DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 966

August 1996

Smoothing Spline Models With Correlated Random Errors[1]

by

Yuedong Wang

[1]Research sponsored in part by NIH Grants R01 EY09946, P60 DK20572 and P30 HD18258.

Smoothing Spline Models With Correlated Random Errors

Yuedong Wang [†]
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

August 30, 1996

Abstract

Spline smoothing is a popular method of estimating the functions in a nonparametric regression model. Its performance greatly depends on the choice of smoothing parameters. Many methods of selecting smoothing parameters such as CV, GCV, UBR and GML are developed under the assumption of independent observations. They fail badly when data are correlated. In this paper, we assume observations are correlated and the correlation matrix depends on a parsimonious set of parameters. We extend the GML, GCV and UBR methods to estimate the smoothing parameters and the correlation parameters simultaneously. We also connect a smoothing spline model with three mixed-effects models. These connections show that the smoothing spline estimates evaluated at design points are BLUP estimates and the GML estimates of the smoothing parameters and the correlation parameters are REML estimates. These connections also suggest a way to fit a spline model with correlated errors using the existing SAS procedure `proc mixed`*. We illustrate our methods with applications to two time series data sets and a spatial data set. Simulations are conducted to evaluate and compare the performance of the GML, GCV, UBR methods and the method proposed in Diggle and Hutchinson (1989). The GML method is recommended since it is stable and works well in all simulations. It performs better than other methods, especially when the sample size is not large.*

Keywords: Best linear unbiased prediction; Generalized cross-validation; Generalized maximum likelihood; Unbiased risk; Mixed-effects model; Restricted maximum likelihood; Smoothing parameters; Smoothing spline; Smoothing spline ANOVA.

1 Introduction

In this paper we consider the general smoothing spline models in Wahba (1990). For an arbitrary index set $\mathcal{T}$ (e.g., $\mathcal{T} = \{1, \cdots, N\}$, $\mathcal{T} = [0, 1]$ or $\mathcal{T} = E^d$, where E^d is the Euclidean

[†]E-mail: yuedong@umich.edu.

d-space), let $\mathcal{H}$ be a reproducing kernel Hilbert space (r.k.h.s.) of real-valued functions on $\mathcal{T}$. See Aronszajn (1950) and Wahba (1990) for details on r.k.h.s. Assume that observations y_i are generated by

$$y_i = L_i f + \epsilon_i, \qquad i = 1, \cdots, n, \tag{1}$$

where $f \in \mathcal{H}_R$, the ϵ_i's are zero mean random errors and the L_i's are bounded linear functionals on $\mathcal{H}_R$. Interesting examples of bounded linear functionals on appropriate spaces are $L_i f = f(t_i)$, $L_i f = a_i f(t_i - b_i)$, $\int_{\mathcal{T}} w_i(t) f(t) dt$ and $L_i f = f'(t_i)$.

Spline smoothing provides a powerful tool for estimating f which does not require specifying a parametric form for f. There is a vast literature on this topic (Eubank 1988, Green and Silverman 1994, Hastie and Tibshirani 1990, Wahba 1990). Almost all of this literature is based on the assumption that the random errors are independent. Often in practice observations are correlated. Examples are time series data, spatial data and clustered data. It is well known that correlation has a great effect on the selection of smoothing parameters, which are critical to the performance of smoothing spline estimates. Popular methods for selecting smoothing parameters such as cross-validation (CV), generalized cross-validation (GCV), generalized maximum likelihood (GML) and unbiased risk (UBR) fail badly when data are positively correlated (see Altman (1990), Diggle and Hutchinson (1989)). To show how standard smoothing parameter selection methods are affected by correlation of the error, we simulated data from the model $y_i = \sin 2\pi i/100 + \epsilon_i$, $i = 1, \cdots, 100$, where ϵ's are generated by a first-order autoregressive process with mean zero, standard deviation 0.1 and first-order correlation 0.55. Figure 1 shows one "typical" (i.e., the first replicate of the simulation) set of simulated data, the true function and three estimates with smoothing parameters chosen by standard GML, GCV and UBR methods respectively. These estimates are wiggly which indicates that the estimated smoothing parameters are too small.

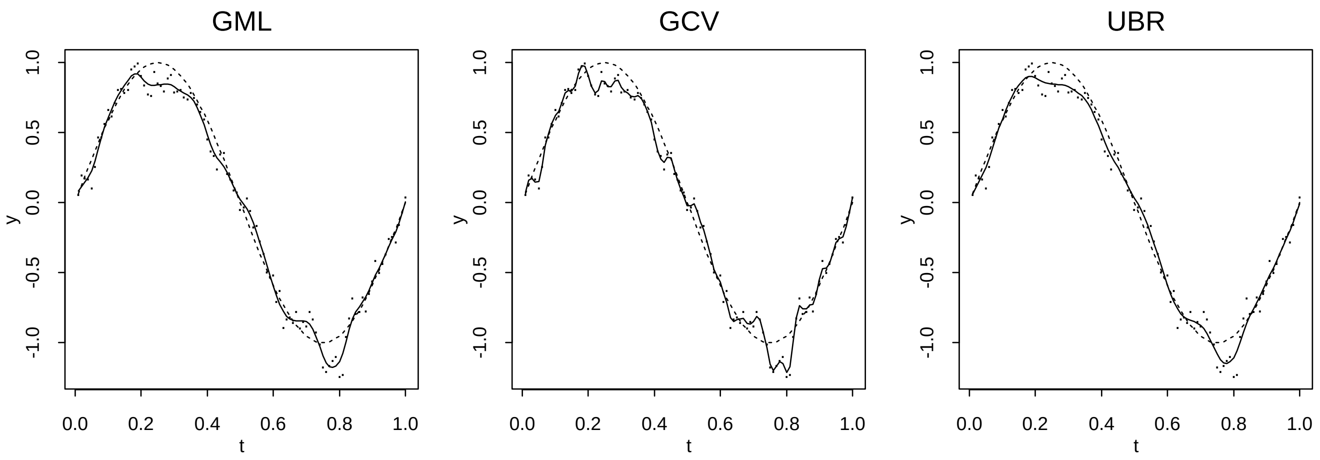

Figure 1: Plots of the true function (dotted lines), observations (dots) and estimates (solid lines) with smoothing parameters chosen by GML (left), GCV (middle) and UBR (right).

Diggle and Hutchinson (1989), Hurvich and Zeger (1990) and Kohn, Ansley and Wong (1992) considered the special case of spline smoothing for time series data. Several authors

(Altman 1990, Hart and Wehrly 1986, Hart 1991, Hart 1994, Herrmann, Gasser and Kneip 1992, Raz, Turetsky and Fein 1989) studied the effects of correlation on the selection of smoothing parameters and developed various methods for kernel regression. Almost all of these methods are developed for time series only and some even requires design points equally spaced.

In this paper we assume that the random errors in (1) are correlated. We extend the usual GML, GCV and UBR methods to these models. The general smoothing spline models on arbitrary domains and smoothing spline ANOVA models are used to present our methods. Therefore these methods can be applied not only to time series data, but also to spatial, spatial-temporal and longitudinal data (Wang 1996). We do not assume any specific error structure. Instead, we assume the correlation matrix depends on a parsimonious set of parameters. We propose to estimate the smoothing and the correlation parameters simultaneously. We also connect a smoothing spline model with three linear mixed-effects models. With these connections, one can easily calculate the smoothing spline estimates using the SAS procedure `proc mixed`.

In section 2, based on a Bayesian model, we propose an extension of the GML method to estimate the smoothing parameters and the correlation parameters simultaneously. In section 3, we provide connections between the smoothing spline models and the mixed-effects models that give another motivation for the GML method and suggest a way to fit a smoothing spline model using the existing SAS procedure. In section 4, we introduce some extensions of the GCV method and the UBR method. In section 5, these methods are extended to fit multivariate functions modeled by smoothing spline ANOVA decompositions. In section 6, we fit data from two time series and an environmental model involving spatial data using the GML method. In section 7, extensive simulations are conducted to evaluate and compare the finite sample performance of the GML, GCV, UBR methods and the method proposed in Diggle and Hutchinson (1989).

2 Generalized Maximum Likelihood Estimates

The r.k.h.s. $\mathcal{H}$ in section 1 is a prechosen model space. See Wahba (1990) for a discussion of different spaces for different index sets. Suppose $\mathcal{H}$ can be decomposed to $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is a finite dimensional space containing terms which are not going to be penalized. Denote $M = \dim(\mathcal{H}_0)$. Let R^1 be the reproducing kernel (r.k.) of $\mathcal{H}_1$. Suppose observations are generated by (1) and

$$\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T \sim \mathrm{N}(0, \sigma^2 W^{-1}). \tag{2}$$

The penalized likelihood estimate of f is the solution to

$$\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} (\boldsymbol{y} - \boldsymbol{f})^T W (\boldsymbol{y} - \boldsymbol{f}) + \lambda \| P_1 f \|^2 \right\}, \tag{3}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ and $\boldsymbol{f} = (L_1 f, \cdots, L_n f)^T$. The first part in (3) is proportional to the negative log likelihood and it measures goodness-of-fit. In the second part, P_1 is the orthogonal projection of f onto $\mathcal{H}_1$ in $\mathcal{H}$. Thus the second part is a penalty to the departure of f from the space $\mathcal{H}_0$. Usually it is a penalty to the roughness of f. For example, $\|P_1 f\|^2 = \int_0^1 (f^{(M)})^2 dt$ when $\mathcal{T} = [0,1]$. λ is a smoothing parameter. It controls the trade-off between goodness-of-fit and the departure of the estimate from the space $\mathcal{H}_0$.

Let $\phi_1, \cdots, \phi_M$ be a basis for $\mathcal{H}_0$ and set $\xi_i(t) = L_{i(\cdot)} R^1(t, \cdot)$. Kimeldorf and Wahba (1971) showed that the solution to (3) has the form

$$\hat{f} = \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{i=1}^{n} c_i \xi_i \ . \tag{4}$$

Let $T_{n \times M} = \{L_i \phi_\nu\}_{i=1}^{n}{}_{\nu=1}^{M}$ and $\Sigma = \{< \xi_i, \xi_j >\}$. For fixed λ and W, it is easy to check that $\boldsymbol{c} = (c_1, \cdots, c_n)^T$ and $\boldsymbol{d} = (d_1, \cdots, d_M)^T$ are solutions to

$$\begin{aligned} T^T W \Sigma \boldsymbol{c} + T^T W T \boldsymbol{d} &= T^T W \boldsymbol{y}, \\ (\Sigma W \Sigma + n\lambda \Sigma) \boldsymbol{c} + \Sigma W T \boldsymbol{d} &= \Sigma W \boldsymbol{y}. \end{aligned} \tag{5}$$

(5) is equivalent to

$$\begin{pmatrix} T^T W T & T^T W \\ \Sigma W T & \Sigma W + n\lambda I \end{pmatrix} \begin{pmatrix} \boldsymbol{d} \\ \Sigma \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} T^T W \boldsymbol{y} \\ \Sigma W \boldsymbol{y} \end{pmatrix}. \tag{6}$$

The system (6) is definite when T is of full column rank, which we assume to be true in this paper. Thus $\hat{\boldsymbol{f}} = (L_1 \hat{f}, \cdots, L_n \hat{f})^T = T\boldsymbol{d} + \Sigma \boldsymbol{c}$ is always unique. In fact the solution to (3) exists and is unique (Gu and Qiu (1993)). We only need *a* solution to (5). It is easy to check that a solution to

$$\begin{aligned} (\Sigma + n\lambda W^{-1}) \boldsymbol{c} + T \boldsymbol{d} &= \boldsymbol{y}, & (7) \\ T^T \boldsymbol{c} &= 0, & (8) \end{aligned}$$

is also a solution to (5). Thus we need to solve (7) and (8) for $\boldsymbol{c}$ and $\boldsymbol{d}$. Let

$$T = (Q_1 \ \ Q_2) \begin{pmatrix} R \\ 0 \end{pmatrix}$$

be the QR decomposition of T. Then it is easy to check that

$$\begin{aligned} \boldsymbol{c} &= Q_2 (Q_2^T (\Sigma + n\lambda W^{-1}) Q_2)^{-1} Q_2^T \boldsymbol{y}, \\ R\boldsymbol{d} &= Q_1^T (\boldsymbol{y} - (\Sigma + n\lambda W^{-1}) \boldsymbol{c}). \end{aligned} \tag{9}$$

Obviously, $\hat{\boldsymbol{f}}$ is a linear function of $\boldsymbol{y}$. Suppose that $\hat{\boldsymbol{f}} = A\boldsymbol{y}$, where A is the "hat" matrix. From (7) and $\hat{\boldsymbol{f}} = T\boldsymbol{d} + \Sigma \boldsymbol{c}$ we have

$$(I - A)\boldsymbol{y} = \boldsymbol{y} - \hat{\boldsymbol{f}} = n\lambda W^{-1} \boldsymbol{c} = n\lambda W^{-1} Q_2 (Q_2^T (\Sigma + n\lambda W^{-1}) Q_2)^{-1} Q_2^T \boldsymbol{y}$$

for all $\boldsymbol{y}$. Thus

$$I - A = n\lambda W^{-1} Q_2 (Q_2^T (\Sigma + n\lambda W^{-1}) Q_2)^{-1} Q_2^T. \tag{10}$$

Note that A is not symmetric, which is different from the independent case.

Consider the following Bayesian model. Let a prior for f be

$$F(t) = \sum_{\nu=1}^{M} \theta_\nu \phi_\nu(t) + b^{1/2} X(t), \qquad t \in \mathcal{T}, \tag{11}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_M)^T \sim \mathrm{N}(0, aI)$, a and b are positive constants. $X(t)$, $t \in \mathcal{T}$ is a zero mean Gaussian stochastic process independent of $\boldsymbol{\theta}$ with covariance $\mathrm{E}X(s)X(t) = R^1(s,t)$. Suppose observations are generated by

$$y_i = L_i F + \epsilon_i, \qquad i = 1, \cdots, n,$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T \sim \mathrm{N}(0, \sigma^2 W^{-1})$. This Bayesian model is the same as the one described in Wahba (1990) except that the errors are correlated. Thus, following arguments in Wahba (1990), one may show that with $\lambda = \sigma^2/nb$, for each fixed t,

$$\lim_{a\to\infty} \mathrm{E}(F(t)|\boldsymbol{y}) = \hat{f}(t), \tag{12}$$

$$\lim_{a\to\infty} \mathrm{Cov}(\boldsymbol{F}|\boldsymbol{y}) = \sigma^2 A W^{-1}, \tag{13}$$

where $\boldsymbol{F} = (L_1 F, \cdots, L_n F)^T$.

The marginal distribution of $\boldsymbol{y}$ is $\mathrm{N}(0, b(\eta T T^T + \Sigma + n\lambda W^{-1}))$, where $\eta = a/b$. Let

$$\begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} Q_2^T \\ \frac{1}{\sqrt{\eta}} T^T \end{pmatrix} \boldsymbol{y}.$$

Then

$$\begin{aligned} \boldsymbol{z} &= Q_2^T \boldsymbol{y} \sim \mathrm{N}(0, b Q_2^T (\Sigma + n\lambda W^{-1}) Q_2), \\ \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{w}) &= \frac{b}{\sqrt{\eta}} Q_2^T (\eta T T^T + \Sigma + n\lambda W^{-1}) T \longrightarrow 0, \quad \eta \to \infty, \\ \mathrm{Var}(\boldsymbol{w}) &= \frac{b}{\eta} T^T (\eta T T^T + \Sigma + n\lambda W^{-1}) T \longrightarrow b(T^T T)(T^T T), \quad \eta \to \infty. \end{aligned}$$

In the following of this paper we assume that the covariance matrix W^{-1} depends on some correlation parameters $\boldsymbol{\tau}$. Interesting examples of covariance structures are first-order autoregressive for time series, compound symmetry or unstructured for repeated measures, and exponential for spatial data (Jennrich and Schluchter 1986, SAS Institute 1992). Let $B(\lambda, \boldsymbol{\tau}) = \Sigma + n\lambda W^{-1}$, where the dependence on parameters λ and $\boldsymbol{\tau}$ is expressed explicitly. As argued in Wahba (1985), the distribution of $\boldsymbol{w}$ is independent of both λ and $\boldsymbol{\tau}$. Therefore the maximum likelihood estimates of λ and $\boldsymbol{\tau}$ should be based on the marginal distribution of $\boldsymbol{z}$ alone. Accordingly, the generalized maximum likelihood (GML) estimates of λ and $\boldsymbol{\tau}$ are maximizers of the log likelihood based on $\boldsymbol{z}$:

$$l_1(\lambda, \boldsymbol{\tau}, b|\boldsymbol{z}) = -\frac{1}{2} \log |b Q_2' B(\lambda, \boldsymbol{\tau}) Q_2| - \frac{1}{2b} \boldsymbol{z}' (Q_2' B(\lambda, \boldsymbol{\tau}) Q_2)^{-1} \boldsymbol{z} + C_1,$$

where C_1 is a constant. Maximizing l_1 with respect to b, we have $\hat{b} = \boldsymbol{z}'(Q_2'B(\lambda,\boldsymbol{\tau})Q_2)^{-1}\boldsymbol{z}/(n-M)$. Then the GML estimates of λ and $\boldsymbol{\tau}$ are maximizers of

$$\begin{aligned} l_2(\lambda,\boldsymbol{\tau}|\hat{b}) &= -\frac{1}{2}\log|Q_2'B(\lambda,\boldsymbol{\tau})Q_2| - \frac{n-M}{2}\log\hat{b} + C_2 \\ &= -\frac{n-M}{2}\log\frac{\boldsymbol{z}'(Q_2'B(\lambda,\boldsymbol{\tau})Q_2)^{-1}\boldsymbol{z}}{[\det(Q_2'B(\lambda,\boldsymbol{\tau})Q_2)^{-1}]^{\frac{1}{n-M}}} + C_2, \end{aligned}$$

where C_2 is another constant. Equivalently the GML estimates of λ and $\boldsymbol{\tau}$ are the minimizers of

$$\begin{aligned} M(\lambda,\boldsymbol{\tau}) &= \frac{\boldsymbol{z}'(Q_2'B(\lambda,\boldsymbol{\tau})Q_2)^{-1}\boldsymbol{z}}{[\det(Q_2'B(\lambda,\boldsymbol{\tau})Q_2)^{-1}]^{\frac{1}{n-M}}} \\ &= \frac{\boldsymbol{y}'W(I-A)\boldsymbol{y}}{[\det^+(W(I-A))]^{\frac{1}{n-M}}}, \end{aligned} \tag{14}$$

where $\det^+$ is the product of the nonzero eigenvalues. An obvious estimate of the variance σ^2 is

$$\hat{\sigma}^2 = \frac{n\hat{\lambda}\boldsymbol{z}'(Q_2'B(\lambda,\boldsymbol{\tau})Q_2)^{-1}\boldsymbol{z}}{n-M} = \frac{\boldsymbol{y}'W(I-A)\boldsymbol{y}}{n-M}. \tag{15}$$

3 Connections Between Smoothing Spline Models and Mixed-Effects Models

In his contribution to the discussion of Robinson (1991), Speed noted the connection between a smoothing spline and a mixed-effects model for the special case that $\mathcal{T} = [0,1]$ and Σ is invertible. In this section, we give more details on this connection for the general smoothing spline models. It provides another motivation for using GML to estimate the smoothing parameter and correlation parameters simultaneously. It also suggests a way to fit a smoothing spline model using the SAS procedure `proc mixed`.

There are three possible ways to connect a smoothing spline to a mixed-effects model. First, the following mixed-effects model was considered by Speed:

$$\boldsymbol{y} = T\boldsymbol{d} + \boldsymbol{u} + \boldsymbol{\epsilon}, \tag{16}$$

where $\boldsymbol{d}$ is fixed, $\boldsymbol{u}$ is random and distributed as $\boldsymbol{u} \sim \mathrm{N}(0, \sigma^2\Sigma/n\lambda)$. $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 W^{-1})$. Note that Σ may or may not be invertible. Write $\Sigma/n\lambda = (I)(\Sigma)(I/n\lambda)$, where I is the identity matrix. It is easy to check that the generalized version of the normal equations (equation (3.3) in Harville (1976)) is the same as equations (7) and (8) for $\boldsymbol{c}$ and $\boldsymbol{d}$. The estimate of $\boldsymbol{u}$ is $\hat{\boldsymbol{u}} = \Sigma\boldsymbol{c}$. Thus the smoothing spline estimate $\hat{\boldsymbol{f}} = T\boldsymbol{d} + \Sigma\boldsymbol{c} = T\boldsymbol{d} + \hat{\boldsymbol{u}}$ is a best linear unbiased prediction (BLUP) estimate.

Second, consider the following mixed-effects model:

$$\boldsymbol{y} = T\boldsymbol{d} + \Sigma\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{17}$$

where $\boldsymbol{d}$ is fixed, $\boldsymbol{u}$ is random and distributed as $\mathrm{N}(0, \sigma^2\Sigma^+/n\lambda)$ with Σ^+ the Moore-Penrose inverse of Σ and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 W^{-1})$. Write $\Sigma^+/n\lambda = (\Sigma^+)(\Sigma)(\Sigma^+/n\lambda)$. It is easy to check that equation (3.3) in Harville (1976) is the same as equations (7) and (8). The estimate of $\boldsymbol{u}$ is $\hat{\boldsymbol{u}} = \Sigma^+\Sigma\boldsymbol{c}$, and again, the smoothing spline estimate $\hat{\boldsymbol{f}} = T\boldsymbol{d} + \Sigma\boldsymbol{c} = T\boldsymbol{d} + \Sigma\hat{\boldsymbol{u}}$ is a BLUP estimate.

Write $\Sigma = ZZ^T$, where Z is a $n \times k$ matrix with $k = \mathrm{rank}(\Sigma)$. Let V be a $n \times k$ matrix such that $Z^T V = I_{k\times k}$. Then we have $V^T\Sigma V = I_{k\times k}$ and $ZV^T\Sigma = \Sigma$. The third possible mixed-effects model is:

$$\boldsymbol{y} = T\boldsymbol{d} + Z\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{18}$$

where $\boldsymbol{u} \sim \mathrm{N}(0, \sigma^2 I/n\lambda)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 W^{-1})$. Writing $I/n\lambda = (V^T)(\Sigma)(V/n\lambda)$, equation (3.3) in Harville (1976) is the same as (7) and (8) for $\boldsymbol{c}$ and $\boldsymbol{d}$. The estimate for $\boldsymbol{u}$ is $\hat{\boldsymbol{u}} = V^T\Sigma\boldsymbol{c}$. Thus, the smoothing spline estimate $\hat{\boldsymbol{f}} = T\boldsymbol{d} + \Sigma\boldsymbol{c} = T\boldsymbol{d} + Z\hat{\boldsymbol{u}}$ is a BLUP estimate.

The smoothing parameter λ depends on the the ratio of two variance components. Let $b = \sigma^2/n\lambda$. Obviously $\lambda = 0$ iff $b = \infty$ and the fit interpolates the observations. $\lambda = \infty$ iff $b = 0$ and the fit is in space $\mathcal{H}_0$. Thus a hypothesis test on the variance component $b = 0$ can be used to test $f \in \mathcal{H}_0$ (see Cox, Koh, Wahba and Yandell (1988)). It is easy to check that the variance and covariance formulas in Theorem 3 of Harville (1976) are exactly the same as the posterior variance and covariance formulas in Wahba, Wang, Gu, Klein and Klein (1995). Thus the covariance matrix of the overall fit $\hat{\boldsymbol{f}}$ based on the mixed-effects model is the same as (13). This covariance matrix can be used to construct Bayesian confidence intervals (Wahba (1990)).

It is a common practice in mixed-effects models to estimate all variance components and correlation parameters using the restricted maximum likelihood (REML) method. Since $\boldsymbol{z}$ are $n - M$ linearly independent contrasts of $\boldsymbol{y}$, the REML estimates of λ, $\boldsymbol{\tau}$ and b is the maximizer of the log likelihood based on $\boldsymbol{z}$. Thus the GML estimates are also REML estimates.

Since a solution to the generalized version of the normal equation is a solution to (5), one can use SAS procedure `proc mixed` to calculate coefficients $\boldsymbol{c}$ and $\boldsymbol{d}$ in (4). Note that the spline estimate $\hat{f}(t)$ is defined on the domain $\mathcal{T}$, while estimate of a linear mixed-effects model is only defined on design points. Our ultimate goal is to get a spline estimate and the connections between a smoothing spline model and mixed-effects models are used to achieve this goal. One may choose any one of the three models that is easiest to use with the SAS procedure `proc mixed`. We will give an example of the SAS program in section 6.

4 Extensions of the UBR and GCV Methods

GCV method is well known for its optimal properties (Wahba 1990). UBR method has been successfully used to select smoothing parameters with non-Gaussian data (Gu 1992, Wahba et al. 1995). In this subsection we describe how to develop analogue of the UBR and GCV methods that can be used with correlated data. For this purpose, define the weighted mean squared errors (WMSE) as

$$T_k = \frac{1}{n}(\hat{\boldsymbol{f}} - \boldsymbol{f})^T W^k (\hat{\boldsymbol{f}} - \boldsymbol{f}) = \frac{1}{n}||W^{k/2}(\hat{\boldsymbol{f}} - \boldsymbol{f})||, \quad k = 0, 1, 2. \tag{19}$$

Then, since $\hat{\boldsymbol{f}} = A\boldsymbol{y}$,

$$\mathrm{E}T_k = \frac{1}{n}\boldsymbol{f}^T(I - A^T)W^k(I - A)\boldsymbol{f} + \frac{\sigma^2}{n}\mathrm{Tr}A^T W^k A W^{-1}, \quad k = 0, 1, 2, \tag{20}$$

an unbiased estimate of $\mathrm{E}T_k$ is

$$U_k = \frac{1}{n}\boldsymbol{y}^T(I - A^T)W^k(I - A)\boldsymbol{y} - \frac{\sigma^2}{n}\mathrm{Tr}W^{k-1} + 2\frac{\sigma^2}{n}\mathrm{Tr}W^{k-1}A, \quad k = 0, 1, 2. \tag{21}$$

Estimates of λ and $\boldsymbol{\tau}$ which minimize U_k are called unbiased risk estimates. This method is an extension of the UBR method for independent observations and is still called the UBR method in this paper. The UBR method needs knowledge or an estimator of σ^2; one possible estimator is given in (15).

Define

$$V_k = \frac{\frac{1}{n}||W^{k/2}(I - A)\boldsymbol{y}||^2}{[\frac{1}{n}\mathrm{Tr}(W^{k-1}(I - A))]^2}, \quad k = 0, 1, 2. \tag{22}$$

If $\mathrm{Tr}W^{k-1}A/\mathrm{Tr}W^{k-1}$ is small, then

$$\begin{aligned}
\mathrm{E}V_k &\approx [\frac{1}{n}\boldsymbol{f}^T(I - A^T)W^k(I - A)\boldsymbol{f} + \frac{\sigma^2}{n}\mathrm{Tr}(I - A^T)W^k(I - A)W^{-1}] \\
&\quad [1 + 2\mathrm{Tr}W^{k-1}A/\mathrm{Tr}W^{k-1} + o(1)]/(\frac{1}{n}\mathrm{Tr}W^{k-1})^2 \\
&\approx (\mathrm{E}T_k + \frac{\sigma^2}{n}\mathrm{Tr}W^{k-1})(1 + o(1))/(\frac{1}{n}\mathrm{Tr}W^{k-1})^2.
\end{aligned}$$

Thus V_0 is a proxy for $\mathrm{E}T_0$ if $\mathrm{Tr}W^{-1}$ does not depend on $\boldsymbol{\tau}$. V_1 is a proxy for $\mathrm{E}T_1$. $\mathrm{E}T_2$ involves both $\mathrm{E}T_2$ and $\mathrm{Tr}W$. Comparing the GML function (14) with the GCV function (4.3.1) and GML function (4.8.4) in Wahba (1990), it is easy to see that V_2 is a direct extension of the usual GCV function to the dependent case. Estimates of λ and $\boldsymbol{\tau}$ which minimize V_k are called the GCV estimates, and the method is still called the GCV method in this paper.

The GCV methods with $k = 0$ and $k = 1$ correspond to the direct and indirect methods in Altman (1990) respectively. The method (b) in Diggle and Hutchinson (1989) used the function $\ln V_1$ plus a term $\ln|W^{-1}|$. As commented in their paper, it does not penalize the degree of freedom $\mathrm{Tr}A$ enough. The method (c) in Diggle and Hutchinson (1989) used the following function

$$L = n \ln \boldsymbol{y}^T(I - A^T)W(I - A)\boldsymbol{y} + \ln|W^{-1}| + (\ln n)\mathrm{Tr}A. \tag{23}$$

We will call estimates of λ and $\boldsymbol{\tau}$ which minimize L as the L estimates and the method as L method in our simulations in section 7.

5 Smoothing Spline ANOVA Models with Correlated Errors

In this section, we extend our methods to fit smoothing spline ANOVA models with correlated errors. Suppose now the index set $\mathcal{T} = \mathcal{T}_1 \otimes \cdots \otimes \mathcal{T}_d$. The model space is an orthogonal decomposition of $\mathcal{H}$ into more than two components:

$$\mathcal{H} = \mathcal{H}_0 \oplus \sum_{\beta=1}^{p} \mathcal{H}_\beta, \tag{24}$$

where $\mathcal{H}_0$ is a finite dimensional space with terms which are not going to be penalized. See Wahba (1990) and Gu and Wahba (1993a) for discussions on how to construct the model space $\mathcal{H}$. Suppose we have observations as in (1). A direct generalization of (3) is

$$\min_{f \in \mathcal{H}} \left\{ \frac{1}{n}(\boldsymbol{y}-\boldsymbol{f})^T W (\boldsymbol{y}-\boldsymbol{f}) + \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} \|P_\beta f\|^2 \right\}, \tag{25}$$

where P_β is the orthogonal projection in $\mathcal{H}$ onto $\mathcal{H}_\beta$. Let $\phi_1, \cdots, \phi_M$ be a basis of $\mathcal{H}_0$. Let $\xi_i^\beta(t) = L_{i(\cdot)} R^\beta(t, \cdot)$, where $R^\beta(t,\cdot)$ is the r.k. of $\mathcal{H}_\beta$. From Wahba (1990), the solution to (25) is

$$\hat{f} = \sum_{\nu=1}^{M} d_\nu \phi_\nu + \sum_{i=1}^{n} c_i \left(\sum_{\beta=1}^{p} \theta_\beta \xi_i^\beta\right). \tag{26}$$

$\boldsymbol{c} = (c_1, \cdots, c_n)^T$ and $\boldsymbol{d} = (d_1, \cdots, d_M)^T$ are solutions to (5) with

$$\Sigma = \sum_{\beta=1}^{p} \theta_\beta \Sigma_\beta, \tag{27}$$

where $\Sigma_\beta = \{< \xi_i^\beta, \xi_j^\beta >\}$. Gu and Wahba (1993b) established the connection between a smoothing spline ANOVA model and a Bayesian model with independent errors. Extension of their model to the correlated errors situation is similar to the previous sections; it is is not repeated here. The GML estimates of the smoothing parameters $\lambda/\theta_1, \cdots, \lambda/\theta_p$ and $\boldsymbol{\tau}$ are the minimizers of (14) with the matrix Σ given by (27).

Next, we connect a smoothing spline ANOVA model with a mixed-effects model. Consider the following mixed-effects model:

$$\boldsymbol{y} = T\boldsymbol{d} + \sum_{\beta=1}^{p} \boldsymbol{u}_\beta + \boldsymbol{\epsilon} = T\boldsymbol{d} + Z\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{28}$$

where $\boldsymbol{d}$ is fixed, $\boldsymbol{u}_\beta$ is random and distributed as $\boldsymbol{u}_\beta \sim \mathrm{N}(0, \sigma^2\theta_\beta\Sigma_\beta/n\lambda)$, $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 W^{-1})$, and $\boldsymbol{u}_\beta$'s and $\boldsymbol{\epsilon}$ are mutually independent. $Z_{n\times np} = (I_{n\times n}, \cdots, I_{n\times n})$. $\boldsymbol{u} = (\boldsymbol{u}_1^T, \cdots, \boldsymbol{u}_p^T)^T$. Let $D = \mathrm{diag}(\theta_1\Sigma_1, \cdots, \theta_p\Sigma_p)$. Then $\mathrm{var}(\boldsymbol{u}) = \sigma^2 D/n\lambda$. Writing $D/n\lambda = (I)(D)(I/n\lambda)$, it may be verified that the equation (3.3) in Harville (1976) is

$$\begin{pmatrix} T^TWT & T^TWZD \\ DZ^TWT & n\lambda D + DZ^TWZD \end{pmatrix} \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{\phi} \end{pmatrix} = \begin{pmatrix} T^TW\boldsymbol{y} \\ DZ^TW\boldsymbol{y} \end{pmatrix}. \tag{29}$$

Let $\boldsymbol{c}$ and $\boldsymbol{d}$ be a solution to (5). Since $ZDZ^T = \Sigma$, it can be shown that $\boldsymbol{d}$ and $\boldsymbol{\phi} = Z^T\boldsymbol{c}$ is a solution to (29) if Σ is invertible. The estimate of $\boldsymbol{u}$ is $\hat{\boldsymbol{u}} = D\boldsymbol{\phi} = DZ^T\boldsymbol{c}$. Thus, $\theta_\beta\Sigma_\beta\boldsymbol{c} = \hat{\boldsymbol{u}}_\beta$, the smoothing spline ANOVA estimate of the component in the subspace $\mathcal{H}_\beta$, is a BLUP. Therefore the smoothing spline ANOVA estimates of the main effects, the interactions and the overall function are BLUP's. If Σ is not invertible, the smoothing spline estimate of the overall function $\hat{\boldsymbol{f}} = T\boldsymbol{d} + \Sigma\boldsymbol{c}$ is still a BLUP since it is unique according to (6).

Extensions of the GCV and the UBR methods in section 4 for fitting smoothing spline ANOVA models are straightforward. The GCV and UBR estimates of the smoothing parameters and the correlation parameters are minimizers of (22) and (21) respectively, with the matrix Σ given by (27).

6 Applications

In this section we apply our method to fit data from two time series and an environmental model involving spatial data. The analysis here are intend to illustrate our methods and does not represent a formal analysis of the data.

6.1 Spline Smoothing of Time Series

Consider the following model:

$$Y(t) = f(t) + Z(t), \qquad t \in [0,1], \tag{30}$$

where f is a deterministic mean function and belongs to the r.k.h.s.

$$W_M = \{f:\ f^{(\nu)}\text{absolutely continuous},\ \nu = 0,\cdots,M-1,\ f^{(M)} \in \mathcal{L}_2[0,1]\},$$

and Z is a stationary error process with mean zero. Suppose we have observations y_i at time t_i of the form:

$$y_i = f(t_i) + Z(t_i), \qquad i = 1,\cdots,n, \tag{31}$$

where $\boldsymbol{\epsilon} = (Z(t_1),\cdots,Z(t_n))^T \sim \mathrm{N}(0, \sigma^2 W^{-1})$. Note that the t_i's are not necessarily equally spaced. Our goal is to estimate the mean function f.

Several authors have studied this problem. Diggle and Hutchinson (1989) considered the case when $Z(t_i)$ are an autocorrelated sequence. They extended the generalized cross-validation method to estimate the smoothing parameter and the autocorrelation parameters simultaneously. Kohn et al. (1992) represented a smoothing spline by a state space model and extended the CV, GCV and GML methods for an autoregressive moving average error sequence. Altman (1990) and Hart (1991) used kernel regression to estimate the mean function with the smoothing parameter estimated from mean squared errors. Hart (1994) used the time-series cross-validation method to estimate the smoothing parameter and the autocorrelation functions simultaneously. Hurvich and Zeger (1990) used a frequency domain cross-validation method to estimate the smoothing parameter.

Theoretically, our method can be used for any error structure. We illustrate our approach using the following two examples which were previously analyzed in the literature.

We use series A in Box and Jenkins (1976) (p.525) as our first example. It was used by Diggle and Hutchinson (1989) to demonstrate their method. This series has 197 measurements of the "uncontrolled" concentration in a continuous chemical process sampled every two hours. We fit this series with a cubic spline for the deterministic mean function ($f \in W_2$) and an AR(1) model for errors. GML method is used in all the examples in this section since it has found to work better than other methods by our simulations in section 7. The left plot in Figure 2 shows the data (points), estimate of f under the AR(1) model for the errors (solid line) and its 95% Bayesian confidence intervals. We also plot the estimate of f under the independence assumption (dotted line). The estimate of f under the independence assumption is more wiggly but within the Bayesian confidence intervals. Our estimates and confidence intervals are visually similar to those in Diggle and Hutchinson (1989). The estimates of the first-order autoregressive parameter, the smoothing parameter λ and the residual variance σ^2 are 0.305, 0.0000048 and 0.098, respectively.

We use the Beveridge data in Anderson (1971) as our second example. It was used by Hart (1994). This series contains the yearly Beveridge index of wheat prices in Europe from 1500 to 1869. We use a cubic spline for the mean function of the log index and an AR(1) model for errors. The right plot in Figure 2 shows the data (points), estimate of f under the AR(1) model (solid line) and its 95% Bayesian confidence intervals. The estimate of f under the independence assumption (dotted line) is more wiggly. The estimates of the first-order autoregressive parameter α, the smoothing parameter λ and the residual variance σ^2 are 0.696, 0.0000014 and 0.051, respectively.

SAS procedure `proc mixed` was employed to fit both data sets using the mixed-effects model (16). Part of the SAS program used for fitting the chemical data is listed below:

```
proc mixed data=samp absolute convf;
    model y = s;
    random col4-col200 / type=lin(1) ldata=ldata;
    repeated / type=ar(1) sub=intercept;
    parms (106) (0.3) (0.1);
```

where data `samp` contains variables y, s with $s(i) = i/197 - 0.5$, $i = 1, \cdots, 197$, and $col4$ to $col200$ which are 197 columns of an 197×197 identity matrix. Data `ldata` contains variables $parm$ with $parm = 1$, row which indicates row number, $col1$ to $col197$ which are 197 columns of the matrix Σ. The covariance matrix of the random-effects Σ is specified by the `random` statement with the new option `type=lin` in SAS release 6.09 (SAS Institute 1994). Some options are taken out to make the program simple. Both programs are available from the author.

6.2 Spline Smoothing of Spatial Data

In this section we analyze a subset of environmental data based on the Eastern Lakes Survey in 1984 by the EPA. This data was used in Gu and Wahba (1993a) and Gu and Wahba (1993b). We only use observations on 112 lakes in the southern Blue Ridge mountains area. The data contains water acidity measurements (surface pH), geographic information (latitude and longitude) and the calcium concentration. Of interest is the dependence of the water

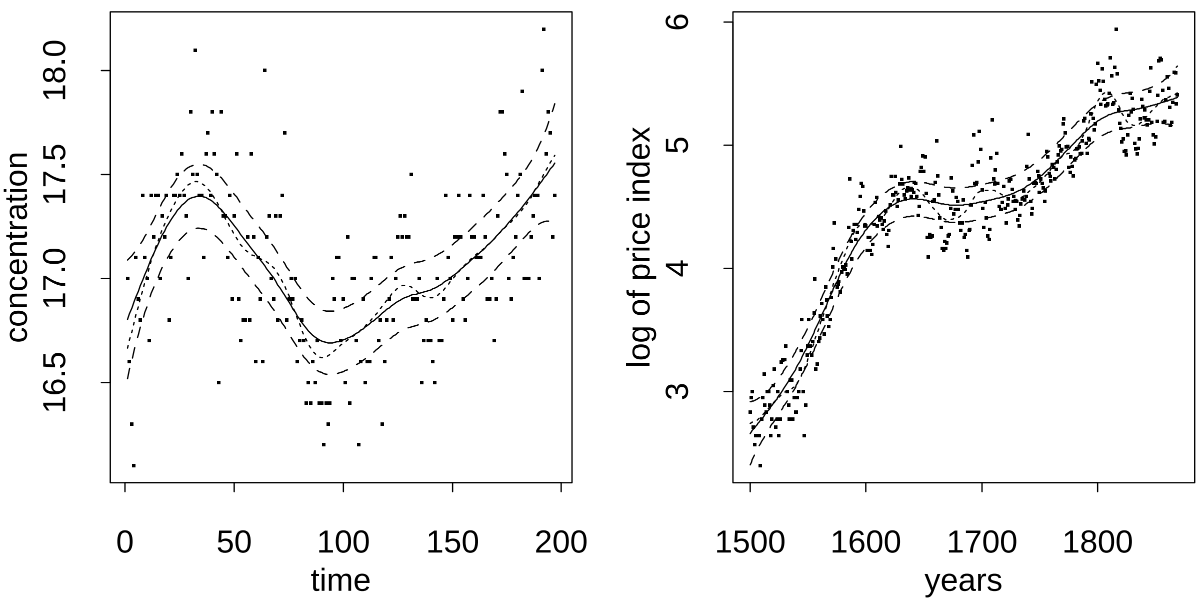

Figure 2: Left: chemical concentration. Right: Beveridge index of wheat price. Dots: observations. Solid lines: estimates with smoothing parameters chosen by GML. Dashed lines: 95% Bayesian confidence intervals. Dotted lines: estimates assuming independence and smoothing parameters are estimated by GML.

acidity on the geographic location and the calcium concentration. Denote $t_1 \in E^1$ as the calcium concentration and $t_2 = (x_1, x_2) \in E^2$ as the latitude and longitude. A smoothing spline ANOVA model of the form

$$y_i = C + f_1(t_1) + f_2(t_2) + f_{12}(t_1, t_2) + \epsilon_i, \quad i = 1, \cdots, 112, \tag{32}$$

was used in Gu and Wahba (1993a), where the ϵ_i's were assumed to be independent. See Gu and Wahba (1993a) for details about the data and the model. The independence assumption may or may not be valid since measurements of lakes close together may be positively correlated. Thus, for the analysis conducted here we will assume the correlation matrix W^{-1} has an exponential structure: $w_{ij}^{-1} = \exp(-d_{ij}/\rho)$, where d_{ij} is the Euclidean distance between the geographic locations $t_2(i)$ and $t_2(j)$. Using model (32) along with the exponential structure for the errors covariances, we fit the data using the SAS procedure `proc mixed`. The estimate of ρ equals 0.02, which is not significantly different from zero. We also tried to fit the data with a power structure: $w_{ij}^{-1} = \sigma^2 \rho^{d_{ij}}$, for the correlation matrix and the estimate of ρ is also near zero. Thus the independence assumption in this case is appropriate. The estimates are similar to Gu and Wahba (1993a). Thus they are not presented here.

7 Simulation Results

A simulation study is conducted to evaluate and compare the performance of several methods presented in previous sections. The model considered is

$$y_i = \sin 2\pi i/n + \epsilon_i, \qquad i = 1, \cdots, n, \tag{33}$$

where the ϵ's are generated by a first-order autoregressive process with mean zero, standard deviation σ and first-order correlation α. Four different sample sizes $n = 50, 100, 200, 400$, two different standard deviations $\sigma = 0.1, 0.3$, and four different correlations $\alpha = 0.3, 0.55, 0.74, 0.86$ are considered. For $n = 400$, we only consider the cases that $\sigma = 0.1$ and $\alpha = 0.3, 0.86$ due to the large CPU time needed. Hence we have $3 \times 2 \times 4 + 2 = 26$ factors in the design of our simulation experiment. Responses are generated for 100 replications of each of these 26 settings. We use the Fortran routine `rnor` of the Core Mathematics Library (Cmlib) from the National Bureau of Standard to generate random numbers in all the simulations. Cubic splines ($m = 2$) are used to fit the mean function.

First, we compare the performance of various methods for estimating the function and the correlation parameter. In previous sections we presented eight different criterion functions: GML(M), GCV(V_0,V_1,V_2), UBR(U_0,U_1,U_2) and Diggle and Hutchinson (1989)'s functions L. In our preliminary simulations with 10 replications, the GCV methods V_0 and V_1 tended to interpolate the data, while U_0 was a decreasing function of α, and it always estimated α as -1. The UBR method U_1 was found to work fine, but not as well as U_2 in terms of WMSE in (19) for all k. Thus we only conducted simulations using functions M, V_2, U_2 and L. They are represented by m, v, u and l in our plots and tables. They are called GML, GCV, UBR and L methods respectively.

The comparison results between methods based on different definitions of WMSE ($k = 0, 1, 2$ in (19)) are the same. The WMSE for different k are highly correlated. We only present the comparison results based on the WMSE with $k = 2$ (T_2 in (19)).

For each of the 100 replications, we calculate $\hat{f}$ and $\hat{\alpha}$ using GML, GCV, UBR, L methods and calculate the WMSE of $\hat{f}$. Figure 3 shows the boxplots of WMSE of $\hat{f}$ for all simulations except for $n = 400$. Figure 4 and 5 show the average WMSE values of $\hat{f}$ on $\log_{10}$ scale for $\sigma = 0.1$ and $\sigma = 0.3$ respectively. Figure 6 shows the boxplots of $\hat{\alpha}$ for all simulations except for $n = 400$. Figure 7 and 8 show the MSE of $\hat{\alpha}$ (MSE $= \sum_{r=1}^{100}(\hat{\alpha}_r - \alpha)^2$) on $\log_{10}$ scale for $\sigma = 0.1$ and $\sigma = 0.3$ respectively.

The GCV and L methods interpolate data in several replications when n is small ($n = 50, 100$). There are no interpolation cases when using GML and UBR methods. The number of replications out of 100 that have a GCV and a L estimate of $n\hat{\lambda}$ smaller than -14 in $\log_{10}$ scale is listed in Table 1. The number decreases with increasing n and decreasing α.

The GML method has the smallest or near the smallest WMSE of $\hat{f}$ and MSE of $\hat{\alpha}$. GCV tends to interpolate the data when n is small and α is large (Table 1). This problem diminishes quickly when n becomes large. This agrees with the conclusions in Wahba and Wang (1993) where observations were assumed to be independent. The WMSE of $\hat{f}$ using GCV converges faster than using GML when the correlation is small ($\alpha = 0.3$), which agrees with the theory in Wahba (1985) for independent observations. GCV method estimates α reasonably well, though not as good as the GML method. The UBR method estimates f

very well, but estimates α poorly. Actually the MSE of $\hat{\alpha}$ does not decrease as n increases. One possible reason for this is that the WMSE is a loss function measuring the difference between $\hat{f}$ and f only and the difference between $\hat{\alpha}$ and α is ignored. A measure of the whole estimation loss should assess the performance of both $\hat{f}$ and $\hat{\alpha}$. The Kullback-Leibler discrepancy may be a more appropriate measure of the whole estimation error. The true variance σ^2 was used in simulations involving the UBR method. The L method interpolates the data when the sample size is small (Table 1). As n becomes large ($n \geq 200$), it has no interpolation and works as well as the GML method in terms of both $\hat{f}$ and $\hat{\alpha}$. We conclude that the GML method works well for all cases in terms of both the estimation of the true function and the estimation of the correlation parameter. Its performance is better than other methods, especially when the sample size is small. This agrees with the conclusions in Kohn, Ansley and Tharm (1991) for independent data. Therefore GML is recommended.

	$\alpha = 0.30$		$\alpha = 0.55$		$\alpha = 0.74$		$\alpha = 0.86$	
Method	v	l	v	l	v	l	v	l
$\sigma = 0.1$								
$n = 50$	2	100	3	100	4	100	8	100
$n = 100$	0	1	0	11	0	35	0	62
$\sigma = 0.3$								
$n = 50$	2	100	3	100	6	100	8	100
$n = 100$	0	1	0	6	0	27	0	51

Table 1: Number of replicates out of 100 total that interpolates. v and l represent GCV and L methods respectively.

For $n = 100$, $\sigma = 0.1$, and $\alpha = 0.55$, for each method, we select the 5th, 25th, 50th, 75th and 95th best estimates of the mean function ordered by their WMSE values. Figure 9 shows these estimates. Again, we can see that the GML method works better.

To look at the shape of the functions of M, V, U and L, for $n = 100$, $\sigma = 0.1$, and $\alpha = 0.55$, we show "typical" (i.e., the first replicate of the simulation) contour plots of these four functions in Figure 10. We can see from these plots that each function has an unique minima inside the search region and it is near the true value of α. The L function has local minimas at $\alpha = \pm 1$. M, V and L functions have local maximas around $\alpha = -1$ and $\log_{10}(n\lambda) = 0$. All functions have nice bowl shapes near their local minima.

Assume the errors in model (33) come from a stationary process with covariance function $\mathrm{Cov}(\epsilon_i, \epsilon_j) = \sigma^2 \rho_n(|i-j|)$, where $\rho_n(k)$ is a correlation function depending on n. Many authors (Altman 1990, Hart 1991) distinguished two different kinds of special error processes: (a) $\rho_n(k) = \rho(k)$ and (b) $\rho_n(k) = \rho(k/n)$. In the first case, the error process is constant regardless of how close together the design points become. Under certain conditions, the MSE of a kernel estimate converges to zero as $n \to \infty$ (Altman 1990, Hart 1991). In the second case, the error process is a realization of a continuous process. There are no consistent linear estimators for this case (Hart and Wehrly 1986). Our simulations involves both cases. First, consider α as fixed and let n increase. This corresponds to the first case. Figure (4)

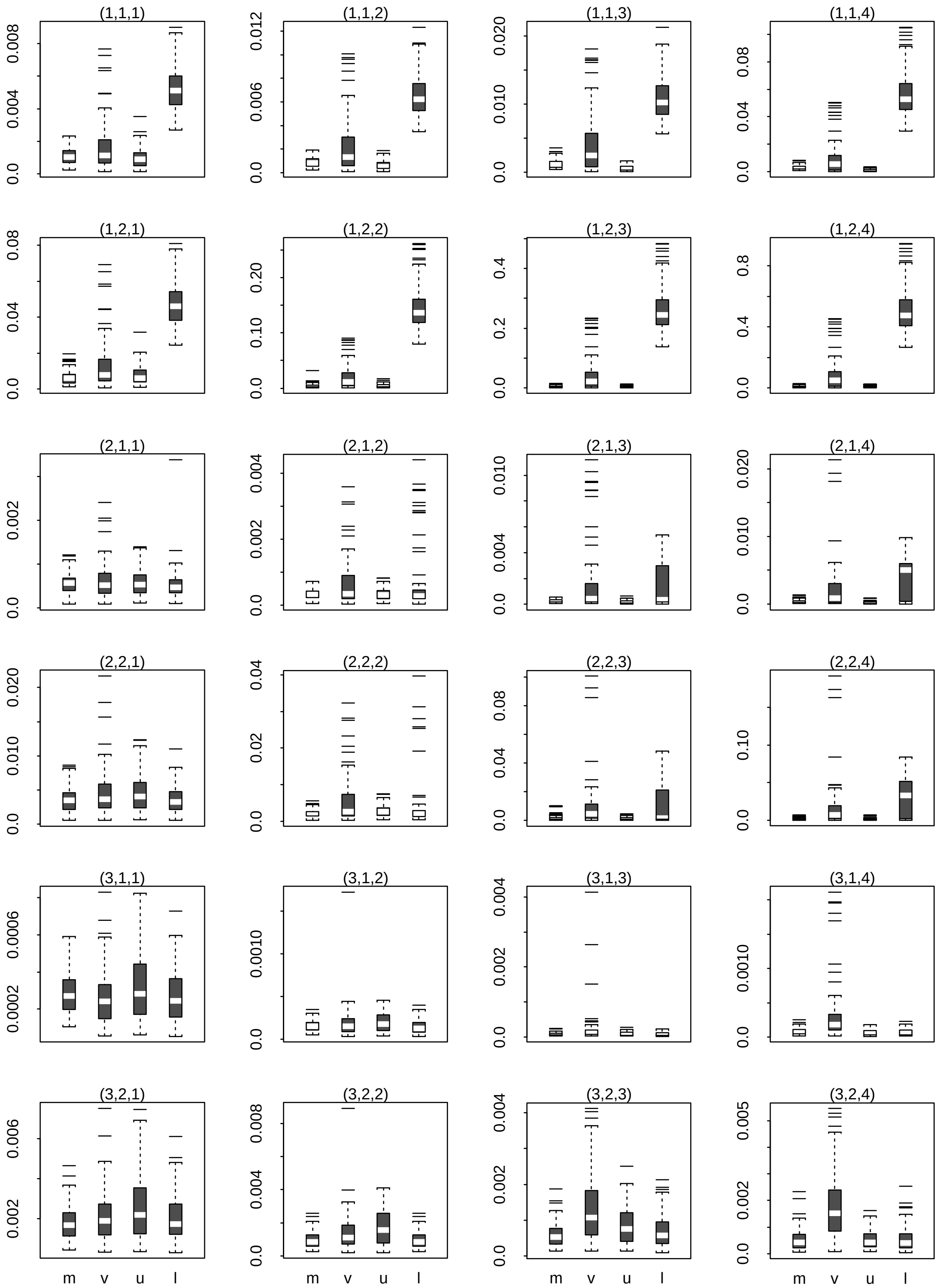

Figure 3: Box plots of WMSE of $\hat{f}$. (i,j,k) on the top of each plots indicates sample size i (i = 1, 2 and 3 correspond to sample sizes 50, 100 and 200 respectively), standard deviation j (j = 1 and 2 correspond to standard deviations 0.1 and 0.3 respectively), correlation k (k = 1, 2, 3 and 4 correspond to correlations 0.3, 0.55, 0.74 and 0.86 respectively). m, v, u and l represent GML, GCV, UBR and L methods respectively.

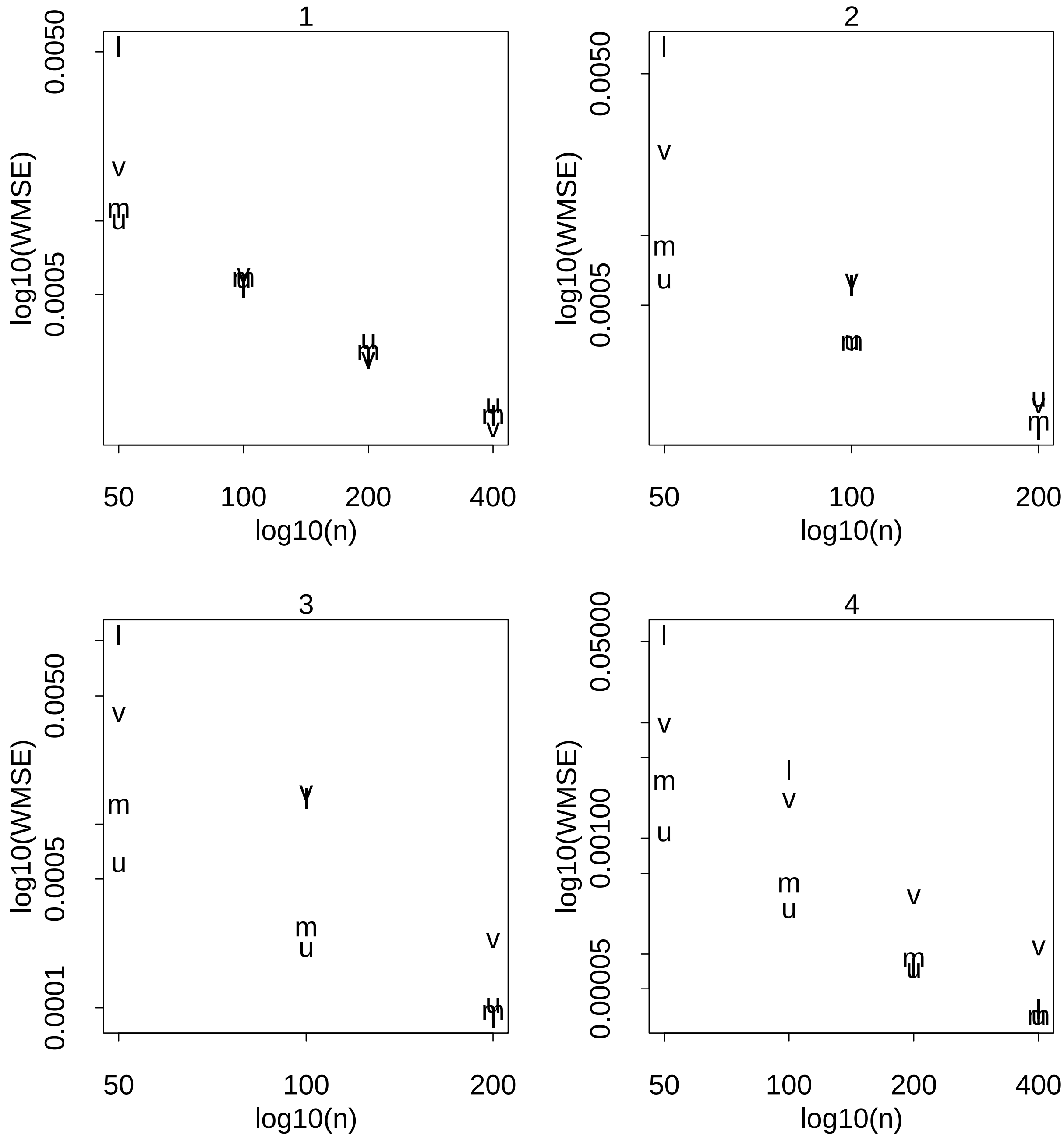

Figure 4: Plots of average WMSE values of $\hat{f}$ vs sample size on $\log_{10}$ scale when $\sigma = 0.1$. 1, 2, 3, and 4 indicate four correlation values at 0.3, 0.55, 0.74 and 0.86 respectively. m, v, u and l represent GML, GCV, UBR and L methods respectively.

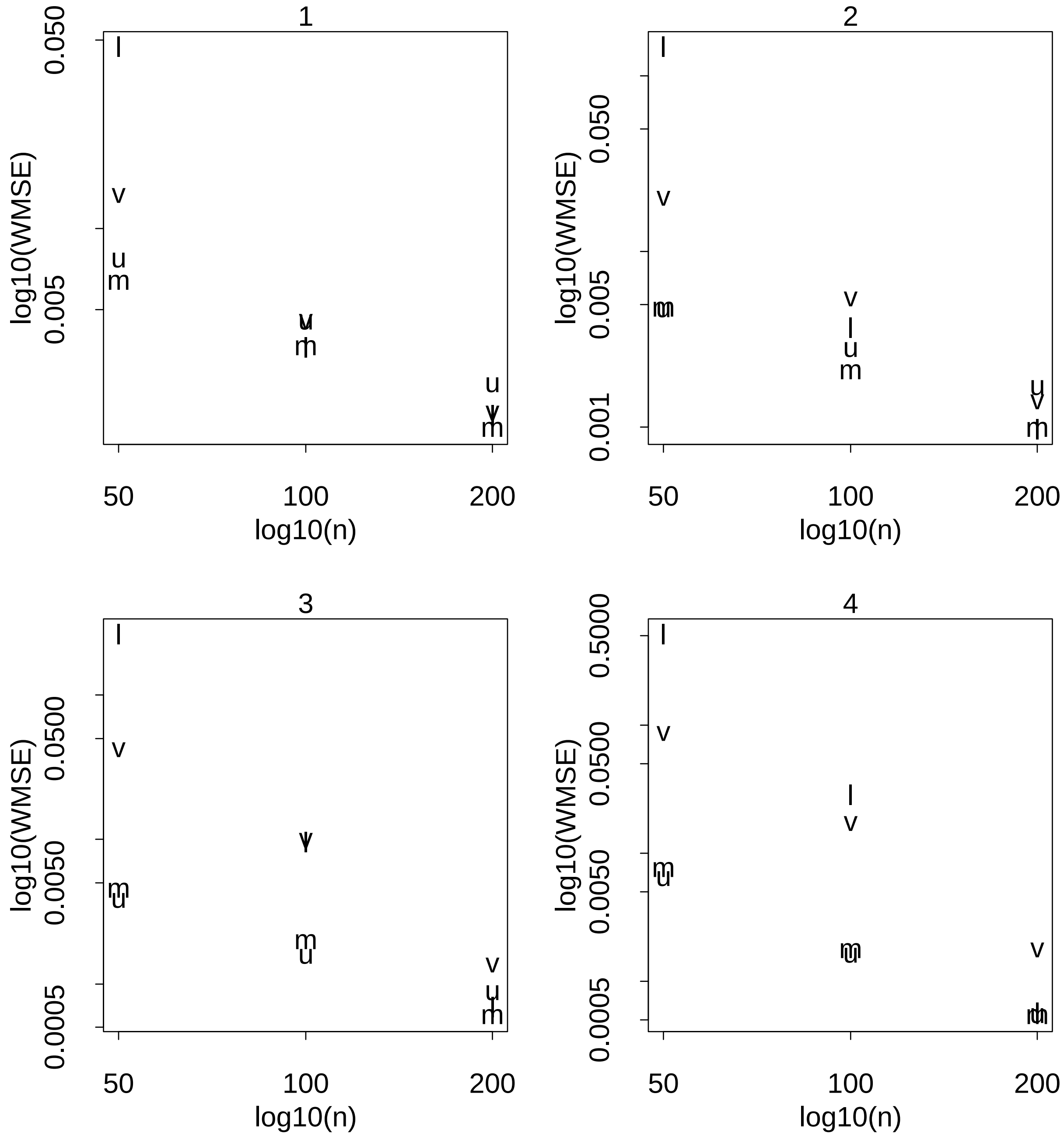

Figure 5: Plots of mean values of WMSE of $\hat{f}$ vs sample size on $\log_{10}$ scale when $\sigma = 0.3$. 1, 2, 3, and 4 indicate four correlation values at 0.3, 0.55, 0.74 and 0.86 respectively. m, v, u and l represent GML, GCV, UBR and L methods respectively.

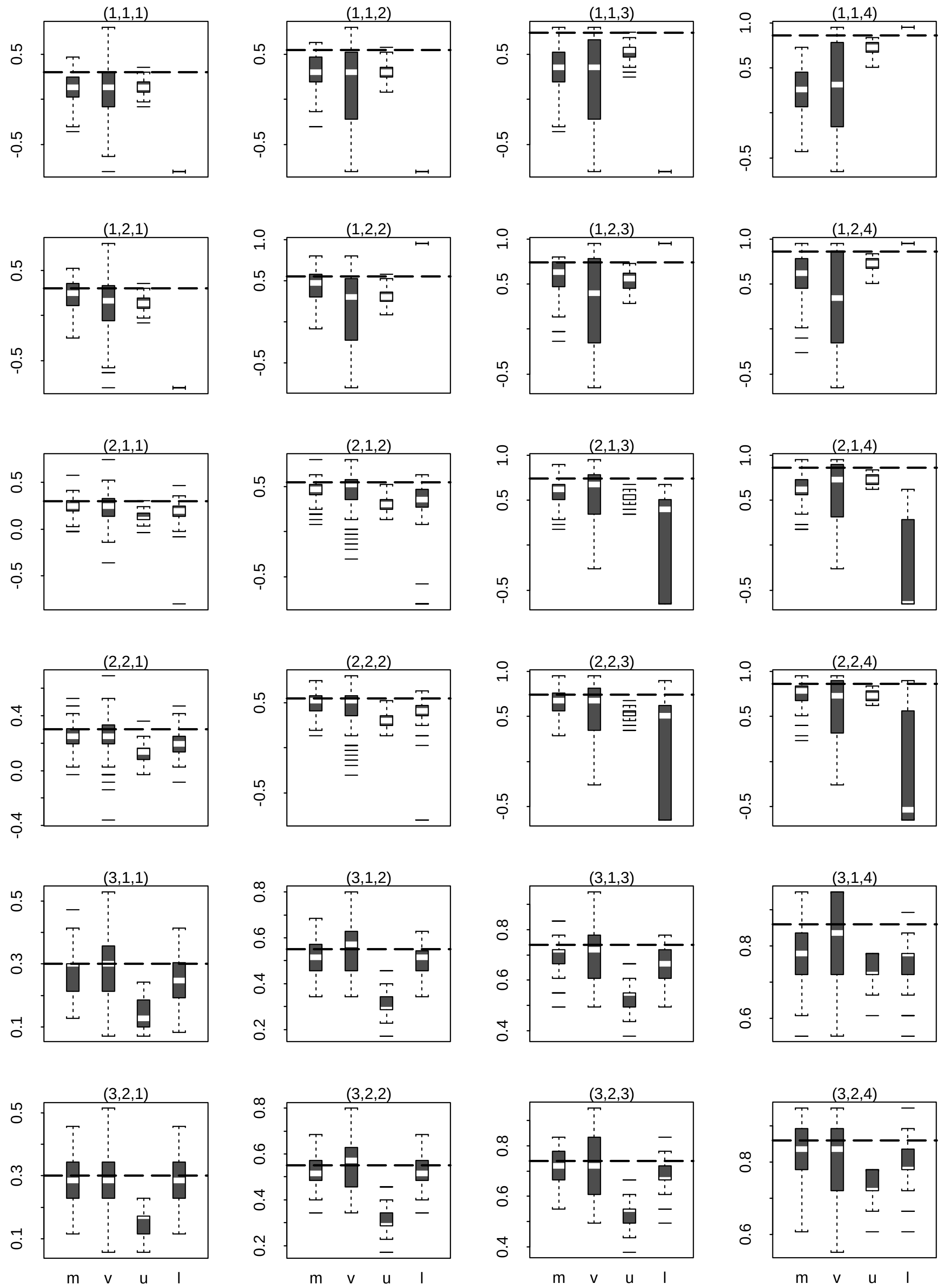

Figure 6: Box plots of $\hat{\alpha}$. (i,j,k) on the top of each plots indicates sample size i (i = 1, 2 and 3 correspond to sample sizes 50, 100 and 200 respectively), standard deviation j (j = 1 and 2 correspond to standard deviations 0.1 and 0.3 respectively), correlation k (k = 1, 2, 3 and 4 correspond to correlations 0.3, 0.55, 0.74 and 0.86 respectively). m, v, u and l represent GML, GCV, UBR and L methods respectively. Dashed lines indicate the true α.

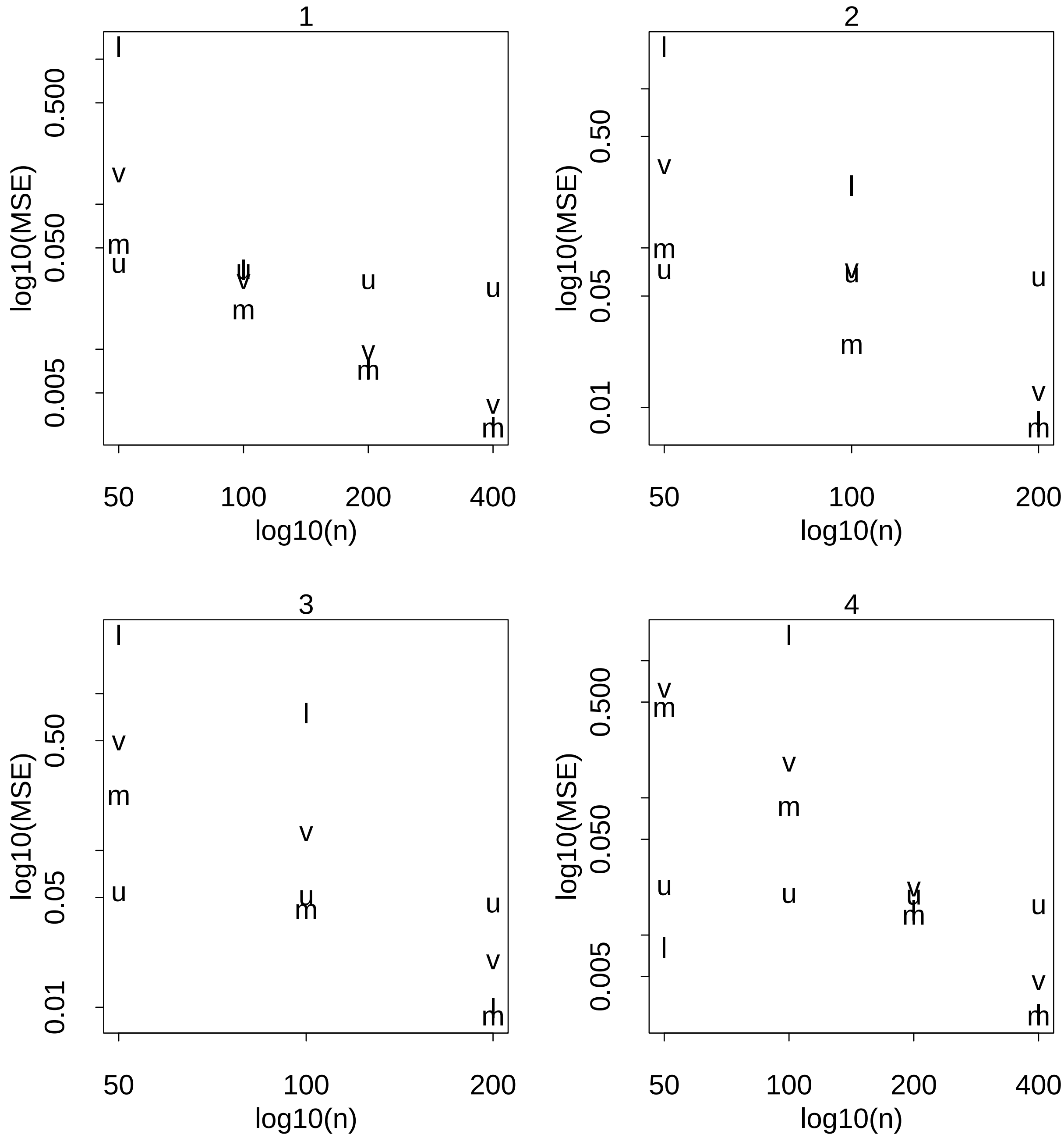

Figure 7: Plots of MSE of $\hat{\alpha}$ vs sample size on $\log_{10}$ scale when $\sigma = 0.1$. 1, 2, 3, and 4 indicate four correlation values at 0.3, 0.55, 0.74 and 0.86 respectively. m, v, u and l represent GML, GCV, UBR and L methods respectively.

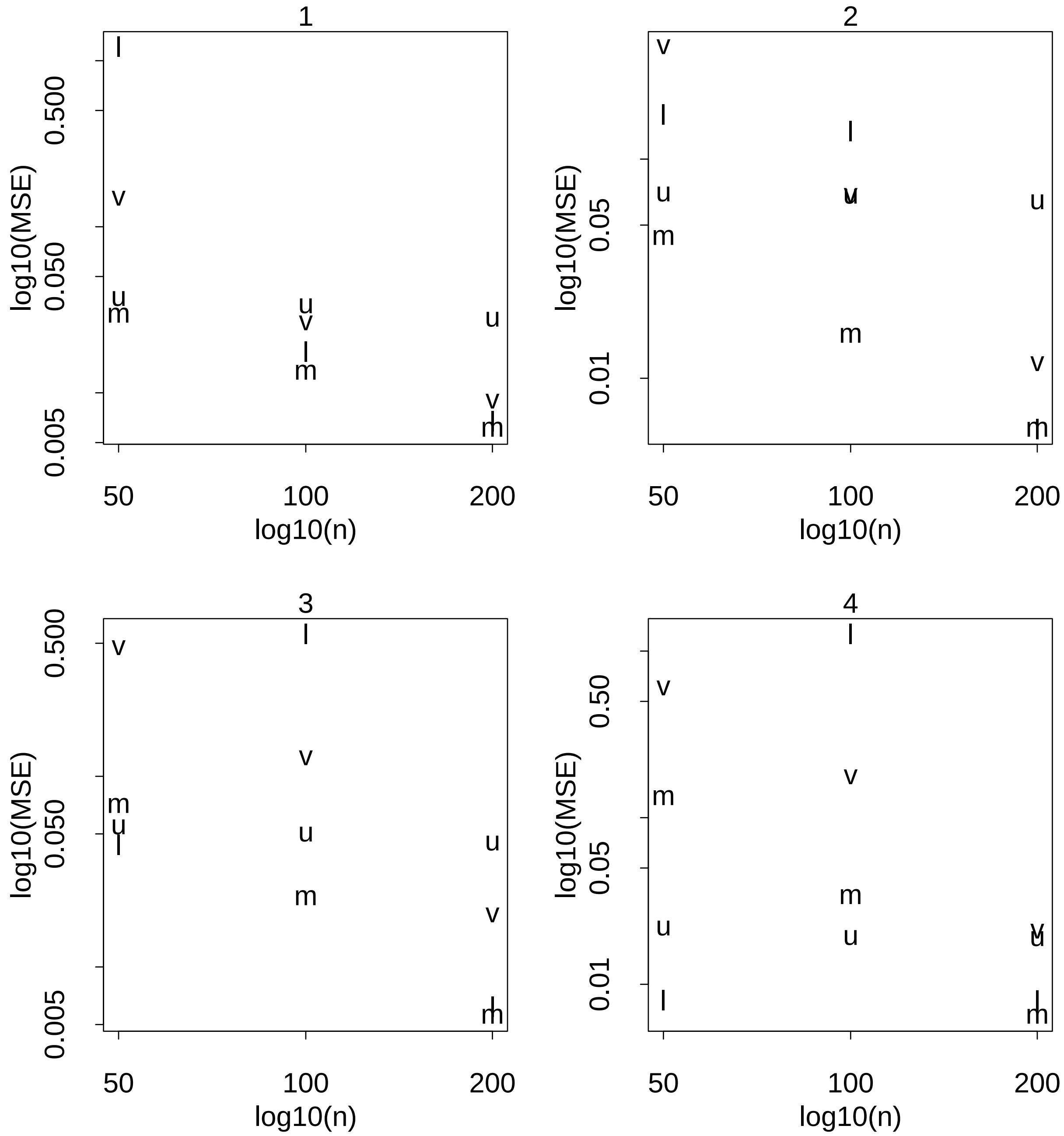

Figure 8: Plots of MSE of $\hat{\alpha}$ vs sample size on $\log_{10}$ scale when $\sigma = 0.3$. 1, 2, 3, and 4 indicate four correlation values at 0.3, 0.55, 0.74 and 0.86 respectively. m, v, u and l represent GML, GCV, UBR and L methods respectively.

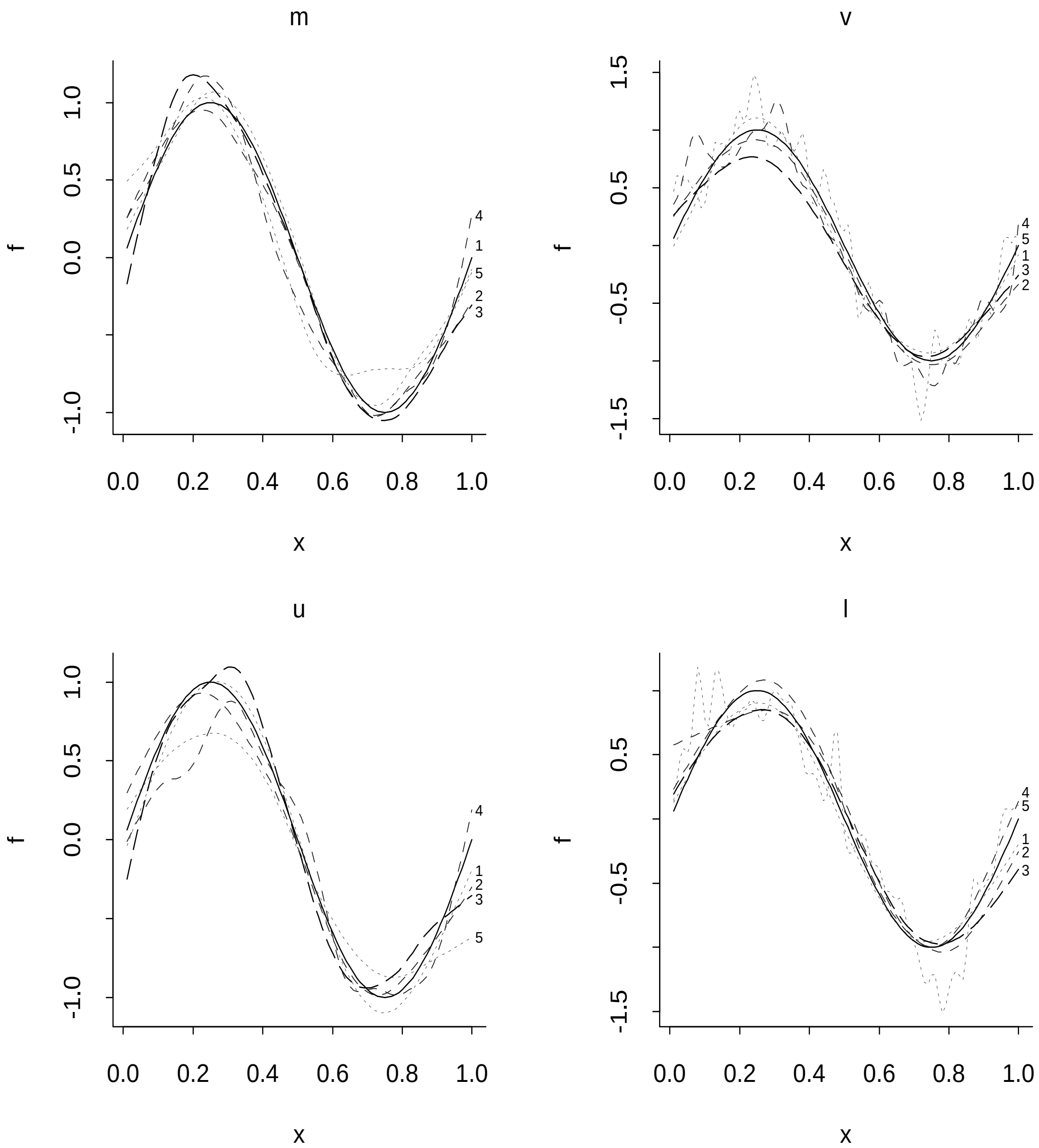

Figure 9: $n = 100$, $\sigma = 0.1$, and $\alpha = 0.55$. Solid lines are true functions. Five dashed lines in each graph are estimates: 1, 2, 3, 4 and 5 are the 5th, 25th, 50th, 75th and 95th best estimates ordered by WMSE. m, v, u and l represent GML, GCV, UBR and L methods respectively.

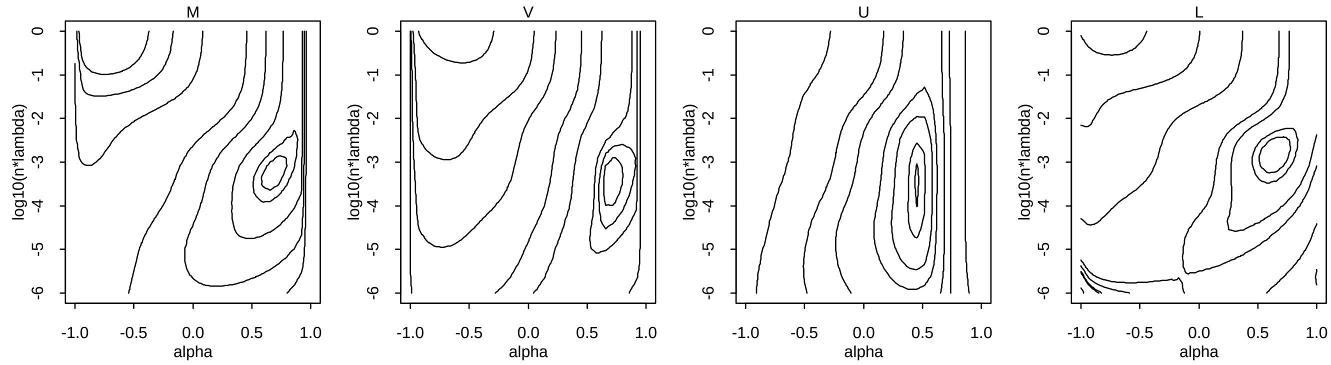

Figure 10: Contour plots of the four criteria functions of the first replicate in the simulation $n = 100$, $\sigma = 0.1$ and $\alpha = 0.55$. M, V, U and L represent GML, GCV, UBR and L functions respectively.

showes that the $\log_{10}(\text{WMSE})$ decreases linearly in $\log_{10}(n)$ and the slopes are near $-4/5$, suggesting the convergence rate of WMSE is $n^{-4/5}$ (see Wahba (1985) for independent case). Second, let α and n increase at the same time. Since the α_i's are selected in such a way that $\alpha_i = \exp(-60.2/n_i)$, this corresponds to the second case. Figure (11) shows that the MSE (WMSE with $k = 0$) does not decrease as n increases. But the WMSE with $k = 1$ and $k = 2$ decrease linearly as n increases. Thus these simulations indicate that the WMSE T_1 and T_2 of the spline estimates converge to zero as $n \to \infty$. Further theoretical research is needed.

We also conducted simulations with other functions such as $\sin 4\pi t$ and $\sin 6\pi t$. The comparison results are similar. When the sample size is small ($n = 50$), the frequency of the deterministic function is high ($\sin 6\pi t$), and/or the correlation parameter is large ($\alpha = 0.86$), the GML method does not work well. This is not surprising since we are trying to model the slowly varying trends by the deterministic mean function and any other regular behavior by the correlation structure. The model becomes less identifiable when the mean function varies quickly and the correlation is large.

8 Discussion

Hart (1991) showed that cross-validation produces kernel estimates that very nearly interpolates the data when the data are sufficiently positively correlated. Thus, an iterative scheme that estimate the covariance function of the errors based on the residuals of a cross-validated kernel estimate is inadvisable, and it is better to estimate the smoothing parameters and the correlation parameters simultaneously. The method proposed here takes the latter approach by assuming a parametric form for the correlation matrix. This is not as restrictive as it seems initially since some assumption has to be made about the error process. Otherwise the deterministic component and the stochastic component are not even identifiable.

We presented our methods using the general spline models on arbitrary domains and smoothing spline ANOVA models. These methods can be applied to time series data ($\mathcal{T} = [0,1]$), spatial data ($\mathcal{T} = E^2$), the functional data and the longitudinal data ($\mathcal{T} =$

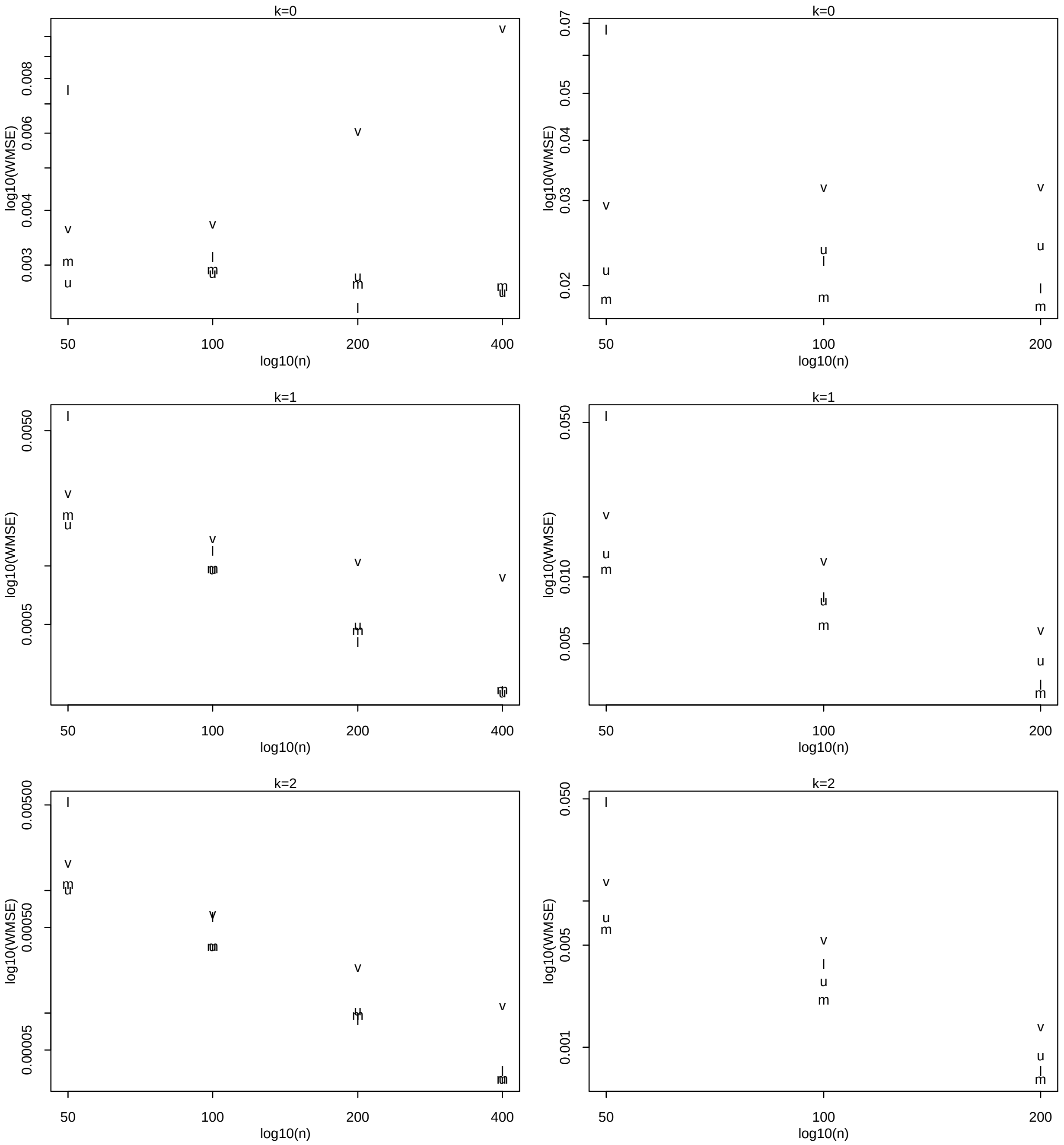

Figure 11: Plots of mean values of WMSE (T_k) on $\log_{10}$ scale for $k = 0$, $k = 1$ and $k = 2$. The correlation parameters for sample sizes 50, 100, 200 and 400 are 0.3, 0.55, 0.74 and 0.86 respectively. Left: $\sigma = 0.1$. Right: $\sigma = 0.3$. m, v, u and l represent GML, GCV, UBR and L methods respectively.

$\{1, \cdots, K\} \otimes [0,1]$, where K is the number of individuals) and spatial-temporal data ($\mathcal{T} = E^2 \otimes [0,1]$). These applications will be discussed elsewhere (Wang 1996).

The proposed GML and GCV methods estimate the correlation parameters very well. Estimation of the correlation parameters are of interest in some applications and is often ignored by previous publications. Our simulations indicate that methods based on WMSE (including MSE as a special case) may estimate the smoothing parameter very well, but they estimates correlation parameters poorly since the WMSE does not measure the whole estimation loss. The Kullback-Leibler discrepancy is a more appropriate measure. Methods of estimating the smoothing parameters and the correlation parameters which minimize the Kullback-Leibler discrepancy will be developed.

9 Acknowledgements

This research was supported in part by NIH Grants R01 EY09946, P60 DK20572 and P30 HD18258. The author is indebted to Grace Wahba for her advice and encouragement.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors, *Journal of the American Statistical Association* **85**: 749–759.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*, New York: Wiley.

Aronszajn, N. (1950). Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68**: 337–404.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis*, Holden-Day.

Cox, D. D., Koh, E., Wahba, G. and Yandell, B. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline model, *Ann. Statist.* **16**: 113–119.

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

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*, New York: Dekker.

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

Gu, C. (1992). Cross-validating non Gaussian data, *Journal of Computational and Graphical Statistics* **2**: 169–179.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory, *The Annals of Statistics* **21**: 217–234.

Gu, C. and Wahba, G. (1993a). Semiparametric ANOVA with tensor product thin plate spline, *Journal of the Royal Statistical Society B* **55**: 353–368.

Gu, C. and Wahba, G. (1993b). Smoothing spline ANOVA with component-wise Bayesian confidence intervals, *Journal of Computational and Graphical Statistics* **2**: 97–117.

Hart, J. D. (1991). Kernel regression estimation with time series errors, *Journal of the Royal Statistical Society B* **53**: 173–187.

Hart, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation, *Journal of the Royal Statistical Society B* **56**: 529–542.

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data, *Journal of the American Statistical Association* **81**: 1080–1088.

Harville, D. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects, *The Annals of Statistics* **4**: 384–395.

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

Herrmann, E., Gasser, T. and Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated, *Biometrika* **79**: 783–795.

Hurvich, C. M. and Zeger, S. L. (1990). A frequency domain selection criterion for regression with autocorrelated errors, *Journal of the American Statistical Association* **85**: 705–714.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics* **42**: 805–820.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions, *J. Math. Anal. Appl.* **33**: 82–94.

Kohn, R., Ansley, C. F. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters, *Journal of the American Statistical Association* **86**: 1042–1050.

Kohn, R., Ansley, C. F. and Wong, C. (1992). Nonparametric spline regression with autoregressive moving average errors, *Biometrika* **79**: 335–346.

Raz, J., Turetsky, B. and Fein, G. (1989). Selecting the smoothing parameter for estimation of slowly changing evoked potential signals, *Biometrics* **45**: 745–762.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects, *Statistical Science* **6**: 15–51.

SAS Institute (1992). *The Mixed Procedure*, SAS Technical Report P-229, SAS Institute, Inc., Cary, North Carolina.

SAS Institute (1994). *Getting Started with PROC MIXED*, SAS Institute, Inc., Cary, North Carolina.

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

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wahba, G. and Wang, Y. (1993). Behavior near zero of the distribution of GCV smoothing parameter estimates for splines, *Statistics and Probability Letters* **25**: 105–111.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy, *The Annals of Statistics* **23**: 1865–1895.

Wang, Y. (1996). Nonparametric mixed effects models, Technical Report 967, University of Wisconsin-Madison, Dept. of Statistics, Available at `http://www.sph.umich.edu/~yuedong`.