Advances and Applications in Statistical Sciences
Volume …, Issue …, January 2009, Pages …
© 2009 Mili Publications

MODELING HETEROSCEDASTICITY IN THE SINGLE-INDEX MODEL WITH THE DIRICHLET PROCESS

GEORGE KARABATSOS

University of Illinois-Chicago
1040 W. Harrison Street (MC 147)
Chicago, IL 60607, U.S.A.
E-mail: georgek@uic.edu

Abstract

The single-index model is a nonparametric regression approach that has seen many applications. The model avoids the curse of dimensionality by reducing the p-dimensional predictor to a univariate single-index (a linear combination of p regression coefficients and covariates), and provides a flexible alternative to ordinary linear regression. In the model, each observed continuous response has mean equal to an unknown (link) function of the single-index, and the errors in regression have common variance under the assumption of homoscedasticity. In this paper, a novel Bayesian heteroscedastic single-index model is introduced, where the link function is modeled by splines, and the distribution of the error variances (over observations) is modeled nonparametrically by a Dirichlet Process prior. The spline coefficients are regularized with a ridge prior with parameter assigned a hyperprior. Methods of Gibbs sampling and adaptive Metropolis-Hastings sampling are presented for posterior inference and goodness-of-fit analysis of the heteroscedastic model. The model is illustrated through the analysis of real data. Compared to the homoscedastic single-index model, the heteroscedastic model demonstrated superior predictive accuracy.

1. Introduction

The single-index regression model, which has seen many applications, especially in econometrics, is represented by

$$y_i = g(\mathbf{x}_i'\boldsymbol{\omega}) + \varepsilon_i, \quad i = 1, \ldots n, \tag{1}$$

$$\varepsilon_1, \ldots, \varepsilon_n | \sigma^2 \sim \text{Normal}\ (0, \sigma^2)$$

2000 Mathematics Subject Classification: Please provide.

Keywords: Bayesian nonparametrics, ridge regression, dimension reduction, Gibbs sampling, adaptive Metropolis-Hastings.

where y_i is an observed response, $\mathbf{x}_i$ is a vector of p covariates, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)'$ is the index vector constrained to unit-norm for identifiability, and $g: \mathbb{R} \to \mathbb{R}$ is an unknown univariate link function [15, 20, 23, 33]. This model is a projection pursuit regression model [4] with a single ridge function the link function (the link function $g(\cdot)$). Since the model applies a nonlinear link function $g(\cdot)$ on a univariate index $\mathbf{x}'\boldsymbol{\omega}$, rather than on the p-dimensional covariate space, it automatically handles interactions among the p covariates, and avoids the "curse of dimensionality" [3]. This model provides a flexible and interpretable compromise between the ordinary linear regression model which is interpretable but often overly restrictive, and the more flexible additive model [17] which also reduces dimensionality but does not automatically handle interactions. Also, when $g(\cdot)$ is monotonic, the vector $\boldsymbol{\omega}$ has the same interpretation as "effect" parameters as in ordinary linear models [23].

Much research on single-index model (1) has focused on the development of approaches for the point-estimation of the index vector $\boldsymbol{\omega}$ and the link function $g(\cdot)$. One group of approaches is based on the method of average derivative estimation [12, 15, 16,19, 27, 30, 33, 34], which exploit the fact that $\boldsymbol{\omega}$ is proportional to the expected value of the gradient $\nabla g(\cdot)$. The average derivative estimate of $\boldsymbol{\omega}$ is then used to estimate $g(\cdot)$, often, using kernel approaches. Among the available methods of average derivative estimation, the method of [19] seems preferable, as it best handles the case of large number of predictors p. A second group of methods is based on the semiparametric efficiency approach [5, 8, 13, 18, 20, 21, 40]. In this approach, an M-estimator of ω is derived on the basis of a nonparametric estimator of $g(\cdot)$ conditioned on design points $\mathbf{x}_i'\boldsymbol{\omega}$, $i = 1, \ldots, n$. Other approaches to estimating $\boldsymbol{\omega}$ include the slice-inverse regression method [23], and the minimum-average variance estimation method [39] which uses ideas of average derivative estimation, the slice-inverse regression method, and locally-linear kernel smoothing. Also, goodness-of-fit tests have been proposed for the single-index model [7, 14, 38].

Antoniadis et al. [1] introduced an empirical Bayes approach to the single-index model, where $g(\cdot)$ is modeled by spline basis functions with

coefficients having a multivariate Normal$(0, v\mathbf{I})$ ridge prior, the index vector $\boldsymbol{\omega}$ is given a Fisher-von Mises prior distribution, and the error variance σ^2 has an inverse-gamma prior. In particular, the location parameter of the Fisher-von Mises prior is defined by a point-estimate of $\boldsymbol{\omega}$ determined under the method of Hristache et al. [19]. Given this point estimate, the ridge parameter v, which regularizes (penalizes) the spline coefficients, is determined through generalized cross-validation. Using Markov Chain Monte Carlo (MCMC), Gibbs samplers are available for generating conditional posterior samples of the spline coefficients and the error variance σ^2, and they implement a random-walk Metropolis Hastings algorithm to generate posterior samples of the index vector $\boldsymbol{\omega}$. A fully Bayesian approach to the single-index model has yet to be developed, which would be based on priors that are not data-dependent. Also, as shown in (1), the single-index model assumes homoscedasticity in errors of regression, which may be an unrealistic assumption. Antoniadis et al. [1] concluded that the development of Bayesian methods for a heteroscedastic single-index model remains an important open problem.

In this paper, we introduce a novel and a fully Bayesian approach to a heteroscedastic single-index model, based on ideas of Bayesian nonparametric inference. In this model, $g(\cdot)$ is modeled by splines, and heteroscedasticity is accounted for by allowing the error variance to depend on the index of the observations, such that the distribution of the error variances $\sigma_1^2, \ldots, \sigma_n^2$ is modeled nonparametrically by a Dirichlet Process (DP) prior. The heteroscedastic single-index model is presented as follows:

$$
\begin{aligned}
f(y \mid \mathbf{x}) &= \int \text{Normal}(g(\mathbf{x}_i'\boldsymbol{\omega}),\, \sigma^2)\, dG(\sigma^2) \\
g(\mathbf{x}_i'\boldsymbol{\omega}) &= \beta_{00} + \beta_{01}\mathbf{x}_i'\boldsymbol{\omega} + \sum\nolimits_{k=1}^{K} \beta_k (\mathbf{x}_i'\boldsymbol{\omega} - t_k)_+ \\
\boldsymbol{\omega} &\sim \text{Uniform}\{\boldsymbol{\omega} : \boldsymbol{\omega}'\boldsymbol{\omega} = 1\} \\
\boldsymbol{\beta},\, \nu \mid \delta_1,\, \delta_2 &\sim \text{Normal}_{K+2}(\mathbf{0},\, \text{diag}(\infty,\, v,\, \ldots,\, v))\text{Ga}(v \mid \delta_1,\, \delta_2) \\
G \mid \alpha,\, G_0 &\sim \text{DP}(\alpha,\, G_0 = \text{Exp}(\mu)),
\end{aligned}
$$

with a $\text{Gamma}(a \to 0, b \to 0)$ hyperprior for α, and a noninformative prior on

μ which is proportional to μ^{-1}. A "default" uniform prior is assigned to the index vector $\boldsymbol{\omega}$. The link function $g(\cdot)$ is modeled by truncated linear splines with coefficients $\boldsymbol{\beta} = (\beta_{00}, \beta_{01}, \beta_1 \ldots, \beta_K)'$, with K knots placed the space of the index values $\mathbf{x}_i'\boldsymbol{\omega}$, $i = 1, \ldots, n$, and where $(x)_+ = \max(0, x)_+$. A multivariate normal ridge prior serves to regularize the spline coefficients $\boldsymbol{\beta}$, this prior being based on a ridge parameter v having a $\text{Gamma}(\delta_1, \delta_2)$ hyperprior. Later in the applications of the heteroscedastic model presented in Section 2, useful choices of hyperprior parameters δ_1 and δ_2 are suggested. Also, while in principle any spline function can be chosen to model $g(\mathbf{x}_i'\boldsymbol{\omega})$, truncated linear splines are chosen because it leads to a flexible function for $g(\mathbf{x}_i'\boldsymbol{\omega})$ that is reasonably smooth, can accommodate sharp changes in the true function, provide a simple representation, and are easy to compute. Furthermore, in typical applications of the single-index model where the true function $g(\cdot)$ is either monotonic or unimodal, 5 to 10 knots are adequate [29, 40]. This, in the above model, $K = 15$ knots are specified, with the knots $t_1, \ldots, t_K$ placed on equally spaced quantiles of the index $\mathbf{x}_i'\boldsymbol{\omega}$, $i = 1, \ldots, n$, these quantiles corresponding to probabilities $0/K$, $1/K$, $\ldots$, $(K-1)/K$, respectively. (For computational convenience, we assume throughout that each of the p observed covariates $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, are centered and standardized so that $\sum_i x_{ik} = 0$ and $\sum_i x_{ik}^2 = 1$.)

The error variances $\sigma_1^2, \ldots, \sigma_n^2$ are modeled nonparametrically by a random distribution G, assumed to follow a Dirichlet Process which depends on precision parameter α and baseline distribution $G_0 = \text{Exponential}(\mu)$ distribution with mean μ. The Dirichlet Process has been studied and applied extensively in the literature; for recent reviews of such priors see, for example [37, 24]. The baseline distribution $G_0 = \text{Exponential}(\mu)$ is the conditional prior mean of G, and the precision parameter α ($\alpha > 0$) measures the amount of faith in G_0. Also, α and μ are given independent "default" non-informative priors, where the $\text{Gamma}(a \to 0, b \to 0)$ prior

corresponds to a uniform prior on $\log(\alpha)$ [10]. A Dirichlet Process for a random distribution G can be described through a stochastic process representation based on a countably-infinite sampling strategy [32]. Let $\theta_1, \ldots, \theta_i, \ldots \overset{iid}{\sim} G_0$, and $b_1, b_2, \ldots, b_i, \ldots \overset{iid}{\sim} \text{Beta}(1, \alpha)(\alpha > 0)$. Then a random distribution function G chosen from a Dirichlet Process prior with parameters (α, G_0) can be constructed via $G(x) = \sum_{i=1}^{\infty} w_i \mathbf{1}(\theta_i \le x)$, where $w_1 = b_1$ and $w_i = b_i \prod_{l<i}(1 - b_l)$ for $i > 1$. From this construction it is obvious that the Dirichlet Process supports only discrete distributions. Thus, ties may occur, and so, in the Bayesian nonparametric single-index model presented above, the error variances $\sigma_1^2, \ldots, \sigma_n^2$ can form clusters among the n observations, and the expected number of clusters under the Dirichlet Process prior is $\sum_{i=1}^{n} \alpha/(\alpha + i - 1)$ [9].

As it turns out, fast Markov Chain Monte Carlo (MCMC) methods can be implemented to sample from the full-conditional posterior distributions for each of the parameters of the Bayesian nonparametric single index model presented above. In particular, an adaptive Metropolis-Hastings algorithm is used to sample the full conditional posterior distribution of the index vector $\boldsymbol{\omega}$ [2], while the full conditional posterior distribution of $\boldsymbol{\beta}$ and the coefficient hyperparameter v is a multivariate normal and a gamma distribution, respectively. Also, the full conditional posterior distributions of $\alpha, \mu,$ and $(\sigma_1^2, \ldots, \sigma_n^2)'$ are sampled using standard MCMC methods for Dirichlet Process models [10, 25]. Moreover, as a simple by-product of this MCMC algorithm, the goodness-of-fit of the heteroscedastic single-index model can be implemented through inference of the posterior predictive distribution of the response residuals $y_i - g(\mathbf{x}_i'\boldsymbol{\omega}), i = 1, \ldots, n$. In the next section the heteroscedastic model is illustrated through the analysis of a few real data sets. For each of the data sets, the heteroscedastic and homoscedastic single-index models are compared for their predictive accuracy, using a model selection criterion computed from the posterior predictive distribution. This accuracy is also evaluated with respect to the choice of gamma prior on the ridge parameter v. Finally, Section 3 presents the details of the MCMC and Section 4 ends with conclusions.

2. Illustrations

This section describes illustrations of the heteroscedastic single-index model through its applications on three real data sets. Each application of the single-index model reported in this entire section is based on 75000 MCMC samples, discarding the first 20000 samples as burn-in, and retaining every 5th sample after burn-in, yielding a total of 11000 samples for posterior inference. After the initial burn-in period of 20000 samples, it was found that the Metropolis proposal parameters λ, τ_ν and τ_σ converged to values that led to the desired acceptance rates for the parameters ω, ν, and $\sigma_i^2, \ldots, \sigma_n^2$, respectively (see Section 4 for details). Also, every application of the single-index model reported in this Section is based on the observed covariates subjected to a normal transformation with unit variance (see Section 1), and is based on a $\text{Gsamma}(1, 1/2)$ prior for the ridge parameter ν, unless otherwise noted, and is conditioned on the specific priors on the rest of the parameters as described in Section 1.

In the first real application, the heteroscedastic single-index model is used to analyze data from the United Nations Organization for Education, Science and Culture (UNESCO), obtained from http://www.uis.unesco.org/. One of the primary aims of UNESCO is to gather data to help countries across the world analyze the efficiency and effectiveness of their educational programs, and to report the global situation with regard to education. This data set consists of $n = 139$ observations, where 33 developed countries worldwide reported on four economic indicators in years 2000 through 2004, with the countries from North America, Europe, Asia, and Australia (not all countries reported in all the 5 years). Three of the indicators include the country's log gross domestic product (GDP) per capita in U.S. dollars, and the country's expenditure per pupil as a log percentage of GDP per capita, for primary schools (SpendPrim), for secondary schools (SpendSec), and for tertiary schools (SpendTert). Using the heteroscedastic single-index model, it was of interest to investigate how the gross domestic product (GDP) depends on three covariates of educational spending (SpendPrim, SpendSec, SpendTert), and on year (Year−2000). Given the data, Table 1 presents the estimate of the posterior means and 95% posterior interval for the parameters of the heteroscedastic single-index model. Note that the posterior

mean estimate of the index vector, $\overline{\boldsymbol{\omega}}$, is obtained by taking the component-wise mean of the MCMC samples of $\boldsymbol{\omega}$, and renormalizing to obtain unit norm. As shown in Table 1, Country spending on primary, secondary and tertiary schools have a significant effect on the level of GDP of the country. Also, the posterior mean estimate of the precision parameter α suggests the presence of heteroscedasticity, and there are on average about 81 distinct clusters among the error variances $\sigma_1^2, \ldots, \sigma_n^2$. Figure 1 presents posterior mean estimate of the link function $g(\cdot)$ under the single-index model, and $g(\cdot)$ is non-monotonically increasing with the single-index $\mathbf{x}'\boldsymbol{\omega}$, with a few sharp changes. The top panel of Figure 2 presents shows that for almost all of the 139 observations in the data, the 99% posterior intervals of the posterior predictive residuals included 0, a result that informally suggests that the single-index model is adequate for these data. The bottom panel of Figure 2 presents the posterior mean estimates of the 139 error variances.

Table 1. Posterior estimates of the heteroscedastic model, for the UNESCO data set.

Parameter	**Posterior**	**2.5%**	**97.5%**
ω, year-2000	.01	-.02	.05
ω, SpendPrim	.66	.63	.68
ω, SpendSec	.51	.48	.55
ω, SpendTert	.56	.53	.57
ν	2.30	1.24	3.86
α	80.62	54.17	112.68
μ	.03	.01	.05
Num. of σ^2 Clusters	80.8	59	96

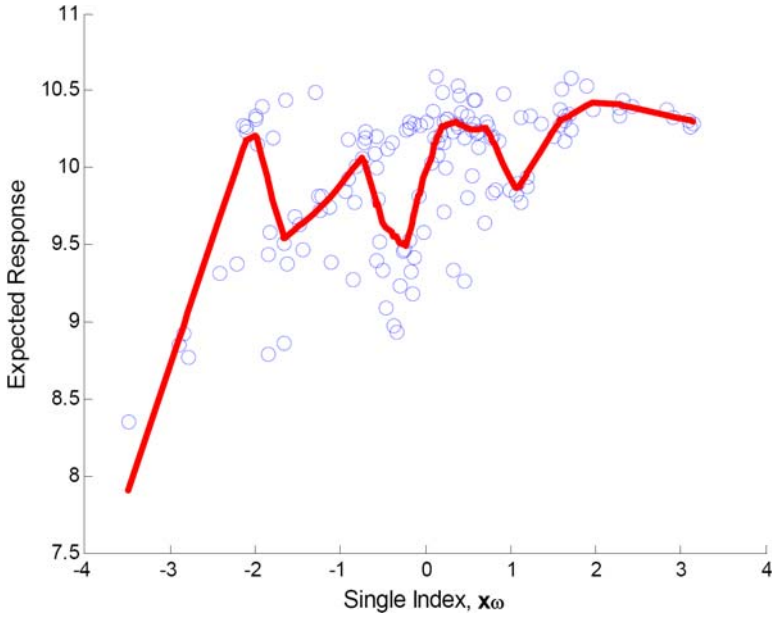

Figure 1. Posterior mean estimate of the link function under the heteroscedastic single-index model, and the 139 observations (circles) of the UNESCO data set.

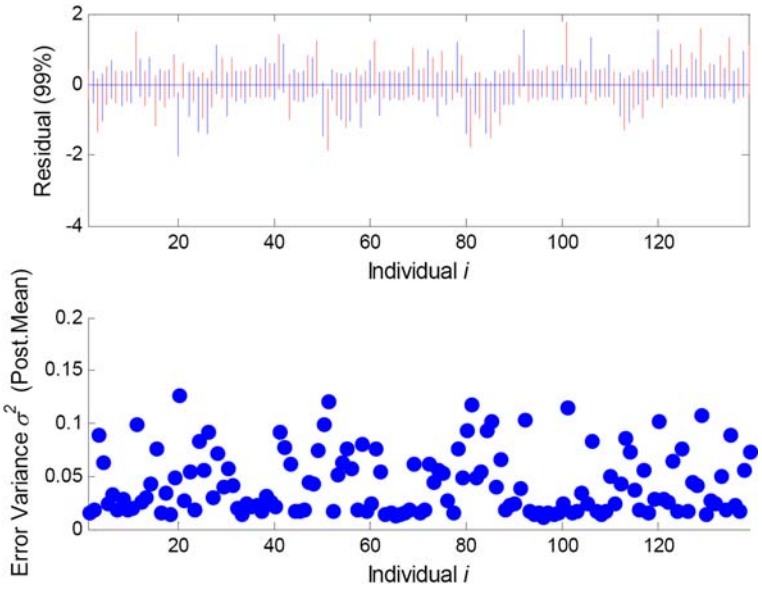

Figure 2. 99% confidence bands of the posterior predictive residuals (top panel) and posterior mean estimates of the error variances (bottom panel), for each of the 139 observations in the UNESCO data set.

It is natural to question whether the heteroscedastic single-index model provides a significant improvement in predictive accuracy over the previously-studied, homoscedastic model. The L^2 criterion [22] can be used to compare predictive accuracy between these two types of single-index models, with the homoscedastic model assuming a uniform prior for the error variance σ^2 instead of a Dirichlet Process prior. The L^2 criterion [22] is defined by the expectation $L^2 = \mathrm{E}[(\mathbf{y}-\mathbf{Z})'(\mathbf{y}-\mathbf{Z})]$ with respect to the posterior predictive distribution of $\mathbf{Z}$, with $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and $\mathbf{y} = (y_1, \ldots, y_2)'$. This criterion can be written as

$$L^2 = \sum_{i=1}^{n} \{(y_1 - \mathrm{E}[Z_i \mid Data, x_i])^2 + \mathrm{var}\,[Z_i \mid Data, x_i]\}, \qquad (2)$$

with expectation $\mathrm{E}[Z_i \mid Data, \mathbf{x}_i]$ and variance $\mathrm{var}[Z_i \mid Data, \mathbf{x}_i]$ taken with respect to the posterior predictive density $p(z_i \mid Data, \mathbf{x}_i)$ in equation (3). Thus, the L^2 criterion is a measure of predictive inaccuracy that is based on the sum-squared error of a model's average predictions (first term in (2)), plus the sum of the variability of the model's predictions (second term in (2)). This criterion can be easily computed as a simple by-product of the MCMC methods for sampling from the posterior predictive density (see Section 4 for details).

The top part of Table 2 presents a comparison of the predictive accuracy between the heteroscedastic and the homoscedastic single-index model for each of the three data sets, according to the L^2 statistic, with the models under various choices of Gamma $(1, \tau_2)$ priors for the coefficient hyperparameter ν. The homoscedastic models were based on gamma priors with various values of the prior parameter $\tau_2 = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, 1\frac{1}{4}, \ldots,$ and Table 2 presents the results of the homoscedastic models with the three lowest values of L^2, for each data set. The same was done for the heteroscedastic model. According to L^2 criterion, the heteroscedastic model has more predictive accuracy than the homoscedastic model, for each of the three data sets. This result indicates that there is evidence of heteroscedasticity in all of

the three data sets. It is worth remarking that the heteroscedastic single-index model outperformed the homoscedastic model even for the in the Petroleum data set where the number of observations is small $(n = 48)$. Also, for each of the three data sets, heteroscedastic models under the $\text{Gamma}(1, 1/2)$ prior for ν had better predictive accuracy than heteroscedastic models under either a $\text{Gamma}(1, 1/4)$ prior or a $\text{Gamma}(1, 3/4)$ prior. In general, however, the different specifications of Gamma prior did not have a large affect on predictive accuracy of the heteroscedastic single-index model.

Table 2. A comparison of the predictive accuracy of the heteroscedastic and homoscedastic single-index models, in each of the three data sets, under different choices of Gamma prior for the ridge parameter.

Data Set	Single-index Model	Prior δ_2	Squared Error	Variance	L^2
UNESCO	Heteroscedastic	.5	15.01	6.75	21.75
	Heteroscedastic	.75	14.86	7.03	21.89
	Heteroscedastic	.25	15.09	7.65	22.75
	Homoscedastic	1.25	15.70	19.39	35.09
	Homoscedastic	1	15.75	19.38	35.13
	Homoscedastic	1.5	15.76	19.54	35.31
Air Pollution	Heteroscedastic	.5	22.35	13.35	35.69
	Heteroscedastic	.75	22.36	13.33	35.70
	Heteroscedastic	.25	22.37	13.36	35.73
	Homoscedastic	.5	22.71	26.04	48.75
	Homoscedastic	.75	22.81	26.17	48.98
	Homoscedastic	.25	160.9	177.6	338.5
Petroleum	Heteroscedastic	.5	25.97	12.85	38.81
	Heteroscedastic	.75	26.12	13.04	39.16
	Heteroscedastic	.25	26.87	13.39	40.26
	Homoscedastic	.75	24.57	33.81	58.39
	Homoscedastic	1	24.41	34.02	58.43
	Homoscedastic	.5	25.53	35.00	60.54

Table 2 also presents comparisons between the heteroscedastic and homoscedastic models on two other data sets which were investigated in the

previous literature on the single-index model. For each of these two other data sets, the heteroscedastic single-index model again displayed superior predictive accuracy over the homoscedastic model. The Air Pollution data set [4] contains 111 daily measurements of ozone levels (Ozone), radiation (SolarRadiation), wind speed (Wind), and temperature (Temp) taken at various sites of the New York metropolitan area, on days between May and September 1973. With these data it was of interest to predict ozone level as a function of the covariates radiation, wind speed, and temperature, and as done in previous research [6], the observed ozone levels were subject to a cube root transformation. The Petroleum data were obtained from a petroleum reservoir study conducted by BP research (from the S-PLUS software; [36]). This data set containing contains a total of $n = 48$ measurements on four cross-sections from each of 12 core samples. Using an image processing procedure performed at the University of Oxford, each sample was measured in the total area in pixels (area), total perimeter in pixels $(\text{peri}), \text{shape} = \text{peri}/\text{area}^{1/2}$, and permeability (perm) measured in milli-Darcies. The aim was to predict log-permeability, as a function of the other three variables, the covariates. The posterior estimates of the parameters for the Air Pollution and the Petroleum data sets are presented in Tables 3 and 4, respectively. Also, for each of these two data sets, the estimated link function was non-monotonic.

Table 3. Posterior estimates of the heteroscedastic model, for the air data set.

Parameter	**Posterior mean**	**2.5%**	**97.5%**
ω, Solar Radiation	.27	.15	.37
ω, Wind	–.34	–.44	–.22
ω, Temp	.90	.83	.96
ν	.56	.07	1.52
α	67.79	45.53	96.05
μ	.09	.06	.14
Num. of σ^2 Clusters	70.0	58	80

Table 4. Posterior estimates of the heteroscedastic model, for the petroleum data set.

Parameter	**Posterior mean**	**2.5%**	**97.5%**
ω, area	.52	.47	.55
ω, peri	–.86	–.88	–.83
ω, shape	.02	–.07	.13
ν	1.56	.55	3.48
α	28.51	15.21	47.88
μ	.19	.07	.36
Num. of σ^2 Clusters	29.2	21	36

3. MCMC Methods

The posterior density of a vector of parameters $\boldsymbol{\theta}$ in the heteroscedastic single-index model is given up to a constant of proportionality by

$$\pi(\boldsymbol{\theta} \mid Data) \propto \prod_{i=1}^{n} \text{normal}(y_i \mid g(\mathbf{x}_i'\boldsymbol{\omega}),\, \sigma_i^2)\pi(\boldsymbol{\theta}),$$

with $\pi(\boldsymbol{\theta})$ the prior density for all the parameters in the model. Using methods of Markov Chain Monte Carlo (MCMC), it is possible to simulate from a Markov Chain $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}, \boldsymbol{\theta}^{4}, \ldots, \boldsymbol{\theta}^{(s)}, \ldots$ until the sequence converges to a sample from the posterior $\pi(\boldsymbol{\theta} \mid Data)$, under mild conditions [34]. This process can be simplified by separating up the parameter vector $\boldsymbol{\theta}$ into 6 blocks, $\boldsymbol{\theta}_1 = \boldsymbol{\omega}$, $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$, $\boldsymbol{\theta}_3 = \nu$, $\boldsymbol{\theta}_4 = \alpha$, $\boldsymbol{\theta}_5 = \mu$ and $\boldsymbol{\theta}_6 = (\sigma_1^2, \ldots, \sigma_n^2)'$, so that for every sampling iteration $s (s = 1, \ldots, S)$, a MCMC sample $\boldsymbol{\theta}^{(s)}$ is generated by sampling from the entire sequence of conditional posterior distributions $f(\boldsymbol{\theta}_q \mid Data, \boldsymbol{\theta}_{r \neq q})$, $q = 1, \ldots, 6$. The following subsections describe MCMC methods for sampling from these conditional posterior distributions, for each state s of the Chain, $s = 1, \ldots, S$. First, it is described how starting values of all the parameters are chosen for state $s = 0$.

3.1. Starting Values

To speed convergence of the MCMC Chain to samples from the posterior distribution of the heteroscedastic single-index model, the starting value $\boldsymbol{\omega}^{(0)}$ is determined by the point estimator of Hristache, et al. [19]. Then the starting values of the knot points $(t_1^{(0)}, \ldots, t_K^{(0)})$ are defined by $K = 15$ equally-spaced quantile values of the index $\mathbf{x}_i'\boldsymbol{\omega}^{(0)}, i = 1, \ldots, n,$ corresponding to the probabilities $0/K, 1/K, \ldots, (K-1)/K,$ respectively. The starting value of the vector of regression coefficients is given by the ridge regression estimator:

$$\boldsymbol{\beta}^{(0)} = ((\mathbf{W}^{-1}))^{(0)} + \mathbf{B}^{(0)}(\mathbf{S}^{-1})^{(0)}\mathbf{B}^{(0)})^{-1}(\mathbf{B}')^{(0)}(\mathbf{S}^{-1})^{(0)}\mathbf{y},$$

with the basis matrix defined by:

$$\mathbf{B}^{(0)} = \begin{pmatrix} 1 & \mathbf{x}_1'\boldsymbol{\omega}^{(0)} & (\mathbf{x}_1'\boldsymbol{\omega}^{(0)} - t_1^{(0)})_+ & \cdots & (\mathbf{x}_1'\boldsymbol{\omega}^{(0)} - t_K^{(0)})_+ \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ 1 & \mathbf{x}_n'\boldsymbol{\omega}^{(0)} & (\mathbf{x}_n'\boldsymbol{\omega}^{(0)} - t_1^{(0)})_+ & \cdots & (\mathbf{x}_n'\boldsymbol{\omega}^{(0)} - t_K^{(0)})_+ \end{pmatrix},$$

along with $\mathbf{W}^{(0)} = \text{diag}_{K+2}(v_0 \to \infty, v^{(0)}, \ldots, v^{(0)})$, $\mathbf{S}^{(0)} = \text{diag}((\sigma_1^2)^{(0)}, \ldots, (\sigma_1^2)^{(0)})$, $\mathbf{y} = (y_1, \ldots, y_n)'$, given $\nu^{(0)} = 1$ and $(\sigma_i^2 = .2)^{(0)}$, $i = 1, \ldots, n$. The starting values of the parameters of the Dirichlet Process prior are set as $\mu^{(0)} = .2$ and $\alpha^{(0)} = 1$. Also, the parameters of the adaptive Metropolis Hastings algorithms, described later in this section, are initialized as $\lambda = 10$ and $\tau_\nu = .2$.

3.2. Sampling ω

To generate a posterior conditional sample of $\boldsymbol{\omega}$ in each state s of the MCMC chain $(s = 1, \ldots, S)$, a special version of [2] adaptive random-walk Metropolis (ARWM) algorithm is implemented. It is known that the choice of variance in the proposal distribution is crucial to the performance of the ordinary random-walk Metropolis algorithm. A choice of variance that is too high causes the algorithm to propose large moves that are likely to be rejected, and a choice of variance that is too low causes the algorithm to

propose small moves that are often accepted. In either case, the algorithm will mix poorly. The ARWM algorithm, an extension of the ordinary random-walk Metropolis algorithm, continuously modifies the proposal variance to obtain a desired acceptance rate. Under mild conditions, including that the size of the change of the proposal variance converges to zero with increasing number of sampling iterations, this adaptive algorithm preserves the ergodicity and stationarity of the specified target posterior distribution (Roberts & Rosenthal, 2005). The AWRM algorithm, as applied to the MCMC sampling of ω, is described as follows.

First, in the ARWM algorithm, the proposal distribution for $\boldsymbol{\omega}$ is specified by the $J_p(\lambda, \boldsymbol{\xi})$ distribution [31], which is a symmetric distribution defined on the unit sphere $S_p = \{\boldsymbol{\omega} \in \mathbb{R}^p : \boldsymbol{\omega}'\boldsymbol{\omega} = 1\}$. This distribution has density [31]:

$$j_p(\boldsymbol{\omega} \mid \lambda, \boldsymbol{\xi}) = \left(\frac{2\pi^{p/2}}{\Gamma(p/2)}\right)^{-1} \exp(-\lambda^2) \sum_{l=0}^{\infty} \left\{\frac{(2\lambda\boldsymbol{\xi}'\boldsymbol{\omega})^l \Gamma([p+l]/2)}{l!\,\Gamma(p/2)}\right\} \mathbf{1}(\lambda \geq 0, \boldsymbol{\omega} \in S_p)$$

where the first term is to the inverse of the surface area of S_p, and $\mathbf{1}(\cdot)$ denotes the indicator function, $\boldsymbol{\xi}$ is the mode parameter, and the parameter λ that is inversely proportional to the variance. A sample $\boldsymbol{\omega}^*$ from $J_p(\lambda, \boldsymbol{\xi})$ is easily obtained by taking $\boldsymbol{\omega}^* = w/(w'w)^{1/2}$, where w is a draw from $\text{Normal}_p(\boldsymbol{\zeta}, \mathbf{I}_p)$, given $\boldsymbol{\xi} = \boldsymbol{\zeta}/\boldsymbol{\zeta}'\boldsymbol{\zeta})^{1/2}$ and $\lambda = (\frac{1}{2}\boldsymbol{\zeta}'\boldsymbol{\zeta})^{1/2}$ ([31], p. 71, property 12). Thus, for the purposes of implementing $J_p(\lambda, \boldsymbol{\xi})$ as a proposal distribution for each state s of the MCMC chain, it is easy to show that a draw $\boldsymbol{\omega}^*$ from the proposal distribution $J_p(\lambda^{(s-1)}, \boldsymbol{\omega}^{(s\text{-}1)})$, is obtained by $\boldsymbol{\omega}^* = w/(w'w)^{1/2}$, where $\text{Normal}_p(\boldsymbol{\omega}^{(\mathbf{s})}2(\lambda^{(s-1)})^2, \mathbf{I}_p)$, and where $\boldsymbol{\omega}^{(s\text{-}1)}$ and $\lambda^{(s-1)}$ denote the old states of $\boldsymbol{\omega}$ and λ at state *s* in the MCMC chain.

With the conditional posterior of $\boldsymbol{\omega}$ proportional to the normal likelihood of the observations, the proposal $\boldsymbol{\omega}^*$ is then accepted as the new state of $\boldsymbol{\omega}$ in the MCMC chain, such that $\boldsymbol{\omega}^* = \boldsymbol{\omega}^{(s)}$ with probability based on a ratio of likelihoods:

$$\rho(\boldsymbol{\omega}^{(s-1)}, \boldsymbol{\omega}^{(s)}) = \min\left\{1, \frac{\prod_{i=1}^{n} \text{normal}\,(y_i \mid g^{(s-1)}(\mathbf{x}_i'\boldsymbol{\omega}^*), (\sigma_i^2)^{(s-1)})}{\prod_{i=1}^{n} \text{normal}\,(y_i \mid g^{(s-1)}(\mathbf{x}_i'\boldsymbol{\omega}^{(s-1)}), (\sigma_i^2)^{(s-1)})}\right\},$$

otherwise $\boldsymbol{\omega}^* = \boldsymbol{\omega}^{(s-1)}$ with probability $1 - \rho(\boldsymbol{\omega}^{(s-1)}, \boldsymbol{\omega}^{(s)})$. In the above equation, the ratio of J_p proposal densities and ratio of uniform prior densities of $\boldsymbol{\omega}$ cancel out, since each of these densities are symmetric. Also, in the notation above,

$$g^{(s-1)}(\mathbf{x}_i'\boldsymbol{\omega}^*) = \beta_0^{(s-1)} + \beta_1^{(s-1)}\mathbf{x}_i'\boldsymbol{\omega}^* + \sum_{k=1}^{K} \beta_k^{(s-1)}(\mathbf{x}_i'\boldsymbol{\omega}^* - t_k^*)_+,$$

where the knot points $(t_1^*, \ldots, t_K^*)$ defined by $K = 15$ equally-spaced quantile values of the index $\mathbf{x}_i'\boldsymbol{\omega}^*, i = 1, \ldots, n,$ and similarly for $g^{(s-1)}(\mathbf{x}_i'\boldsymbol{\omega}^{(s-1)})$, and for $g^{(s)}(\mathbf{x}_i'\boldsymbol{\omega}^{(s)})$. The new state $\boldsymbol{\omega}^{(s)}$ corresponds to a new state for the knot points, $(t_1^{(s)}, \ldots, t_K^{(s)})$.

Then, to conclude state s in the MCMC chain, the proposal parameter $\lambda^{(s-1)}$ of the proposal distribution J_p is updated to

$$\lambda^{(s)} = \max\{0, \lambda^{(s-1)} + s^{-1/2}(.234 - \rho(\boldsymbol{\omega}^{(s-1)}, \boldsymbol{\omega}^{(s)}))\},$$

where, recall that λ is inversely proportional to the variance. Thus, the ARWM algorithm entails the use of a Robbins-Monro algorithm to approximate the proposal variance in such a way to achieve the acceptance rate of about .234. This is the optimal rate for multidimensional parameters such as $\boldsymbol{\omega}$ [28]. To ensure the ergodicity and stationarity of the specified target posterior distribution, the proposal parameter λ is updated only in the first set of initial "burn-in" samples generated from the entire MCMC algorithm (say, the first 20000 MCMC samples).

3.3. Sampling β and ν

Based on standard results for the Bayesian analysis of linear models (O'Hagan and Forster, 2002, p. 330), a sample $\boldsymbol{\beta}^{(s)}$ (and a sample $g^{(s)}(\cdot)$) is

generated from the conditional posterior distribution defined by the multivariate normal distribution $\text{Normal}_{K+2}(\mathbf{m}, \mathbf{V}^{(s-1)})$, where

$$\mathbf{V}^{(s-1)} = ((\mathbf{W}^{-1})^{(s-1)} + \mathbf{B}^{(s)}(\mathbf{S}^{-1})^{(s-1)}\mathbf{B}^{(s)})^{-1}$$

$$\mathbf{m} = \mathbf{V}^{(s)}(\mathbf{B}')^{(s)}(\mathbf{S}^{-1})^{(s-1)}\mathbf{y}$$

$$\mathbf{W}^{(s-1)} = \text{diag}_{K+2}(v_0 \to \infty, v^{(s-1)}, \ldots, v^{(s-1)}),$$

$$(\mathbf{S})^{(s-1)} = \text{diag}((\sigma_1^2)^{(s-1)}, \ldots, (\sigma_n^2)^{(s-1)}),$$

$$\mathbf{y} = (y_1, \ldots, y_n)',$$

with basis matrix:

$$\mathbf{B}^{(s)} = \begin{pmatrix} 1 & \mathbf{x}_1'\boldsymbol{\omega}^{(s)} & (\mathbf{x}_1'\boldsymbol{\omega}^{(s)} - t_1^{(s)})_+ & \cdots & (\mathbf{x}_1'\boldsymbol{\omega}^{(s)} - t_K^{(s)})_+ \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ 1 & \mathbf{x}_n'\boldsymbol{\omega}^{(s)} & (\mathbf{x}_n'\boldsymbol{\omega}^{(s)} - t_1^{(s)})_+ & \cdots & (\mathbf{x}_n'\boldsymbol{\omega}^{(s)} - t_K^{(s)})_+ \end{pmatrix},$$

along with $\boldsymbol{\omega}^{(s)}$ and $(t_1^{(s)}, \ldots, t_K^{(s)})$, the current values of the respective parameters of the MCMC chain.

The ARWM algorithm is implemented to update the ridge parameter ν for state s of the MCMC chain $(s = 1, \ldots, S)$. With the conditional posterior distribution of ν proportional to a normal-gamma density, a proposal ν^* is generated from a $\text{Normal}(\log(\nu^{(s-1)}), \tau_\nu^{(s-1)})$ proposal distribution, and is accepted as the current state $\nu^{(s)} = \nu^*$ with probability:

$$\rho(\nu^{(s-1)}, \nu^{(s)}) =$$

$$\min\left\{1, \frac{\prod_{i=1}^{n} \text{normal}_{K+2}(\boldsymbol{\beta}^{(s)} \mid \mathbf{0}, \underset{K+2}{\text{diag}}((\infty, v^*, \ldots, v^*)))\,\text{gamma}(\nu^* \mid \delta_1, \delta_2)}{\prod_{i=1}^{n} \text{normal}_{K+2}(\boldsymbol{\beta}^{(s)} \mid \mathbf{0}, \underset{K+2}{\text{diag}}((\infty, v^{(s-1)}, \ldots, v^{(s-1)})))\,\text{gamma}(\nu^{(s-1)} \mid \delta_1, \delta_2)}\right\},$$

otherwise $\nu^{(s)} = \nu^{(s-1)}$ with probability $1 - \rho(\nu^{(s-1)}, \nu^{(s)})$. From the ratio above, the ratio of proposal densities cancels out, since these densities are

symmetric. Using Atchadé & Rosenthal's algorithm, the proposal variance is updated to $\tau_{\nu}^{(s)} = \max(10^{-10}, \rho(\nu^{(s-1)}, \nu^{(s)}) - .44)$ using a Robbins-Monro algorithm, in order to approximate the acceptance rate of .44. This is the optimal acceptance rate for univariate parameters [28]. Also, this updating of the proposal variance occurs only in the first set of initial "burn-in" samples of the MCMC algorithm. Finally, as an alternative to the ARWM algorithm, Gibbs sampling can be used to draw $\nu^{(s)}$ from the conditional posterior distribution $\text{Gamma}(\delta_1 + \frac{1}{2}(K+1), [\delta_2^{-1} + \frac{1}{2}\boldsymbol{\beta}_*'\boldsymbol{\beta}_*]^{-1})$, with $\boldsymbol{\beta}_* = (\boldsymbol{\beta}_{01}^{(s)}, \boldsymbol{\beta}_1^{(s)}, \ldots, \boldsymbol{\beta}_K^{(s)})'$.

3.4. Sampling α, μ, and $\sigma_1^2, \ldots, \sigma_n^2$

Using the method by [10], a conditional posterior sample of the Dirichlet Process precision parameter α and exponential mean parameter μ is generated using Gibbs sampling. Recall that α has a noninformative, $\text{Gamma}(a \to 0, b \to 0)$ prior. Let $(\sigma_1^2)^{(s-1)}, \ldots, (\sigma_n^2)^{(s-1)}$ denote the current state of the error variance parameters in the MCMC chain, having C distinct values (clusters) denoted by $(\sigma_c^2)^{(s-1)}, c = 1, \ldots, C^{(s-1)}$, with $C^{(s-1)} \le n$. Then at state s of the MCMC chain $(s = 1, \ldots, S)$, a conditional posterior sample $\alpha^{(s)}$ is generated by a mixture of two gamma densities, defined by:

$$\pi_\eta \text{Gamma}(C^{(s-1)}, \log(\eta)) + (1 - \pi_\eta)\text{Gamma}(C^{(s-1)} - 1, \log(\eta)),$$

where η is a draw from $\text{Beta}(1, n)$, and $\pi_\eta / (1 - \pi_\eta) = (C^{(s-1)} - 1) / \{n(-\log(\eta))\}$. A Gibbs sample $\mu^{(s)}$ of μ is obtained by drawing $1/\mu^{(s)}$ from a gamma distribution with shape $C^{(s-1)}$ and scale $\left(\sum_{c=1}^{C} (\sigma_c^2)^{(s-1)}\right)^{-1}$.

To generate a conditional posterior sample of the error variances $\sigma_1^2, \ldots, \sigma_n^2$ for state s of the MCMC chain, Algorithm 8 of [25] is implemented, under $m = 1$. This algorithm involves two steps, first, sample the clusters of the error variances, and second, sample the error variances given the

updated clusters. In the first step, for $i = 1, \ldots, n$, a sample $(\sigma^2_{C+1})^{(s)}$ is generated from Exponential $(\mu^{(s)})$, setting $(\sigma^2_{C+1})^{(s)} = (\sigma^2_i)^{(s-1)}$ if $(\sigma^2_i)^{(s-1)}$ is unique among all the error variance parameters $(\sigma^2_i)^{(s-1)}, i = 1, \ldots, n$. Then, for $i = 1, \ldots, n$, subject i is assigned to a cluster $c_i^{(s)}$ using the following probabilities, which are given up to constant of proportionality by:

$$P(c_i^{(s)} = c \,|\, y_i, c^{(s)}_{h \neq i}, (\sigma^2_c)^{(s-1)}, c = 1, \ldots, C^{(s-1)}, (\sigma^2_{C+1})^{(s)})$$

$$\propto \begin{cases} \dfrac{n^{(s-1)}_{-i,\,c}}{n-1+\alpha^{(s)}} \text{normal}\,(y_i \,|\, g^{(s)}(\mathbf{x}_i'\boldsymbol{\omega}^{(s)}), (\sigma^2_c)^{(s-1)}), \; c = 1, \ldots, C^{(s-1)}, \\ \dfrac{\alpha^{(s)}}{n-1+\alpha^{(s)}} \text{normal}\,(y_i \,|\, g^{(s)}(\mathbf{x}_i'\boldsymbol{\omega}^{(s)}), (\sigma^2_{C+1})^{(s)}), \end{cases}$$

with $n^{(s-1)}_{-i,\,c}$ the number of the n observations sharing the same value of the error variance $(\sigma_c)^{(s-1)}$ excluding case i. At the conclusion of this first step, a new set of distinct clusters of the n error variances is established for state s of the MCMC chain, and they are denoted by $(\sigma^2_c)^{(s)}, c = 1, \ldots, C^{(s)}$.

Then in the second step, for each cluster c, the error variances are updated to $(\sigma^2_1)^{(s)}, \ldots, (\sigma^2_n)^{(s)}$, by σ^2_i generating a sample of each σ^2_i from a conditional posterior density that is proportional to:

$$\prod_{h \in c} \text{normal}\,(y_h \,|\, g^{(s)}(\mathbf{x}_h'\boldsymbol{\omega}^{(s)}), (\sigma^2_c)^{(s)}) \,\text{ex}\,((\sigma^2_c)^{(s)} \,|\, \mu^{(s)}),$$

assuming i is a member of cluster c, with $\text{ex}(\cdot \,|\, \mu)$ denoting the exponential density function with mean μ. An adaptive random walk Metropolis Hastings algorithm is implemented to generate a sample of the error variances, which is based on a normal proposal distribution with mean $\log((\sigma^2_c)^{(s)})$, and variance $\tau^{(s-1)}_\sigma$ that is adapted to approximate a .44 acceptance rate over the $C^{(s)}$ distinct values of the error variances. At the conclusion of each state s of the MCMC chain $(s = 1, \ldots, S)$, the proposal variance is updated to $\tau^{(s)}_\sigma = \exp\,(\log(\tau^{(s-1)}_\sigma) - \min\{.01, s^{-1/2}\})$ if less than a

proportion .44 of the $C^{(s)}$ distinct error variances were accepted, otherwise, the update is $\tau_{\sigma}^{(s)} = \exp\,(\,\log(\tau_{\sigma}^{(s-1)}) + \min\{.01,\, s^{-1/2}\})$ when the acceptance rate is at least .44 (see [28]). As before, this updating occurs only in the first set of initial "burn-in" samples of the MCMC algorithm.

3.5. Sampling from the posterior predictive distribution

An informal but informative approach to evaluating the goodness-of-fit of the heteroscedastic single-index model is based on the inference of the residual posterior predictive density

$$p((y_i - z_i)\,|\,Data, \mathbf{x}_i) = \int (y_i - z_i) f(z_i\,|\,Data, \mathbf{x}_i)\,\mathrm{d}z_i$$

for each observation $y_i (i = 1, \ldots, n),$ with

$$f(z_i\,|\,Data, \mathbf{x}_i) = \int \mathrm{normal}(z_i\,|\,g(\mathbf{x}_i'\boldsymbol{\omega}), \sigma_i^2) f(\boldsymbol{\theta}\,|\,Data)\,\mathrm{d}\boldsymbol{\theta} \tag{3}$$

denoting the posterior predictive density of $y,$ and $f(\boldsymbol{\theta}\,|\,Data)$ is the posterior density of the vector of parameters of the single-index model. In particular, the 99% confidence region of the residual posterior predictive density $f((y_i - z_i)\,|\,Data, \mathbf{x}_i)$ can be compared against 0, such that the response y_i is viewed as fitting the single-index model if this interval contains zero, and the response is viewed as an outlier when this interval does not include zero. Inference of the residual posterior predictive density $f((y_i - z_i)\,|\,Data, \mathbf{x}_i)$ can be obtained by adding one simple step to the MCMC algorithm described above. In particular, for each state s of the MCMC chain $(s = 1, \ldots, S),$ and each observation indexed by $i = 1, \ldots, n,$ a MCMC sample $(y_i - z_i^{(s)})$ from the residual posterior predictive density is based on a sample $z_i^{(s)}$ from $\mathrm{Norrmal}\,(g^{(s)}(\mathbf{x}_i'\boldsymbol{\omega}^{(s)}), (\sigma_i^2)^{(s)}),$ given the current state of the parameters $\boldsymbol{\omega}^{(s)}, \boldsymbol{\beta}^{(s)},$ and $(\sigma_i)^{(s)}$ in the MCMC chain. Of course, the MCMC samples $z_i^{(s)}$ from $\mathrm{Norrmal}\,(g^{(s)}(\mathbf{x}_i'\boldsymbol{\omega}^{(s)}), (\sigma_i^2)^{(s)}), s = 1, \ldots, S,$ enable the inference of other quantities that may be of interest from the posterior predictive density, such as the expectation $\mathrm{E}\,[Z_i\,|\,Data, \mathbf{x}_i]$ and the variance $\mathrm{var}\,[Z_i\,|\,Data, \mathbf{x}_i].$

4. Conclusions

In this paper we introduced a heteroscedastic single-index model, where the link function is modeled by splines, and heteroscedasticity is modeled nonparametrically by a Dirichlet Process prior. Straightforward MCMC methods are available for sampling the posterior distribution of this model, and this model has demonstrated improved predictive accuracy over the homoscedastic single-index model which has received much attention in previous research. An important feature of the heteroscedastic model is the $\text{Gamma}(\tau_1, \tau_2)$ hyperparameter for the ridge parameter ν, useful for regularizing the spline coefficients. In the current paper, it was shown that the $\text{Gamma}(1, 1/2)$ worked rather well for simulated data and some real data sets. Of course, the $\text{Gamma}(1, 1/2)$ prior may not work for all data sets, and so in practice, it is important to evaluate the impact of the choice of hyperparameters τ_1 and τ_2 on the predictive accuracy of the heteroscedastic single-index model.

References

[1] A. Antoniadis, G. Grégoire, and I. McKeague, bayesian estimation in single-index models, Statistica Sinica 14 (2004), 1147-1164.

[2] Y. Atchadé, and J. Rosenthal, On adaptive Markov Chain Monte Carlo algorithms, Bernoulli 11 (2005), 815-828.

[3] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, 1961.

[4] S. Bruntz, W. Cleveland, B. Kleiner and J. Warner, The dependence of ambient ozone on solar radiation, temperature, and mixing height, Proceedings of the Symposium on Atmospheric Diffusion and Air Pollution, American Meterological Society, Boston, 1974, pp. 125-128.

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

[6] W. Cleveland, S. Devlin and E. Grosse, Regression by local fitting: methods, properties, and computing, Journal of Econometrics 37 (1988), 87-114.

[7] M Delecroix, P. Hall and V.-R, Test des modéles á direction révélatrice unique, Abstracts of the 34th meeting of the French Statistical Society, 2002, 1999, pp. 364-365.

[8] M. Delecroix and M. Hristache, M-estimateurs semiparamétriques dans les modèles à direction révélatrice unique, Bulletin of the Belgian Mathematical Society 6 (1999), 161-185.

[9] M. Escobar, nonparametric Bayesian methods in hierarchical models, Journal of Statistical Planning and Inference 43 (1995), 97-106.

[10] M. Escobar and M. West, Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association 90 (1995), 577-588.

[11] J. Friedman and W. Stuetzle, projection pursuit regression, Journal of the American Statistical Association 76 (1981), 817-823.

[12] W. Härdle, , M. J. and A. Tsybakov, , bandwidth choice for average derivative estimation, Journal of the American Statistical Association 87 (1992), 218-226.

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

[14] W. Härdle, V. Spokoiny and S. Sperlich, Semiparametric single index versus fixed function modelling, Annals of statistics 25 (1997), 212-243.

[15] W. Härdle and T. Stoker, Investigating smoothing multiple regression by the method of average derivatives, Journal of the American Statistical Association 84 (1989), 986-995.

[16] W. Härdle and A. Tsybakov, How sensitive are average derivatives? Journal of Econometrics 58(1993), 31-48.

[17] T. Hastie and R. Tibshirani, Generalized additive models (with discussion), Statistical Science 1 (1986), 297-318.

[18] J. Horowitz and W. Härdle, Direct semiparametric estimation of single-index models with discrete covariates, Journal of the American Statistical Association 91 (1996), 1632-1640.

[19] M. Hristache, A. Juditsky and V. Spokoiny, Direct estimation of the index coefficients in a single-index model, Annals of Statistics 29 (2001), 595623.

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

[21] R. Klein and R. Spady, An efficient semiparametric estimator for binary response models, Econometrica 61 (1993), 387-412.

[22] P. Laud and J. Ibrahim, Predictive model selection, Journal of the Royal Statistical Society, Series B 57 (1995), 247-262.

[23] K. Li, sliced inverse regression for dimension reduction, Journal of the American Statistical Association 86 (1991), 316-342.

[24] P. Müller and F. Quintana, Nonparametric Bayesian data analysis, Statistical Science 19 (2004), 95-110.

[25] R. Neal, Markov chain sampling methods for Dirichlet process mixture models, Journal of Computational and Graphical Statistics 9 (2000), 249-265.

[26] A. O'Hagan and J. Forster, Kendall's Advanced Theory of Statistics: Bayesian Inference, Vol. 2B, Arnold, London, 2004.

[27] J. Powell, J. Stock and T. Stoker, semiparametric estimation of index coefficients, Econometrica 57 (1989), 1403-1430.

[28] G. Roberts and J. Rosenthal, Optimal scaling of various Metropolis-Hastings algorithms, Statistical Science 16 (2001), 351-367.

–, Examples of adaptive MCMC, Tech. Rep., University of Toronto, Department of Statistics, 2006.

–, Coupling and ergodicity of adaptive MCMC, Journal of Applied Probability 44 (2007), 458-475.

[29] D. Ruppert, Selecting the number of knots for penalized splines, Journal of Computational and Graphical Statistics 11 (2002), 735-757.

[30] A. Samarov, Exploring regression structure using nonparametric functional estimation, Journal of the American Statistical Association 88 (1993), 836-847.

[31] J. Saw, A family of distributions on the *M*-sphere and some hypothesis tests, Biometrika 65 (1978), 69-73.

[32] J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica 4 (1994), 639-650.

[33] T. Stoker, Consistent estimation of scaled coefficients, Econometrica 54 (1986), 1461-1481.

[34] L. Tierney, Markov chains for exploring posterior distributions, Annals of Statistics 22 (1994), 1701-1728.

[35] B. Turlach, Fast implementation of density-weighted average derivative estimation," in Computing Science and Statistics, proceedings of the symposium on the interface (26th), computationally intensive statistical methods, eds. E. Wegman, J. S. and A. Lehman, fairfax station, VA: Interface foundation of North America vol. 26 (1994), pp. 28-33.

[36] W. Venables and B. Ripley, Modern applied statistics with S-PLUS, Springer, New York, 1999.

[37] S. Walker, P. Damien, P. Laud and A. Smith, Bayesian nonparametric inference for random distributions and related functions, Journal of the Royal Statistical Society, Series B 61 (1999), 485-527.

[38] Y. Xia, W. K. Li, H. Tong and D. Zhang, A goodness-of-fit test for single-index models (with discussion), Statistica Sinica 14 (2004), 1-39.

[39] Y. Xia, H. Tong, W. Li and L. Zhu, An adaptive estimation of dimension reduction space (with discussion), Journal of the Royal Statistical Society Series B 64 (2002), 363-410.

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