introduced by Mallat (1989) in the general context of a multiresolution analysis (MRA), i.e. as a decomposition of the space $L_2(\mathbb{R})$ into a sequence of linear closed subspaces $\{V_j, j \in \mathbb{Z}\}$. For any given j, the family of scaling functions $\{\phi_{j,k}(x), k \in \mathbb{Z}\}$ is an orthonormal basis in V_j, while the family of wavelets $\{\psi_{j,k}(x), j, k \in \mathbb{Z}\}$ forms an orthonormal basis in $L_2(\mathbb{R})$. Any function f in that space can therefore be represented by a wavelet series as

$$f(x) = \sum_{j,k\in\mathbb{Z}} w_{j,k}\psi_{j,k}(x) = \sum_{k\in\mathbb{Z}} c_{j_0,k}\phi_{j_0,k}(x) + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} w_{j,k}\psi_{j,k}(x) \qquad (3)$$

for any j_0, with wavelet coefficients defined as $w_{j,k} = \int f(x)\psi_{j,k}(x)dx$ and scaling coefficients as $c_{j,k} = \int f(x)\phi_{j,k}(x)dx$. The second equality in equation (3) results from the definition of the scaling functions $\phi_{j,k}$ and the properties of the MRA, see Mallat (1989) for more details. When truncated, the wavelet expansion (3) is an orthogonal projection of f into a V-subspace that can be expressed in terms of scaling functions only as

$$P_m f(x) = \sum_{k\in\mathbb{Z}} c_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{m-1}\sum_{k\in\mathbb{Z}} w_{j,k}\psi_{j,k}(x) = \sum_{k\in\mathbb{Z}} c_{m,k}\phi_{m,k}(x). \qquad (4)$$

The MRA construction implies that $\lim_{m\to\infty} P_m f(x) = f(x)$. In the examples we will employ Daubechies (1992) wavelets, extensively used in statistical applications because of their nice properties of orthogonality, compact support, different degrees of smoothness and maximum number of vanishing moments.

2.2 A brief review of wavelet series expansions for nonparametric function estimation

Nonparametric wavelet estimators have now been extensively used in the statistical literature, mainly for density and regression estimation. For density estimation, classical linear wavelet estimators use empirical coefficients defined as $\hat{c}_{j,k} = \frac{1}{n}\sum \phi_{j,k}(X_i)$ based on a random sample $X_1, \ldots, X_n$ from density f; see, for example, Walter (1992), Kerkyacharian and Picard (1993) and Vannucci and Vidakovic(1997). Thresholded wavelet density estimators, that apply thresholding or shrinkage techniques to the empirical coefficients, were proposed by Donoho *et al.* (1996) and Hall and Patil(1995), while Bayesian approaches, that impose mixture priors on the wavelet coefficients of the density expansion, were investigated by Müller and Vidakovic (1999).

As for regression models, the majority of the contributions in the literature have focused on the case of equally spaced data, i.e. by assuming the covariate values to be on a regular grid, following the seminal work of Donoho and Johnstone (1994). Several papers have been published since then, on modelling issues and extensions, using both classical and Bayesian methods. Rather than give a partial list of references, we refer readers to the recent paper of Antoniadis *et al.* (2001) that presents an exhaustive

review of wavelet methods for the equispaced design. Less work has been done for non-equispaced data, the setup we deal with in this paper. Classical approaches proposed so far mainly rely on reducing the design to the equispaced case, see the binning methods of Antoniadis *et al.* (1997), and the transformation methods of Cai and Brown (1998). As for Bayesian methods, when the design is non-equispaced inference cannot rely on settings that imply the *a posteriori* independence of the coefficients, unlike for the case of equispaced data. Mixture prior models can still be applied to the coefficients of the wavelet expansion but appropriate inferential procedures need to be developed, see Müller and Vidakovic (1999).

3 The model

3.1 Wavelet representation of single-index models

Let r be the regression function of equation (1) and let $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta}$. A nonparametric linear wavelet estimator of r can be constructed from the orthogonal projection of the function into a subspace of $L_2(\mathbb{R})$ as

$$\hat{r}(z) = \sum_{k \in \mathbb{Z}} \hat{c}_{m,k} \phi_{m,k}(z), \tag{5}$$

where $\hat{c}_{m,k}$ estimates $c_{m,k}$. Alternatively, a nonlinear *thresholded* wavelet estimator is defined as

$$\hat{r}(z) = \sum_{k \in \mathbb{Z}} \hat{c}_{j_0,k} \phi_{j_0,k}(z) + \sum_{j=j_0}^{m-1} \sum_{k \in \mathbb{Z}} s_{j,k} \hat{w}_{j,k} \psi_{j,k}(z), \tag{6}$$

with $j_0 \leq m-1$ and smoothing coefficients $s_{j,k}$ typically in $\{0, 1\}$, and where $\hat{c}_{j_0,k}$ and $\hat{w}_{j,k}$ estimate $c_{j_0,k}$ and $w_{j,k}$, respectively. A thresholded estimator

can be seen as a coarse approximation at scale j_0 plus a nonlinear part that adapts to local fluctuations, such as discontinuities or high-frequency oscillations. Notice that, due to (3) and (4), the linear estimator (5) is equivalent to the thresholded estimator (6) if no thresholding is done on the wavelet coefficients, i.e. if $s_{j,k} = 1$ for every j, k. Without loss of generality we will assume $j_0 = 0$ in the sequel. We refer readers to Hall and Patil (1995) and Nason (2002) for theoretical and practical investigations on the choice of the resolution level.

3.2 Likelihood function and prior distributions

Bayesian methods require prior distributions for all unknown parameters of the model. Let us define $m_0 = m - 1$. From (1), (2) and (4) the likelihood function is

$$P(Y|\sigma^2, \{c_{0,k}\}, \{w_{j,k}\}, \{s_{j,k}\}, \boldsymbol{\theta}, m_0, \mathbf{X}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} Q^2(Y_i)\right) \quad (7)$$

where

$$Q(Y_i) = Y_i - \sum_{k\in\mathbf{Z}} c_{0,k}\phi_{0,k}(\mathbf{X}_i T(\boldsymbol{\theta})) - \sum_{j=0}^{m_0}\sum_{k\in\mathbf{Z}} s_{j,k} w_{j,k}\psi_{j,k}(\mathbf{X}_i T(\boldsymbol{\theta})). \quad (8)$$

If we regard the truncation parameter as fixed, the unknown parameters of the model are $(\sigma^2, \{c_{0,k}\}, \{w_{j,k}\}, \{s_{j,k}\}, \boldsymbol{\theta})$.

Hierarchical mixture models, with a probability mass at zero for thresholding and with level-dependent variances, have been used extensively in the wavelet literature as prior distributions for wavelet coefficients, see Antoniadis *et al.* (2001) for references and Morris *et al.*(2003) for recent extensions to models with hierarchical functions. These priors take into account

the parsimony of the wavelet representation, for which many coefficients tend to be very small, particularly at finer levels, by implementing shrinkage rules that shrink small coefficients significantly stronger than larger ones. For our model we choose

$$\begin{aligned}
\sigma^2 &\sim IG(a_v, b_v), && (9)\\
c_{0,k}|\tau &\sim N(0, \tau), && (10)\\
w_{j,k}|\tau, s_{j,k} = 1 &\sim N(0, \tau 2^{-j}), && (11)\\
s_{j,k}|\alpha &\sim Bernoulli(\alpha^j), && (12)\\
\alpha &\sim Beta(a_\alpha, b_\alpha), && (13)\\
\tau &\sim IG(a_\tau, b_\tau) && (14)
\end{aligned}$$

with IG indicating the inverse gamma distribution. In addition, we define the prior for the direction $\boldsymbol{\theta}$ to be uniform, i.e.,

$$P(\boldsymbol{\theta}) = \frac{1}{2\pi}(\frac{1}{\pi})^{p-2}, \qquad (15)$$

for $0 < \theta_1 < 2\pi, -\frac{\pi}{2} < \theta_i < \frac{\pi}{2}, \quad i = 2, \ldots, p-1$, and 0 otherwise. Scaling factors of the type 2^{-j} in (11) and geometrically decreasing prior probabilities as in (12) take into account information on the rate of decay in the magnitude of the wavelet coefficients. See Müller and Vidakovic (1999) for a detailed explanation of the role of these parameters. Alternative exponential prior distributions for σ^2, allowing marginal distribution of the wavelet coefficients to be peaked, are investigated in Vidakovic and Ruggeri (2001).

The prior model implies a variable dimension of the parameter space, in that for $s_{j,k} = 0$ the corresponding wavelet coefficient is not included in the likelihood. Here we follow Müller and Vidakovic (1999) and define pseudo

priors for the case $w_{j,k}|s_{j,k} = 0$. Pseudo priors were first proposed for variable selection in regression by Carlin and Chib (1995). Alternatively, a reversible jump MCMC can be implemented, Green (1995). From (9)-(15) the prior model including the pseudo priors is

$$\begin{aligned} P(\sigma^2, \{c_{0,k}\}, \{s_{j,k}\}, \{w_{j,k}\}, \tau, \alpha, \boldsymbol{\theta}) &= P(\sigma^2|a_v, b_v) \cdot \prod_k P(c_{0,k}|\tau) \\ &\times \prod_{j=0}^{m_0} \prod_k P(w_{j,k}|\tau, s_{j,k} = 1) \\ &\times \prod_{j,k} h(w_{j,k}|s_{j,k} = 0) \cdot \prod_{j,k} P(s_{j,k}|\alpha) \\ &\times P(\alpha|a_\alpha, b_\alpha) \cdot P(\tau|a_\tau, b_\tau) \cdot P(\boldsymbol{\theta}) \quad (16) \end{aligned}$$

where we specify, for each j and k, the pseudo prior $h(w_{j,k}|s_{j,k} = 0)$ to be a Gaussian distribution with mean $\hat{w}_{j,k}$ and variance $\hat{\sigma}^2_{j,k}$, Müller and Vidakovic (1999).

4 Computational issues

4.1 MCMC procedure

For given $\mathbf{X}$ and m_0 the joint distribution of $\mathbf{Y}$ and the unknown parameters can be computed from the likelihood and the prior model. An MCMC scheme can then be implemented for inference on this joint distribution, see Gilks *et al.* (1996) for a collection of methods. Indeed, the full conditional distributions can be easily derived for all parameters of the model except for α and $\boldsymbol{\theta}$, see Appendix.

Let $\boldsymbol{\Omega} = \left\{\sigma^2, \{c_{0,k}\}, \{w_{j,k}\}, \{s_{j,k}\}, \tau, \alpha, \boldsymbol{\theta}\right\}$ and let $\boldsymbol{\Omega}(-\xi)$ denote the parameter vector without the parameter ξ, where ξ could be any one of the

parameters. Given initial values, at a generic step of the MCMC the parameters are updated according to the following scheme:

1. Generate σ^2 from the inverse gamma distribution

$$P\Big(\sigma^2|\mathbf{\Omega}(-\sigma^2),\mathbf{Y},\mathbf{X},m_0\Big)=IG\left[\frac{n}{2}+a_v,\left(\frac{1}{b_v}+\frac{1}{2}\sum_{i=1}^{n}Q(Y_i)\right)^{-1}\right]. \quad (17)$$

2. Generate τ from the inverse gamma distribution

$$P\Big(\tau|\mathbf{\Omega}(-\tau),\mathbf{Y},\mathbf{X},m_0\Big)=IG\left[\frac{S}{2}+a_\tau,\frac{1}{b_\tau}+\frac{1}{2}\sum_{k=a_0}^{b_0}c_{0,k}^2+\frac{1}{2}\underbrace{\sum_{j=0}^{m_0}\sum_{k=a_j}^{b_j}}_{j,k\in\{s_{j,k}=1\}}\frac{w_{j,k}^2}{2^{-j}}\right]$$

where $S=b_0-a_0+1+\sum_{j=0}^{m_0}\sum_{k=a_j}^{b_j}s_{j,k}$. Here $[a_0,b_0]$ and $[a_j,b_j]$ represent the range of the translation parameter k in $\phi_{0,k}(\cdot)$ and $\psi_{j,k}(\cdot)$, respectively, see Section 4.3 for details.

3. Generate the scaling coefficients $c_{0,k}$ from the Gaussian distributions

$$P\Big(c_{0,k}|\mathbf{\Omega}(-c_{0,k}),\mathbf{Y},\mathbf{X},m_0\Big)=N(\mu_k,\sigma_k^2), \quad (18)$$

where μ_k and σ_k^2 are given in the Appendix.

4. For those $s_{j,k}=1$ generate $w_{j,k}$ from the Gaussian distributions

$$P\Big(w_{j,k}|\mathbf{\Omega}(-w_{j,k}),\mathbf{Y},\mathbf{X},m_0\Big)=N(\mu_{j,k},\sigma_{j,k}^2), \quad (19)$$

where $\mu_{j,k}$ and $\sigma_{j,k}^2$ are given in the Appendix. For those $s_{j,k}=0$ generate $w_{j,k}$ from the pseudo-prior $h(w_{j,k})$.

5. Generate $s_{j,k}$ from Bernoulli distributions with probabilities $Pr(s_{j,k} = 0) = pr_0/(pr_0 + pr_1)$ and $Pr(s_{j,k} = 1) = pr_1/(pr_0 + pr_1)$ where

$$\begin{cases} Pr(s_{j,k} = 0) \propto pr_0 = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} Q(Y_i)\right)(1-\alpha^j)h(w_{j,k}), \\ \\ Pr(s_{j,k} = 1) \propto pr_1 = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} Q(Y_i)\right)(\alpha^j)P(w_{j,k}|s_{j,k} = 1, \tau). \end{cases} \tag{20}$$

6. Update α by a Metropolis algorithm. We used a Gaussian proposal distribution centered at the previous α value and with constant standard deviation .1. The new value is accepted with probability

$$A(\alpha_{old}, \alpha_{new}) = \min\left[1, \frac{\alpha_{new}^{a_\alpha}(1-\alpha_{new})^{b_\alpha}}{\alpha_{old}^{a_\alpha}(1-\alpha_{old})^{b_\alpha}} \prod_{j,k}\left(\frac{\alpha_{new}^j}{\alpha_{old}^j}\right)^{s_{j,k}}\left(\frac{1-\alpha_{new}^j}{1-\alpha_{old}^j}\right)^{1-s_{j,k}}\right]$$

7. Update $\boldsymbol{\theta}$. A new value is sampled from a proposal distribution $q(\cdot|\cdot)$ and accepted with probability

$$A\Big(\boldsymbol{\theta}_{old}, \boldsymbol{\theta}_{new}\Big) = \min\left[1, \frac{P\Big(\boldsymbol{\theta}_{new}|\Omega_1(-\boldsymbol{\theta})\Big)\cdot q\Big(\boldsymbol{\theta}_{old}|\boldsymbol{\theta}_{new}\Big)}{P\Big(\boldsymbol{\theta}_{old}|\Omega_1(-\boldsymbol{\theta})\Big)\cdot q\Big(\boldsymbol{\theta}_{new}|\boldsymbol{\theta}_{old}\Big)}\right]. \tag{21}$$

We investigated an independence-type Metropolis-Hastings (M-H) sampler, described in the next section.

4.2 Independence-type sampler for the direction parameter

In order to implement Step 7 above, for each direction component θ we selected a Gaussian proposal distribution with constant variance s^2 and centered at the mode of the target posterior distribution $P(\theta|\boldsymbol{\Omega}(-\theta), \mathbf{Y}, \mathbf{X}, m_0)$.

We used the following fast bisection method to search for the mode, see also Figure 1. Let θ_{old} be the current value of the direction parameter and let $P(\theta)$ indicate a function of θ that is proportional to the target full conditional distribution. The slope of the line connecting two points, $(\theta_0, P(\theta_0))$ and $(\theta_1, P(\theta_1))$ is $\gamma = \frac{P(\theta_1)-P(\theta_0)}{\theta_1-\theta_0}$.

Step 1: To see where the current θ_{old} is located relative to the mode, compare the two slopes obtained from the 3 points $(\theta_{old} - s, P(\theta_{old} - s))$, $(\theta_{old}, P(\theta_{old}))$, and $(\theta_{old}+s, P(\theta_{old}+s))$. If both slopes are negative, and the target distribution is unimodal, then the current direction value is to the right of the mode. If they are positive, the value is to the left of the mode.

Step 2: If both slopes are negative (positive), shift θ_{old} to the left (right) by defining $\theta_{new} = \theta_{old} - s$ ($\theta_{new} = \theta_{old} + s$). Repeat Step 1 and shift until the signs of the two slopes are different.

Step 3: Consider now the three points $(L, P(L))$, $(M, P(M))$ and $(R, P(R))$, where $L = M - s$, $R = M + s$ and where the signs of the two slopes are different, see Figure 1.

Step 4: If $P(L)$ is greater (smaller) than $P(R)$, then the middle point M is on the right (left) side of the mode. Calculate the slope γ of the line corresponding to the two points that are on the same side of the mode. Construct a line passing through the point on the opposite side of the mode and having slope $-\gamma$.

Step 5: Define the candidate mode to be the abscissa at which the two lines

intersect.

4.3 Range of the translation parameter k in $\phi_{j,k}, \psi_{j,k}$

We consider minumum phase Daubechies wavelets, see Daubechies (1992), which have compact support and maximum number of vanishing moments. Compact support ensures finite summations over the translation parameter k of the wavelet expansions. Let N indicate the number of vanishing moments of the wavelets. Then the support of $\phi(x)$ is $[0, 2N-1]$ and the support of $\psi(x)$ is $[-N, N-1]$. The supports of $\phi_{j,k}(x)$ and $\psi_{j,k}(x)$ are $[k2^{-j}, (2N-1+k)2^{-j}]$ and $[(1-N+k)2^{-j}, (N+k)2^{-j}]$, respectively. Thus, given a function with support in the interval $[a, b]$, one has to calculate only the coefficients for those values of k for which the supports of $\phi_{j,k}(x)$ and $\psi_{j,k}(x)$ intersect $[a, b]$. Simple calculations give the range of k as

$$\left[\lceil a2^j \rceil - 2N + 1, \lfloor b2^j \rfloor\right], \quad \left[\lceil a2^j \rceil - N, \lfloor b2^j \rfloor + N - 1\right] \tag{22}$$

for scaling and wavelet coefficients, respectively, where $\lfloor x \rfloor = \max\{n \in \mathbf{Z}; n \leq x\}$, and $\lceil x \rceil = \min\{n \in \mathbf{Z}; n \geq x\}$. In our examples we will use $[a, b]$ as the sample range of $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta}$, for given $\boldsymbol{\beta}$.

Daubechies wavelets cannot be computed in analytic form, except for the case of Haar wavelets. A cascade algorithm can be used, see Daubechies (1992) for a description, that computes the values of the wavelet and scaling functions in dyadic points. MCMC inferential procedures like the one we implemented here can be made considerably faster if, before running the MCMC, the values of the scaling and wavelet functions in a fine grid of dyadic points are computed and stored in a table. In our implementation, since $\boldsymbol{\beta}$

is unknown, we chose a grid that covers the maximum range of $\mathbf{X}T(\boldsymbol{\theta})$. It is well known that $|\mathbf{X}T(\boldsymbol{\theta})| \leq (\mathbf{X}^T\mathbf{X})^{1/2}$, Rao(1973), page 60, and therefore, regardless of the choice of $T(\boldsymbol{\theta})$, we have $|X_i^T T(\boldsymbol{\theta})| \leq \max_i (X_i^T X_i)^{1/2}$. During the MCMC procedure, values of scaling and wavelet functions at arbitrary points can be computed by interpolation or simply by considering the value at the closest point in the grid.

5 Applications

5.1 Hyperparameter setting and initialization

It is important to select a good initial value for the direction parameter $\boldsymbol{\theta}$ because this value affects the initialization of the other parameters. We found an initial estimate of $\boldsymbol{\theta}$ by implementing a preliminary step with parallel chains starting from several different values and choosing the posterior mean estimate that minimizes the residual sum of squares $\sum_i (Y_i - \hat{r}(X_i\boldsymbol{\beta}))^2$. During this preliminary step we kept all $s_{j,k} = 1$, without updating them. This was done also during the burnin period of the MCMC's.

Motivated by the orthogonality of the wavelet functions we propose initial values of the scaling and wavelet coefficients based on "quadrature" type estimates, see for example Hart (1997),

$$\begin{aligned} \hat{c}_{0,k} &= \sum_{i=1}^{n} Y_i (s_i - s_{i-1}) \phi_{0,k}\left(\frac{s_i + s_{i-1}}{2}\right) \\ &\approx \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} \phi_{0,k}(z) dz, \end{aligned} \tag{23}$$

and

$$\begin{aligned}\hat{w}_{j,k} &= \sum_{i=1}^{n} Y_i(s_i - s_{i-1})\psi_{j,k}\left(\frac{s_i + s_{i-1}}{2}\right) \\ &\approx \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} \psi_{j,k}(z)dz, \end{aligned} \tag{24}$$

where $z_{(i)}$ is the ith smallest of the ordered $\mathbf{X}\boldsymbol{\beta}$, $s_0 = z_{(1)}$, $s_i = \frac{z_{(i)}+z_{(i+1)}}{2}$, $i = 1, \ldots, n-1$, $s_n = z_{(n)}$, and $\frac{s_i+s_{i-1}}{2} = \frac{z_{(i-1)}+2z_{(i)}+z_{(i+1)}}{2}$ for $z_{(1)} < \cdots < z_{(n)}$.

We chose a non-informative prior on α by setting $a_\alpha = b_\alpha = 1$. Pseudo priors $h(w_{j,k}|s_{j,k} = 0)$ were specified as Gaussian distributions with mean $\hat{w}_{j,k}$ and variance $\hat{\sigma}^2_{j,k}$ estimated as ergodic mean and variance of $w_{j,k}$, after burnin, based on a preliminary MCMC run where we kept all $s_{j,k} = 1$, without updating them. We set $a_\nu = a_\tau = 1/2$ and $b_\nu = b_\tau = 1$ to obtain vague priors on σ^2 and τ. Initial values for τ and α were sampled from the corresponding prior distributions. A Gaussian proposal distribution was used to update the direction parameter $\boldsymbol{\theta}$. For the preliminary step we determined an initial choice of the variance of the proposal distribution as a value proportional to $(1/(n-1))\sum_i (Y_i - \hat{r}_i)^2$ with $\hat{r}$ an estimate based on the initial direction. Once a suitable initial value of the direction was determined, as described above, the variance of the proposal was updated to give an acceptance ratio of around 70% to ensure good mixing.

5.2 Simulation studies

In order to assess performances of our method we performed simulation studies with two different functions, a simple cosine function and the *Doppler* function. We focused on $p = 2$. For both examples we used three different

values of the direction parameter θ and three values of the error variance σ^2. We used $n = 200$ and computed bias and mean squared errors of the estimates based on 20 replications. We ran the preliminary MCMC steps with 1,000 iterations and the MCMC's with 1,000 burnin followed by 10,000 iterations. We also compared performances of the proposed independence-type Metropolis-Hastings (M-H) sampler with the more standard Metropolis algorithm that uses a Gaussian proposal distribution centered at the previous θ value and with variance s^2. Results did not appear to be dramatically affected by the regularity of the wavelet family we picked. We present here results obtained using Daubechies wavelets with four vanishing moments.

Cosine function: As a first example we used the simple cosine function, i.e., $r(z) = \cos(z)$. We simulated $(Z_i, Y_i), i = 1, \ldots, n$, as independent and identically distributed observations from $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a bivariate vector with $X_j \overset{iid}{\sim} N(0, \sigma_c^2),\ j = 1, 2$, with $\sigma_c = 1.5$, and where $Y_i = r(Z_i) + \epsilon_i$, $Z_i = X_{i1}t_1(\theta) + X_{i2}t_2(\theta)$ and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Given the smoothness of the cosine function we computed the wavelet estimator in the form (5), i.e. using only scaling coefficients, and did not perform shrinkage on the coefficients. We monitored the chains for convergence. Figures 2 and 3 show the MCMC traces for the direction parameter and the scaling coefficients, respectively, for one of the simulated datasets. Estimates of all parameters were computed as posterior means of the MCMC samples and corresponding estimates of the function r were obtained by computing the truncated wavelet expansion. Figure 4 shows the estimated function, for both Metropolis and the independence-type M-H sampler, together with the true function, for one of the datasets we simulated. A scatter plot of the estimated and the true

design points is also given. All estimation results are summarized in Table 1, which reports bias and MSE's of the estimates of θ and σ and the integrated mean squared error and bias of the estimates of the function r. The wavelet procedure appears to do a very good job at estimating both the function and the other parameters of the model, for all directions and noise levels considered in the study. Although there is no apparent difference in the estimates of the parameters, the independence-type method led to greater acceptance ratios and had faster convergence.

Doppler function: We also looked at a much less regular function, the Doppler function, i.e., $r(z) = 2\sqrt{z(1-z)}\sin\left[\frac{2.1\pi}{z+0.05}\right]$, for $0 \leq z \leq 1$. We simulated the two covariates independently from uniform distributions in $[0, 0.45]$. Here we used the wavelet estimator in the form (6) and performed shrinkage on the wavelet coefficients. Table 2 reports bias and MSE's of the estimates of θ and σ and the integrated mean squared error and bias of the estimates of the function r. Results are given for the case $\theta = .35$ and for three different noise levels, for both truncation levels $m_0 = 5$ and $m_0 = 6$. We notice that there appears to be a trade-off between the estimates of the direction parameter θ and the estimates of the function r. For given θ and σ, a larger m_0 value leads to a better estimation of the direction parameter but also to a worse integrated MSE for the link function. Indeed, boxplots of the estimated posterior means for the 20 datasets we simulated, given in Figure 5, show that we underestimate the direction parameter for $m_0 = 5$. On the other hand, the estimated functions, for both Metropolis and the independence-type M-H sampler, are reasonably good for $m_0 = 5$ but worsen for $m_0 = 6$, see Figure 6. Improvements could be

obtained by modeling the uncertainty about the truncation parameter via a prior distribution. We consider the results here presented as satisfactory, considering that Doppler represents a challenging estimation problem. The independence-type M-H method led again to greater acceptance ratios and faster convergence. Moreover, estimates of the direction parameter obtained with the independence-type M-H sampler appear to have smaller standard deviations (see Figure 5).

Comparison with existing methods: We also looked at comparing our results with other existing methods. In particular, we used the Bayes-splines method of Antoniadis *et al.* (2004) and a more traditional kernel-type method, briefly described here. Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be the observations. Given a direction $\boldsymbol{\beta}$ satisfying the constraint $\sum_{i=1}^{p} \beta_i^2 = 1$, let $Z_1(\boldsymbol{\beta}) < \cdots < Z_n(\boldsymbol{\beta})$ be the ordered values of $\mathbf{X}_i^T \boldsymbol{\beta}$, $i = 1, \ldots, n$. Let $Y_1(\boldsymbol{\beta}), \ldots, Y_n(\boldsymbol{\beta})$ be the concomitant response values. For a given $\boldsymbol{\beta}$, consider a Gaussian kernel local linear smooth based on the data $(u_1, Y_1(\boldsymbol{\beta})), \ldots, (u_n, Y_n(\boldsymbol{\beta}))$, where $u_i = (i - 1/2)/n$, $i = 1, \ldots, n$. We used one-sided cross-validation (OSCV) to choose the bandwidth h of the local linear smooth, Hart and Yi (1998). The transformation to evenly spaced values on $(0, 1)$ was used since it tends to stabilize the bandwidth selection process. Let $\hat{Y}_1(\boldsymbol{\beta}), \ldots, \hat{Y}_n(\boldsymbol{\beta})$ be the predicted values obtained from the OSCV smooth, and define

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i(\boldsymbol{\beta}) - \hat{Y}_i(\boldsymbol{\beta}))^2. \tag{25}$$

In principle, we then choose $\boldsymbol{\beta}$ to minimize $RSS(\boldsymbol{\beta})$. The procedure we used to find the approximate minimizer is a stochastic one, akin to MCMC. We generated directions $\boldsymbol{\beta}$ from a distribution that is uniform over the unit

3-dimensional sphere. For each $\boldsymbol{\beta}$ so generated we performed OSCV as described above and then computed $RSS(\boldsymbol{\beta})$ for the selected smooth. Results here reported were obtained by generating 10,000 values.

In order to be consistent with the simulation study of Antoniadis *et al.* (2004) we summarize results in terms of the two error criteria

$$\text{angle}(\hat{\beta}, \beta) = \cos^{-1}(\hat{\beta}'\beta), \quad \sup_{1\leq j\leq d}|\hat{\beta}'_j\beta_j|. \tag{26}$$

Table 3 reports the results for our Independence-type M-H, the kernel-type smoother and the splines-based Bayesian method of Antoniadis *et al.* (2004), for cosine and Doppler (case $m_0 = 6$ only) functions. For the splines-based Bayesian method we specified the concentration parameters of the Fisher-von Mises prior on $\boldsymbol{\beta}$ and of the proposal distribution of the Metropolis as 100, we used a vague prior on σ^2 by chosing $A = 0.0001$ and $B = 1,000$ and B-splines with 20 knots and 2 degrees of freedom. See Antoniadis *et al.* (2004) for more details on these hyperparameters and their specifications. Performance of our method appears to be comparable with those of the kernel-type smoother in terms of the angle and sup-norm error criteria, and slightly better in terms of integrated MSE for the link function. Our method outperforms the splines-based Bayesian approach in the estimation of both the direction parameter and the link function. We notice that the orders of magnitude of the angle and sup-norm errors we obtained for the splines-based method are consistent with those reported by Antoniadis *et al.* (2004) in their simulation study.

5.3 Air pollution data

We conclude the paper by presenting an application to real data from an environmental study. We use a benchmark dataset on the relationship between the concentration of the air pollutant ozone (Y) and three meteorological variables, solar radiation (x_1), windspeed (x_2), and temperature (x_3). Measurements of daily ozone concentration are taken in parts per billion (ppb), solar radiations in Langleys (langleys), wind speeds in miles/hour (mph), and daily maximum temperatures in degrees Fahrenheit (F). There are 111 days of observations, from May to September 1973, taken in New York. The dataset was recently analysed by Yu and Ruppert (2002) who compared five different models: a linear model, a single-index model using a 10-knot cubic P-spline, an additive model, a partially linear single-index model, with radiation in the linear term, fitted using P-splines, a partially linear additive model, and a fully nonparametric model using LOESS. Their findings show that single-index models and additive models perform much better than the linear model.

We applied our wavelet-based Bayesian estimation procedure to the air pollution data. We used the smooth part of the wavelet expansion. Adding the detail part did not lead to a substantial improvement in the estimates. Table 4 summarizes the results on the estimates of the index parameters. Boxplots of the sampled values, after burnin, are also given. Results agree well with the findings of Yu and Ruppert (2002). Figure 7 shows the estimated curve together with the 95% posterior confidence interval, confirming the existence of curvature in the data. Figure 8 gives some residuals diagnostics. The error terms exhibit no obvious pattern and their distribution

appears to be approximately normal.

We also compared our results with those from the more traditional kernel approach previously used in the simulation study. Ten thousand i.i.d. variates were generated, resulting in 10,000 values of $RSS(\boldsymbol{\beta})$. These are shown in Figure 9. The direction that minimized $RSS(\boldsymbol{\beta})$ and the corresponding residual sum of squares were $\hat{\boldsymbol{\beta}}^T = (0.027, -0.829, 0.559)$ and $RSS(\hat{\boldsymbol{\beta}}) = 22.727$. The estimated direction is in close agreement with that obtained in our Bayesian wavelet analysis. As a point of reference, the residual sum of squares for the fitted linear model $Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$ is 27.848. So, indeed the single index model explains more of the variation in the data than does the simple linear model. Estimates of the link function are shown in Figure 10. The local linear bandwidth chosen by OSCV for the data $(u_1, \hat{Y}_1(\hat{\boldsymbol{\beta}})), \ldots, (u_n, \hat{Y}_n(\hat{\boldsymbol{\beta}}))$ was 0.168, using a standard normal kernel. The smooth on the left in Figure 10 was obtained by simply plotting the OSCV local linear estimate against $Z_1(\hat{\boldsymbol{\beta}}), \ldots, Z_n(\hat{\boldsymbol{\beta}})$, as opposed to $u_1, \ldots, u_n$. This estimate inherits the lack of smoothness of the back transformation from u_i's to quantiles of the single index. If a smoother estimate is desired, this is easily obtained by smoothing the smooth slightly, which was done to produce the estimate on the right in Figure 10. As was true for the direction, these estimates agree well with the Bayesian link estimate.

6 Discussion

We have investigated a wavelet-based Bayesian approach to modeling and estimation for single-index models. The developed methodology allows an

unequally spaced design and employs pseudo-priors for the wavelet coefficients. We have used MCMC techniques for posterior inference, producing simultaneous estimates of the index parameters and of the link function. We have proposed an independence-type M-H algorithm to sample from the distribution of the direction parameter. We have used fairly automatic or vague specifications of the prior model and have investigated performances of the methods on simulated and real data. Our method has compared favorably with respect to more traditional kernel and spline-based Bayesian approaches.

Mixture prior models that use a Dirac measure at zero can be used as an alternative modelling to what we have done here. In addition, results on wavelet smoothing have shown that when shrinkage is applied the choice of the truncation parameter is influential, inducing a trade-off between the estimate of the direction parameter and the estimate of the link function. Improvements may be obtained by treating the truncation parameter as unknown and imposing a prior distribution on it. This is left to future research.

Acknowledgments

Vannucci's research is supported by National Science Foundation, CAREER award number DMS-0093208.

Appendix: Calculations of Posterior distributions

Let $\mathbf{\Omega} = \Big\{\sigma^2, \{c_{0,k}\}, \{w_{j,k}\}, \{s_{j,k}\}, \tau, \alpha, \boldsymbol{\theta}\Big\}$. The joint distribution of Y and the parameters of the model, conditional on $\mathbf{X}$ and on the resolution m_0, is obtained by multiplying likelihood and priors as

$$
\begin{aligned}
P(\mathbf{Y}, \Omega | \mathbf{X}, m_0) &= P(\mathbf{Y} | \{c_{0,k}\}, \{s_{j,k}\}, \{w_{j,k}\}, \sigma^2, \boldsymbol{\theta}, \mathbf{X}, m_0) \cdot P(\Omega) \\
&\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} Q^2(Y_i)\right) \cdot (\sigma^2)^{-(a_v+1)} \exp\left(-\frac{1}{\sigma^2 b_v}\right) \\
&\quad \times \tau^{-(a_\tau+1)} \exp\left(-\frac{1}{\tau b_\tau}\right) \cdot \left[\prod_{k\in\mathbb{Z}} \tau^{-\frac{1}{2}} \exp\left(-\frac{c_{0,k}^2}{2\tau}\right)\right] \\
&\quad \times \left[\underbrace{\prod_{j=0}^{m_0}\prod_{k=a_j}^{b_j}}_{j,k\in\{s_{j,k}=1\}} \tau^{-\frac{1}{2}} \exp\left(-\frac{w_{j,k}^2}{2\cdot 2^{-j}\tau}\right)\right] \cdot \left[\underbrace{\prod_{j=1}^{m_0}\prod_{k=a_j}^{b_j}}_{j,k\in\{s_{j,k}=0\}} h(w_{j,k})\right] \\
&\quad \times \left[\prod_{j=1}^{m_0}\prod_{k=a_j}^{b_j} (\alpha^j)^{s_{j,k}}(1-\alpha^j)^{1-s_{j,k}}\right] \cdot \alpha^{(a_\alpha-1)}(1-\alpha)^{(b_\alpha-1)} \\
&\propto (\sigma^2)^{-(\frac{n}{2}+a_v+1)} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} Q^2(Y_i) - \frac{1}{\sigma^2 b_v}\Big)\tau^{-(\frac{S}{2}+a_\tau+1)} \\
&\quad \times \exp\Big(-\frac{1}{\tau b_\tau} - \sum_{k\in\mathbb{Z}}\frac{c_{0,k}^2}{2\tau} - \underbrace{\sum_{j=0}^{m_0}\sum_{k=a_j}^{b_j}}_{j,k\in\{s_{j,k}=1\}} \frac{w_{j,k}^2}{2\cdot 2^{-j}\tau}\Big) \\
&\quad \times \left[\prod_{j=1}^{m_0}\prod_{k=a_j}^{b_j} (\alpha^j)^{s_{j,k}}(1-\alpha^j)^{1-s_{j,k}}\right] \cdot \alpha^{(a_\alpha-1)}(1-\alpha)^{(b_\alpha-1)}
\end{aligned}
$$

$$\times \left[\underbrace{\prod_{j=1}^{m_0} \prod_{k=a_j}^{b_j}}_{j,k \in \{s_{j,k}=0\}} h(w_{j,k}) \right],$$

where $Q(y_i) = Y_i - \sum_{k=a_0}^{b_0} c_{0,k}\phi_{0,k}(\mathbf{X}_i T(\boldsymbol{\theta})) - \sum_{j=0}^{m_0} \sum_{k=a_j}^{b_j} s_{j,k} w_{j,k} \psi_{j,k}(\mathbf{X}_i T(\boldsymbol{\theta}))$ and $S = b_0 - a_0 + 1 + \sum_{j=0}^{m_0} \sum_{k=a_j}^{b_j} s_{j,k}$.

- The conditional distribution of σ^2 is

$$\begin{aligned} P(\sigma^2 | \Omega(-\sigma^2), \mathbf{Y}, \mathbf{X}, m_0) &\propto (\sigma^2)^{-\nu_1} \exp\left[-\frac{\nu_2}{\sigma^2}\right] \\ &\propto IG(\nu_1 - 1, 1/\nu_2), \end{aligned}$$

where $\nu_1 = \frac{n}{2} + a_v + 1$ and $\nu_2 = \frac{1}{b_v} + \frac{1}{2}\sum_{i=1}^{n} Q(Y_i)$.

- The conditional distribution of τ is

$$\begin{aligned} P(\tau | \Omega(-\tau), \mathbf{Y}, \mathbf{X}, m_0) &\propto \tau^{-\nu_3} \exp\left[-\frac{\nu_4}{\tau}\right] \\ &\propto IG(\nu_3 - 1,\ 1/\nu_4), \end{aligned}$$

where $\nu_3 = \frac{S}{2} + a_\tau + 1$ and $\nu_4 = \frac{1}{b_\tau} + \frac{1}{2}\sum_{k=a_0}^{b_0} c_{0,k}^2 + \frac{1}{2}\sum_{j=0}^{m_0} \sum_{k=a_j}^{b_j} \frac{s_{j,k} w_{j,k}^2}{2^{-j}}$.

- The conditional distribution of $c_{0,k}$ is

$$P(c_{0,k} | \Omega(-c_{0,k}), \mathbf{Y}, \mathbf{X}, m_0) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} Q(Y_i) - \frac{1}{2} \sum_{k'=a_0}^{b_0} \frac{c_{0,k'}^2}{\tau} \right)$$

where

$$\begin{aligned} & -\frac{1}{2\sigma^2} \sum_{i=1}^{n} Q(Y_i) - \frac{1}{2} \sum_{k'=a_0}^{b_0} \frac{c_{0,k'}^2}{\tau} \\ = \ & -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(E_{-(k)}^i - c_{0,k}\phi_{0,k}(\mathbf{X}_i T(\boldsymbol{\theta})) \right)^2 - \frac{1}{2\tau} c_{0,k}^2 - C \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\left[\frac{1}{\sigma^2}\left(c_{0,k}^2\sum_{i=1}^{n}\phi_{0,k}^2(\mathbf{X}_iT(\boldsymbol{\theta})) - 2c_{0,k}\sum_{i=1}^{n}\phi_{0,k}(\mathbf{X}_iT(\boldsymbol{\theta}))E^i_{-(k)}\right) + \frac{1}{\tau}c_{0,k}^2\right] - C \\
&= -\frac{1}{2}\left[c_{0,k}^2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\phi_{0,k}^2(\mathbf{X}_iT(\boldsymbol{\theta})) + \frac{1}{\tau}\right) - \frac{2}{\sigma^2}c_{0,k}\sum_{i=1}^{n}\phi_{0,k}(\mathbf{X}_iT(\boldsymbol{\theta}))E^i_{-(k)}\right] - C \\
&= -\frac{1}{2\sigma_k^2}(c_{0,k}-\mu_k)^2 - C,
\end{aligned}
$$

with C constant, $\mu_k = \frac{\sigma_k^2}{\sigma^2}\sum_{i=1}^{n}\phi_{0,k}(\mathbf{X}_iT(\boldsymbol{\theta}))\cdot E^i_{-(k)}$ and $\frac{1}{\sigma_k^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\phi_{0,k}^2(\mathbf{X}_iT(\boldsymbol{\theta})) + \frac{1}{\tau}$, and where

$$
E^i_{-(k)} = Y_i - \sum_{\substack{k'=a_0\\k'\neq k}}^{b_0} c_{0,k'}\phi_{0,k'}(\mathbf{X}_iT(\boldsymbol{\theta})) - \sum_{j=0}^{m_0}\sum_{k'=a_j}^{b_j} s_{j,k'}w_{j,k'}\psi_{j,k'}(\mathbf{X}_iT(\boldsymbol{\theta})) \text{ for } k\in[a_0,b_0].
$$

Therefore

$$
P(c_{0,k}|\Omega(-c_{0,k}),\mathbf{Y},\mathbf{X},m_0) \propto N(\mu_k,\sigma_k^2). \tag{27}
$$

- The conditional distribution of $w_{j,k}|s_{j,k}=1$ is

$$
P(w_{j,k}|\Omega(-w_{j,k}),s_{j,k}=1,\mathbf{Y},\mathbf{X},m_0) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}Q(Y_i) - \frac{1}{2}\sum_{j'=0}^{m_0}\sum_{k'=a_{j'}}^{b_{j'}}\frac{w_{j',k'}^2}{2^{-j'}\tau}\right)
$$

where

$$
\begin{aligned}
&-\frac{1}{2\sigma^2}\sum_{i=1}^{n}Q(Y_i) - \frac{1}{2}\sum_{j'=0}^{m_0}\sum_{k'=a_{j'}}^{b_{j'}}\frac{w_{j',k'}^2}{2^{-j'}\tau} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(E^i_{-(j,k)} - w_{j,k}\psi_{j,k}(\mathbf{X}_iT(\boldsymbol{\theta}))\right)^2 - \frac{1}{2\cdot 2^{-j}\tau}w_{j,k}^2 - C \\
&= -\frac{1}{2}\left[\frac{1}{\sigma^2}\left(w_{j,k}^2\sum_{i=1}^{n}\psi_{j,k}^2(\mathbf{X}_iT(\boldsymbol{\theta})) - 2w_{j,k}\sum_{i=1}^{n}\psi_{j,k}(\mathbf{X}_iT(\boldsymbol{\theta}))E^i_{-(j,k)}\right) + \frac{1}{2^{-j}\tau}w_{j,k}^2\right] - C
\end{aligned}
$$

$$= -\frac{1}{2}\left[w_{j,k}^2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\psi_{j,k}^2(\mathbf{X}_iT(\boldsymbol{\theta}))+\frac{1}{2^{-j}\tau}\right)-\frac{2}{\sigma^2}w_{j,k}\sum_{i=1}^{n}\psi_{j,k}(\mathbf{X}_iT(\boldsymbol{\theta}))E^i_{-(j,k)}\right]-C$$

$$= -\frac{1}{2\sigma_{j,k}^2}(w_{j,k}-\mu_{j,k})^2-C,$$

with C constant, $\mu_{j,k}=\frac{\sigma_{j,k}^2}{\sigma^2}\sum_{i=1}^{n}\psi_{j,k}(\mathbf{X}_iT(\boldsymbol{\theta}))\cdot E^i_{-(j,k)}$ and $\frac{1}{\sigma_{j,k}^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}\psi_{j,k}^2(\mathbf{X}_iT(\boldsymbol{\theta}))+\frac{1}{2^{-j_2}\tau}$, and where

$$E^i_{-(j,k)}=Y_i-\sum_{k'=a_0}^{b_0}c_{0,k'}\phi_{0,k'}(\mathbf{X}_iT(\boldsymbol{\theta}))-\sum_{\substack{j'=0\\j'\neq j}}^{m_0}\sum_{\substack{k'=a_{j'}\\k'\neq k}}^{b_{j'}}s_{j',k'}w_{j',k'}\psi_{j',k'}(\mathbf{X}_iT(\boldsymbol{\theta})),$$

for $j\in[0,m_0]$ and $k\in[a_j,b_j]$. Therefore

$$P(w_{j,k}|\Omega(-w_{j,k}),\mathbf{Y},\mathbf{X},m_0)\propto N(\mu_{j,k},\sigma_{j,k}^2). \tag{28}$$

• The conditional probability of $s_{j,k}$ is Bernoulli with

$$P(s_{j,k}|\Omega(-s_{j,k}),\mathbf{Y},\mathbf{X},m_0)$$

$$\propto\begin{cases}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}Q(Y_i)\right)(1-\alpha^j)h(w_{j,k}), & s_{j,k}=0,\\ \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}Q(Y_i)\right)(\alpha^j)\exp\left(-\frac{w_{j,k}^2}{2\cdot 2^{-j}\tau}\right), & s_{j,k}=1,\end{cases} \tag{29}$$

for $j\in[1,m_0]$ and $k\in[a_j,b_j]$.

• The conditional probabilities of α and $\boldsymbol{\theta}$ cannot be determined in closed form

$$P(\alpha|\Omega(-\alpha),\mathbf{Y},\mathbf{X},m_0)\ \propto\ \left[\prod_{j=1}^{m_0}\prod_{k=a_j}^{b_j}(\alpha^j)^{s_{j,k}}(1-\alpha^j)^{1-s_{j,k}}\right]\alpha^{(a_\alpha-1)}(1-\alpha)^{(b_\alpha-1)}$$

$$P(\boldsymbol{\theta}|\Omega(-\boldsymbol{\theta}),\mathbf{Y},\mathbf{X},m_0)\ \propto\ \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}Q(Y_i)\right).$$

References

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. *Journal of Statistical Software*, **6,** 1–83.

Antoniadis, A., Gregoire, G. and McKeague, I.W. (2004). Bayesian estimation in single-index models. *Statistica Sinica* to appear.

Antoniadis, A., Gregoire, G. and Vial, P. (1997). Random design wavelet curve smoothing. *Statistics and Probability Letters*, **35,** 235–232.

Cai, T. and Brown, L.D. (1998). Wavelet shrinkage for nonequispaced samples. *Annals of Statistics*, **26(5),** 1783–1799.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo. *Journal of the Royal Statistical Society,* B, **57,** 473–484.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92(438),** 477–489.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, volume 61. SIAM, CBMS-NSF Conference Series.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaption via wavelet shrinkage. *Biometrika*, **81,** 425–455.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics*, **24(2),** 508–539.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in Practice*. Chapman and Hall. London.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82,** 711–732.

Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Annals of Statistics*, **23(3),** 905–928.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21(1),** 157–178.

Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, **84,** 986–995.

Hart, J.D. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer Verlag.

Hart, J.D. and Yi, S. (1998). One-sided cross-validation. *Journal of the American Statistical Association*, **93,** 620–631.

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

Kerkyacharian, G. and Picard, D. (1993). Density estimation by kernel and wavelet methods: Optimality of Besov spaces. *Statistics & Probability Letters*, **18,** 327–336.

Li, K.C. and Duan, N. (1989). Regression analysis under link violation. *Annals of Statistics*, **17,** 1009–1052.

Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11(7),** 674–693.

Morris, J.S., Vannucci, M., Brown, P.J. and Carroll, R.J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association*, **98**.

Müller, P. and Vidakovic, B. (1999). MCMC methods in wavelet shrinkage: Non-equally spaced regression, density and spectral density estimation. In *Bayesian Inference in Wavelet based Models* (eds B. Vidakovic and P. Müller), pp. 1–18. Springer Verlag.

Nason, G.P. (2002). Choice of wavelet smoothness, primary resolution and threshold in wavelet shrinkage. *Statistics and Computing*, **12,** 219–227.

Rao, C.R. (1973). *Linear statistical inference and its applications*. John Wiley and Sons. New York (2nd edition).

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

Vannucci, M. and Vidakovic, B. (1997). Preventing the dirac disaster: wavelet based density estimation. *Journal of the Italian Statistical Society*, **6(2),** 145–159.

Vidakovic, B. and Ruggeri, F. (2001). Bams method: Theory and simulations. *Shankya, Series* B, **63(2),** 234–253.

Walter, G.G. (1992). Approximation of Delta function by wavelets. *Journal of Approximation Theory*, **71,** 329–343.

Xia, Y., Tong, H. and Li, W.K. (1999). On extended partially linear single-index models. *Biometrika*, **86,** 831–842.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, **97,** 1042–1054.

	Metropolis					
$(\theta,\ \sigma)$	θ		σ		r	
true value	bias	mse	bias	mse	abias	amse
(0.35, 0.02)	2.12e-4	3.31e-6	8.52e-2	7.26e-3	3.63e-4	5.43e-5
(0.35, 0.5)	3.00e-3	2.29e-3	1.10e-2	8.12e-4	1.36e-2	1.26e-2
(0.35, 1)	-6.24e-3	7.60e-3	2.37e-2	2.68e-3	-2.52e-3	5.79e-2
(2.54, 0.02)	4.26e-4	1.80e-6	8.52e-2	7.26e-3	3.40e-4	8.98e-5
(2.54, 0.5)	1.47e-2	1.39e-3	1.08e-2	7.84e-4	8.29e-3	1.43e-3
(2.54, 1)	1.78e-2	6.68e-3	1.87e-2	2.45e-3	-2.09e-2	5.54e-2
(4.72, 0.02)	-2.34e-4	3.64e-6	8.50e-2	7.22e-3	3.30e-4	5.06e-5
(4.72, 0.5)	-1.11e-2	1.66e-3	9.33e-3	8.21e-4	1.24e-2	1.37e-2
(4.72, 1)	8.76e-4	7.91e-3	2.76e-2	2.92e-3	-9.40e-4	6.01e-2
	Independence-type M-H					
(0.35, 0.02)	1.98e-4	1.68e-6	8.53e-2	7.28e-3	-2.10e-5	3.48e-5
(0.35, 0.5)	4.55e-3	2.28e-3	7.06e-3	6.91e-4	9.00e-3	1.09e-2
(0.35, 1)	-1.24e-2	6.15e-3	9.51e-3	2.09e-3	-2.58e-3	5.78e-2
(2.54, 0.02)	7.46e-5	2.12e-6	8.51e-2	7.24e-3	-3.51e-5	3.89e-5
(2.54, 0.5)	1.21e-2	1.21e-3	6.13e-3	6.86e-4	8.40e-3	1.16e-6
(2.54, 1)	1.32e-2	7.26e-3	6.45e-3	2.00e-3	6.19e-3	4.77e-2
(4.72, 0.42)	-1.69e-4	2.77e-6	8.50e-2	7.23e-3	-2.86e-5	3.09e-5
(4.72, 0.5)	-1.24e-2	1.71e-3	5.64e-3	7.31e-4	8.60e-3	1.19e-2
(4.72, 1)	-5.66e-3	6.93e-3	1.17e-2	2.16e-3	1.17e-2	4.32e-2

Table 1: Simulation results for cosine function.

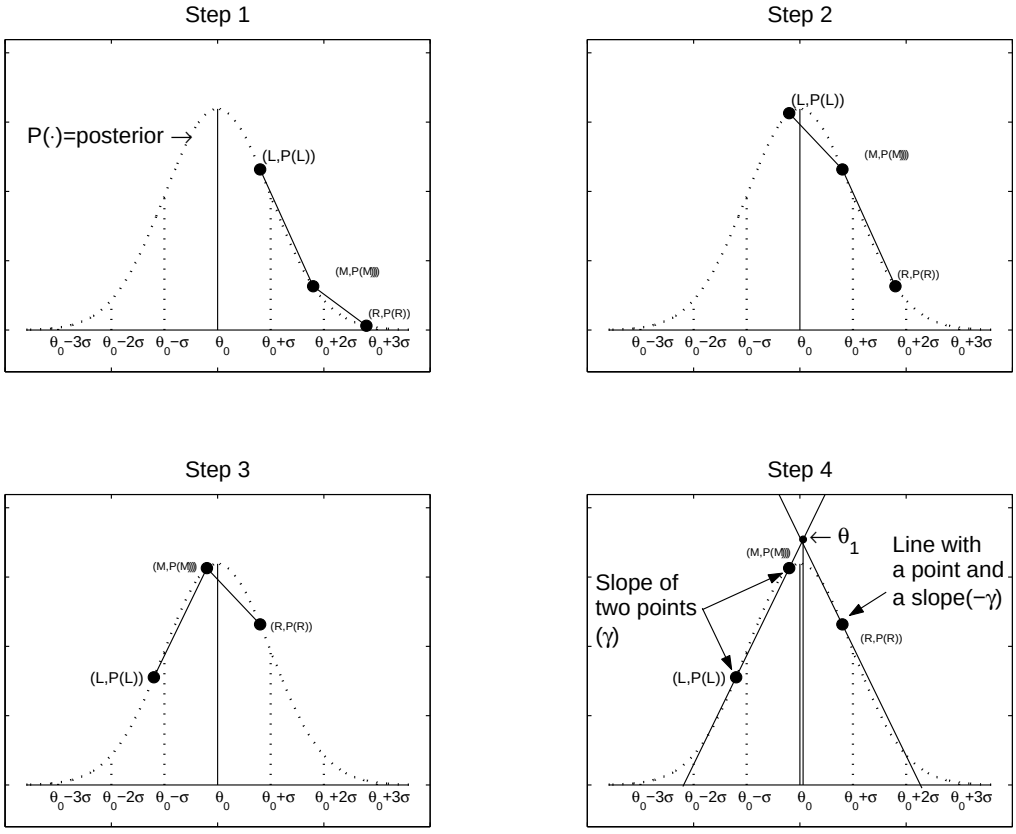

Figure 1: The three-point method for mode location of a target distribution.

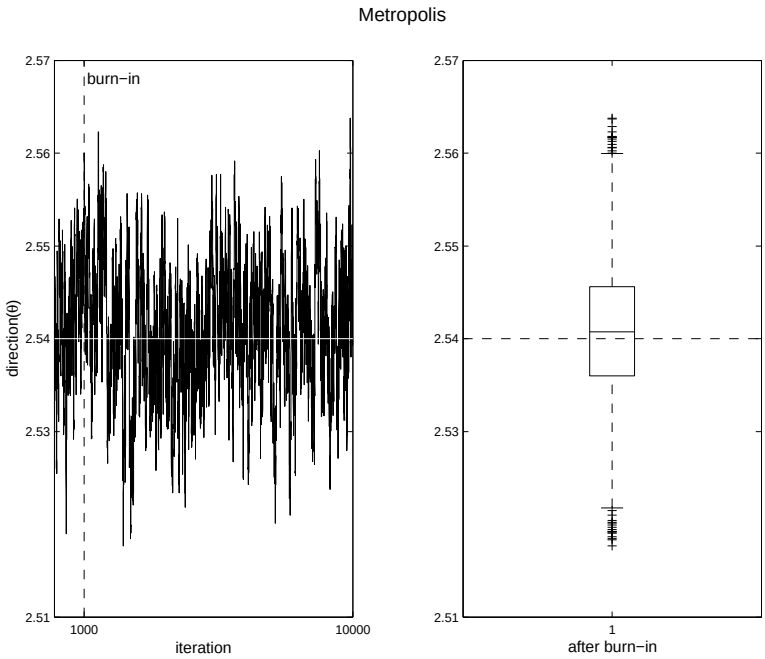

Figure 2: Cosine function: MCMC trace for the direction parameter (left) and boxplots of the sampled values after burnin (right).

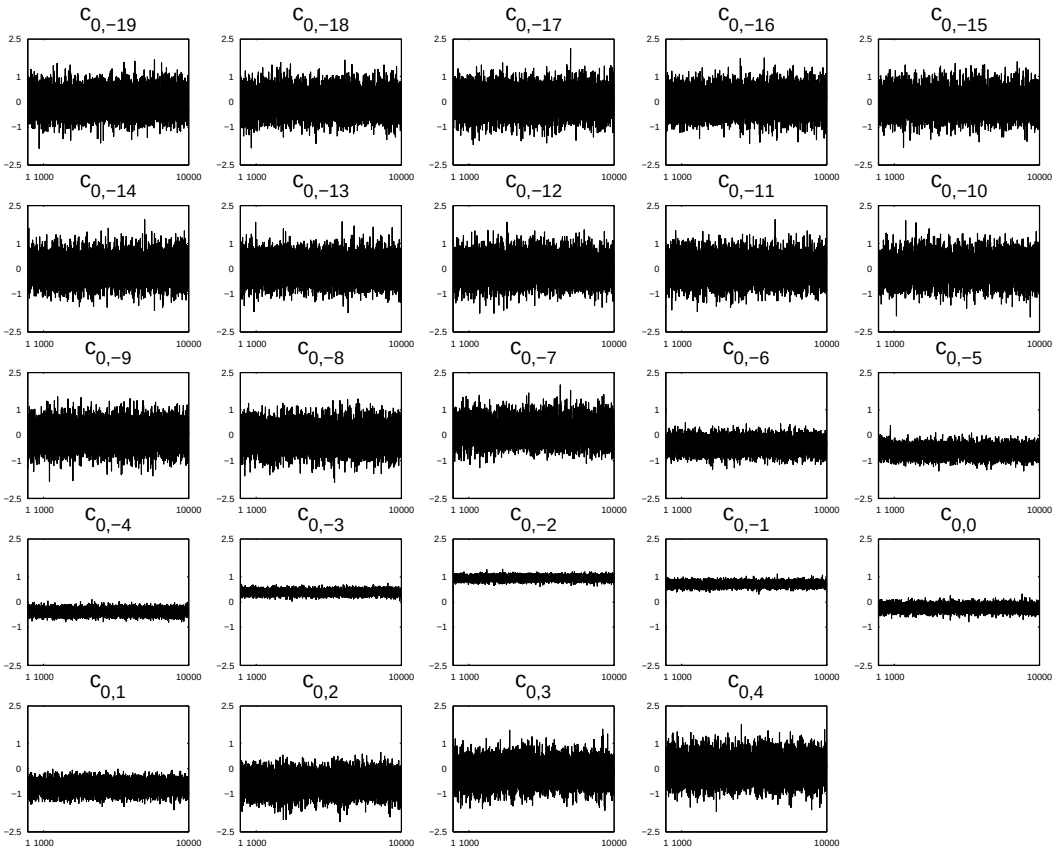

Figure 3: Cosine function: MCMC traces for the scaling coefficients.

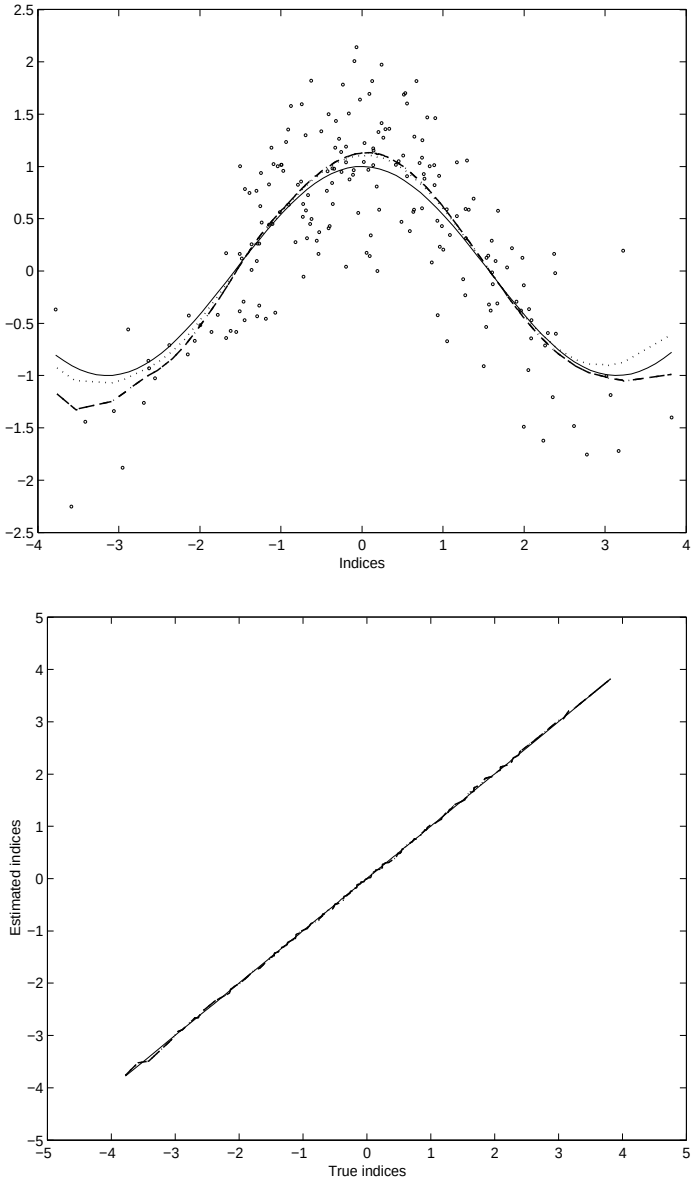

Figure 4: Cosine function. Upper plot: posterior estimated mean regression functions (dotted and dashed lines) by Metropolis and Independence-type M-H, respectively. The true function is indicated by the solid line. The '+' symbols indicate the noisy data. Lower plot: a scatter plot of the estimated and the true design points.

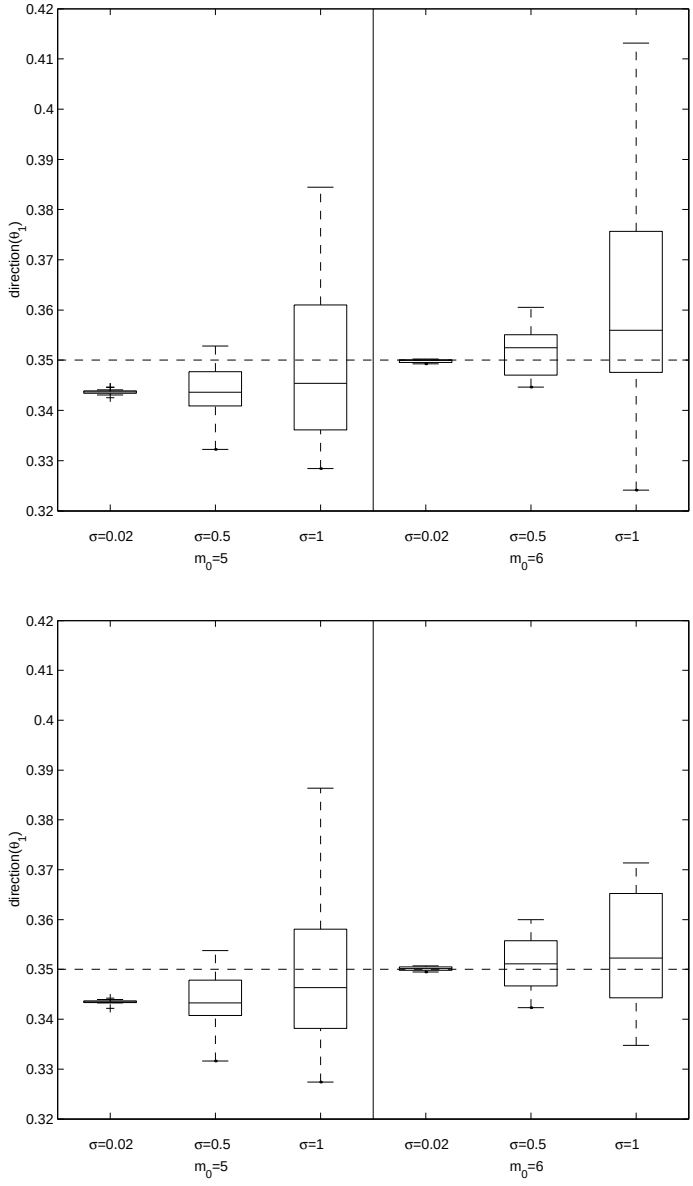

Figure 5: Doppler function. Boxplots of the estimates of the direction parameter for the 20 simulated datasets, for both Metropolis (upper plot) and the independence-type M-H (lower plot), obtained with truncation levels $m_0 = 5$ and $m_0 = 6$.

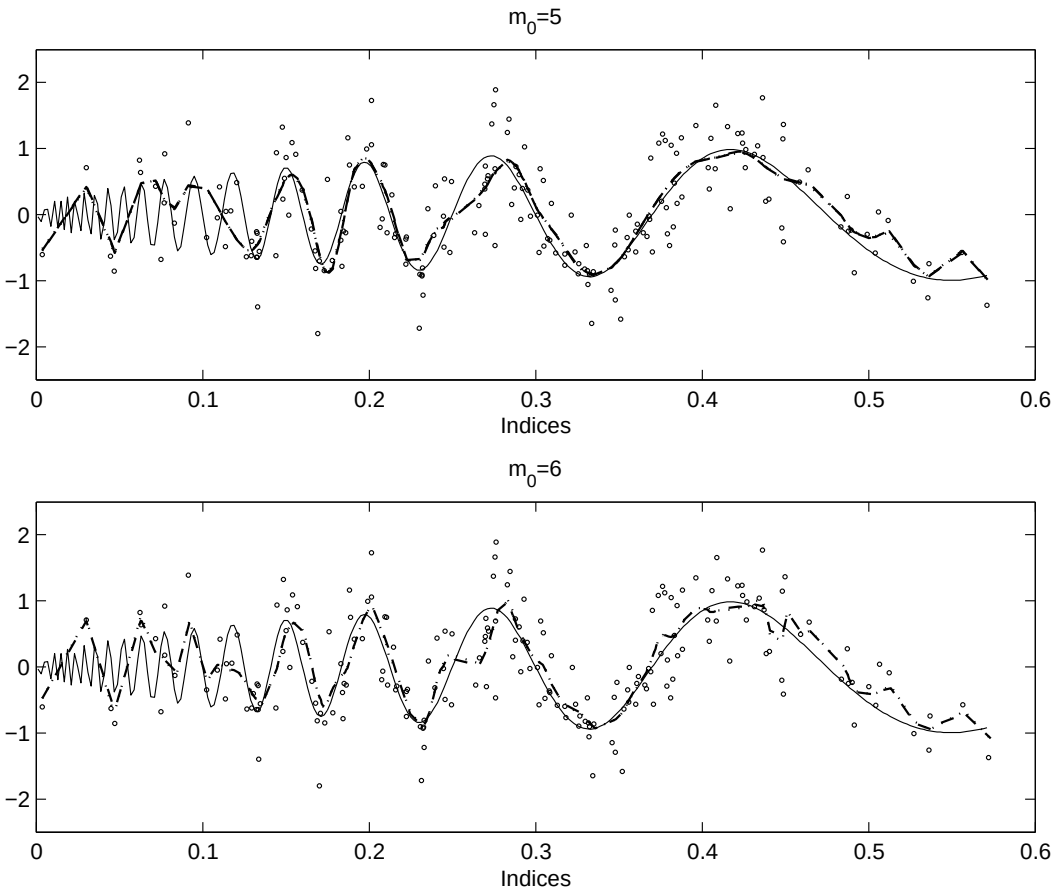

Figure 6: Doppler function. Posterior estimated mean regression functions by Metropolis and Independence-type M-H (dotted and dashed lines, respectively). The true function is indicated by the solid line. The '+' symbols indicate the noisy data. Estimates are obtained with truncation level $m_0 = 5$ (upper plot) and $m_0 = 6$ (lower plot).

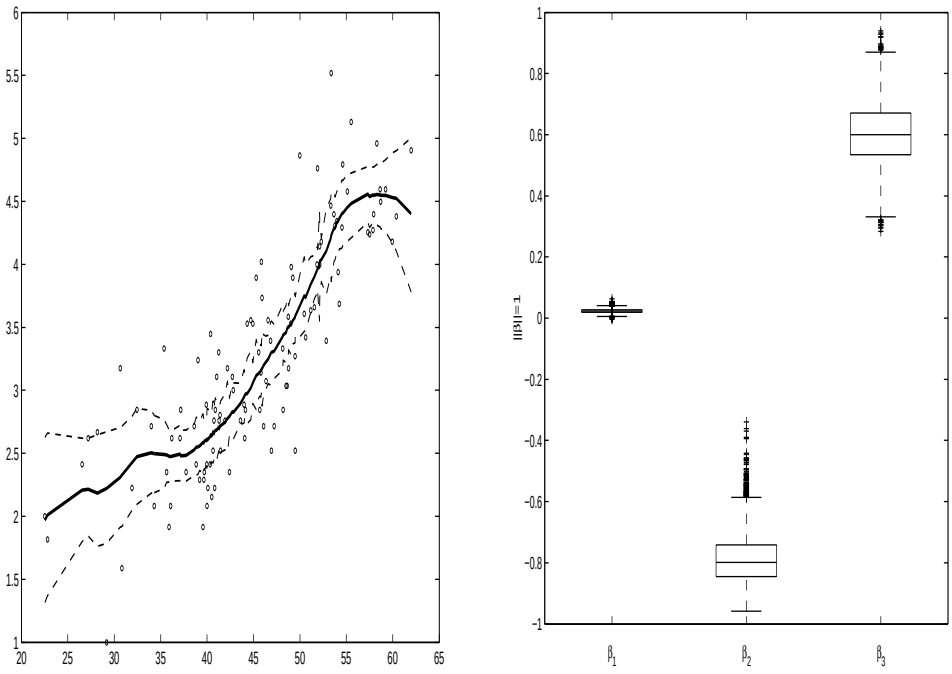

Figure 7: Air pollution data. *Left:* Posterior estimated mean regression function with the 95% posterior confidence interval. *Right:* Boxplots of the sampled values of the direction parameters, after burnin.

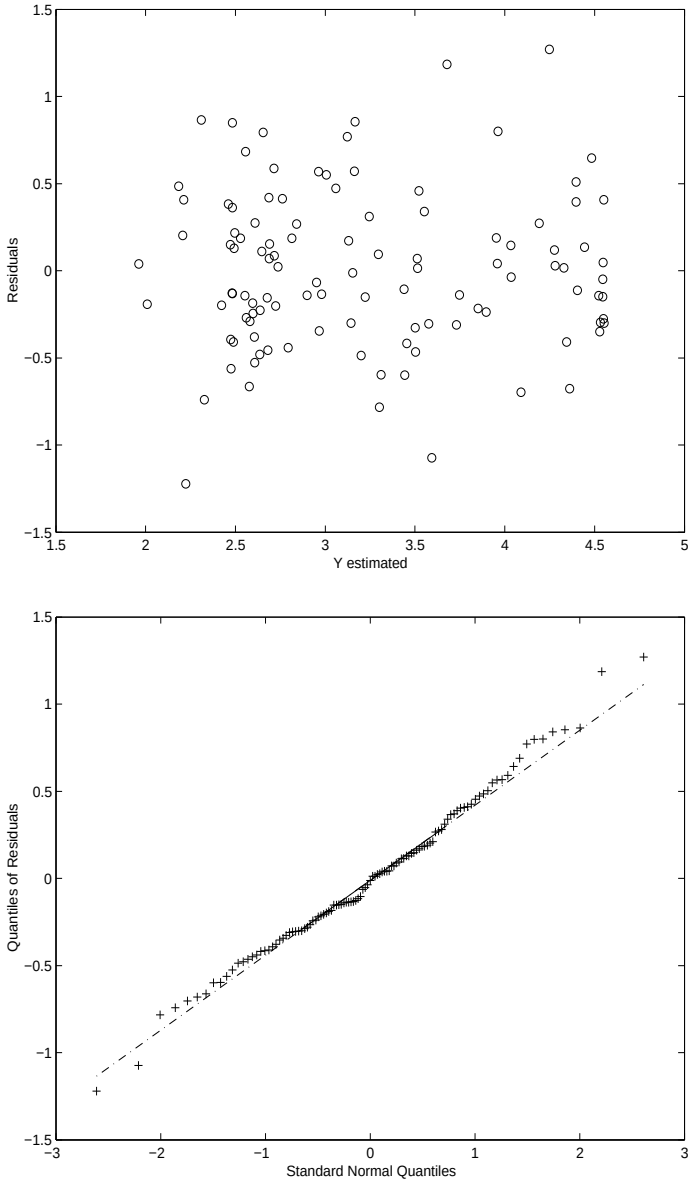

Figure 8: Air pollution data. Upper plot: scatter plot of the residual versus the predicted value. Lower plot: Normal probability plot of the residuals.

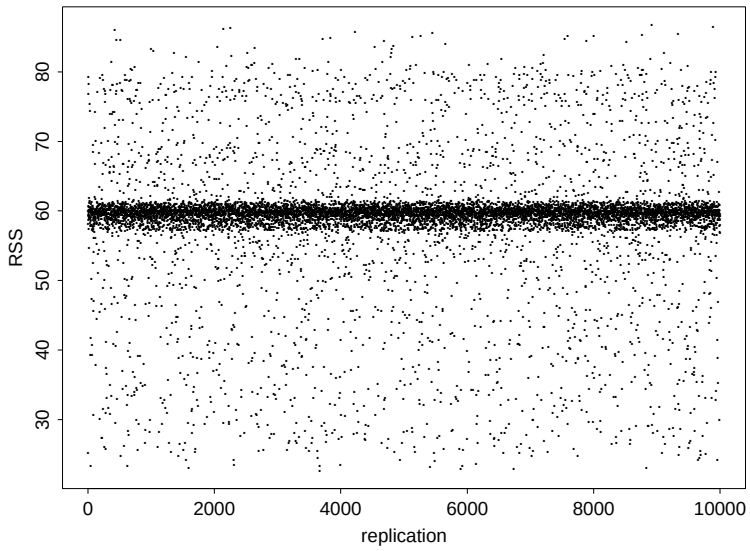

Figure 9: Air pollution data. RSS for 10,000 randomly selected directions.

Metropolis

resolution	$(\theta,\ \sigma)$ true value	θ bias	θ mse	σ bias	σ mse	r abias	r amse
$m_0 = 5$	(0.35, 0.02)	-6.32e-3	4.02e-5	1.71e-1	2.92e-2	4.64e-4	1.80e-2
	(0.35, 0.5)	-5.95e-3	6.17e-5	3.62e-2	1.87e-2	8.45e-3	5.36e-2
	(0.35, 1)	-3.93e-4	2.68e-4	1.42e-2	2.65e-3	7.94e-3	1.43e-1
$m_0 = 6$	(0.35, 0.02)	-1.76e-4	1.25e-7	1.46e-1	2.14e-2	-2.50e-4	2.84e-4
	(0.35, 0.5)	1.76e-3	2.48e-5	9.68e-2	1.00e-2	6.82e-3	6.15e-2
	(0.35, 1)	1.11e-2	5.76e-4	1.02e-1	1.37e-2	4.74e-3	2.06e-1

Independence-type M-H

$m_0 = 5$	(0.35, 0.02)	-6.51e-3	4.25e-5	1.70e-1	2.90e-2	7.94e-4	1.82e-2
	(0.35, 0.5)	-5.77e-3	6.13e-5	3.67e-2	1.91e-3	8.50e-3	5.39e-2
	(0.35, 1)	-4.09e-4	2.36e-4	1.28e-2	2.52e-3	7.75e-3	1.45e-1
$m_0 = 6$	(0.35, 0.02)	1.61e-4	1.77e-7	1.45e-1	2.11e-2	-2.20e-4	2.77e-4
	(0.35, 0.5)	1.37e-3	2.98e-5	9.98e-2	1.11e-2	6.99e-3	6.03e-2
	(0.35, 1)	1.07e-2	9.89e-4	1.09e-1	1.61e-2	5.08e-3	2.08e-1

Table 2: Simulation results for Doppler function.

Cosine function

$(\theta,\ \sigma)$	Indep-type M-H		Kernel-type method			Bayes-splines method		
true value	angle	snorm	angle	snorm	amse(r)	angle	snorm	amse(r)
(0.35, 0.02)	1.08e-3	2.47e-3	4.55e-3	4.27e-3	3.43e-4	1.75e-1	4.27e-1	9.19e-2
(0.35, 0.5)	3.52e-2	1.13e-1	4.78e-2	4.48e-2	1.11e-2	6.58e-2	2.15e-1	7.42e-3
(0.35, 1)	6.44e-2	1.62e-1	7.90e-2	7.30e-2	4.43e-2	8.11e-2	2.56e-1	1.15e-2
(2.54, 0.02)	1.30e-3	2.75e-3	3.30e-3	2.72e-3	6.45e-3	2.10e-1	4.10e-1	9.61e-2
(2.54, 0.5)	2.99e-2	7.06e-2	4.11e-2	3.40e-2	2.40e-2	1.40e-1	4.06e-1	1.06e-1
(2.54, 1)	6.77e-2	1.97e-1	8.79e-2	7.10e-2	4.94e-2	1.99e-1	7.40e-1	8.03e-2
(4.72, 0.02)	1.41e-3	3.36e-3	1.34e-2	1.34e-2	1.34e-4	2.01e-1	3.87e-1	1.70e-1
(4.72, 0.5)	3.35e-2	9.23e-2	3.06e-2	3.06e-2	1.07e-2	1.75e-1	5.47e-1	1.37e-1
(4.72, 1)	6.97e-2	1.62e-1	5.80e-2	5.78e-2	3.55e-2	2.82e-1	6.80e-1	1.35e-1

Doppler function

(0.35, 0.02)	6.45e-3	7.72e-3	3.99e-3	3.74e-3	4.66e-2	9.97e-2	3.46e-1	2.20e-1
(0.35, 0.5)	6.48e-3	1.84e-2	7.99e-3	7.50e-3	9.14e-2	1.39e-1	4.31e-1	2.65e-1
(0.35, 1)	1.15e-2	3.47e-2	2.44e-2	2.28e-2	1.84e-1	1.40e-1	4.25e-1	3.29e-1

Table 3: Simulation results for the Independence-type M-H, the kernel-type and the Bayes-splines methods, for cosine and Doppler functions.

Radiation (β_1)		Wind (β_2)		Temperature (β_3)	
mean	std.	mean	std.	mean	std.
0.0236	0.0072	-0.7860	0.0831	0.6036	0.1017

Table 4: Results for air pollution data

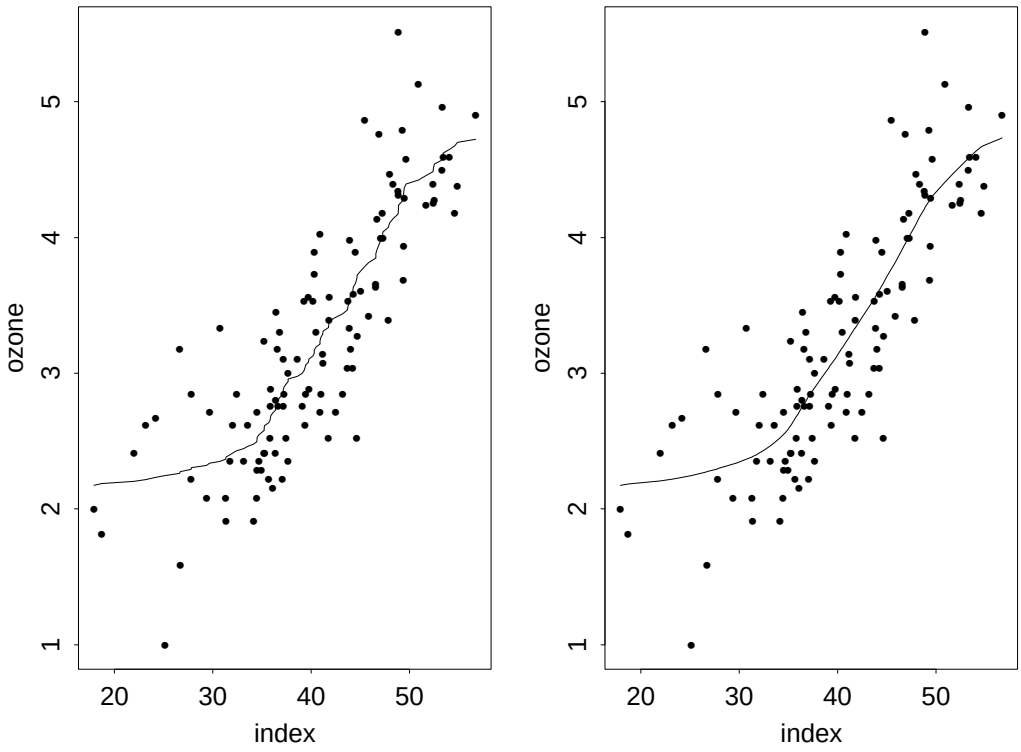

Figure 10: Air pollution data. Local linear estimates of the link function. *Left:* OSCV smooth, *right:* local linear smooth of the OSCV smooth.