Statistica Sinica **27** (2017), 43-69
doi:http://dx.doi.org/10.5705/ss.202015.0096

BAYESIAN ANALYSIS OF SHAPE-RESTRICTED FUNCTIONS USING GAUSSIAN PROCESS PRIORS

Peter J. Lenk and Taeryon Choi

University of Michigan and Korea University

Abstract: This paper proposes a Bayesian method to estimate shape-restricted functions using Gaussian process priors. The proposed model enforces shape-restrictions by assuming that the derivatives of the functions are squares of Gaussian processes. The resulting functions, after integration, are monotonic, monotonic convex or concave, U–Shaped, and S–shaped. The latter two allow estimation of extreme points and inflection points. The Gaussian process's covariance function has hyper parameters to control the smoothness of the function and the tradeoff between the data and the prior distribution. The Bayesian analysis of these hyper parameters provides a data–driven method to identify the appropriate amount of smoothing. The posterior distributions of the proposed models are consistent. We modify the basic model with a spike-and-slab prior that improves model fit when the true function is on the boundary of the constraint space. We also examine Bayesian hypothesis testing for shape restrictions and discuss its potentials and limitations. We contrast our approach with existing Bayesian regression models with monotonicity and concavity and illustrate the empirical performance of the proposed models with synthetic and actual data.

Key words and phrases: Adaptive Markov chain Monte Carlo, isotonic regression, Karhunen-Loève expansion, lasso, model choice, semiparametric regression, shape restriction, smoothing, spectral representation.

1. Introduction

Shape constrained regression models arise naturally in a wide variety of applications: children grow taller; star light intensity decreases with distance given fixed luminosity; demand for electricity increases as temperatures depart from 68°F; and an occupation's prestige tends to increase with its salary and educational requirements. Researchers often assume that such shape restrictions as monotonicity and convexity are known *a priori* or plausible in theory. Semiparametric Bayesian models express these *a priori* shape constraints in the prior distribution of the unknown function. Neelon and Dunson (2004) and Cai and Dunson (2007) developed methods for Bayesian isotonic regression using piecewise linear models and monotone splines based on order-restricted inference. Other approaches to shape constraints are Brezger and Steiner (2008) with penalized splines, Wang (2008) with free-knot monotone cubic splines, Bornkamp and Ickstadt (2009) with mixtures of cumulative distribution functions, Shively, Sager and Walker (2009) and Shively, Walker and Damien (2011) with restricted splines, Curtis and Ghosh (2011) with Bernstein polynomials, and Lin and Dunson (2014) using the projection of the posterior under the Gaussian process. Recently, Bayesian shape-restricted regression models have been extended to generalized partial linear models where regression functions are decomposed into a nonparametric function with a shape constraint and a parametric function. Meyer, Hackstadt and Hoeting (2011) proposed a Bayesian approach to partial linear models using regression splines with assumptions about shape and smoothness based on the shape-restricted regression splines of Ramsay (1998) and Meyer (2008).

Most of the existing literature on Bayesian monotone regression puts constraints on the coefficients of the basis functions through prior distributions. This paper enforces shape restrictions by assuming that the derivatives of the functions are squared, Gaussian processes. This representation naturally incorporates monotone, monotone convex or concave, U–shaped, and S–shaped restrictions. The proposed model results in consistent posterior distributions. We use the spectral representation of the Gaussian process prior to simplify the posterior analysis. The partially linear model can include multiple functions with different shape restrictions. We also examine a spike-and-slab prior to model functions on the boundary of the constraint space. We illustrate the empirical performance of the proposed model based on simulation studies and two data applications. In addition, we examine Bayesian hypothesis testing for shape restrictions based on the marginal distribution of the data and discuss its potentials and limitations.

The frequentist literature on isotonic regression is vast. Ayer et al. (1955) and Brunk (1955) formulated isotonic regression as a constrained optimization problem: minimize weighted sums-of-squares error subject to ordering of the function at the observations. This early work inspired research into constrained optimization algorithms (Barlow and Brunk (1972), Dykstra and Robertson (1982), Robertson, Wright and Dykstra (1988), Groeneboom and Wellner (2014), and Luss and Rosset (2014)). Two alternative approaches are isotonic splines (Ramsay (1998) and Wang and Li (2008)) and kernel methods (Mammen (1991), Mukarjee and Stern (1994), Hall and Huang (2001), and Dette and Pilz (2006)). Bhattacharya and Lin (2010, 2011) propose data adaptive methods, and Bhattacharya and Lin (2013) perform small sample simulations that compare their data-adaptive method to splines and kernels. Other variations are isotonic median regression (Menendez and Salvador (1987)), local averaging with isotonic regression (Friedman and Tibshirani (1984)), nearly isotonic regression (Tibshirani, Hoefling and Tibshirani (2011)), LASSO applied to multiple isotonic regression functions (Fang and Meinshausen (2012)), and multivariate isotonic regression (Sasabuchi, Inutsuka and Kulatunga (1983)). Wu, Meyer and Opsomer (2015) impose a penalty on the range of the regression function to mitigate "spiking" at the end of the end of the estimation interval. The differences and similarity between frequentist and Bayesian approaches are too numerous to list here and depend on where to draw the line between the two. In a narrow sense, the shape restrictions are part of the likelihood function and the Bayesian and frequentist only differ on the estimation method of the parameters. In a wide sense, shape restrictions are *a priori* beliefs imposed by the researcher, and frequentists become more similar to Bayesians.

The remainder of the paper is organized as follows. Section 2 reviews Gaussian process priors and their spectral representation, which we call "Bayesian spectral analysis regression" (BSAR), and discusses smoothing priors for BSAR. Section 3 imposes functional constraints by assuming that the positive or negative square roots of the first or second derivatives have Gaussian process priors. Section 4 illustrates of the performance of the proposed method with simulation studies and applications. The empirical results compare existing methods with the proposed approach. Section 5 concludes with a discussion.

2. Bayesian Spectral Analysis Regression

The observational equation is a partially linear, semi–parametric model:

$$Y_i = \boldsymbol{w}_i^\intercal \boldsymbol{\beta} + \sum_{k=1}^{K} f_k(x_{i,k}) + \epsilon_i,\ \ i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{\beta}$ are $p+1$–dimensional vectors of covariates and coefficients; f_k is an unknown function of the scalar $x_{i,k}$; and the error terms $\{\epsilon_i\}$ are a random sample from a normal distribution, $N(0, \sigma^2)$. Without loss of generality, we assume that $0 \leq x_{i,k} \leq 1$. The covariates $\boldsymbol{w}$ do not include functions of the x's, which are also not functions of each other. The Y–intercept β_0 is included in $\boldsymbol{\beta}$. The model is unidentified because constants can be added and subtracted to β_0 and the f_k's without changing the distribution of Y. We assume that the f_k are mean centered and orthogonal to the constant function to identify the model,

$$\int_0^1 f_k(x)dx = 0 \text{ for } k = 1, \ldots, K. \tag{2.2}$$

This identification restriction reduces the posterior correlation between β_0 and f_k and can significantly reduce the posterior uncertainty of f_k.

Without shape–restrictions, f_k has a Gaussian process prior, and the different shape–restricted models alter this prior specification. We first review nonparametric regression with Gaussian process priors. Section 2.1 presents the

Karhunen-Loève representation for Gaussian processes. The representation linearizes the covariance function and simplifies the posterior analysis. Section 2.2 discusses smoothing priors and their derivation from the Gaussian process and contrasts them to an exchangeable Lasso prior. In Sections 2 to 4 we restrict our attention to one ($K = 1$) unknown function to simplify the presentation. The model with multiple unknown functions is a straightforward extension, and we use it for multiple components f_k, $k = 1, \ldots, K(> 1)$ in an empirical example in Section 5.

2.1. Spectral analysis of Gaussian process priors

Gaussian processes provide a natural method to specify distributions on the space of functions for nonparametric regression (O'Hagan (1978), Wahba (1978), and Rasmussen and Williams (2006)). Based on the observational equation (2.1) without shape restrictions, the prior for f is $f(x) = Z(x)$ where Z is a second-order Gaussian process with mean function equal to zero and covariance function $\nu(s,t) = E[Z(s)Z(t)]$ for $s,t \in [0,1]$. The covariance function acts as a smoothing kernel, and the posterior distribution of Z requires inverting $n \times n$ matrices with entries $\nu(x_i, x_j)$. Lenk (1999) linearizes the covariance kernel with the Karhunen-Loève representation for Z (Grenander (1981) and Wahba (1990)). The Bayesian spectral analysis regression (BSAR) model expresses the Gaussian process as an infinite series expansion with the Karhunen-Loève representation

$$Z(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x), \tag{2.3}$$

where $\{\varphi_j\}$ forms an orthonormal basis on $[0,1]$. The implied prior distribution of the spectral coefficients are normal and mutually independent. Their means are zero, and their variances are: $\nu_j^2 = \int_0^1 \int_0^1 \nu(s,t)\varphi_j(s)\bar{\varphi}_j(t) ds\, dt$ where $\sum_{j=0}^{\infty} \nu_j^2 < \infty$. Conversely, the covariance function is $\nu(s,t) = \sum_{j=0}^{\infty} \nu_j^2 \varphi_j(s)\bar{\varphi}_j(t)$ provided $\sum_{j=0}^{\infty} \nu_j^2 < \infty$.

Our choice of orthonormal system is the cosine basis function on $[0,1]$: $\varphi_0(x) = 1$ and $\varphi_j(x) = \sqrt{2}\cos(\pi j x)$ for $j \geq 0$. By excluding sines, f does not have to be periodic on $[0,1]$, and all piecewise continuous function can be expressed as an infinite series of $\{\varphi_j\}$. The cosine basis has a natural ordering based on their frequencies. This ordering relates to nonparametric estimation: smooth functions have small weights on high–frequency components. If the coefficients decay at rate $o(c^j)$ for some $c > 1$, then the function is j times differentiable almost everywhere (Katznelson (2004)). For unrestricted f, θ_0 is confounded with the Y–intercept β_0, and we will drop it from the representation for f.

More generally, if the support of x is S, then the orthonormal basis can be defined as: $\varphi_0(x) = \sqrt{q(x)}$ and $\varphi_j(x) = \sqrt{2q(x)}\cos[\pi j Q(x)]$ where Q is a cumulative distribution function with support S, and q is its density. We use the uniform distribution on $S = [a, b]$ for the empirical analysis in Section 4. Other orthonormal basis functions can be used without loss of generality, but they may not have a natural ordering that relates to a smoothing prior.

In implementations we need to truncate the infinite series to a finite sum $Z_J(x) = \sum_{j=0}^{J} \theta_j \varphi_j(x)$. The mean integrated squared error (MISE) between Z and Z_J decreases in J:

$$\mathrm{MISE}(J) = E\left(\int_0^1 [Z(x) - Z_J(x)]^2 \, dx\right) = \sum_{j=J+1}^{\infty} \nu_j^2, \tag{2.4}$$

and can be made as small as desired because the sum of the variances are finite. An important point is that if the prior distribution of θ_j is inherited from Z by the spectral representation, then the choice of J does not make a material difference to the accuracy of estimating f for sufficiently large J because the truncation error becomes negligible. However, if the spectral coefficients are *a priori* exchangeable, then Z_J is a proper distribution but a limiting second order Gaussian process does not exist. In this case, the choice of J is critical for estimating f. The next subsection takes a closer look at these issues and the specification of prior distributions of BSAR.

2.2. Smoothing priors

Our approach is similar to that of Young (1977) who puts a prior distribution on the coefficients of polynomials such that the coefficients tend to decrease as the power increases. The rate at which their prior variances $\{\nu_j^2\}$ converge to zero determines the smoothness of the sample paths of Z. Lenk (1991, 1993, 2003) for nonparametric density estimation and Lenk (1999) and Choi, Lee and Roy (2009) for nonparametric regression use exponentially decreasing variances. Then the sample paths of Z are piecewise analytic. We use scale–invariant priors:

$$\theta_j|\sigma, \tau, \gamma \sim \mathrm{N}(0, \sigma^2\tau^2 \exp[-j\gamma]) \text{ for } j \geq 1 \text{ and } \gamma > 0. \tag{2.5}$$

The prior distribution on the spectral coefficients shrinks their estimators towards zero, and the amount of shrinkage increases for higher frequency terms. Increasing τ puts more weight on the likelihood function and less on the prior, and increasing γ increases the rate that the prior variances approach zero. The scale-invariant prior facilitates implementation of the methodology. Multiplying Y by a constant does not alter these prior specifications, and τ has the interpretation of a signal–to–noise ratio. In most applications, the signal–to–noise ratio is likely to be between 1 than 10. Without scale-invariant priors, reasonable priors for τ depend on the variation in the data. The MISE between Z and Z_J in Equation (2.4) decreases exponentially in J: $\mathrm{MISE}(J) \propto \exp[-(J + 1)\gamma]$.

We specify two prior distributions for τ and γ. The T Smoother uses the the inverse Gamma distribution($X \sim IG(a, b)$ has mean $\mu = b/(a-1)$ and variance $\mu^2/(a-2)$ for $a > 2$ and $b > 0$) for τ^2 and the exponential prior for γ:

$$\text{T Smoother: } \tau^2 \sim \text{IG}\left(\frac{r_{0,\tau}}{2}, \frac{s_{0,\tau}}{2}\right) \text{ and } \gamma \sim \text{Exp}(w_0). \tag{2.6}$$

After integrating over τ^2, the spectral coefficients θ_j have a multivariate T-distribution. The Lasso Smoother uses exponential prior distributions for both τ^2 and γ:

$$\text{Lasso Smoother: } \tau^2 \sim \text{Exp}(u_0) \text{ and } \gamma \sim \text{Exp}(w_0). \tag{2.7}$$

After integrating over τ, the spectral coefficients have a multivariate, double exponential distribution (Eltoft, Kim and Lee (2006)), that is the Bayesian Lasso model (Park and Casella (2008)). To complete the model specification, we use the conjugate prior distributions for $\boldsymbol{\beta}$ and σ:

$$\boldsymbol{\beta}|\sigma \sim \text{N}(\boldsymbol{m}_{0,\beta}, \sigma^2 \boldsymbol{V}_{0,\beta}) \text{ and } \sigma^2 \sim \text{IG}\left(\frac{r_{0,\sigma}}{2}, \frac{s_{0,\sigma}}{2}\right).$$

Next, we compare the T and Lasso Smoothers to an exchangeable Lasso prior to demonstrate the importance of deriving the variance specification of the spectral coefficients from the Gaussian process covariance function. Given Lasso's ability to handle $p >> n$, one may ask if we need smoothing priors at all. The exchangeable Lasso prior is:

$$\text{Lasso Prior: } \tau^2 \sim \text{Exp}(u_0) \text{ and } \gamma = 0. \tag{2.8}$$

Then Z_J is a Gaussian process prior but does not converge to a second order Gaussian process. A simulation demonstrates that the Lasso Prior over–fits the data because it does not impose smoothing constraints on the spectral coefficients, while both the T Smoother and Lasso Smoother correctly identify f. We generated 100 observations from:

$$Y = 120 + 5x + 10\sqrt{x}\left\{1 + 2\exp\left[-0.1(x-2)^2\right] - \exp\left[-0.5(x-5)^2\right]\right\} + \epsilon,$$

where $\epsilon \sim$ N(0,100) and $x \sim U(0, 10)$, the uniform distribution. We used the uniform cdf transform, $Q(x) = x/10$, to map $[0, 10]$ to $[0, 1]$, and set $J = 100$. Table 1 presents fit statistics between the posterior means and true values using the three prior distributions. We used Gelfand and Dey (1994) to approximate the Log Integrated Likelihood (LIL). The RMISE for f and RMSE for the spectral coefficients for the Lasso Prior are around four times larger than the T and Lasso Smoothers. The exchangeable Lasso Prior has the largest R-Square, which indicates that the Lasso Prior over-fits the data.

Table 1. Fit statistics for smoothing and Lasso priors.

	T Smoother	Lasso Smoother	Lasso Prior
Log Integrated Likelihood	-507.6471	-607.4359	-520.5975
R-Square	0.7354	0.7229	0.8910
RMISE for f	5.8524	9.5384	37.561
RMSE for Spectral Coefficients	0.5889	0.9550	3.7375

RMISE is the root mean integrated squared error between the true f and its posterior mean. RMISE is approximated with Simpson's rule using 201 intervals. RMSE is the root mean squared error between the true θ and their posterior means.

Figure 1 confirms that the Lasso prior mistakes noise for signal. The posterior mean for the Lasso Prior almost connects the observations, while the T and Lasso Smoothers effectively recover the true f. The figure also plots the true and estimated spectral coefficients for frequencies $1 \leq j \leq 50$. The behavior for $j > 50$ extends the trends in the graphs. The true coefficients rapidly approach zero for frequencies larger than 10. The additional uncertainty in the Lasso prior for high frequency spectral coefficients allows the posterior mean of f to over-fit the data. Due to the slightly better performance of the T Smoother over the Lasso Smoother, we will present results for the T Smoother in the rest of the paper. Also, the Lasso Smoother requires an additional Metropolis step.

3. Bayesian Shape-Restricted Spectral Analysis Regression

Section 3.1 considers monotonic functions and monotonic convex or concave functions, and Section 3.2 develops "S"–shaped and "U"–shaped functions. Section 3.3 modifies the model to include functions on the boundary of the constraint space by using spike-and-slab priors.

3.1. Restrictions on sample path derivatives

The q^{th} derivative of f is the square of a Gaussian process:

$$f^{(q)}(x) = \delta Z^2(x) \text{ for } \delta = 1 \text{ or } -1 \text{ and } q = 1 \text{ or } 2, \tag{3.1}$$

where δ and q are given by the user. Higher order derivatives ($q > 2$) are possible, but have limited application. The marginal prior distribution of $Z^2(x)$ is a scaled, Chi-squared distribution with one degree of freedom because the prior mean is zero. To distinguish this model from the spike-and-slab prior in Section 3.3, we call it the "Gamma Prior."

When q is 1, f is monotone:

$$f(x) = \delta \left[\int_0^x Z^2(s)ds - \int_0^1 \int_0^x Z^2(s)ds\, dx \right], \tag{3.2}$$

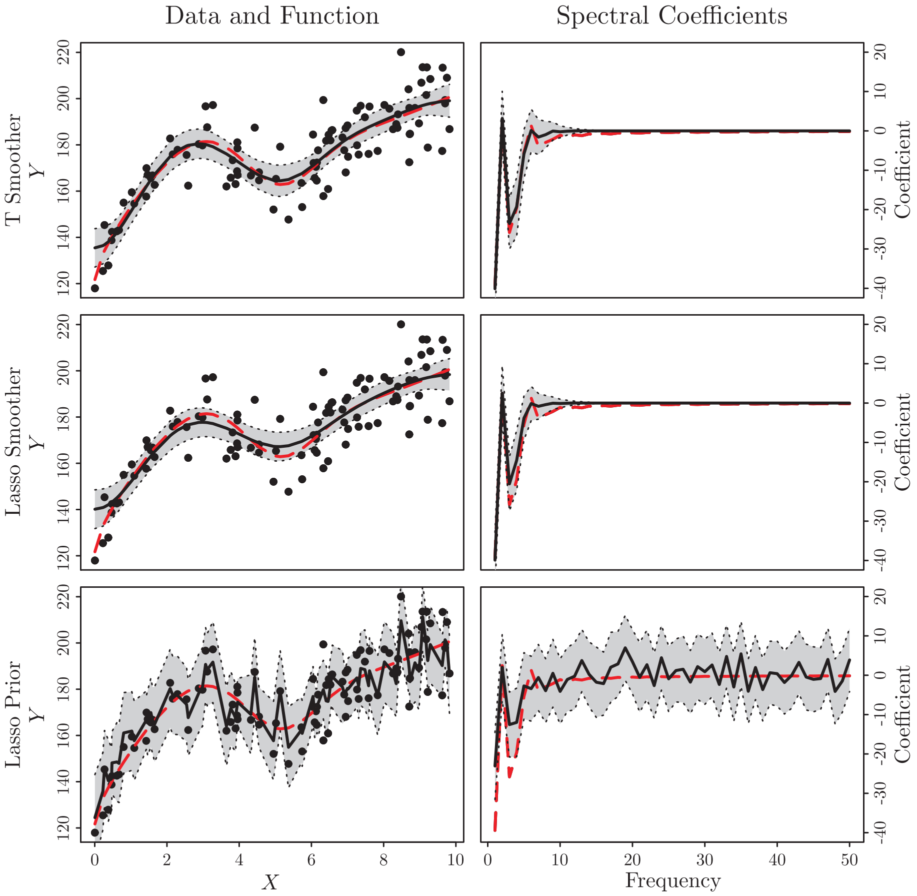

Figure 1. Impact of priors on posterior distributions. Dots are Y observations; dashed lines are the true values; solid lines are posterior means; and dotted lines and shaded areas are 95% credible intervals.

where the last term is the constant of integration that satisfies the mean–centering condition at (2.2). The function is non-decreasing when $\delta = 1$ and is non-increasing when $\delta = -1$.

When q is 2, f is a non-decreasing and convex function when $\delta = 1$ or a non-increasing and concave function when $\delta = -1$:

$$f(x) = \delta \left[\int_0^x \int_0^s Z^2(t)dt\, ds - \int_0^1 \int_0^x \int_0^s Z^2(t)dt\, ds\, dx \right] + \alpha(x - 0.5). \quad (3.3)$$

Here the second term and α are constants of integration and make f satisfy the mean–centering condition. To ensure monotonicity, $\delta\alpha \geq 0$.

The first and second derivatives of (3.3) have the same sign. Reversing the range of x in the integrals produces functions where the first and second derivatives have opposite signs. The model for non-decreasing and concave functions ($\delta = 1$) or non-increasing and convex functions ($\delta = -1$) is

$$f(x) = -\delta \left[\int_0^{1-x} \int_0^s Z^2(t) dt\, ds - \int_0^1 \int_0^{1-x} \int_0^s Z^2(t) dt\, ds\, dx \right] + \alpha(x - 0.5), \tag{3.4}$$

where $\delta\alpha \geq 0$.

Our theorem provides a class of functions with sample paths of the q^{th} derivative almost surely piecewise continuous and positive. Theorem 1 is proven using an argument similar to Shively, Sager and Walker (2009) and Ramsay (1998); it is given in the Supplementary Material.

Theorem 1. *Let q be a positive integer. Let the class of $C^{(q)+}$ be the class of functions*

$$C^{(q)+} = \{f(x) |\ f^{(q)}(x) \textit{ exists},\ f^{(q)}(x) \textit{ is piecewise continuous}, \\ f^{(q)}(x) \geq 0,\ x \in [0,1]\},$$

and let $C^1[0,1]$ denote the class of piecewise continuous functions on $[0,1]$. Then, $f(x) \in C^{(q)+}$ if and only if there exists $u(x) \in C^1[0,1]$ such that $f^{(q-1)}(x) = a + \int_0^x u^2(t)dt$ where $f^{(0)} \equiv f$.

The cases of negative derivatives or opposite signs for the first and second derivatives follow by trivial modifications.

Next, we consider the spectral representation for Z. Unlike the unconstrained case of BSAR at (2.3), we include the constant function and θ_0 because the effect of θ_0 can be more complex than shifting the sample path of Z^2, and it is not confounded with the intercept β_0. The spectral coefficients have a sign indeterminacy because multiplying them by minus one results in the same f. We identify the model by assuming that $\theta_0 \geq 0$. We modify the scale invariant prior for θ_j at (2.5) by replacing σ^2 with σ. The priors for the spectral coefficients are

$$\theta_0|\sigma \sim N(0, \sigma v_{\theta_0}^2) I(\theta_0 \geq 0), \text{and } \theta_j|\sigma, \tau, \gamma \sim N(0, \sigma\tau^2 \exp[-j\gamma]). \tag{3.5}$$

The prior for θ_0 is the truncated normal distribution. In addition, to ensure that $f^{(1)}$ is positive or negative for the convex or concave cases, we truncate the prior distribution of α,

$$\alpha|\sigma \sim N(m_{0,\alpha}, \sigma^2 v_{0,\alpha}^2) I(\delta\alpha \geq 0).$$

Our empirical experience has been that estimating the linear term α often does not improve the accuracy of the estimator for f. In the empirical section, we set $\alpha = 0$ to focus attention on the Gaussian process.

Using the spectral representation of Z at (2.3) gives BSAR with monotone constraints (BSARM) at (3.2):

$$\begin{aligned} f(x) &= \delta \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \theta_j \theta_k \varphi_{j,k}^{a}(x), \qquad (3.6)\\ \varphi_{j,k}^{a}(x) &= \int_0^x \varphi_j(s)\varphi_k(s)ds - \int_0^1 \int_0^s \varphi_j(t)\bar{\varphi}_k(t)dt\,ds \text{ for } j,k \geq 0. \end{aligned}$$

Replacing Z with the truncated Z_J gives a quadratic form for f_J:

$$f_J(x) = \delta \boldsymbol{\theta}_J^{\intercal} \boldsymbol{\Phi}_J^{a}(x) \boldsymbol{\theta}_J, \qquad (3.7)$$

where $\boldsymbol{\theta}_J$ is the $J+1$ vector of spectral coefficients, and $\boldsymbol{\Phi}_J^a(x)$ is a $J+1 \times J+1$ matrix with (j,k) entry $\varphi_{j,k}^a(x)$. The online appendix displays the integrated, mean–centered basis functions $\varphi_{j,k}^a$.

BSAR with monotone convexity or concavity (BSARMC) at (3.3) where the first and second derivatives have the same sign becomes

$$\begin{aligned} f(x) &= \delta \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \theta_j \theta_k \varphi_{j,k}^{b}(x) + \alpha(x-0.5), \qquad (3.8)\\ \varphi_{j,k}^{b}(x) &= \int_0^x \int_0^s \varphi_j(t)\varphi_k(t)dt\,ds - \int_0^1 \int_0^x \int_0^s \varphi_j(t)\varphi_k(t)dt\,ds\,dx. \end{aligned}$$

Truncating the infinite sum at J terms gives

$$f_J(x) = \delta \boldsymbol{\theta}_J^{\intercal} \boldsymbol{\Phi}_J^{b}(x) \boldsymbol{\theta}_J + \alpha(x-0.5), \qquad (3.9)$$

where $\boldsymbol{\Phi}_J^b(x)$ is the $J+1 \times J+1$ matrix with (j,k) entries $\varphi_{j,k}^b(x)$. The online appendix displays the integrated, mean–centered, basis functions $\varphi_{j,k}^b$.

When the first and second derivative of BSARMC at (3.4) have opposite signs, the integrated basis functions are given by

$$\begin{aligned} f(x) &= -\delta \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \theta_j \theta_k \varphi_{j,k}^{c}(x) + \alpha(x-0.5), \qquad (3.10)\\ \varphi_{j,k}^{c}(x) &= \varphi_{j,k}^{b}(1-x). \qquad (3.11) \end{aligned}$$

The finite sum representation for f is

$$f_J(x) = -\delta \boldsymbol{\theta}_J^{\intercal} \boldsymbol{\Phi}_J^{c}(x) \boldsymbol{\theta}_J + \alpha(x-0.5), \qquad (3.12)$$

where $\boldsymbol{\Phi}_J^c(x) = \boldsymbol{\Phi}_J^b(1-x)$.

The proposed approach at (3.1) does not introduce restrictions on the spectral coefficients. Thus, posterior consistency follows from the Kullback-Leibler property of Gaussian process prior by the integral representation of f and the existence of uniformly consistent tests, summarized in Theorem 2. To discuss posterior consistency, we consider the nonparametric regression model $Y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, for simplicity, neighborhoods H_ϵ and $W_{\epsilon,n}$, the Hellinger neighborhood and the $L_1(Q_n)$-neighborhood of the true value of parameter $\vartheta_0 = (f_0, \sigma_0)$ respectively, defined as follows. For every $\epsilon > 0$, $H_\epsilon = \{\vartheta = (f, \sigma) \ : \ d_H(p_\vartheta, p_{\vartheta_0}) < \epsilon\}$, where $d_H(p_\vartheta, p_{\vartheta_0})$ is a version of Hellinger distance given by $d_H(p_\vartheta, p_{\vartheta_0}) = 1 - \int \sqrt{p_\vartheta, p_{\vartheta_0}} d\mu(x)$, and

$$W_{\epsilon,n} = \left\{ (f, \sigma) \ : \ \int \big| f(x) - f_0(x) \big| dQ_n(x) < \epsilon, \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\},$$

where Q_n is the empirical measure based on the design points, $Q_n = n^{-1} \sum_{i=1}^n I_{x_i}(x)$. We show that posterior probabilities of H_ϵ and $W_{\epsilon,n}$ converge to one with probability tending to one respectively, in other words, posterior probabilities of the complements of H_ϵ and $W_{\epsilon,n}$ converge to zero with probability tending to one.

Theorem 2. *Let $\vartheta_0 = (f_0, \sigma_0)$, and let $p^n_{\vartheta_0}$ denote the true distribution of data $Y^n \equiv (Y_1, \ldots, Y_n)$ given the covariates $x^n \equiv (x_1, \ldots, x_n)$. Assume that f_0 and f are q-times continuously differentiable and that f is uniformly bounded. Then, the posterior distribution of f and σ is consistent under the $L^1(Q_n)$ norm and the Hellinger norm,*

$$\Pi \left\{ W^C_{\epsilon,n} | (x_1, Y_1), \ldots, (x_n, Y_n) \right\} \to 0 \ \ \textit{in } p^n_{\vartheta_0} \textit{ probability.}$$

$$\Pi \left\{ H^C_{\epsilon} | (X_1, Y_1), \ldots, (X_n, Y_n) \right\} \to 0 \ \ \textit{in } p^n_{\vartheta_0} \textit{ probability.}$$

Theorem 2 can be easily established by adapting consistency theorems of Choi and Schervish (2007) and Shively, Sager and Walker (2009) based on the characterization in Theorem 1. It is possible to extend the posterior consistency result in Theorem 2 to the partial linear model of (2.1) without much difficulty. Similar results as in Theorem 2 can be obtained by modifying the proof in Section 5 of Shively, Sager and Walker (2009); this is an alternative asymptotic technique that adapts the approach of combining the consistency property of the maximum likelihood estimator and a Bayesian component, originally proposed in Walker and Hjort (2001). The detailed proof of Theorem 2 can be found in the Supplementary Material.

3.2. Restriction on derivatives with a unique root

In this section we consider U–shaped and S–shaped functions by forcing the first or second derivative of f to change signs at a unique point. Suppose that $f^{(q)}$ is piecewise continuous and that it has a unique root at $x = \omega$:

$$\delta f^{(q)}(x) > 0, \ 0 < x < \omega, \ \delta f^{(q)}(\omega) = 0, \ \text{and} \ \delta f^{(q)}(x) < 0, \ \omega < x < 1 \qquad (3.13)$$

for $\delta = -1$ or 1 and $q = 1$ or 2. In order to force the signs of $f^{(q)}$ to switch at ω, we introduce the "squish" function h, a decreasing logistic function between 1 and -1, into the model for f:

$$f^{(q)}(x) = \delta Z^2(x)h(x) \text{ for } \delta = 1 \text{ or } -1 \text{ and } q = 1 \text{ or } 2, \tag{3.14}$$

$$h(x) = \frac{1 - \exp[\psi(x-\omega)]}{1 + \exp[\psi(x-\omega)]} \text{ for } \psi > 0 \text{ and } 0 < \omega < 1, \tag{3.15}$$

where ω is unique zero of h, and the slope ψ controls the steepness of h at ω. The squish function h has several attractive features. First, $f^{(q)}(\omega) = 0$, and h flips the sign of the q^{th} derivative after ω. Second, the q^{th} derivatives of f are continuous at ω. Third, h is nearly plus or minus one outside a small neighborhood of ω when ψ is large.

When $q = 1$, ω is the maximum for inverted U–shaped functions ($\delta = 1$) or the minimum for U–shaped functions ($\delta = -1$). The model for f is:

$$f(x) = \delta \left[\int_0^x Z^2(s)h(s)ds - \int_0^1 \int_0^x Z^2(t)h(t)dtds \right]. \tag{3.16}$$

The second term is the constant of integration and satisfies the mean–centering constraint at (2.2).

When $q = 2$, ω is the inflection point of f, and the model for f is

$$f(x) = \delta \int_0^x \int_0^s Z^2(t)h(t)dt + c_1 x + c_2,$$

where c_1 and c_2 are constants of integration. We select c_2 to satisfy the mean–centering constraint. S–shaped functions require a second condition on the first derivative to ensure monotonicity of f, which imposes condition on c_1. We consider four cases that are specified by a combination of δ and a second indicator ζ: increasing and convex-to-concave ($\delta = 1, \zeta = 1$), decreasing and concave-to-convex ($\delta = -1, \zeta = 1$), increasing and concave-to-convex ($\delta = 1, \zeta = -1$), and decreasing and convex-to-concave ($\delta = -1, \zeta = -1$). The model for f is

$$\begin{aligned} f(x) = \delta\zeta & \left[\int_0^x \int_0^s Z^2(t)h(t)dt\, ds - \int_0^1 \int_0^x \int_0^s Z^2(t)h(t)dt\, ds\, dx \right] \\ & + (\alpha - \delta\xi)(x - 0.5), \\ \xi = \min & \left[0, \min_{x \in [0,1]} \zeta \int_0^x Z^2(s)h(s)ds \right], \end{aligned} \tag{3.17}$$

where $\delta\alpha > 0$. It is easy to check that the $f^{(2)}$ satisfies (3.13), that $f^{(1)}$ is positive or negative, and that f is mean centered.

The class of function represented by the model in (3.14) characterizes the sample paths of the q^{th} derivative to be almost surely piecewise continuous and to have a unique root at ω. This leads to a complete class theorem for the representation: this is proven in the Supplementary Material, and the cases for negative derivatives follow similar arguments.

Theorem 3. *If h is as in* (3.15)*, with $\psi > 0$ and $0 < \omega < 1$, and for $x \in [0,1]$,*

$$f_1(x) = \int_0^x u^2(s)h(s)ds + a, \quad \textit{and} \quad f_2(x) = ax + b + \int_0^x \int_0^s u^2(t)h(t)dt\ ds,$$

then $f_1^{(1)}(x)$ and $f_2^{(2)}(x)$ are piecewise continuous with unique zeros at $x = \omega$. Conversely, if $f^{(q)}(x)$ is continuously differentiable on $[0,1]$ and has a unique zero at $x = \omega$, then there exists $u(x) \in C^1[0,1]$ such that $f^{(q)}(x) = u^2(x)h_{\psi,\omega}(x)$.

The prior distributions for ψ and ω are truncated normal distributions:

$$\psi \sim N(m_{0,\psi}, v_{0,\psi}^2)I(\psi > 0) \text{ and } \omega \sim N(m_{0,\omega}, v_{0,\omega}^2)I(\omega \in S),$$

where S is the support of x. As ψ goes to infinity, the squish function h is 1 for $x < \omega$, 0 for $x = \omega$ and -1 for $x > \omega$. In our simulation studies, we find that the likelihood function is fairly flat in the slope ψ for sufficiently large ψ. Intuitively, if ω is between two order statistics; $x_{(k)} < \omega < x_{(k+1)}$, then the likelihood for ψ is nearly constant for $\psi > \psi_k$ where $h(x_{(k)}|\psi_k) \approx 1-\epsilon$ and $h(x_{(k+1)}|\psi_k) \approx -1+\epsilon$ for small $\epsilon > 0$. One way to specify ψ is to find the value such that $h(\omega-\lambda) = 1-\epsilon$ for small, positive λ and ϵ. Then $\psi = \ln[(2-\epsilon)/\epsilon]/\lambda$. Therefore, we will sometimes fix ψ to a large value (1000) in our empirical studies.

We use numerical integration because the integral of the product of the cosines and squish functions does not have a closed form. We find the trapezoid rule simple to implement, and fast, since we need integrals at each observation and each value of a fine grid of $[0,1]$ to plot f. Typically, the fine grid has 200 intervals.

3.3. Spike-and-slab priors

A limitation of the previous models for f is that the posterior mean will not include functions on the boundary of the constraint space for different regions of its support. If $Z^2(x) = 0$ over an interval, then f is on the boundary of the constraint space over that region. Even though Z^2 puts positive mass on neighborhoods of 0, its posterior mean will be positive. Neelon and Dunson (2004) examine this issue for monotone plines. They ameliorate bias in the posterior mean by truncating the slopes to 0 if their absolute value is less than a constant. The implied, spike–and–slab prior is a mixture of a point mass at zero and a truncated normal distribution. We adapt Neelon and Dunson's method to Z^2 instead of its coefficients. We define the "latent" $\tilde{Z}$ to be the Gaussian process Z in the previous sections. The actual Z that is used in the models for f truncates the latent $\tilde{Z}$:

$$Z(x) = \tilde{Z}(x) \text{ if } \tilde{Z}^2(x) > \chi \text{ and } Z(x) = 0 \text{ if } \tilde{Z}^2(x) \le \chi,$$

where χ is a non-negative constant. For instance, (3.1) becomes $f^{(q)}(x) = \delta\tilde{Z}^2(x)I(\tilde{Z}^2(x) > \chi)$. The spike-and-slab truncation is imposed on the sample paths of $\tilde{Z}^2$ not on the prior distribution for the spectral coefficients. By replacing small values of $\tilde{Z}^2$ by zero, the model puts positive probability on the boundary of the restriction space. We treat χ as an unknown parameter. Its prior distribution is truncated normal on the non-negative numbers:

$$\chi \sim N(\mu_\chi, \sigma_\chi^2)I(\chi \ge 0).$$

The prior and posterior distributions of $Z(x)$ are equivalent to a mixture of 0 and $\tilde{Z}(x)$ where the prior and posterior probability of 0 depends on x:

$$\pi_0(x) = \mathrm{E}\left[I\left(\tilde{Z}^2(x) < \chi\right)\right], \text{ and } \pi_n(x) = \mathrm{E}\left[I\left(\tilde{Z}^2(x) < \chi\right)|Y_1, ..., Y_n\right].$$

Plotting π_n versus x provides information about the domain of f that is on the boundary.

The integrals of Z^2 no longer have analytical expressions, and we use numerical integration: the trapezoid rule on a very fine grid. When we need to distinguish among the models, we call the models in Sections 3.1 and 3.2 by "Gamma Prior," and the models of Section 3.3 by "Spike-and-Slab Prior." Based on our experience, the MCMC for the Spike-and-Slab Prior does not mix as well as the Gamma Prior.

The MCMC algorithm for both the Gamma and Spike–and–Slab Priors is presented in the online appendix. They use an adaptive Metropolis procedure (Haario, Saksman and Tamminen (2001) and Atchadé and Rosenthal (2005)), that improves the mixing and convergence compared to random walk Metropolis.

4. Empirical Analysis of Shaped–Restricted BSAR

This section examines the empirical analysis of shape–restricted BSAR based on simulation studies and two applications. Our software implementations of BSAR are written in GAUSS and R. Both are available from the authors upon request. We are developing a user-friendly R package for BSAR, and a preliminary version of R package is available from `http://statlab2.korea.ac.kr/software/bsar`. Table 2 gives the abbreviations for the different methods.

Table 2. Abbreviation of methods.

Abbreviation	Method	Source
BSAR	Bayesian Spectral Analysis Regression	Equation (2.3)
BSARM	BSAR with Monotone Constraint	Equation (3.6)
BSARMC	BSARM with Convex or Concave Constraints	Equations (3.8) or (3.10)
BSARU	BSAR with U–Shaped Constraint	Equations (3.16)
BSARS	BSAR with S–Shaped Constraints	Equation (3.17)
BRSM	Bayesian Regression Spline with Monotone Constraint	Meyer et al. (2011)
BRSMC	BRSM with Convex or Concave Constraints	Meyer et al. (2011)
BBMP	Bayesian Bernstein Polynomial with Monotone Constraint	Curtis and Ghosh (2011)
Gamma Prior	Models that do not truncate Z^2	Sections 3.1 and 3.2
Spike–and–Slab Prior	Models that truncate Z^2	Section 3.3

4.1. Simulation studies for curve fitting

This section compares shaped–restricted BSAR with Bayesian shape- restricted regression splines (BRSM and BRSMC) of Meyer, Hackstadt and Hoeting (2011) and Bayesian Bernstein Polynomial Monotone regression (BBPM) of Curtis and Ghosh (2011) through simulations. BRSM is Bayesian regression splines with monotone restrictions, and BRSMC is Bayesian regression splines with monotone, convex or concave restrictions. For BRSM and BRSMC, the R code is available from the author's website of Meyer, Hackstadt and Hoeting (2011), `http:/www.stat.colostate.edu/~meyer/bayescode.htm`, and R package `bisoreg` is used for BBPM. We generated 50 data sets using low and high information conditions. The low information condition had 50 observations and 30 basis functions, and the high information condition had 200 observations and 50 basis functions.

The focal performance metric is the RMISE between the true $\beta_0 + f$ and its posterior mean. We simulated data from five functions: Linear and Sinusoid are monotone. Exponential (Expo) and the sum of four cosines (QuadCos) are monotone and convex. Logarithm (LogX) is monotone and concave. The online appendix gives the models. BSARMC sets $\alpha = 0$ to focus attention on the spectral representation.

Table 3 summarizes the RMISE's, and the bold numbers in Table 3 indicate the smallest, average RMISE. The BSAR estimates use the Gamma Prior of Sections 3.1 and 3.2. BSARM has the smallest, average RMISE for the monotone function, and BSARMC has the smallest, average RMISE for the convex or concave functions. They did not uniformly dominated the other Bayesian estimators for all data sets. Overall, the simulation results indicate that the shape–restricted BSAR provides a competitive fit.

Table 3. Average RMISE for monotone/convex/concave functions.

Function	n	BSARM	BSARMC	BRSM	BRSMC	BBPM
Linear	50	**0.1741**	0.1865	0.2181	NA	0.2087
Monotone	(s.e.)	(0.012)	(0.0102)	(0.0121)	NA	(0.0121)
	200	**0.1085**	0.1211	0.1356	NA	0.1362
	(s.e.)	(0.0050)	(0.0037)	(0.0055)	NA	(0.0055)
Sinusoid	50	**0.3051**	0.5786	0.3499	NA	0.3052
Monotone	(s.e.)	(0.0116)	(0.0058)	(0.0129)	NA	(0.0115)
	200	**0.1726**	0.4758	0.1959	NA	0.1870
	(s.e.)	(0.0073)	(0.0021)	(0.0075)	NA	0.0070)
Expo	50	0.3134	**0.2840**	0.3188	0.3730	0.3329
Monotone	(s.e.)	(0.0115)	(0.0121)	(0.0123)	(0.0123)	(0.0126)
Convex	200	0.1884	**0.1571**	0.1731	0.2007	0.2149
	(s.e.)	(0.0053)	(0.0050)	(0.0064)	(0.0064)	(0.0058)
QuadCos	50	0.3204	**0.2492**	0.3826	0.3213	0.3092
Monotone	(s.e.)	(0.0136)	(0.0120)	(0.0176)	(0.0176)	(0.0133)
Convex	200	0.1865	**0.1481**	0.1876	0.1885	0.2028
	(s.e.)	(0.0068)	(0.0060)	(0.0084)	(0.0084)	(0.0062)
LogX	50	0.2178	**0.1811**	0.2607	0.2177	0.2322
Monotone	(s.e.)	(0.0116)	(0.0130)	(0.0128)	(0.0126)	(0.0120)
Concave	200	0.1357	**0.1149**	0.1646	0.1347	0.1579
	(s.e.)	(0.0061)	(0.0054)	(0.0066)	(0.0057)	(0.0061)

Table 4. Estimating the maximum of a function with BSARU: 50 data sets per condition.

Abscissa at Maximum, $\omega = 2$						
Statistic	n	Mean	Standard Deviation	Q2	Median	Q3
Posterior Mean	50	2.099	0.156	2.011	2.085	2.165
	200	2.006	0.099	1.949	1.992	2.070
Posterior Standard Deviation	50	0.153	0.055	0.114	0.145	0.178
	200	0.081	0.035	0.061	0.077	0.092

To test the estimating of upside-down U–shaped functions and their maxima, we considered the model: $Y = 1.67 + 0.5(x+2)^4 \exp(-x) + \epsilon$ where $\epsilon \sim N(0,1)$ and $x \in [0,10]$. The maximizer ω of f is 2. Table 4 reports the simulation statistics based on 50 and 200 observations using 50 simulated data sets. The results indicate that the posterior distribution of ω correctly recovers the value of x that maximizes f. As far as we know, Bayesian regression splines or Bernstein polynomials currently do not have U–shaped options for comparison.

The online appendix continues the simulation study to demonstrate BSARS and the Spike-and-Slab Prior. The simulation studies confirm that BASRS accurately estimates the inflection point of S–shaped functions. The simulation for the Spike-and-Slab Prior uses a test function that is on the boundary of the constraint space in two regions of its support. The fit statistics favor the Spike-and-Slab Prior over the Gamma Prior in this case. Please refer to the online appendix for more details.

4.2. Testing the adequacy of shape restrictions

We test the adequacy of shape restrictions by computing the marginal likelihoods of competing models (Jeffreys (1961) and Kass and Raftery (1995)). Our situation differs from Bayesian model testing for variable selection or for comparing a parametric model to a nonparametric model (Lenk (1999)). In our application the likelihood functions for the different models are the same if one treats the unknown f as a parameter. The shape–restrictions belong to the prior distribution of f, and Bayesian hypothesis testing is selecting the "best" prior distribution.

When a model incorrectly imposes a shape restriction, then the marginal likelihood for this model should suffer because the sums-of-squares error (SSE) between the observed Y and estimated regression function tends to be larger than the SSE for models with the correct restrictions or no restrictions. In the case where two or more models are consistent with the true function, the SSEs may be nearly equivalent. This case is more nuanced, and model choice is driven by at least two factors. First, constraints help the model to separate noise from signal. For example, the unrestricted model is consistent with monotone functions but may mistake noise for signal by introducing small wobbles in the estimated function. The monotone model correctly recognizes these wobbles as noise. The SSE for the restricted model will be slightly larger than that of the unrestricted model because adding a constraint does not improve fit. However, the monotone model has less posterior uncertainty about the function, which can lead to larger LILs. Second, the prior distributions have greater influence on model choice. In the simulation study, we used the same prior parameters across the models; however, the BSARU and BSARS have additional parameters, ω and ψ. The marginal likelihood may prefer a simpler model if the reduction in uncertainty does not compensate for higher model complexity.

Our simulation study considered five true models: one for each set of constraints, and used the Gamma Prior of Sections 3.1 and 3.2. Model M1 does not have a constraint and is consistent with only BSAR. Model M2 is monotone and is consistent with BSARM and BSAR. Model M3 is montone and convex and is consistent with BSARMC, BSARM, and BSAR. Model M4 is upside-down U–Shaped and is consistent with BSARU and BSAR. Model M5 is S–Shaped and is consistent with BSARS, BSARM, and BSAR. Also, BSARU and BSARS could mimic monotonic models by sending ω towards the endpoints 0 or 1. The

Table 5. Average log integrated likelihood over 50 data sets per condition.

50 Observations	BSAR	BSARM	BSARMC	BSARU	BSARS
M1: Free	**-168.114**	-192.767	-201.994	-220.503	-209.099
M2: Monotone	-159.346	**-145.007**	-158.733	-159.162	-153.302
M3: Monotone Convex	-161.931	-142.352	**-140.618**	-152.399	-147.947
M4: U–Shaped	-160.708	-203.013	-220.222	**-159.440**	-203.926
M5: S–Shaped	-215.488	-173.730	-215.466	-179.728	**-178.649**
200 Observations	BSAR	BSARM	BSARMC	BSARU	BSARS
M1: Free	**-501.123**	-655.949	-701.773	-690.688	-710.554
M2: Monotone	-492.962	**-480.329**	-550.351	-499.800	-526.203
M3: Monotone Convex	-490.013	-474.379	**-465.555**	-488.631	-485.809
M4: U–Shaped	**-490.729**	-705.068	-776.406	-509.745	-720.649
M5: S–Shaped	-569.465	**-535.112**	-690.668	-580.578	-576.267

Bold face is maximum in row.

online appendix gives the model specifications. The simulations had a low information condition with 50 observations and a high information condition with 200 observations. Fifty data sets were generated for each function and information condition.

Table 5 summaries the average LIL over the 50 data sets within each condition. The online appendix presents more complete results with standard deviations of the LIL, the proportion of times each model had maximum LIL, and the RMISE between the true function and the posterior mean. When none of the restrictions were appropriate (M1), the LIL correctly selected BSAR: BSAR had maximum LIL for all 50 samples. Similarly, the average LIL correctly favored BSARM for M2, with choice rates of 84% and 92% on 50 and 200 observations and BSARMC for M3, with choice rates of 54% and 74% on 50 and 200 observations.

The LIL found it more challenging to identify M4 (U–shaped) and M5 (S–shaped). With 50 observations, the LIL correctly favored BSARU for M4, and BSAR was a strong contender. The choice rates were 64% for BSARU and 36% for BSAR. With 200 observations, the LIL picked BSAR 94% of time for M4. For M5 the average LIL slightly favored BASRM over BASRS with 50 observations, with choice rates of 44% for BSARM and 40% for BSARS. With 200 observations, LIL favored BSARM for M5, with choice rates of 54% for BSARM and 18% for BSARU.

In the high-information condition the reduction in estimation error did not compensate for the extra parameters for U and S–shaped models. Because the LILs for these models are relative close to each other, it would be possible to adjust the priors to shift the results in a different direction. One could make the prior distributions equivalent by adjusting their parameters to equalize the prior information as determined by a measure such as the Kullback-Leibler information. These results do not rule out using BSARU and BSARS because they allow for the explicit estimation and inference of the extreme or the point of inflection, a challenging problem in unconstrained nonparametric models.

4.3. Prestige of occupations

This example illustrates use of two, additive BSAR functions. The data consists of the prestige of 102 occupations in Canada and are from Fox and Weisber (2011), available at `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/data.html`. The dependent variable is the Pineo-Porter prestige score from a social survey conducted in mid-1960s. Three independent variables are from the 1971 Canadian Census: 1) the average years of education for workers in the occupation, 2) the average yearly income for workers in the occupation, and 3) the percentage of workers in the occupation who are women. One may reasonably expect that prestige increases with both education and income, and exploratory data analysis indicates this to be true. The relation between prestige and the percentage of women is less clear. Women-dominated occupations tend to have lower incomes than men-dominated ones, so their prestige may be lower. However, women are nearly equal to men in education in 1971, so their prestige may be higher than expected based on income alone. We will treat percentage of women as a linear covariate.

Figure 2 plots the BSAR functions for two models. Panels A and B used unrestricted BSAR for both Income and Education. This model has a LIL of –756.50. BSAR over-fits the data due to the "clumpiness" of Education and Income, and the functions are difficult to interpret. Panels C and D fit S–shaped BSARS to Education and increasing, convex BSARMC to Income. The shape–restrictions help to smooth the estimated function, and its LIL increases to –618.38. The estimated inflection point for Education is 12.23 years with a posterior standard deviation of 1.25. The marginal benefit of education is increasing up to 12 years and declining afterward. The estimated coefficient for Percentage of Women is 0.0614 with a posterior standard deviation of 0.0274 in the BSARS/BSARMC model. Interestingly, women–dominated occupations, such as nurses and primary school teachers, have higher prestige than men–dominated professions, such as computer programmers and typesetters, given similar income and education. Unfortunately, women dominated occupations tend to have lower incomes given education. Fortunately the difference, though highly significant, is relatively small (64 Canadian dollars per year) at less than 1% of the average, yearly income. Canada in 1971 seems to have been more progressive on gender equity in pay

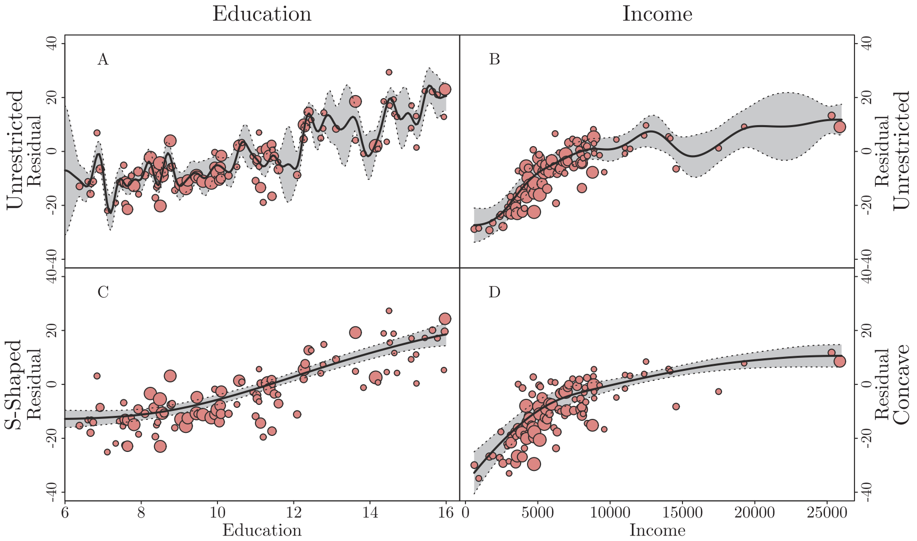

Figure 2. Occupational prestige. Dots are partial residuals, which are scaled relative to the percentage of women. Solid lines are posterior means. Shaded areas between dashed lines are 95% credible intervals. Panel A: unrestricted education using BSAR. Panel B: unrestricted income using BSAR. Panel C: increasing, S–shaped education using BSARS. Panel D: increasing, concave income using BSARMC.

than its southern neighbor where the gender pay gap is commonly reported to be around 20% in 2014 (Institute for Women's Policy Research).

4.4. Electricity demand

Demand for electricity depends on its price, on economic activity, and on temperature. Because many homes do not use heating or cooling when the ambient temperature is around 68° Fahrenheit, temperature is often coded as the number of heating or cooling degree days relative to a reference temperature. Engle et al. (1986) proposed a cubic spline model for the relation between demand and temperature. We used the data in Yatchew (2003), Section 4.6.3, which consists of 288 quarterly observations in Ontario from 1971 to 1994. Yatchew (2003) found that the demand for electricity is co-integrated with gross domestic product (GDP). Consequently, he uses the log of the ratio of electricity demand to GDP as the dependent variable. A covariate W is the log price ratio of electricity to natural gas. As W increases, demand for electricity should fall off as customers switch from electricity to natural gas. The focal, independent variable "Temperature" is the number of heating and cooling degree days relative to 68°F. "Temperature" is positive when the average temperature is above 68°F (more cooling days) and negative below 68°F (more heating days). We fitted our shape-restricted models to Yatchew's data.

Table 6. Electricity demand estimated models.

	BSAR	BSARM	BSARM	BSARMC	BSARMC	BSARS
			Spike & Slab		Spike & Slab	
LIL	195.6015	194.5515	152.6503	146.7671	125.7278	157.3315
R-Square	0.8080	0.8002	0.8076	0.7918	0.8000	0.8054
Error Standard Deviation	0.1004	0.1005	0.0991	0.0998	0.0987	0.0991
	(0.0042)	(0.0041)	(0.0041)	(0.0042)	(0.0041)	(0.0041)
Intercept	-1.5802	-1.5813	-1.579	-1.5769	-1.579	-1.5792
	(0.0309)	(0.0306)	(0.0303)	(0.0302)	(0.0301)	(0.0303)
Log Price Ratio	-0.0754	-0.0719	-0.0751	-0.0712	-0.0720	-0.0741
	(0.0245)	(0.0244)	(0.0242)	(0.0240)	(0.0240)	(0.0240)
Inflection Point						-338.5744
						(1.7782)
Slope of Squish Function						12.0828
						(1.9388)
Cutoff			0.00024		2.2E-06	
			(0.00012)		(1.6E-06)	

Posterior standard deviations are in parentheses.

Figure 3 plots the posterior means and 95% HPD intervals for the estimated models. Panels A, C, and E include the parametric residuals $y_i - \hat{\beta} w_i$. Panel A is the unrestricted model, and Panel B compares the S–shaped model with the unrestricted model. Except for the right most end of Temperature, the posterior mean of the unrestricted model is within the HPD intervals of the S–shaped model. Panels C and E plot the monotonically decreasing model and monotonically decreasing and convex models, respectively, with Gamma Priors. Panels D and F plot these models using Spike-and-Slab Priors and compares them to the posterior means with Gamma Priors. The posterior means with Gamma Priors are almost entirely contained in the HPD intervals for the Spike-and-Slab Priors, except at the right endpoint in Panel D.

Table 6 gives the fit statistics and estimated parameters. The unrestricted model has marginally better LIL than the decreasing function models. All of the models confirm that demand decreases as the price of electricity increases relative to the price of natural gas.

5. Conclusion

In this paper, we considered a Bayesian method for shape-restricted functions by modeling derivatives of the functions with squared, Gaussian processes. The proposed representation expresses monotone and convex/concave restrictions as well U-shaped functions that have extrema and S–shaped functions that have inflection points. The spectral analysis of the Gaussian processes facilitates inference. We illustrated the empirical performance of the proposed model with

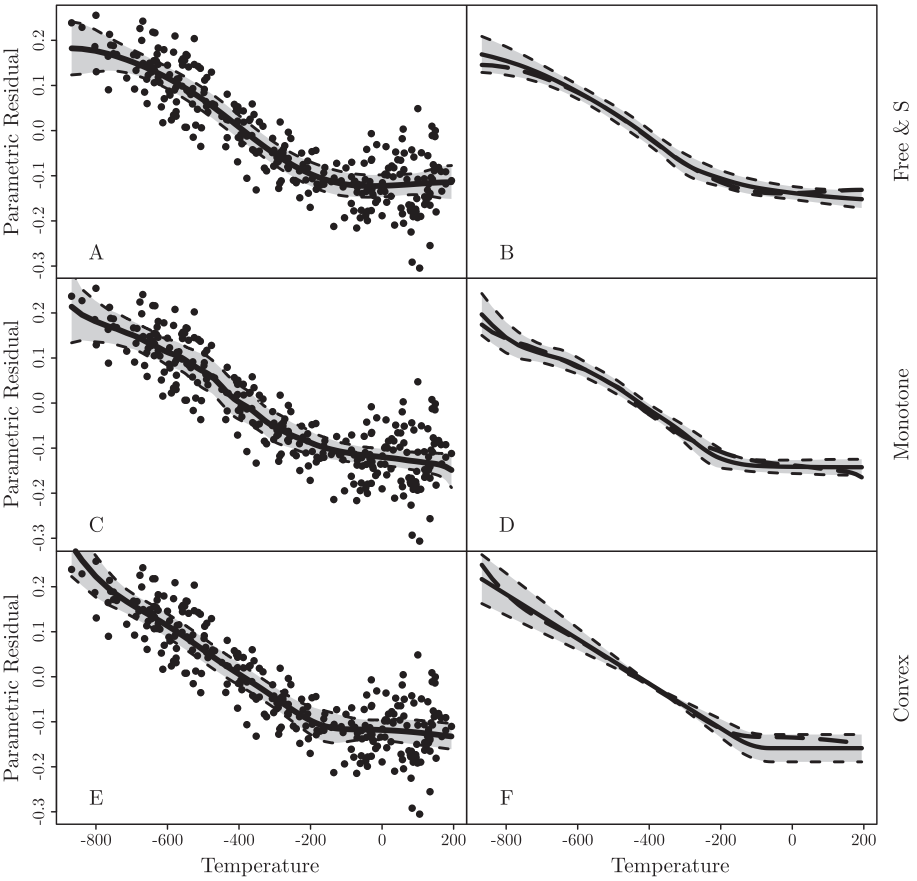

Figure 3. Estimated electric demand. Dots are parametric residuals; solid lines are posterior means, and shaded areas between dashed lines are 95% credible intervals. Panel A: unrestricted demand. Panel B: monotonically decreasing S-shaped demand, and the long-dashed line is the posterior mean for the unrestricted model from Panel A. Panel C: monotonically decreasing demand. Panel D: monotonically decreasing demand with Spike-and-Slab Prior, and the long-dashed line is the posterior mean of for the monotonically decreasing demand of Panel C. Panel E: monotonically decreasing convex demand. Panel F: monotonically decreasing convex demand with a Spike-and-Slab Prior, and the long-dashed line is the posterior mean of the monotonically decreasing convex demand of Panel E.

simulations and data applications, and we compared the method to other existing methods. The simulation studies favored the proposed method, though all contenders performed well.

We also considered Spike-and-Slab Priors to estimate functions on the boundary of the constraint space and found improvements in the posterior mean at the cost of increased computations and degraded mixing of the MCMC. If the researcher's decisions do not depend on the function being exactly on the boundary, then we recommend using the Gamma Prior to improve the estimation of f. However, there are situations where testing boundary conditions are meaningful for theoretical reasons. Then the Spike–and–Slab Prior is beneficial.

We also tested the ability of Bayesian hypothesis testing to confirm restrictions and demonstrated its ability to select the correct model when the models have the sample parameter space. Least surprising, models with incorrect constraints were dominated by models with constraints that are consistent with the true function. In addition, among models with the correct constraints, Bayesian hypothesis testing tended to select the model that imposed the greatest number of correct constraints provided that the prior distributions are the same. However, when a correct constraint expands the parameter space by adding parameters, Bayesian hypothesis testing tends to select the correct model with a smaller parameter space in high–information conditions. This deficiency may be overcome by judicious selection of prior distributions; for example, adjusting the prior variances so that information measures are equal across prior distributions may handicap priors on smaller parameters spaces.

There are several issues and extensions. For theoretical aspects, we have not discussed the large sample properties of the BSAR except for posterior consistency but believe that other asymptotic properties such as posterior convergence rates and Bernstein-von Mises theorem for the BSAR could also be established by verifying general sufficient conditions (see, e.g., Ghosal and van der Vaart (2007)). One challenging part would be in dealing with the function space under shape restriction and the non-normality as well as non-orthogonality due to the squared Gaussian process in our characterization. For methodological developments, we could extend the proposed BSAR to cases with multivariate predictors and non-Gaussian errors function for instance. For multivariate predictors, other basis functions, such as radial basis function or Gaussian kernel function (e.g., Konishi, Ando and Imoto (2004) and Chakraborty, Ghosh and Mallick (2012)), may be more suitable than cosine basis functions. For the non-Gaussian error distribution, and more generally the unobserved errors from an unknown distribution function, we can consider Bayesian quantile regression (e.g., Koenker (2005) and Yue and Rue (2011)) with shape-restriction and the shape-restricted regression with unknown error distribution by using a Dirichlet process mixture of normals (e.g., Chib and Greenberg (2010)) as well as the shape-restricted density regression (Wang and Dunson (2011)). Furthermore, the proposed method can certainly be used in practical applications such as dose-response functions, utility functions, or risk aversion modeling, of which in theory the shape restriction needs to be incorporated for estimation.

Supplementary Materials

Supplementary materials available at Statistica Sinica online contains technical details on the equations for the integrated, mean–centered, cosine basis, the MCMC algorithms, the proofs of Theorems 1–3, additional information about the simulation studies, and simulations for BSARS and Spike–and–Slab priors.

Acknowledgement

Research of Taeryon Choi was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2012R1A1A2041370). Assistance of Seongil Jo for the numerical work with the R package development is greatly appreciated. The review team's comments greatly improved the paper and in particular, one reviewer suggested the use of the spike and slab prior. We wish to thank the editor and reviewers for their helpful comments that improved the style and substance of the paper.

References

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11**, 815-828.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641-647.

Barlow, R. E. and Brunk, H. D. (1972). The isotonic regression problem and its dual. *J. Amer. Statist. Assoc.* **67**, 337, 140-147.

Bhattacharya, R. and Lin, L. (2010). An adaptive nonparametric method in benchmark analysis for bioassay and environmental studies. *Statist. Probab. Lett.* **80**, 1947-1953.

Bhattacharya, R. and Lin, L. (2011). Nonparametric benchmark analysis in risk assessment: a comparative study by simulation and data analysis. *Sankhyā B* **73**, 144-163.

Bhattacharya, R. and Lin, L. (2013). Recent progress in the nonparametric estimation of monotone curves – Sith applications to bioassay and environmental risk assessment. *Comput. Statist. Data Anal.* **63**, 63-80.

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics* **65**, 198-205.

Brezger, A. and Steiner, W. (2008). Monotonic regression based on Bayesian P-splines: an application to estimating price response functions from store-level scanner data. *J. Bus. Econom. Statist.* **26**, 90-104.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26**, 607-616.

Cai, B. and Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *J. Amer. Statist. Assoc.* **102**, 1158-1171.

Chakraborty, S., Ghosh, M. and Mallick, B. K. (2012). Bayesian nonlinear regression for large p small n problems. *J. Multivariate Anal.* **108**, 28-40.

Chib, S. and Greenberg, E. (2010). Additive cubic spline regression with Dirichlet process mixture errors. *J. Econom.* **156**, 322-336.

Choi, T., Lee, J. and Roy, A. (2009). A note on the Bayes factor in a semiparametric regression model. *J. Multivariate Anal.* **100**, 1316-1327.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal.* **98**, 1969-1987.

Curtis, S. M. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials. *J. Appl. Statist.* **38**, 961-976.

Dette, H. and Pilz, K. (2006). A comparative study of monotone nonparametric kernel estimates. *J. Statist. Comput. Simulation* **76**, 41-56.

Dykstra, R. L. and Robertson, T. (1982). An algorithm for isotonic regression for 2 or more independent variables. *Ann. Statist.* **10**, 708-716.

Eltoft, T., Kim, T. and Lee, T.-W. (2006). On the multivariate laplace distribution. *IEEE Signal Process. Lett.* **13**, 300-303.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81**, 310-320.

Fang, Z. and Meinshausen, N. (2012). LASSO isotone for high-dimensional additive isotonic regression. *J. Comput. Graph. Statist.* **1**, 72-91.

Fox, J. and Weisber, S. (2011). *An R Companion to Applied Regression*. Sage Publications, Thousand Oaks, CA.

Friedman, J. and Tibshirani, R. (1984). The monotone smoothing of scatterplots. *Technometrics* **26**, 234-250.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56**, 501-514.

Ghosal, S. and van der Vaart (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35**, 192-223.

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Groeneboom, P. and Wellner, J. A. (2014). *Nonparametric Estimation under Shape Constraints.* Cambridge University Press, Cambridge.

Haario, H., Saksman, E. and Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7**, 223-242.

Hall, P. and Huang, L. S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.* **29**, 624-647.

Jeffreys, H. (1961). *Theory of probability.* 3rd edition. Clarendon Press, Oxford.

Kass, R. E. and Raftery, A. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Katznelson, Y. (2004). *An Introduction to Harmonic Analysis.* Cambridge University Press, Cambridge.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, Cambridge.

Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91**, 27-43.

Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78**, 531-543.

Lenk, P. J. (1993). A Bayesian nonparametric density estimator. *J. Nonparametr. Statist.* **3**, 53-69.

Lenk, P. J. (1999). Bayesian inference for semiparametric regression using a fourier representation. *J. Roy. Statist. Soc. Ser. B* **61**, 863-879.

Lenk, P. J. (2003). Bayesian semiparametric density estimation and model verification using a logistic-gaussian process. *J. Comput. Graph. Statist.* **12**, 548-565.

Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101**, 303-317.

Luss, R. and Rosset, S. (2014). Generalized isotonic regression, *J. Comput. Graph. Statist.* **23**, 192-210.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19**, 724-740.

Menendez, J. A. and Salvador, B. (1987). An algorithm for isotonic median regression, *Comput. Statist. Data Anal.* **5**, 399-406.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *Ann. Appl. Statist.* **2**, 1013-1033.

Meyer, M. C., Hackstadt, A. J. and Hoeting, J. A. (2011). Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *J. Nonparametr. Statist.* **23**, 867-884.

Mukarjee, H. and Stern, S. (1994). Feasible nonparametric estimation of multiargument monotone functions. *J. Amer. Statist. Assoc.* **89**, 425, 77-80.

Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60**, 398-406.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. Ser. B* **40**, 1-42.

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103**, 681-686.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *J. Roy. Statist. Soc. Ser. B* **60**, 365-375.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* MIT Press, Cambridge, MA.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference.* Wiley, New York.

Sasabuchi, S., Inutsuka, M. and Kulatunga, D. (1983). A multivariate version of isotonic regression. *Biometrika* **70**, 465-472.

Shively, T., Sager, T. W. and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. Roy. Statist. Soc. Ser. B* **71**, 159-175.

Shively, T. S., Walker, S. G. and Damien, P. (2011). Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *J. Econom.* **161**, 166-181.

Tibshirani, R. J., Hoefling, H. and Tibshirani, R. (2011). Nearly-isotonic regression. *Technometrics* **53**, 54-61.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40**, 364-372.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *J. Roy. Statist. Soc. Ser. B* **63**, 811-821.

Wang, L. and Dunson, D. B. (2011). Bayesian isotonic density regression. *Biometrika* **98**, 537-551.

Wang, X. (2008). Bayesian free-knot monotone cubic spline regression. *J. Comput. Graph. Statist.* **17**, 373-387.

Wang, X. and Li, F. (2008). Isotonic smoothing spline regression. *J. Comput. Graph. Statist.* **17**, 21-37.

Wu, J., Meyer, M. C. and Opsomer, J. D. (2015). Penalized isotonic regression. *J. Statist. Plann. Inference* **161**, 12-24

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician.* Cambridge University Press, Cambridge.

Young, A. S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika* **64**, 309-317.

Yue, Y. R. and Rue, H. (2011). Bayesian inference for additive mixed quantile regression models. *Comput. Statist. Data Anal.* **55**, 84-96.

Stephen M. Ross Business School, The University of Michigan, Ann Arbor, MI, U.S.A.

E-mail: plenk@umich.edu

Department of Statistics, Korea University, Seoul, Republic of Korea.

E-mail: trchoi@korea.ac.kr

(Received March 2015; accepted December 2015)