Diagnosing non-linear regression structure with power-additive smoothing splines

Wataru Sakamoto
Osaka University

1 Introduction

Nonparametric regression, a powerful tool for exploratory data analysis, explores unknown nonlinear regression structure underlying data from a flexible and extensive functional space. It plays an important role in choosing one appropriate model from some possible models, or diagnosing whether a parametric model is valid. The smoothing spline is one of the most useful methods of the nonparametric regression. It is represented as the optimal solution in a penalized approach and so it is easy to extend to a variety of models. It is related to Bayesian approaches and it results in the best prediction for a linear mixed model, and so it has a good estimation property. Moreover, a number of programs and softwares for computation on splines have been developed.

If continuous responses are observed with some explanatory variables, we often try to apply an additive regression model to explore nonlinear regresssion structure. The additive regression model is discussed in Hastie and Tibshirani (1986, 1990, 2000) in detail, and is adopted in some statistical packages such as the S language (Chambers and Hastie, 1992). However, it assumes the following requirements: (i) additivity among explanatory variables, (ii) homoscedasticity, (iii) normality (implicitly assumed in penalized least squares estimation), and (iv) independency. It is difficult to satisfy all these requirements simultaneously, so these *rigid* assumption might give poor estimation of regression functions,

Relaxing the assumptions in the additive regression model would be helpful in diagnosing validity of assuming these requirements. A variety of models in which the sides of explanatory variables are modified have been developed to extend from the additivity assumption, especially by adding some interaction terms, such as SS-ANOVA (Wahba, 1990; Wahba *et al.*, 1995), thin plate splines (Green and Silverman, 1994; Eubank, 1999), CART (Breiman *et al.*, 1984) and MARS (Friedman, 1991). However, extremely complicated models might make interpretation of the models difficult. The assumption of additivity has a practical advantage that it is easier to interpret. Stone (1985) described that a statistical model has three fundamental aspects: flexibility, dimensionality and interpretability, and proved the dimensionality reduction principle in which the best estimator (in terms of mean square errors) of the additive regression model is a good approximation to the true regression function.

Transforming response variables is also a useful method to diagnose the validity of the requirements (i)–(iii). Some nonparametric transformations have been developed, which estimate the functions transforming responses using some smoothing methods, such as ACE (Breiman and Friedman, 1985), AVAS (Tibshirani, 1988), and nonparametric both-sides transformation (Wang and Ruppert, 1995). However, in practical applications, measured responses often

have a particular meaning such as the concentration or the mass, so it is desired that the functions transforming responses should be monotone and meaningful for themselves. From this point of view, parametric transformations would give more helpful suggestion in interpreting results of analysis. The power transformation proposed by Box and Cox (1964),

$$y^{(\phi)} = \begin{cases} (y^{\phi} - 1)/\phi, & \phi \neq 0, \\ \log y, & \phi = 0, \end{cases} \tag{1.1}$$

is a typical choice of the parametric transformations, where $y > 0$, and ϕ is the power parameter. Hastie and Tishirani (1990) and Linton *et al.*, (1997) assumes that the responses after the power transformation satisfy the requirements (i)–(iv) simultaneously. However, it is impossible to obtain a single transformation that satisfies some different criteria simultaneously (Bartlett, 1947; Draper and Hunter, 1969), so it is necessary to separate the assumptions before the transformation from the requirements that should be satisfied after the transformation (Goto, 1992, 1995).

Sakamoto (2004) proposed the power weighted smoothing spline (PWSS), which aims the diagnosis of homoscedasticity by assuming it on power-transformed responses. In this paper we aim to diagnose additivity. We extend the idea of power additive transformation in linear regression models (Draper and Hunter, 1969; Goto, 1992, 1995). to the nonparametric additive regression model. We call it a power additive smoothing spline (PASS) model, which assumes the requirements (ii)–(iv) on original responses and then assumes the additivity on power-transformed responses in the additive regression model. The PASS model aims to extract unknown nonlinear regression structure that is comparatively easy to understand. It includes an additive model (in the case $\phi = 1$, that is equivalent to the identical transformation)

$$y \simeq f_1(t_1) + \cdots + f_r(t_r) + \text{error}$$

and a multiplicative model (in the case $\phi = 0$, that is the log transformation)

$$y \simeq g_1(t_1) \times \cdots \times g_r(t_r) \times \text{error}$$

as special cases, where y is the response variable, $f_j(t_j)$ and $g_j(t_j)$ are functions of the explanatory variable t_j $(j = 1, \ldots, r)$, and $\simeq$ shows some regression relationship. Estimation of the power parameter, which gives a response transformation that influences global regression structure, and the smoothing parameters, which control smoothness of the functions of explanatory variables, are very important. We apply the maximum marginal likelihood based on the empirical Bayes method. In general the marginal likelihood involves a high-dimensional integral, and so it is impossible to compute it exactly. We use a Laplace approximation based on the second-order Taylor expansion of the penalized log-likelihood.

Section 2 overviews the additive regression model and its estimation with smoothing splines. Section 3 proposes the PASS model and develop its estimation, including the estimation of the power and the smoothing parameters. Section 4 provides results of application to some data in the literature.

2 Additive regression model and smoothing spline

2.1 Additive regression model

Suppose that each of the responses y_i $(i = 1, \ldots, n)$ is independently observed with the explanatory variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{ir})^{\mathrm{T}}$. An ordinary additive regression model assumes that the response y_i has the mean that is an additive function of $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$,

$$\mu_i \equiv \mathrm{E}[y_i|\boldsymbol{x}_i, \boldsymbol{t}_i] = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + f_1(t_{i1}) + \cdots + f_r(t_{ir}), \quad i = 1, \ldots, n \tag{2.1}$$

and a constant variance $\mathrm{var}[y_i|\boldsymbol{x}_i, \boldsymbol{t}_i] = \sigma^2$, where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ is a vector of regression parameters, and $f_1, \ldots, f_r$ are smooth functions, both of which are unknown and estimated. We give constraints that $\sum_{i=1}^n f_j(t_{ij}) = 0$ $(j = 1, \ldots, r)$ to keep uniqueness of the estimates.

The backfitting algorithm (Hastie and Tibshirani, 1986) is well known as a method of computing estimates of parameters and smooth functions. The estimates are obtained as a set of solutions when the following updating equation converges:

$$\hat{\boldsymbol{\beta}}^{\mathrm{new}} = (X_L^{\mathrm{T}} X_L)^{-1} X_L^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{f}}_1 - \cdots - \hat{\boldsymbol{f}}_r), \tag{2.2a}$$

$$\hat{\boldsymbol{f}}_j^{\mathrm{new}} = S_j(\boldsymbol{y} - X_L\hat{\boldsymbol{\beta}} - \textstyle\sum_{k\neq j} \hat{\boldsymbol{f}}_k), \qquad j = 1, \ldots, r, \tag{2.2b}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{f}_j = (f_j(t_{1j}), \ldots, f_j(t_{nj}))^{\mathrm{T}}$, $X_L = [\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n]^{\mathrm{T}}$ and $\tilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}$ $(i = 1, \ldots, n)$, and the notation $\hat{}$ means estimates. S_j is a smoother for the j-th nonparametric explanatory variable, which corresponds to one of different smoothing methods such as kernel smoothing, local smoothing, and spline smoothing.

It is known that, if we use a spline smoother as S_j, the estimates are optimal solutions for a penalized approach, that is, we obtain smoothing splines as the estimates of f_j $(j = 1, \ldots, r)$ by minimizing the ordinary residual sum of squares plus penalty terms in the additive regression model (2.1). If we assume that each original response distributes independently and normally, this is equivalent to maximizing the penalized log-likelihood

$$l_{\mathrm{P}}(\boldsymbol{\beta}, f_1, \ldots, f_r; \boldsymbol{y}) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu_i)^2 - \frac{1}{2}\sum_{j=1}^r \lambda_j J(f_j) + \mathrm{const.} \tag{2.3}$$

with respect to $\boldsymbol{\beta}$ and $f_1, \ldots, f_r$ for given σ^2 and λ_j $(j = 1, \ldots, r)$. The positive numbers λ_j $(j = 1, \ldots, r)$ are called smoothing parameters, and they control smoothness of estimated functions. $J(f_j)$ is the roughness penalty for f_j. If we use the squared integral of the m-th derivative of f_j, $J(f_j) = \int\{f_j^{(m)}(t_j)\}^2 dt_j$ (where the integral is taken over a finite interval including $\{t_{ij}\}_{i=1,\ldots,n}$), we obtain natural splines of the degree $(2m-1)$ as maximum penalized likelihood estimates (MPLE) (Green and Silverman, 1994; Eubank, 1999). Using natural cubic splines $(m = 2)$ is a common choice.

It is convenient to introduce expression with basis functions for computation (and sometimes for approximation) of smoothing splines. Suppose that each f_j

$(j = 1, \ldots, r)$ is represented as a linear combination of q_j basis functions $\varphi_{jk}(t_j)$ $(k = 1, \ldots, q_j)$ such as the B-splines, that is, $f_j(t_j) = \sum_{k=1}^{q_j} \xi_{jk}\varphi_{jk}(t_j)$. Defining the matrix $B_j = \{\varphi_{jk}(t_{ij})\}_{i=1,\ldots,n;\, k=1,\ldots,q_j}$ and the vector $\boldsymbol{\xi}_j = (\xi_{j1}, \ldots, \xi_{jq_j})$, we can write as $\boldsymbol{f}_j = B_j\boldsymbol{\xi}_j$, Moreover, suppose that the roughness penalty for f_j is given by a quadratic form $J(f_j) = \boldsymbol{\xi}_j^{\mathrm{T}} K_j \boldsymbol{\xi}_j$, where K_j is a non-negative definite symmetric matrix depending on $(t_{1j}, \ldots, t_{nj})$. In the case $J(f_j) = \int \{f_j^{(m)}(t_j)\}^2 dt_j$, the (k, l)-th component of K_j is $\int \varphi_{jk}^{(m)}(t_j)\, \varphi_{jl}^{(m)}(t_j)\, dt_j$. See also Eubank (1999). (In the case of cubic smoothing splines, we can also use incidence matrices in place of B_j, and the penalty matrices K_j as defined in Green and Silverman (1994).) Then the penalized log-likelihood (2.3) becomes

$$l_{\mathrm{P}}(\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r; \boldsymbol{y}) = -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}) - \frac{1}{2}\sum_{j=1}^{r} \lambda_j \boldsymbol{\xi}_j^{\mathrm{T}} K_j \boldsymbol{\xi}_j + \text{const.}, \quad (2.4)$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\boldsymbol{\mu} = X_L\boldsymbol{\beta} + B_1\boldsymbol{\xi}_1 + \cdots + B_r\boldsymbol{\xi}_r$. By putting $\partial l_{\mathrm{P}}/\partial\boldsymbol{\beta} = \boldsymbol{0}$ and $\partial l_{\mathrm{P}}/\partial\boldsymbol{\xi} = \boldsymbol{0}$, we obtain the updating equations (2.2) for the backfitting algorithm, where $S_j = B_j(B_j^{\mathrm{T}}B_j + \lambda_j^* K_j)^{-1}B_j^{\mathrm{T}}$ $(\lambda_j^* = \sigma^2\lambda_j)$ is the spline smoother matrix.

2.2 Bayesian approaches for smoothing splines

We introduced the penalized likelihood approach in the frequentist's context in the previous subsection, and it is also explained in the Bayesian context. Introducing the roughness penalty is equivalent to considering the prior density of f_j which is proportional to $\exp\{-\frac{1}{2}\lambda_j J(f_j)\}$, and the smoothing parameter λ_j shows prior variability of f_j (Silverman, 1985).

We illustrate the Bayesian approach using the basis function representation. Suppose that the prior density of $\boldsymbol{\xi}_j$ is $p(\boldsymbol{\xi}_j; \lambda_j) \propto \exp(-\frac{1}{2}\lambda_j \boldsymbol{\xi}_j^{\mathrm{T}} K_j \boldsymbol{\xi}_j)$ $(j = 1, \ldots, r)$, where the density is defined only over $D_j = \{\boldsymbol{\xi}_j : \boldsymbol{1}^{\mathrm{T}} B_j \boldsymbol{\xi}_j = \boldsymbol{0}\}$ ($\boldsymbol{1}$ is the vector with all components 1) because of the constraint $\sum_i f_j(t_{ij}) = 0$, and that $\boldsymbol{\beta}$ has an improper prior $p(\boldsymbol{\beta}) \propto 1$. Let the conditional density of $\boldsymbol{y}$ for given $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r$ be denoted by $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r; \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu})\right\}$. By using the Bayes theorem, the joint posterior density of $\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r$ is proportional to the joint density of $\boldsymbol{y}$, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r$:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r | \boldsymbol{y}; \boldsymbol{\lambda}, \sigma^2) &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r; \sigma^2) p(\boldsymbol{\beta}) \prod_{j=1}^{r} p(\boldsymbol{\xi}_j; \lambda_j) \\ &\propto \exp l_{\mathrm{P}}(\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r; \boldsymbol{y}), \end{aligned} \quad (2.5)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}}$. Hence obtaining the mode of the joint posterior density of $\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r$ is equivalent to maximizing the penalized log-likelihood.

Another useful idea related to the Bayesian approach is to represent the smoothing spline as a linear mixed model. If the prior distribution of f_j is represented as

$$\boldsymbol{f}_j = X_{Sj}\boldsymbol{\delta}_j + Z_{Sj}\boldsymbol{a}_j, \quad (2.6)$$

the posterior mean and mode of f_j is a smoothing spline of the degree $(2m-1)$ (Green, 1987), where X_{Sj} is a $n \times (m-1)$ matrix, Z_{Sj} is a $n \times (q_j - m)$ matrix, $\boldsymbol{\delta}_j$ is a $(m-1)$-vector of fixed effect parameters, and $\boldsymbol{a}_j$ is a $(q_j - m)$-vector of random parameters whose components distribute independently and normally with the mean 0 and the variance λ_j^{-1}. It should be noted that, because of the constraint $\sum_i f_j(t_{ij}) = 0$, X_{Sj} has no column corresponding to the constant term. The additive regression model is then represented as a linear mixed model

$$\boldsymbol{y} = X_L\boldsymbol{\beta} + X_S\boldsymbol{\delta} + Z_S\boldsymbol{a} + \boldsymbol{\epsilon}, \quad \boldsymbol{a}_j \sim \mathrm{N}(\boldsymbol{0}, \lambda_j^{-1}\mathrm{I}_{q_j-m}), \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\mathrm{I}_n), \quad (2.7)$$

where $X_S = [X_{S1} \ \cdots \ X_{Sr}]$, $Z_S = [Z_{S1} \ \cdots \ Z_{Sr}]$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\delta}_r^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{a} = (\boldsymbol{a}_1^{\mathrm{T}}, \ldots, \boldsymbol{a}_r^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$, and I_k is a k-dimensional unit matrix. The MPLEs $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \tilde{\boldsymbol{a}})$ give the best linear unbiased predictors (BLUP) of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{a})$ (Robinson, 1991; Speed, 1991; Zhang *et al.*, 1998).

2.3 Selecting the smoothing parameters

The most popular and standard procedure of choosing smoothing parameter in the context of smoothing splines is cross-validation, represented by the generalized cross-validation (Craven and Wahba, 1979). However, the idea of cross-validation is to optimize on predicting responses, which does not match to a primary objective of nonparametric regression, that is, to explore nonlinear regression structure. Moreover, a distance between response variables might be difficult to take information on complicated regression structure into account. Some authors indicate that cross-validation often lead to undersmoothing (Diggle and Hutchinson, 1989; Simonoff, 1996, Wang, 1998). Optimization of some information criteria have been also considered (Hastie and Tibshirani, 1990; Eilers and Marx, 1996; Imoto and Konishi, 1999).

Another procedure of selecting the smoothing parameters is to maximize the marginal likelihood, which is proposed and called type-II likelihood by Good (1965). The marginal density of $\boldsymbol{y}$ is defined by integrating out the parameters $\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r$ constructing the additive regression model (2.7) from the joint density of $(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r)$:

$$p(\boldsymbol{y}; \boldsymbol{\lambda}, \sigma^2) = \int \cdots \int_D p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r; \sigma^2) p(\boldsymbol{\beta}) \prod_{j=1}^{r} p(\boldsymbol{\xi}_j; \lambda_j) d\boldsymbol{\beta}\, d\boldsymbol{\xi}_1 \cdots d\boldsymbol{\xi}_r,$$

where $D = \mathbf{R}^{p+1} \times D_1 \times \cdots \times D_r$. Then we obtain estimates of $\boldsymbol{\lambda}$ and σ^2 that maximize the marginal log-likelihood $l_{\mathrm{M}}(\boldsymbol{\lambda}, \sigma^2) = \log p(\boldsymbol{y}; \boldsymbol{\lambda}, \sigma^2)$. We call the procedure maximum marginal likelihood (MML). The procedure is also called empirical Bayes approach, and is discussed in Akaike (1980), Ishiguro and Arahata (1982), Kitagawa and Gersch (1984), Hastie and Tibshirani (2000) and so on. A specific form of the marginal log-likelihood has a special one for the PASS model, described in the next section. If we assume that the responses follow a normal distribution, the generalized maximum likelihood (GML) (Wahba, 1985) and the restricted maximum likelihood (REML) (Patterson and Thompson, 1971; Zhang *et al.*, 1998; Sakamoto, 2002) are equivalent to the MML.

3 Power additive smoothing splines

3.1 Power additive transformation in linear regression models

Here we describe the power additive transformation for linear regression models. Suppose that each response y_i $(i = 1, \ldots, n)$ is independently observed with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. We assume that y_i is normally distributed with a mean μ_i and a constant variance σ^2, and that the power-transformed response $y_i^{(\phi)}$ defined by (1.1) has a linear form

$$\eta_i \equiv \mathrm{E}[y_i^{(\phi)} | \boldsymbol{x}_i] = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1, \quad i = 1, \ldots, n. \tag{3.1}$$

Draper and Hunter (1969) considered choosing ϕ that attain the maximum mean square ratio in linear regression models. Goto (1992, 1995) called the method power additive transformation (PAT), in which the transformed expectation $\mu_i^{(\phi)}$ is considered to approximate η_i given by (3.1), that is,

$$\mu_i \approx \begin{cases} (\phi \eta_i + 1)^{1/\phi}, & \phi \neq 0 \\ \exp(\eta_i), & \phi = 0, \end{cases} \tag{3.2}$$

and the parameters $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$, ϕ and σ^2 are estimated with the maximum likelihood.

The estimates are computed using the Newton–Raphson or the Fisher scoring algorithm. For given ϕ and σ^2, the updating equation for $\boldsymbol{\beta}$ in the Fisher scoring algorithm (see, for example, McCullagh and Nelder (1989)) is given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{new}} = (X_L^{\mathrm{T}} W X_L)^{-1} X_L^{\mathrm{T}} W \boldsymbol{z}, \tag{3.3}$$

where X_L is the same as in Section 2.1, $W = \mathrm{diag}(w_1, \ldots, w_n)$ is the diagonal matrix composed of working weights $w_i = \hat{\mu}_i^{2-2\phi}$, and $\boldsymbol{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$ is the vector of working responses $z_i = (y_i - \hat{\mu}_i)/\sqrt{w_i} + \hat{\eta}_i$.

3.2 Power additive smoothing splines

We return to the situation where each responses y_i $(i = 1, \ldots, n)$ is independently observed with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{ir})^{\mathrm{T}}$. We assume that y_i is normally distributed with a mean μ_i and a constant variance σ^2, and that the power-transformed response $y_i^{(\phi)}$ has an additive form

$$\eta_i \equiv \mathrm{E}[y_i^{(\phi)} | \boldsymbol{x}_i, \boldsymbol{t}_i] = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_1 + f_1(t_{i1}) + \cdots + f_r(t_{ir}), \quad i = 1, \ldots, n, \tag{3.4}$$

where, as in Section 2, we give constraints $\sum_{i=1}^n f_j(t_{ij}) = 0$ $(j = 1, \ldots, r)$ for uniqueness. Both the parameters $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ and the smooth functions $f_1, \ldots, f_r$ are estimated together with the power parameter ϕ.

As in the PAT for the linear regression models (3.1), $\mu_i^{(\phi)}$ is considered to approximate η_i given by (3.4), and $\boldsymbol{\beta}$ and $f_1, \ldots, f_r$ are estimated by maximizing the penalized log-likelihood

$$l_{\mathrm{P}}(\boldsymbol{\beta}, f_1, \ldots, f_r; \boldsymbol{y}) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 - \frac{1}{2} \sum_{j=1}^r \lambda_j J(f_j) + \mathrm{const.} \tag{3.5}$$

for given ϕ, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^{\mathrm{T}}$ and σ^2, where $J(f_j) = \int \{f_j^{(m)}(t_j)\}^2 dt_j$ is the roughness penalty for f_j as in Section 2. We call the MPLEs of $f_1, \ldots, f_r$ power additive smoothing splines (PASS), which are natural splines of the degree $(2m-1)$.

We introduce the basis function expression for f_j as in section 2.1. Then the penalized log-likelihood has the same form as (2.4), where $\boldsymbol{\mu}^{(\phi)} = X_L\boldsymbol{\beta} + B_1\boldsymbol{\xi}_1 + \cdots + B_r\boldsymbol{\xi}_r$. The updating equation for the MPLEs $\hat{\boldsymbol{\beta}}$ and $\hat{f}_j$ in the Fisher scoring are given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{new}} = (X_L^{\mathrm{T}} W X_L)^{-1} X_L^{\mathrm{T}} W (\boldsymbol{z} - \hat{\boldsymbol{f}}_1 - \cdots - \hat{\boldsymbol{f}}_r), \tag{3.6a}$$

$$\hat{\boldsymbol{f}}_j^{\mathrm{new}} = S_{Wj}(\boldsymbol{z} - X_L\hat{\boldsymbol{\beta}} - \textstyle\sum_{k\neq j} \hat{\boldsymbol{f}}_k), \qquad j = 1, \ldots, r, \tag{3.6b}$$

where $W = \mathrm{diag}(w_1, \ldots, w_n)$, $w_i = \hat{\mu}_i^{2-2\phi}$, $\boldsymbol{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$, $z_i = (y_i - \hat{\mu}_i)/\sqrt{w_i} + \hat{\eta}_i$, and S_{Wj} is the weighted spline smoother $S_{Wj} = B_j(B_j^{\mathrm{T}} W B_j + \lambda_j^* K_j)^{-1} B_j^{\mathrm{T}} W$. For given ϕ and $\lambda_j^* = \sigma^2\lambda_j$, the working weights w_i and the working responses z_i are updated in each time $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{f}}_j$ are once updated, and the updating iteration is continued until these values converge. The algorithm is a reweighted version of the backfitting algorithm, which is also called local scoring by Hastie and Tibshirani (1986, 1990). In the case $\phi = 1$, where $W = \mathrm{I}$ and $\boldsymbol{z} = \boldsymbol{y}$, it results in the ordinary backfitting (2.2).

3.3 Selecting the power and the smoothing parameters

An objective of the PASS model (3.4) is to diagnose additivity, that is, to find the most appropriate power transformation that satisfies the requirement of additivity. The power parameter ϕ indicates the additive transformation directly, and so its estimation is essential. Box and Hill (1974) applied the maximum marginal likeihood to select a power parameter for power weighted transformation (PAT) in linear regression models to diagnose homoscedasticity. As we have seen in Section 2.3, the maximum marginal likelihood is also applied to the selection of the smoothing parameter, and Sakamoto (2004) proposed to estimate the power and the smoothing parameters jointly in power weighted smoothing spline models (PWSS). Here, we propose to estimate the power and the smoothing parameters jointly in the PASS model with the maximum marginal likelihood.

We introduce the prior density of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_j$ $(j = 1, \ldots, r)$ as in Section 2.2. Then the marginal density of $\boldsymbol{y}$ becomes

$$\begin{aligned} p(\boldsymbol{y}; \phi, \boldsymbol{\lambda}, \sigma^2) &= \int \cdots \int_D p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r; \phi, \sigma^2) p(\boldsymbol{\beta}) \prod_{j=1}^{r} p(\boldsymbol{\xi}_j; \lambda_j) d\boldsymbol{\beta}\, d\boldsymbol{\xi}_1 \cdots d\boldsymbol{\xi}_r \\ &\propto \int \cdots \int_D \exp l_{\mathrm{P}}(\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_r; \boldsymbol{y}) d\boldsymbol{\beta}\, d\boldsymbol{\xi}_1 \cdots d\boldsymbol{\xi}_r. \end{aligned} \tag{3.7}$$

We consider (3.7) as a function of $(\phi, \boldsymbol{\lambda}, \sigma^2)$ and maximize it with respect to these parameters. However, it is impossible to compute the integral in the right-hand side exactly except when $\phi = 1$.

Some approaches of computing the marginal density of $\boldsymbol{y}$ approximately have been discussed. One approach is to use the Markov chain Monte Carlo methods, such as the Gibbs sampling (Zeger and Karim, 1991) and the Monte Carlo filter (Kitagawa, 1996). However, the numerical results with them depend on random numbers and so they require numerous time of computation. Another approach is to derive approximation forms without using the integral. We adopt a Laplace approximation (Tierney and Kadane, 1986; Davison, 1986) because of its easy computation.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\xi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\xi}_r^{\mathrm{T}})^{\mathrm{T}}$ for simplicity of notation. We consider Taylor expansion of the penalized log-likelihood $l_{\mathrm{P}}(\boldsymbol{\theta}; \boldsymbol{y})$ around its maximum point, that is, the MPLEs $\hat{\boldsymbol{\theta}}$, and approximate the Hessian of the penalized log-likelihood to its expectation. Then we obtain the Laplace approximation

$$l_{\mathrm{P}}(\boldsymbol{\theta}; \boldsymbol{y}) \approx l_{\mathrm{P}}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \tag{3.8}$$

where

$$\mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}}) = \left\{ \mathrm{E}\left(-\frac{\partial^2 l_{\mathrm{P}}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right) \right\}_{\hat{\boldsymbol{\theta}}}.$$

By substituting (3.8) into (3.7), an approximated marginal density of $\boldsymbol{y}$ becomes

$$\begin{aligned} p(\boldsymbol{y}; \phi, \boldsymbol{\lambda}, \sigma^2) &\approx \exp l_{\mathrm{P}}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) \int \cdots \int_D \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right\} d\boldsymbol{\theta} \\ &\propto |\mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})|_+^{-1/2} \exp l_{\mathrm{P}}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}), \end{aligned}$$

where $|\mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})|_+$ is the product of non-zero eigenvalues of $\mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})$. The approximation is exact if $\phi = 1$. Hence we obtain an approximated marginal log-likelihood

$$l_{\mathrm{M}}(\phi, \boldsymbol{\lambda}, \sigma^2; \boldsymbol{y}) = l_{\mathrm{P}}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) - \frac{1}{2} \log |\mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})|_+ \mathrm{const.}, \tag{3.9}$$

and we maximize (3.9) with respect to $(\phi, \boldsymbol{\lambda}, \sigma^2)$ to obtain MML estimates.

After reparametering as $\boldsymbol{\lambda}^* = \sigma^2 \boldsymbol{\lambda}$ and some computation, we obtain the MML estimate of $\hat{\sigma}^2$ explicitly as

$$\hat{\sigma}^2 = \frac{1}{n-d} \boldsymbol{z}^{\mathrm{T}} W (\boldsymbol{z} - \hat{\boldsymbol{\eta}}), \tag{3.10}$$

where $d = 1 + p + r(m-1)$, $\boldsymbol{z}$ and W are working responses and working weights, respectively, when the Fisher scoring algorithm (3.6) converges, and $\hat{\boldsymbol{\eta}} = X_L \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{f}}_1 + \cdots + \hat{\boldsymbol{f}}_r$. By substituting (3.10) into (3.9) and conducting some matrix operation using a BLUP equation in the linear mixed model (2.7), we obtain a specific form of the approximated marginal likelihood

$$l_{\mathrm{M}}(\phi, \boldsymbol{\lambda}^*, \hat{\sigma}^2; \boldsymbol{y}) = -\frac{n-d}{2}(1 + \log \hat{\sigma}^2) + \sum_{j=1}^{r} \frac{q_j - m}{2} \log \lambda_j^* - \frac{1}{2} \log |C_W| + \mathrm{const.}, \tag{3.11}$$

where

$$C_W = \begin{bmatrix} X_L^{\mathrm{T}} W X_L & X_L^{\mathrm{T}} W X_S & X_L^{\mathrm{T}} W Z_S \\ X_S^{\mathrm{T}} W X_L & X_S^{\mathrm{T}} W X_S & X_S^{\mathrm{T}} W Z_S \\ Z_S^{\mathrm{T}} W X_L & Z_S^{\mathrm{T}} W X_S & Z_S^{\mathrm{T}} W Z_S + \Lambda^* \end{bmatrix}$$

and

$$\Lambda^* = \begin{bmatrix} \lambda_1^* \mathrm{I}_{q_1-m} & & O \\ & \ddots & \\ O & & \lambda_r^* \mathrm{I}_{q_r-m} \end{bmatrix}.$$

In the case $r = 1$, we obtain its more specific form

$$\begin{aligned} l_{\mathrm{M}}(\phi, \boldsymbol{\lambda}^*, \hat{\sigma}^2; \boldsymbol{y}) = &-\frac{n-(p+m)}{2}(1+\log\hat{\sigma}^2) + \frac{q_1-m}{2}\log\lambda_1^* \\ &- \frac{1}{2}\log|B_1^{\mathrm{T}} W B_1 + \lambda_1^* K_1| - \frac{1}{2}\log|X_L^{\mathrm{T}} W (I - S_1) X_L| + \text{const.}, \end{aligned}$$

which is also derived in the case $\phi = 1$ ($W = \mathrm{I}$) by Tanabe and Sagae (1992) as the improper prior Bayes information criterion (IPBIC).

3.4 The algorithm of PASS

The algorithm of PASS is summarized as follows.

1. For given ϕ,
   (a) Setting initial weights:
      (i) Consider the PAT model $y_i \sim \mathrm{N}(\mu_i, \sigma^2)$, $\mu_i^{(\phi)} = \beta_0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{\beta}_2$. First set the initial weights as $w_i = y_i^{2-2\phi}$, for example, and find regression estimates $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ in the way described in Section 3.1.
      (ii) Set the initial weights in (b) as $w_i = \hat{\mu}_i^{2-2\phi}$, where $\hat{\mu}_i^{(\phi)} = \hat{\beta}_0 + \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{t}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_2$.
   (b) For given $\boldsymbol{\lambda}^*$, fit the PASS model (3.4) in the way described in Section 3.2.
      (i) Update the weights as $w_i = \hat{\mu}_i^{2\phi-2}$.
      (ii) Update $\hat{\boldsymbol{\beta}}$ with (3.6a).
      (iii) For $j = 1, \ldots, r$, update $\hat{\boldsymbol{f}}_j$ with (3.6b).
      (iv) Decide whether the estimates have converged or not. If converged go to (c), otherwise return to (i).
   (c) Compute the MML estimate of σ^2 with (3.10), and the approximated log-likelihood $l_{\mathrm{M}}(\phi, \boldsymbol{\lambda}^*, \hat{\sigma}^2)$ with (3.11), described in Section 3.3.
2. Find $(\phi, \boldsymbol{\lambda}^*)$ that maximizes $l_{\mathrm{M}}(\phi, \boldsymbol{\lambda}^*, \hat{\sigma}^2)$ in the following way.
   - Grid search, or
   - Newton–Raphson algorithm, in which the derivarives of l_{M} are approximated by its differences.

The Bayesian posterior covariance matrix of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_r)$ is obtained as $\hat{\sigma}^2 \mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})^-$, where $\mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})^-$ is the generalized inverse of $\mathcal{I}_{\mathrm{P}}(\hat{\boldsymbol{\theta}})$. Especially, the

posterior covariance matrices of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{f}}_j$ are

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \hat{\sigma}^2 C_W^{(L)}, \\ \mathrm{var}(\hat{\boldsymbol{f}}_j) &= \hat{\sigma}^2 [X_{Sj}\, Z_{Sj}] C_W^{(j)} \begin{bmatrix} X_{Sj}^{\mathrm{T}} \\ Z_{Sj}^{\mathrm{T}} \end{bmatrix}, \quad j = 1, \ldots, r,\end{aligned}$$

where $C_W^{(L)}$ and $C_W^{(j)}$ are corresponding linear and the j-th smooth components of C_W^{-1}, respectively.

4 Application to some data

We describe some application of the PASS models (3.4) to four data sets listed in Table 1, comparing with the power additive transformation in linear regression (PATLR) models (3.1) and the ordinary additive regression models (2.1). If we set $\phi = 1$ in PASS, we get the ordinary additive model, whie if the smoothing parameters are taken to infinity, the PASS model tends to the PATLR model. For the PASS and the additive regression models we also compute the equivalent degree of freedom (EDF) for each $\hat{f}_j$ (Hastie and Tibshirani, 1990), where we define it as the trace of the smoother matrix S_j (or S_{Wj}) minus 1. If it is close to 1, $\hat{f}_j$ is almost linear, whie if it is larger, $\hat{f}_j$ becomes rough. The results of analysis are shown in Table 2.

For the data set (1), the estimate of the power parameter ϕ with PASS indicates almost log transformation, which suggests multiplicative relationship. There is small difference between PASS and the additive model in the smoothness of $\hat{f}_1$ and $\hat{f}_2$. For the data sets (2) and (3), extremely large values of ϕ are indicated with PASS, which suggests that the data might involve some complicated relationship that can not be explained by the additive model. For the data set (4), we could not obtain convergence in the backfitting algorithm if ϕ is larger than 1. Most of the failure in convergence is caused by obtaining non-positive $\hat{\mu}_i$ with the relation (3.2).

Table 1. Four data sets.

Data	Variables	Sample size
(1) Engine exhaust for burnings of ethanol Chambers and Hastie (1992)	y : NOx concentration t_1: compression ratio t_2: equivalence ratio	$n = 88$
(2) Volume of cherry blossom Bowman and Azzalini (1997) Schimek and Turlach (2000)	y : volume t_1: diameter t_2: height	$n = 31$
(3) Diabetes data Hastie and Tibshirani (1990)	y : log concentration of C-peptide t_1: age t_2: base deficit (acidity)	$n = 43$
(4) New York ozone concentration Chambers and Hastie (1992)	y : ozone concentration t_1: solar radiation t_2: temperature t_3: wind speed	$n = 111$

Table 2. Result of analysis with PASS and so on.

データ	PATLR (3.1)	PASS (3.4)		Additive (2.1)
	$\hat{\phi}$	$\hat{\phi}$	EDF	EDF
(1)	(3.515)	0.228	$\hat{f}_1$: 2.44	$\hat{f}_1$: 2.74
			$\hat{f}_2$: 9.06	$\hat{f}_2$: 10.03
(2)	0.521	(5.139)	$\hat{f}_1$: 8.26	$\hat{f}_1$: 3.41
			$\hat{f}_2$: 1.04	$\hat{f}_2$: 1.01
(3)	(4.000)	(10.00)	$\hat{f}_1$: 2.38	$\hat{f}_1$: 2.52
			$\hat{f}_2$: 2.97	$\hat{f}_2$: 2.02
(4)	(0.400)	(1.000)		$\hat{f}_1$: 1.72
				$\hat{f}_2$: 3.54
				$\hat{f}_3$: 3.47

Note: each estimate in parantheses indicates that the maximum of the marginal likelihood was found at the boundary of the region where we obtained convergence in the backfitting algorithm.

5 Concluding remarks

We have proposed the PASS model to diagnose validity of assuming additivity in the additive regression model. The regression parameter and smooth functions are estimated with the penalized likelihood, in which we have constructed the backfitting algorithm based on Fisher scoring. The power and the smoothing parameters, which govern global nonlinear regression structure, are estimated with the maximum marginal likelihood, in which we have considered the Laplace approximation of the marginal likeihood. We have examined application of the PASS model to some data sets. One problem is to investigate how we obtain convergence in the backfitting algorithm easily.

We intend to evaluate the performance of the PASS model through some simulation experiment, both when an additive relationship holds and when non-additive components exist. Moreover, from a practical point of view, we should examine application to longitudinal data, functional data and so on.

Acknowledgement

This research is supported by MEXT Grant-in-Aid for Young Scientists (B) (2002–2004, No. 14780171).

References

Akaike, H. (1980). Likelihood and Bayes procedure. In *Bayesian Statistics* (Bernardo, J. M. *et al.*, eds.), pp. 143–166, University Press, Valencia.

Bartlett, M. S. (1947). The use of transformations. *Biometrics*, **3**, 39–52.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis.* Oxford University Press, Oxford.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, California.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580–619.

Box. G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B*, **26**, 211–243.

Box. G. E. P. and Hill, W. J. (1974). Correcting inhomogenity of variance with power transformation weighting. *Technometrics*, **16**, 385–389.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Pacific Grove, California.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika*, **73**, 323–332.

Diggle, P, J. and Hutchinson, M. F. (1989). On spline smoothing with autocorrelated errors. *Austral. J. Statist.*, **31**, 166–182.

Draper, N. R. and Hunter, W. G. (1969). Transformation: some examples revisited. *Technometrics*, **11**, 23–40.

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

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing* (2nd Edition). Marcel Dekker, New York.

Friedman, J. H. (1991). Multiple adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–144.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press.

Goto, M. (1992). Extensive views of power transformation: some recent developments. *Invited paper at Honolulu Conference on Computational Statistics as a memorial of the fifth anniversary of JSCS*, 1-5.

Goto, M. (1995). Double power transformations and their performances. *Proceedings of the International Conference on Statistical Methods and Statistical Computing for Quality and Productivity Improvement (Seoul, Korea), Vol. I*, 386–397.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Statist. Rev.*, **55**, 245–259.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Sci.*, **1**, 297–318.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (2000). Bayesian backfitting (with discussion). *Statist. Sci.*, **15**, 196–223.

Imoto, S. and Konishi, S. (1999). Nonlinear regression models using B-spline and information criteria. *Proceedings of the Institute of Statistical Mathematics*, **47**, 359–373 (in Japanese)

Ishiguro, M. and Arahata, E. (1982). A Bayesian spline regression. *Proceedings of the Institute of Statistical Mathematics*, **30**, 29–36 (in Japanese).

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian non-linear state-space models. *J. Comput. Graph. Statist.*, **5**, 1–25.

Kitagawa, G. and Gersch, W. (1984). A smoothness prior — state space modeling of time series with trend and seasonality. *J. Amer. Statist. Assoc.*, **79**, 378–389.

Linton, O. B., Chen, R., Wang, N. and Härdle, W. (1997). An analysis of transformations for additive nonparametric regression. *J. Amer. Statist. Assoc.*, **92**, 1512–1521.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). Chapman and Hall, London.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Robinson, G. K. (1991). BLUP is a good thing: the estimation of random effects (with discussion). *Statist. Sci.*, **6**, 15–51.

Sakamoto, W. (2002). Selecting smoothing parameters with restricted maximum likelihood estimation: a procedure for efficient computation and its apprication. *Bulletin Comput. Statist. Japan*, **15**, 19–45 (in Japanese).

Sakamoto, W. (2004). Diagnosing homoscedasticity with the power-weighted smoothing spline. *Japanese J. Appl. Statist*, **33**, 27–49 (in Japanese).

Schimek, M. G. and Turlach, B. A. (2000). Additive and generalized additive models. In *Smoothing and Regression: Approaches, Computation, and Application* (Schimek, M. G., ed.), Chapter 10 (pp. 277–327). John Wiley and Sons, New York.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B*, **47**, 1–52.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer-Verlag, New York.

Speed, T. (1991). Comment on Robinson (1991). *Statist. Sci.*, **6**, 42–44.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, **13**, 689–705.

Tanabe, K. and Sagae, M. (1992). How to incorporate inconsistent prior information safely in Bayesian lienar models by adjusting hyperparameter via data-based method. *Institute of Statistical Mathematics Research Memorandom*, No. 442.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82–86.

Tibshirani, R. (1988). Estimating optimal transformations for regression via additivity and variance stabilization. *J. Amer. Statist. Assoc.*, **83**, 394–405.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameters in the generalized spline smoothing problem. *Ann. Statist.*, **4**, 1378–1402.

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

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.*, **23**, 1865–1895.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Amer. Statist. Assoc.*, **93**, 341–348.

Wang, N. and Ruppert, D. (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. *J. Amer. Statist. Assoc.*, **90**, 522–534.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.*, **86**, 79–86.

Zhang, Z., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *J. Amer. Statist. Assoc.*, **93**, 710–719.