Ann Inst Stat Math (2009) 61:291–308
DOI 10.1007/s10463-007-0145-1

On regression model selection for the data with correlated errors

Wen Hsiang Wei

Received: 18 November 2005 / Revised: 1 May 2007 / Published online: 11 August 2007

Abstract A class of regression model selection criteria for the data with correlated errors is proposed. The proposed class of selection criteria is an estimator of weighted prediction risk. In addition, the proposed selection criteria are the generalizations of several commonly used criteria in statistical analysis. The theoretical and asymptotic properties for the class of criteria are established. Further, in the medium-sample case, the results based on a simulation study are quite consistent with the theoretical ones. The proposed criteria perform well in the simulations. Several applications are also given for a variety of statistical models.

Keywords Generalized cross-validation · Model selection · Nonparametric regression · Penalized likelihood · Smoothing splines

1 Introduction

Consider first the model

$$y_i = f(\boldsymbol{t}_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where y_i are observations at design points $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{id})$, $f(t)$ is a function and ϵ_i are zero mean, uncorrelated random errors with common variance σ^2. Let the fitted values $\hat{\boldsymbol{f}}(\boldsymbol{\lambda}) = \boldsymbol{H}(\boldsymbol{\lambda})\boldsymbol{y}$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ is a set of parameters associated with the selection of the model, $\boldsymbol{H}(\boldsymbol{\lambda})$ is an $n \times n$ matrix and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^t$. The parameter λ_j could be the subset of the discrete index set $\{1, 2, \ldots, p_j\}$ (see Li 1987) or the selection parameter in multivariate nonparametric regression, for examples, the

W. H. Wei (✉)
Department of Statistics, Tung Hai University, Taichung, Taiwan, ROC
e-mail: wenwei@thu.edu.tw

bandwidth in kernel-based method or the smoothing parameter in smoothing splines. Therefore, λ_j is assumed to be non-negative.

Obtaining a good selection parameter estimate is very crucial in the fitting process. One approach is to use the prediction risk

$$P(\boldsymbol{\lambda}) = E\left\{\frac{\left\|\boldsymbol{y}^\star - \hat{\boldsymbol{f}}(\boldsymbol{\lambda})\right\|^2}{n}\right\} = \frac{E\left\{[\boldsymbol{y}^\star - \hat{\boldsymbol{f}}(\boldsymbol{\lambda})]^t[\boldsymbol{y}^\star - \hat{\boldsymbol{f}}(\boldsymbol{\lambda})]\right\}}{n}, \tag{1}$$

as an object function (see Eubank 1988, p. 17), where $\boldsymbol{y}^\star = (y_1^\star, y_2^\star, \ldots, y_n^\star)^t$ is a vector of n new observations, $y_i^\star = f(\boldsymbol{t}_i) + \epsilon_i^\star$, and where $\epsilon_i^\star$, uncorrelated with $\epsilon_1, \ldots, \epsilon_n$, are zero mean random errors. The estimator of the prediction risk can be used as a criterion for selecting a sensible value of $\boldsymbol{\lambda}$. For instance, the commonly used GCV (generalized cross validation) criterion (Craven and Wahba 1979),

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{\hat{\sigma}^2(\boldsymbol{\lambda})}{[1-\mu_1(\boldsymbol{\lambda})]^2},$$

is nearly an unbiased estimator of the prediction risk in some cases (see Eubank 1988, Theorem 2.1), where $\hat{\sigma}^2(\boldsymbol{\lambda}) = \boldsymbol{y}^t[\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{\lambda})]^t[\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{\lambda})]\boldsymbol{y}/n$ is the variance estimate, $[1-\mu_1(\boldsymbol{\lambda})]^2$ is a penalty function for the smoothness of the fit, $\mu_1(\boldsymbol{\lambda}) = \mathrm{Tr}[\boldsymbol{H}(\boldsymbol{\lambda})]/n$, and $\mathrm{Tr}(\boldsymbol{A})$ is the trace of the matrix $\boldsymbol{A}$. The other predictive mean square error (see Wahba 1990, p. 55), closely related to the previous one, is

$$T(\boldsymbol{\lambda}) = \frac{\left\|\hat{\boldsymbol{f}}(\boldsymbol{\lambda}) - \boldsymbol{f}\right\|^2}{n} = \frac{\left[\hat{\boldsymbol{f}}(\boldsymbol{\lambda}) - \boldsymbol{f}\right]^t\left[\hat{\boldsymbol{f}}(\boldsymbol{\lambda}) - \boldsymbol{f}\right]}{n}, \tag{2}$$

where $\boldsymbol{f} = [f(\boldsymbol{t}_1), f(\boldsymbol{t}_2), \ldots, f(\boldsymbol{t}_n)]^t$. As indicated by the weak GCV theorem in the paper of Craven and Wahba, the minimizers of expected values of the predictive mean square error $T(\boldsymbol{\lambda})$ and GCV criterion are asymptotically equal in term of the prediction risk $E[T(\boldsymbol{\lambda})]$. Since $E[T(\boldsymbol{\lambda})] = P(\boldsymbol{\lambda}) - \sigma^2$, the minimizers of the two prediction risks, $E[T(\boldsymbol{\lambda})]$ and $P(\boldsymbol{\lambda})$, are the same provided that σ^2 is known. In addition to GCV, other related criteria have the form, $\hat{\sigma}^2(\boldsymbol{\lambda})/\phi[\mu_1(\boldsymbol{\lambda})]$, where $\phi(\cdot)$ is a penalty function for the smoothness of the fit. Commonly used selection criteria can be obtained by employing different choices of ϕ, including

1. (Craven and Wahba 1979) GCV: $\phi(\mu_1) = (1-\mu_1)^2$, (3)
2. (Akaike 1974) AIC: $\phi(\mu_1) = exp(-2\mu_1)$, (4)
3. (Rice 1984) T: $\phi(\mu_1) = 1 - 2\mu_1$, (5)
4. (Akaike 1970) FPE: $\phi(\mu_1) = \dfrac{1-\mu_1}{1+\mu_1}$, (6)

5. (Shibata 1981) $nS(\boldsymbol{\lambda})$: $\phi(\mu_1) = \dfrac{1}{1+2\mu_1}$, (7)

6. (Hocking 1976) $U(\boldsymbol{\lambda})$: $\phi(\mu_1) = \dfrac{(1-\mu_1)(n-1-n\mu_1)}{n-1}$, (8)

(also see Eubank 1988, pp. 38–40). The selection parameter might be sensitive to the presence of correlation in the errors. The breakdown of several popular data-driven smoothing parameter selection methods was indicated by Opsomer et al. (2001) in nonparametric regression. Further, the prediction risk given in the expression (1) might not be sensible for random errors with unequal variances. For example, suppose the variance of $\epsilon_i^\star$ is much larger than the one of $\epsilon_j^\star$. This implies the error in predicting $y_j^\star$ by $f(\boldsymbol{t}_j)$ might be more "predictable" than the one in predicting $y_i^\star$ by $f(\boldsymbol{t}_i)$, owing to the smaller variation. Thus, it seems to be sensible to assign different weights, which are related to the variances of the random errors, to the predictive errors. In next section, a weighted prediction risk and its estimator are proposed for cases where the errors of assumed model are correlated. The theoretical properties of the proposed estimator are also provided. Several examples, including a simulation study to justify the theoretical results and the applications of the proposed estimator to a variety of statistical models, are presented in Sects. 3 and 4, respectively. Finally, a concluding discussion is given in Sect. 5. The computational details for the proposed criteria and some supplementary material can be found at **http://web.thu.edu.tw/wenwei/www/papers/aismSupplement.pdf/**.

2 Weighted selection criteria

2.1 Weighted predictive mean square error and weighted selection criteria

Assume the $n \times n$ variance–covariance matrix of the correlated errors $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^t$ is $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}^{-1}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ is a set of correlation parameters. Let the fitted values $\hat{\boldsymbol{f}}_v = \boldsymbol{H}_v(\boldsymbol{h})\boldsymbol{y}$, where $\boldsymbol{h} = (\boldsymbol{\alpha}, \boldsymbol{\lambda}) = (h_1, \ldots, h_{m+k})$. Define the weighted prediction risk,

$$\begin{aligned} \mathrm{WP}(\boldsymbol{h}) &= E\left\{ \frac{\left|\left|\boldsymbol{y}^\star - \hat{\boldsymbol{f}}_v\right|\right|^2_{\boldsymbol{V}(\boldsymbol{\alpha})}}{\psi[\boldsymbol{V}(\boldsymbol{\alpha})]} \right\} \\ &= \frac{E\left\{\left[\boldsymbol{y}^\star - \boldsymbol{H}_v(\boldsymbol{\lambda})\boldsymbol{y}\right]^t \boldsymbol{V}(\boldsymbol{\alpha})\left[\boldsymbol{y}^\star - \boldsymbol{H}_v(\boldsymbol{\lambda})\boldsymbol{y}\right]\right\}}{\psi[\boldsymbol{V}(\boldsymbol{\alpha})]}, \end{aligned} \tag{9}$$

where $\psi[\boldsymbol{V}(\boldsymbol{\alpha})]$ is a positive function. The choice of $\psi(\cdot)$ reflects the effect of the matrix $\boldsymbol{V}(\boldsymbol{\alpha})$ on the weighted prediction risk. Except the sample size, another sensible choice is $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = \mathrm{Tr}[\boldsymbol{V}(\boldsymbol{\alpha})]$. Thus, when $\boldsymbol{V}(\boldsymbol{\alpha})$ is an identity matrix, the weighted prediction risk is also the prediction risk given in the expression (1). When $\boldsymbol{V}(\boldsymbol{\alpha})$ is a diagonal matrix with all diagonal elements equal to 1 except that the ith element is

equal to 0, i.e., no contribution from observation i to the weighted prediction risk, $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = n - 1$ in this situation might be a better choice than $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = n$. The other possible choice of $\psi(\cdot)$ is the retained number of principal components of $\boldsymbol{V}(\boldsymbol{\alpha})$. The weighted predictive mean square error, which generalizes the one given in the expression (2), is

$$\mathrm{WT}(\boldsymbol{h}) = \frac{\left|\left|\boldsymbol{f} - \hat{\boldsymbol{f}}_v\right|\right|^2_{\boldsymbol{V}(\boldsymbol{\alpha})}}{\psi[\boldsymbol{V}(\boldsymbol{\alpha})]} = \frac{(\boldsymbol{f} - \hat{\boldsymbol{f}}_v)^t \boldsymbol{V}(\boldsymbol{\alpha})(\boldsymbol{f} - \hat{\boldsymbol{f}}_v)}{\psi[\boldsymbol{V}(\boldsymbol{\alpha})]}. \tag{10}$$

When $\boldsymbol{V}(\boldsymbol{\alpha})$ is a diagonal matrix and $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = n$, $\mathrm{WT}(\boldsymbol{h})$ is the weighted mean square error discussed in O'Sullivan et al. (1986). If only $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = n$, $\mathrm{WT}(\boldsymbol{h})$ is the weighted mean square error in Wang (1998) with the order of the matrix $\boldsymbol{V}(\boldsymbol{\alpha})$ equal to one.

The proposed estimator of $\mathrm{WP}(\boldsymbol{h})$ is

$$W(\boldsymbol{h}) = \frac{\hat{\sigma}^2_v(\boldsymbol{h})}{\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]}, \tag{11}$$

where

$$\hat{\sigma}^2_v(\boldsymbol{h}) = \frac{\boldsymbol{y}^t\left[\boldsymbol{I} - \boldsymbol{H}_v(\boldsymbol{h})\right]^t \boldsymbol{V}(\boldsymbol{\alpha})\left[\boldsymbol{I} - \boldsymbol{H}_v(\boldsymbol{h})\right]\boldsymbol{y}}{\psi[\boldsymbol{V}(\boldsymbol{\alpha})]},$$

is the variance estimate,

$$\mu_{1v}(\boldsymbol{h}) = \frac{\mathrm{Tr}[\boldsymbol{H}_v(\boldsymbol{h})]}{\psi[\boldsymbol{V}(\boldsymbol{\alpha})]},$$

is a bounded function, and $\phi_W(\cdot)$ is a penalty function for the smoothness of the fit satisfying

$$\phi_W(x) = 1 - 2x + p(x), \quad \lim_{x\to 0}\frac{p(x)}{x^2} = c,$$

and where c is a finite constant. When $\boldsymbol{V}(\boldsymbol{\alpha})$ is an identity matrix and $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = n$, the selection criteria given in expressions from (3) to (8) take the form of the proposed class of weighted selection criteria $W(\boldsymbol{h})$. In addition, when $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = n$ and $\phi_W(x) = (1-x)^2$, $W(\boldsymbol{h})$ is the GCV function considered in Wang's paper, which corresponds to the one used in Altman (1990).

2.2 Properties of weighted selection criteria

The following theorem and corollary provide theoretical supports for the proposed selection criteria. The proofs largely follow the GCV theorem (Eubank 1988, Theorem 2.1) and the weak GCV theorem given by Craven and Wahba.

Theorem 1 *Let* $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{\epsilon}$, $\boldsymbol{f} = [f(\boldsymbol{t}_1), \ldots, f(\boldsymbol{t}_n)]^t$, *and* $Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}^{-1}(\boldsymbol{\alpha})$ *is an* $n \times n$ *variance–covariance matrix. Denote a positive function*

$$\mu_{2v}(\boldsymbol{h}) = \frac{\mathrm{Tr}\left[\boldsymbol{H}_v^t(\boldsymbol{h})\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{H}_v(\boldsymbol{h})\boldsymbol{V}^{-1}(\boldsymbol{\alpha})\right]}{\psi[\boldsymbol{V}(\boldsymbol{\alpha})]}.$$

Then,

$$\frac{|\mathrm{WP}(\boldsymbol{h}) - \{E[W(\boldsymbol{h})] - r(\boldsymbol{h})\}|}{E[WT(\boldsymbol{h})]} \leq e(\boldsymbol{h}),$$

where

$$e(\boldsymbol{h}) = \frac{1}{\phi_W[\mu_{1v}(\boldsymbol{h})]}\left\{\left(\left|\frac{p[\mu_{1v}(\boldsymbol{h})]}{\mu_{1v}(\boldsymbol{h})}\right| + 2\right)|\mu_{1v}(\boldsymbol{h})| + \frac{1_v p[\mu_{1v}(\boldsymbol{h})]}{\mu_{2v}(\boldsymbol{h})}\right\},$$

$r(\boldsymbol{h}) = 2(1_v - 1)\mu_{1v}(\boldsymbol{h})\sigma^2/\phi_W[\mu_{1v}(\boldsymbol{h})]$, *and where* $1_v = n/\psi[\boldsymbol{V}(\boldsymbol{\alpha})]$.

If the values of $e(\boldsymbol{h})$ and $r(\boldsymbol{h})$ are small, the theorem implies that the difference between $E[W(\boldsymbol{h})]$ and $\mathrm{WP}(\boldsymbol{h})$ is small relative to the other weighted prediction risk $E[\mathrm{WT}(\boldsymbol{h})]$. Further, if $E[\mathrm{WT}(\boldsymbol{h})]$ is bounded, $W(\boldsymbol{h})$ is nearly an unbiased estimator of $\mathrm{WP}(\boldsymbol{h})$ in this situation. Also, the minimum values of $E[\mathrm{WT}(\boldsymbol{h})]$ and $E[W(\boldsymbol{h})]$ have similar properties in term of minimizing the weighted prediction risk $E[\mathrm{WT}(\boldsymbol{h})]$, as indicated by the following corollary.

Corollary 1 *If* $E[\mathrm{WT}(\boldsymbol{h})]$ *has (at least) one minimizer* $\boldsymbol{h}_n^*$ *and* $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = O(n)$, *there exists a sequence of minimizers* $\hat{\boldsymbol{h}}_n$ *of* $E[W(\boldsymbol{h})]$ *such that*

$$\lim_{n\to\infty} \frac{E\left[\mathrm{WT}(\hat{\boldsymbol{h}}_n)\right]}{E\left[\mathrm{WT}(\boldsymbol{h}_n^*)\right]} = 1,$$

under the assumptions that

$$\lim_{n\to\infty} \mu_{1v}(\boldsymbol{h}_n) = 0, \quad \lim_{n\to\infty} \frac{\mu_{1v}^2(\boldsymbol{h}_n)}{\mu_{2v}(\boldsymbol{h}_n)} = 0, \quad \lim_{n\to\infty} \frac{(1_v - 1)\mu_{1v}(\boldsymbol{h}_n)}{\mu_{2v}(\boldsymbol{h}_n^*)} = 0,$$

where $\boldsymbol{h}_n = \hat{\boldsymbol{h}}_n$ *or* $\boldsymbol{h}_n = \boldsymbol{h}_n^*$.

Both the proofs of Theorem 1 and Corollary 1 are given in the Appendices A and B, respectively. The condition

$$\lim_{n\to\infty} \frac{(1_v - 1)\mu_{1v}(\boldsymbol{h}_n)}{\mu_{2v}(\boldsymbol{h}_n^*)} = 0,$$

is not required if either $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = n$ or $E[\mathrm{WT}(\boldsymbol{h}_n^*)]$ does not tend to zero. The corollary indicates the selection criteria $W(\boldsymbol{h})$ given in the expression (11) are weighted

predictive mean-square error (the one given in the expression (10)) criteria. When $\boldsymbol{V}(\boldsymbol{\alpha})$ is an identity matrix and $\phi_W(x) = (1-x)^2$, the corollary is the weak GCV theorem. Since $E[\mathrm{WT}(\boldsymbol{h})] = \mathrm{WP}(\boldsymbol{h}) - 1_v\sigma^2$, the minimizers of the two prediction risks, $E[\mathrm{WT}(\boldsymbol{h})]$ and $\mathrm{WP}(\boldsymbol{h})$, are the same given known values of σ^2 and correlation parameters.

To have the optimal properties, it is not necessary to define $\mu_{1v}(\boldsymbol{h})$ as a function of $\boldsymbol{H}_v(\boldsymbol{h})$ through its trace. For instance, if the function $\mu_{1v}(\boldsymbol{h})$ satisfies $\mu_{1v}(\boldsymbol{h}) - \mathrm{Tr}[\boldsymbol{H}_v(\boldsymbol{h})]/\psi[\boldsymbol{V}(\boldsymbol{\alpha})]$ converging uniformly to 0 or more relaxed condition $\lim_{n\to\infty}\{\mu_{1v}(\boldsymbol{h}_n) - \mathrm{Tr}[\boldsymbol{H}_v(\boldsymbol{h}_n)]/\psi[\boldsymbol{V}(\boldsymbol{\alpha}_n)]\} = 0$, the theorem and corollary still hold, where $\boldsymbol{h}_n = (\boldsymbol{\alpha}_n, \boldsymbol{\lambda}_n)$. Thus, more flexible choices of $\mu_{1v}(\boldsymbol{h})$ can be made than only through the trace of $\boldsymbol{H}_v(\boldsymbol{h})$.

3 Simulations

The purpose of the following simulations is to illustrate that all the selection criteria given in the expressions from (3) to (8) perform well. In addition, the numerical results are consistent with the theoretical ones, even in medium-sample case. A range of scenarios, including different choices of $\boldsymbol{V}(\boldsymbol{\alpha})$ and noise levels, have been set up for the simulation study.

3.1 Weighted linear regression

In the simulation, the values of four input variables, X_1, X_2, X_3, X_4, were in [0, 1] and 100 observations were generated from the model

$$y_i = 1 + 2x_{i1} + 4x_{i3} + \epsilon_i, \quad i = 1, \ldots, 100,$$

where ϵ_i are zero mean random errors. The errors were generated from both Gaussian AR(1) and MA(1) processes with the standard deviations of uncorrelated Gaussian errors, σ_g, equal to 0.2, 1, and 2. The autocorrelation values at lag 1, $\rho(1)$, for the Gaussian AR(1) process were -0.8, -0.2, 0.2, and 0.8, while -0.4, -0.2, 0.2, and 0.4 for the Gaussian MA(1) process. For simplicity, assume $\boldsymbol{V}(\boldsymbol{\alpha}) = \boldsymbol{V}$ is known. Two choices for $\psi(\boldsymbol{V})$ were the sample size and retained number of principal components of the matrix $\boldsymbol{V}$ by including just enough components to explain 90% amount of the variance. 500 replicates of random errors were generated. For each sample, the proposed selection criteria and associated weighted predictive mean-square error given in the expression (9) were computed for all possible models with at least one input variable, i.e., total 15 possible models. The averages of these quantities can be used to estimate the expected values of the proposed selection criteria and weighted predictive mean-square error. The results for the averages corresponding to the true model, i.e., the estimates of $\mathrm{WP}(\hat{\boldsymbol{h}}_n)$ and $E\big[W(\hat{\boldsymbol{h}}_n)\big]$, can be obtained. The selection criteria and weighted predictive mean-square error produce very similar results, which provide numerical support for Theorem 1. Since the results under different settings are quite consistent, Table 1 summarizes parts of these results for Gaussian AR(1) process. The first number in the parenthesis is the average with $\psi(\boldsymbol{V})$ equal to the sample size and

Table 1 Weighted linear regression and smoothing with Gaussian AR(1) and MA(1) random errors, respectively

	AR(1)		MA(1)	
	$\rho(1) = -0.8$ $\sigma_g = 1$	$\rho(1) = 0.8$ $\sigma_g = 1$	$\rho(1) = -0.4$ $\sigma_g = 1$	$\rho(1) = 0.4$ $\sigma_g = 1$
WP($\boldsymbol{h}$)	(1.030,1.688)	(1.030,1.688)		
GCV	(1.036,1.768)	(1.032,1.760)	(1.001,1.057)	(1.002,1.106)
AIC	(1.035,1.764)	(1.031,1.756)	(1.000,1.023)	(1.002,1.106)
T	(1.037,1.773)	(1.033,1.765)	(1.001,1.057)	(1.002,1.106)
FPE	(1.035,1.764)	(1.031,1.756)	(1.000,1.023)	(1.002,1.106)
nS	(1.034,1.756)	(1.029,1.748)	(1.000,1.023)	(1.054,1.106)
U	(1.037,1.769)	(1.032,1.761)	(1.001,1.057)	(1.002,1.106)

the second number otherwise. On the other hand, since the averages of the proposed selection criteria and weighted predictive mean-square error attain their minimums as the postulated model being the true model, the estimate of $E[\mathrm{WT}(\hat{\boldsymbol{h}}_n)]/E[\mathrm{WT}(\boldsymbol{h}_n^*)]$ is equal to 1.

3.2 Weighted smoothing

In the simulation, 100 observations were generated from the model

$$y_i = \sin(2\pi i/n), \quad i = 1, \ldots, 100,$$

where ϵ_i are zero mean random errors. The function was utilized in the simulation study (Wang 1998, p. 344). The errors were generated from both Gaussian AR(1) and MA(1) processes with the same variances of uncorrelated Gaussian errors and autocorrelation values at lag 1 as the ones in the simulation study given in Sect. 3.1. Also, there were two choices for $\psi(\boldsymbol{V})$, as given in the previous simulation study. 100 replicates of random errors were generated. For each sample, the smoothing spline fit using the B-spline of degree 3 and a second order penalty (see Eilers and Marx 1996) was computed. The chosen knots divided the domain of t $(0.05\pi - 2\pi)$ into 20 intervals of equal width. Then, the proposed selection criteria and weighted predictive mean-square error given in the expression (9) were computed for the smoothing parameter with values 10, 20, …, 1000. The averages of the proposed selection criteria and weighted predictive mean-square error were computed. Evaluated at their minimums, the results can be obtained. Similar to Sect. 3.1, the numbers in the parenthesis are the averages based on different choices of $\psi(\boldsymbol{V})$. The values for different selection criteria and the weighted predictive mean-square error are quite close. This provides numerical support for Theorem 1. In addition, the estimated values of $E\big[\mathrm{WT}(\hat{\boldsymbol{h}}_n)\big]/E\big[\mathrm{WT}(\boldsymbol{h}_n^*)\big]$ are quite close to 1. Note that similar results can be obtained when the domain of t was divided into 50 intervals of equal width. Since these results under different settings are quite

consistent, parts of these results, the estimated values of $E\left[\mathrm{WT}(\hat{\boldsymbol{h}}_n)\right]/E\left[\mathrm{WT}(\boldsymbol{h}_n^*)\right]$ for Gaussian MA(1) process, are provided in Table 1.

4 Applications

In the following examples, the proposed weighted selection criteria are applied to a variety of statistical models.

4.1 Weighted linear regression

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is an $n \times p$ design matrix and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ is a $p \times 1$ coefficient vector. For simplicity, assume $\boldsymbol{V}(\boldsymbol{\alpha}) = \boldsymbol{V}$ is known. Let the discrete index set $\Lambda = \{1, 2, \ldots, p\}$ and $\boldsymbol{X}_\lambda = (\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_l})$, where $\lambda = \{i_1, \ldots, i_l\}$ is a subset of the index set Λ. Thus, $\boldsymbol{H}_v(\lambda) = \boldsymbol{X}_\lambda[\boldsymbol{X}_\lambda^t \boldsymbol{V} \boldsymbol{X}_\lambda]^{-1}\boldsymbol{X}_\lambda^t \boldsymbol{V}$ and $\mu_{1v}(\lambda) = \mu_{2v}(\lambda) = l/\psi(\boldsymbol{V})$. Further, both $e(\lambda)$ and $r(\lambda)$ tend to 0 as n tends to infinity. Therefore, $W(\lambda)$ is nearly an unbiased estimator of WP(λ) implied by Theorem 1. As $\lim_{n\to\infty}(1_v - 1) = 0$, Corollary 1 also holds.

4.2 Weighted smoothing

Suppose the weighted (or generalized) smoothing spline estimate $\hat{\boldsymbol{f}}_v$ is the minimizer of

$$(\boldsymbol{y} - \boldsymbol{f})^t \boldsymbol{V}(\boldsymbol{y} - \boldsymbol{f}) + \sum_{j=1}^{k} \lambda_j \left|\left|J_j(f)\right|\right|^2, \tag{12}$$

over the class of all twice differentiable functions (also see Opsomer et al. 2001, p. 148), where J_j are some operators, for examples, the orthogonal projectors given in chapter 10 of Wahba (1990) or $\left|\left|J_j(f)\right|\right|^2 = \int [\frac{\partial^2 f(\boldsymbol{t})}{\partial t_j^2}]^2 dt_j$ for additive splines (see Hastie and Tibshirani 1990). Let $f(t) = \sum_{j=1}^{p_n} a_j B_j(\boldsymbol{t})$, where p_n is the number of suitably chosen basis functions, usually at least large enough to ensure the accuracy of the approximation and $B_j(\boldsymbol{t})$ are basis functions, for example, the commonly used B-splines (see Green and Silverman 1994, pp. 155–159). Thus, when $\left|\left|J_j(f)\right|\right|^2 = \boldsymbol{a}^t \boldsymbol{P}_j \boldsymbol{a}$, the expression (12) reduces to

$$(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a})^t \boldsymbol{V}(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}) + \sum_{j=1}^{k} \lambda_j \boldsymbol{a}^t \boldsymbol{P}_j \boldsymbol{a},$$

where $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)^t = [B_j(\boldsymbol{t}_i)]_{ij}$ is an $n \times p_n$ matrix, $\boldsymbol{a} = (a_1, \ldots, a_{p_n})^t$ and $\boldsymbol{P}_j$ are $p_n \times p_n$ penalty matrices. The estimate of the coefficient $\boldsymbol{a}$ is $\hat{\boldsymbol{a}}(\boldsymbol{\lambda}) = (\boldsymbol{B}^t\boldsymbol{V}\boldsymbol{B} + \sum_{j=1}^{k} \lambda_j \boldsymbol{P}_j)^{-1}\boldsymbol{B}^t\boldsymbol{V}\boldsymbol{y}$. Then, the minimizer is

$$\hat{\boldsymbol{f}}_v = \left[\hat{f}_v(\boldsymbol{t}_1, \boldsymbol{\lambda}), \ldots, \hat{f}_v(\boldsymbol{t}_n, \boldsymbol{\lambda})\right]^t = \boldsymbol{B}\left(\boldsymbol{B}^t\boldsymbol{V}\boldsymbol{B} + \sum_{j=1}^{k} \lambda_j \boldsymbol{P}_j\right)^{-1} \boldsymbol{B}^t\boldsymbol{V}\boldsymbol{y} = \boldsymbol{H}_v(\boldsymbol{\lambda})\boldsymbol{y},$$

where $\boldsymbol{H}_v(\boldsymbol{\lambda}) = [h_{is}(\boldsymbol{\lambda})]_{is} = \boldsymbol{B}(\boldsymbol{B}^t\boldsymbol{V}\boldsymbol{B} + \sum_{j=1}^{k} \lambda_j \boldsymbol{P}_j)^{-1}\boldsymbol{B}^t\boldsymbol{V}$ is the hat matrix, $i = 1, \ldots, n,\ s = 1, \ldots, n$.

Suppose $\boldsymbol{P}_j$ are positive-definite matrices with eigenvalues $0 < h^*_{j,p_n} \leq \cdots \leq h^*_{j,2} \leq h^*_{j,1}$ and $0 \leq h_{\min(n,p_n)} \leq \cdots \leq h_2 \leq h_1$ are the eigenvalues of $\boldsymbol{B}^t\boldsymbol{V}\boldsymbol{B}$. If $h_i = O(i^{-q_1}),\ q_1 > 1$ and the fastest decay rate of the eigenvalues $h^*_{j,\cdot}$ is $h^*_{l,p_n} = 0(n^{-q_2}),\ q_2 < 1/2$, Theorem 1 and Corollary 1 hold. The justifications are given in Appendix C. When $\boldsymbol{B}^t\boldsymbol{V}\boldsymbol{B}$ is nonsingular and there exist generalized eigenvalues $h_{l,i}$ of $(\boldsymbol{B}^t\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{x} = \lambda^*\boldsymbol{P}_l^{-1}\boldsymbol{x}$ with decay rate $O(i^{-q_3}),\ q_3 > 1$ (see Golub and Van Loan 1993, pp. 466–472), Theorem 1 and Corollary 1 also hold. The detailed justifications were delegated to the supplementary material. When the matrix $\boldsymbol{V}$ involves a set of correlation parameters, Theorem 1 and Corollary 1 can be justified alone the lines given in Appendix C by specifying the decay rate of the eigenvalues of $\boldsymbol{B}^t\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{B}$.

4.3 Model selection in multivariate linear regression

Consider the linear model

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, k,$$

where $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{jn_j})^t$, $\boldsymbol{X}_j = (\boldsymbol{x}_{j1}, \ldots, \boldsymbol{x}_{jp_j})$ is an $n_j \times p_j$ design matrix, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})^t$ and $\boldsymbol{\epsilon}_j = (\epsilon_{j1}, \epsilon_{j2}, \ldots, \epsilon_{jn_j})^t$. Let the discrete index sets $\Lambda_j = \{1, 2, \ldots, p_j\}$ and $\boldsymbol{X}_{\lambda_j} = (\boldsymbol{x}_{ji_1}, \ldots, \boldsymbol{x}_{ji_{l_j}})$, where $\lambda_j = \{i_1, \ldots, i_{l_j}\}$ is a subset of Λ_j. Thus,

$$\boldsymbol{H}_v(\boldsymbol{h}) = \boldsymbol{X}_{\boldsymbol{\lambda}}\left[\boldsymbol{X}^t_{\boldsymbol{\lambda}}\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{X}_{\boldsymbol{\lambda}}\right]^{-1}\boldsymbol{X}^t_{\boldsymbol{\lambda}}\boldsymbol{V}(\boldsymbol{\alpha}),$$

and $\mu_1(\boldsymbol{h}) = \mathrm{Tr}[\boldsymbol{H}_v(\boldsymbol{h})]/\psi[\boldsymbol{V}(\boldsymbol{\alpha})]$, where $\boldsymbol{X}_{\boldsymbol{\lambda}} = \mathrm{Diag}(\boldsymbol{X}_{\lambda_1}, \ldots, \boldsymbol{X}_{\lambda_k})$ is a $\sum_{j=1}^{k} n_j \times \sum_{j=1}^{k} l_j$ matrix and $\sigma^2\boldsymbol{V}^{-1}(\boldsymbol{\alpha})$ is the variance–covariance matrix of $(\boldsymbol{\epsilon}^t_1, \ldots, \boldsymbol{\epsilon}^t_k)^t$.

4.4 Generalized ridge regression

Let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{X}$ is an $n \times p_n$ design matrix and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^t$, and where $p_n \leq n$. The weighted (or generalised) ridge regression estimate of $\boldsymbol{\beta}$, the

minimizer of

$$(\boldsymbol{y}-\boldsymbol{X\beta})^t\boldsymbol{V}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{X\beta})+\sum_{j=1}^{k}\lambda_j\boldsymbol{\beta}^t\boldsymbol{P}_j\boldsymbol{\beta},$$

is $[\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{X}+\sum_{j=1}^{k}\lambda_j\boldsymbol{P}_j]^{-1}\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{y}$, where $\boldsymbol{P}_j$ are nonsingular matrices. Thus,

$$\boldsymbol{H}_v(\boldsymbol{h})=\boldsymbol{X}\left[\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{X}+\sum_{j=1}^{k}\lambda_j\boldsymbol{P}_j\right]^{-1}\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha}).$$

4.5 Partial splines

Let

$$y_i=\boldsymbol{x}_i\boldsymbol{\beta}+f(\boldsymbol{t}_i)+\epsilon_i,\quad i=1,\ldots,n,$$

(see Eubank 1988, pp. 292–293; Wahba 1990, Chapt. 6) and $f(\boldsymbol{t})=\sum_{j=1}^{p_n}a_jB_j(\boldsymbol{t})$. Then, the weighted (or generalized) partial spline estimate $\hat{\boldsymbol{f}}_v$ is the minimizer of

$$(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{f})^t\boldsymbol{V}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{f})+\sum_{j=1}^{k}\lambda_j\left|\left|J_j(f)\right|\right|^2, \tag{13}$$

where J_j are some operators. When $\left|\left|J_j(f)\right|\right|^2=\boldsymbol{a}^t\boldsymbol{P}_j\boldsymbol{a}$, the expression (13) reduces to

$$(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Ba})^t\boldsymbol{V}(\boldsymbol{\alpha})(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Ba})+\sum_{j=1}^{k}\lambda_j\boldsymbol{a}^t\boldsymbol{P}_j\boldsymbol{a},$$

where $\boldsymbol{B}=(\boldsymbol{b}_1,\ldots,\boldsymbol{b}_n)^t=[B_j(\boldsymbol{t}_i)]_{ij}$ is an $n\times p_n$ matrix, $\boldsymbol{a}=(a_1,\ldots,a_{p_n})^t$, and $\boldsymbol{P}_j$ are $p_n\times p_n$ penalty matrices. The estimates of the coefficients $\boldsymbol{a}$ and $\boldsymbol{\beta}$ are

$$\hat{\boldsymbol{a}}(\boldsymbol{h})=\left(\boldsymbol{B}^t\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{B}+\sum_{j=1}^{k}\lambda_j\boldsymbol{P}_j\right)^{-1}\boldsymbol{B}^t\boldsymbol{V}(\boldsymbol{\alpha})\left[\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{h})\right],$$

and

$$\hat{\boldsymbol{\beta}}(\boldsymbol{h})=\left\{\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha})\left[\boldsymbol{I}-\boldsymbol{H}_2(\boldsymbol{h})\right]\boldsymbol{X}\right\}^{-1}\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha})\left[\boldsymbol{I}-\boldsymbol{H}_2(\boldsymbol{h})\right]\boldsymbol{y},$$

respectively, where

$$\boldsymbol{H}_2(\boldsymbol{h}) = \boldsymbol{B}\left(\boldsymbol{B}^t\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{B} + \sum_{j=1}^{k}\lambda_j\boldsymbol{P}_j\right)^{-1}\boldsymbol{B}^t\boldsymbol{V}(\boldsymbol{\alpha}).$$

Then, the minimizer is

$$\begin{aligned}\hat{\boldsymbol{f}}_v &= \left[\hat{f}_v(\boldsymbol{t}_1,\boldsymbol{h}),\ldots,\hat{f}_v(\boldsymbol{t}_n,\boldsymbol{h})\right]^t = \boldsymbol{H}_v(\boldsymbol{h})\boldsymbol{y}\\ &= \left[\boldsymbol{H}_1(\boldsymbol{h}) + \boldsymbol{H}_2(\boldsymbol{h}) - \boldsymbol{H}_1(\boldsymbol{h})\boldsymbol{H}_2(\boldsymbol{h})\right]\boldsymbol{y},\end{aligned}$$

where

$$\boldsymbol{H}_1(\boldsymbol{h}) = \boldsymbol{X}\left\{\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha})\left[\boldsymbol{I}-\boldsymbol{H}_2(\boldsymbol{h})\right]\boldsymbol{X}\right\}^{-1}\boldsymbol{X}^t\boldsymbol{V}(\boldsymbol{\alpha})\left[\boldsymbol{I}-\boldsymbol{H}_2(\boldsymbol{h})\right].$$

4.6 Spline smoothing in generalized linear models

Consider the standard generalized linear model in which each component of the response vector has a distribution taking the form

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - m(\theta_i)}{u(\phi)} + c(y_i,\phi)\right],$$

where θ_i and ϕ are scalar parameters, and $m(\cdot)$, $u(\cdot)$ and $c(\cdot)$ are specific functions. The dependence of the response y_i on the associated explanatory variable $\boldsymbol{t}_i$ can be modeled through the link function $d(\cdot)$, where $\theta_i = d(\alpha + \boldsymbol{\beta}^t\boldsymbol{t}_i)$, and where α and $\boldsymbol{\beta}$ are some parameters. The natural link and $u(\phi) = 1$ are assumed hereafter. In addition, let $\theta_i = f(\boldsymbol{t}_i)$ and $\boldsymbol{f} = \boldsymbol{B}\boldsymbol{a}$. The estimate of f is the minimizer of the penalized negative logarithm of the likelihood,

$$\begin{aligned}&\sum_{i=1}^{n}\left\{m\left[f(\boldsymbol{t}_i)\right] - y_i f(\boldsymbol{t}_i)\right\} + \frac{1}{2}\sum_{j=1}^{k}\lambda_j\left|\left|J_j(f)\right|\right|^2\\ &= \sum_{i=1}^{n} m\left[(\boldsymbol{B}\boldsymbol{a})_i\right] - \boldsymbol{y}^t\boldsymbol{B}\boldsymbol{a} + \frac{1}{2}\sum_{j=1}^{k}\lambda_j\boldsymbol{a}^t\boldsymbol{P}_j\boldsymbol{a},\end{aligned}$$

where $(\boldsymbol{B}\boldsymbol{a})_i$ is the ith element of the vector $\boldsymbol{B}\boldsymbol{a}$. The estimate of the coefficient vector $\boldsymbol{a}$ can be written as a weighted penalized least squares estimate,

$$\hat{\boldsymbol{a}}(\boldsymbol{\lambda}) = \left(\boldsymbol{B}^t\ddot{\boldsymbol{M}}\boldsymbol{B} + \sum_{j=1}^{k}\lambda_j\boldsymbol{P}_j\right)^{-1}\boldsymbol{B}^t\ddot{\boldsymbol{M}}\boldsymbol{z},$$

where $\ddot{\boldsymbol{M}} = \text{diag}\{m''[\hat{\boldsymbol{a}}^t(\boldsymbol{\lambda})\boldsymbol{b}_1], \ldots, m''[\hat{\boldsymbol{a}}^t(\boldsymbol{\lambda})\boldsymbol{b}_n]\}$ and $z = \boldsymbol{B}\hat{\boldsymbol{a}}(\boldsymbol{\lambda}) + \ddot{\boldsymbol{M}}^{-1}(\boldsymbol{y} - \dot{\boldsymbol{m}})$, and where $\dot{\boldsymbol{m}} = \{m'[\hat{\boldsymbol{a}}^t(\boldsymbol{\lambda})\boldsymbol{b}_1], \ldots, m'[\hat{\boldsymbol{a}}^t(\boldsymbol{\lambda})\boldsymbol{b}_n]\}^t$. Let

$$\boldsymbol{H}_v(\boldsymbol{\lambda}) = \ddot{\boldsymbol{M}}^{1/2}\boldsymbol{B}\left(\boldsymbol{B}^t\ddot{\boldsymbol{M}}\boldsymbol{B} + \sum_{j=1}^{k}\lambda_j\boldsymbol{P}_j\right)^{-1}\boldsymbol{B}^t\ddot{\boldsymbol{M}}^{1/2},$$

and $\mu_{1v}(\boldsymbol{\lambda}) = \text{Tr}[\boldsymbol{H}_v(\boldsymbol{\lambda})]/\psi(\ddot{\boldsymbol{M}})$. The smoothing parameter estimate is the minimizer of the following function, $\|\ddot{\boldsymbol{M}}^{-1/2}(\boldsymbol{y} - \dot{\boldsymbol{m}})\|^2/\phi_W[\mu_{1v}(\boldsymbol{\lambda})]$. When $\phi_W(x) = (1-x)^2$ and $\psi(\ddot{\boldsymbol{M}}) = n$, the above function is the GCV function (see Wahba 1990, p. 113; O'Sullivan et al. 1986).

4.7 Nonparametric regression incorporating the information provided by derivatives

The derivatives of a function can provide useful information for data analysis. To incorporate the information provided by the derivatives, consider first the model

$$y_i^{(d)} = f^{(d)}(t_i) + \epsilon_i^{(d)}, \quad i = 1, \ldots, n; \ \ d = 0, \ldots, k,$$

where the data, $y_i^{(d)}$, associated with the first k'th derivatives of the function $f(\cdot)$ are available, $f^{(d)}(\cdot)$ is the dth order of derivative of the function with $f(\cdot) = f^{(0)}(\cdot)$, and where $\epsilon_i^{(d)}$ are zero mean random errors. Let $f(t) = \sum_{j=1}^{p_n} a_j B_j^{(0)}(t)$. The weighted sum of squares are

$$\sum_{d=0}^{k}\lambda_d\left(\boldsymbol{y}^{(d)} - \boldsymbol{B}^{(d)}\boldsymbol{a}\right)^t\boldsymbol{V}^{(d)}(\boldsymbol{\alpha})\left(\boldsymbol{y}^{(d)} - \boldsymbol{B}^{(d)}\boldsymbol{a}\right),$$

where $\lambda_0 = 1$, $\boldsymbol{y}^{(d)} = (y_1^{(d)}, \ldots, y_n^{(d)})^t$, $\boldsymbol{B}^{(d)} = \{B_j^{(d)}(t_i)\}_{ij}$, $\sigma^2\boldsymbol{V}^{(d)}(\boldsymbol{\alpha})$ is the variance–covariance matrix of the data $\boldsymbol{y}^{(d)}$, and where $B_j^{(d)}$ is the dth order of derivative of the basis function $B_j^{(0)}$ with respect to t. The parameters λ_d play a key role in controlling the trade-off between zero order derivative information represented by the weighted residual sum of squares and the information provided by other order of derivatives. Then,

$$\begin{aligned}\boldsymbol{H}_v(\boldsymbol{h})\boldsymbol{y} &= (\oplus_{d=0}^{k}\boldsymbol{B}^{(d)})\left\{\oplus_{i=0}^{k}\left\{\boldsymbol{B}^t\boldsymbol{V}^{(0)}(\boldsymbol{\alpha})\boldsymbol{B} + \left[\sum_{d=1}^{k}\lambda_d(\boldsymbol{B}^{(d)})^t\boldsymbol{V}^{(d)}(\boldsymbol{\alpha})\boldsymbol{B}^{(d)}\right]\right\}^{-1}\right\} \\ &\quad \times\left[\oplus_{d=0}^{k}\boldsymbol{B}^{(d)}\boldsymbol{V}^{(d)}(\boldsymbol{\alpha})\right]\boldsymbol{y},\end{aligned}$$

where $\boldsymbol{y} = [(\boldsymbol{y}^{(0)})^t, \ldots, (\boldsymbol{y}^{(k)})^t]^t$ and $\oplus_{i=1}^{k}A_i$ is the Kronecker sum of the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_k$.

5 Concluding discussion

As indicated by Theorem 1, all the selection criteria given in the expressions from (3) to (8) have similar optimal properties in term of the difference between their expected values and the prediction risk WP($\boldsymbol{h}$). Further, Corollary 1 indicates that the minimizers of $E[W(\boldsymbol{h})]$ and $E[\mathrm{WT}(\boldsymbol{h})]$ also have similar properties in term of minimizing the weighted prediction risk. In Sect. 3.1, the correct model can be selected by both the averages of these selection criteria and the prediction risk WP($\boldsymbol{h}$). As illustrated in Table 1, the numerical results based on the simulated data are quite consistent with the theoretical results, even in the medium-sample case. In addition, the proposed criteria have wide applications, as presented in Sect. 4.

When the correlation is parametrically specified, commonly used structures, such as the ARIMA model, might be employed to fit the data. Thus, the proposed criteria can be used to estimate both the selection parameters and correlation parameters. For the correlation not parametrically specified, the analogue criteria can be obtained by replacing $\boldsymbol{V}(\alpha)$ in the proposed criteria given in the expression (11) with a sensible non-parametric estimate $\hat{\boldsymbol{V}}$. The minimum values of $E[\mathrm{WT}(\boldsymbol{\lambda})]$ and the expected value of the analogue criteria still have similar properties in term of minimizing the weighted prediction risk $E[\mathrm{WT}(\boldsymbol{\lambda})]$, i.e., a result analogous to Corollary 1. The arguments in proving the result are similar to the ones given in Corollary 1 of Wei (2005).

There is still room for future research in regression model selection with correlated errors. Nonparametric modeling and semi-parametric modelling have been widely used techniques in recent years (see Härdle et al. 2004). The proposed criteria could be applied to some of these models. The proposed criteria have different sensitivities to the changes of the selection parameters or correlation parameters. A thorough robust and sensitivity analysis for different selection criteria could be helpful. The other issue is about the comparison of different criteria. The characterization of efficiency associated with a selection criterion is still unclear.

6 Appendix: Proofs

6.1 Appendix A: Proofs of Theorem 1

Based on the formula for the mean of a quadratic form, the expected value of the weighted predictive mean-square error is

$$E\left[\mathrm{WT}(\boldsymbol{h})\right] = b^2(\boldsymbol{h}) + \sigma^2\mu_{2v}(\boldsymbol{h}),$$

where $b^2(\boldsymbol{h}) = \{1/\psi[\boldsymbol{V}(\boldsymbol{\alpha})]\}\boldsymbol{f}^t[\boldsymbol{I} - \boldsymbol{H}_v(\boldsymbol{h})]^t\boldsymbol{V}(\boldsymbol{\alpha})[\boldsymbol{I} - \boldsymbol{H}_v(\boldsymbol{h})]\boldsymbol{f}$ is the bias term and $\sigma^2\mu_{2v}(\boldsymbol{h})$ is the variance term. Similarly,

$$E\left[W(\boldsymbol{h})\right] = \frac{b^2(\boldsymbol{h}) + \sigma^2\left[1_v - 2\mu_{1v}(\boldsymbol{h}) + \mu_{2v}(\boldsymbol{h})\right]}{\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]}.$$

The following shows that the difference between the expected values of the selection criteria and WP($\boldsymbol{\lambda}$) is close to 0. Denote

$$r(\boldsymbol{h}) = 2(1_v - 1)\mu_{1v}(\boldsymbol{h})\sigma^2/\phi_W[\mu_{1v}(\boldsymbol{h})].$$

Then,

$$\begin{aligned} & E\left[W(\boldsymbol{h})\right] - 1_v\sigma^2 - r(\boldsymbol{h}) \\ &= \frac{b^2(\boldsymbol{h}) + \sigma^2\left[1_v - 2\mu_{1v}(\boldsymbol{h}) + \mu_{2v}(\boldsymbol{h})\right] - 1_v\sigma^2\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]}{\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]} - r(\boldsymbol{h}) \\ &= \frac{b^2(\boldsymbol{h}) + \sigma^2\mu_{2v}(\boldsymbol{h}) - 1_v\sigma^2 p\left[\mu_{1v}(\boldsymbol{h})\right]}{\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]}. \end{aligned}$$

Further, since $\mathrm{WP}(\boldsymbol{h}) = E[\mathrm{WT}(\boldsymbol{h})] + 1_v\sigma^2$,

$$\begin{aligned} & \frac{WP(\boldsymbol{h}) - \{E\left[W(\boldsymbol{h})\right] - r(\boldsymbol{h})\}}{E\left[\mathrm{WT}(\boldsymbol{h})\right]} \\ &= \frac{b^2(\boldsymbol{h}) + \sigma^2\mu_{2v}(\boldsymbol{h}) - \left\{b^2(\boldsymbol{h}) + \sigma^2\mu_{2v}(\boldsymbol{h}) - 1_v\sigma^2 p\left[\mu_{1v}(\boldsymbol{h})\right]\right\}/\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]}{b^2(\boldsymbol{h}) + \sigma^2\mu_{2v}(\boldsymbol{h})} \\ &= \frac{-2\mu_{1v}(\boldsymbol{h}) + p\left[\mu_{1v}(\boldsymbol{h})\right]}{\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]} + \frac{1_v\sigma^2 p\left[\mu_{1v}(\boldsymbol{h})\right]}{\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]\left[b^2(\boldsymbol{h}) + \sigma^2\mu_{2v}(\boldsymbol{h})\right]}. \end{aligned}$$

Finally, since $|\mu_{1v}(\boldsymbol{h})| \le M$ and $b^2(\boldsymbol{h}) \ge 0$,

$$\begin{aligned} & \frac{|\mathrm{WP}(\boldsymbol{h}) - \{E\left[W(\boldsymbol{h})\right] - r(\boldsymbol{h})\}|}{E\left[\mathrm{WT}(\boldsymbol{h})\right]} \\ &\le \frac{1}{\phi_W\left[\mu_{1v}(\boldsymbol{h})\right]}\left\{\left(\left|\frac{p\left[\mu_{1v}(\boldsymbol{h})\right]}{\mu_{1v}(\boldsymbol{h})}\right| + 2\right)|\mu_{1v}(\boldsymbol{h})| + \frac{1_v p\left[\mu_{1v}(\boldsymbol{h})\right]}{\mu_{2v}(\boldsymbol{h})}\right\} \\ &= e(\boldsymbol{h}), \end{aligned}$$

where M is some constant.

6.2 Appendix B: Proofs of Corollary 1

By Theorem 1, the following inequality can be obtained,

$$E\left[\mathrm{WT}(\boldsymbol{h})\right]\left[1 - e(\boldsymbol{h})\right] \le E\left[W(\boldsymbol{h})\right] - 1_v\sigma^2 - r(\boldsymbol{h}) \le E\left[\mathrm{WT}(\boldsymbol{h})\right]\left[1 + e(\boldsymbol{h})\right],$$

for all $\boldsymbol{h}$. Note that $r(\boldsymbol{h})$ tends to 0 as $\mu_{1v}(\boldsymbol{h})$ tends to 0 and $\psi[\boldsymbol{V}(\boldsymbol{\alpha})] = O(n)$. Thus, the difference between the expected values of $W(\boldsymbol{h})$ and the predictive mean-square error $\mathrm{WT}(\boldsymbol{h})$ is approximately $1_v\sigma^2$. Intuitively, this implies the minimizers of $E[\mathrm{WT}(\boldsymbol{h})]$ and $E[W(\boldsymbol{h})]$ should be very close and further the corollary holds. The rigorous justifications are as follows. Since $E[W(\boldsymbol{h}_n^*)] - 1_v\sigma^2 - r(\boldsymbol{h}_n^*) \le E[\mathrm{WT}(\boldsymbol{h}_n^*)][1 + e(\boldsymbol{h}_n^*)]$

and $r(\boldsymbol{h}_n^*)$ tends to 0, there exists N such that at least one minimizer $\hat{\boldsymbol{h}}_n$ of $E[W(\boldsymbol{h})]$ is in the nonempty set

$$\left\{\boldsymbol{h} : E\left[W(\boldsymbol{h})\right] - 1_v\sigma^2 - r(\boldsymbol{h}) \le E\left[W(\boldsymbol{h}_n^*)\right] - 1_v\sigma^2 - r(\boldsymbol{h}_n^*)\right\},$$

for $n > N$. Thus,

$$\begin{aligned} & E\left[\mathrm{WT}(\hat{\boldsymbol{h}}_n)\right]\left[1 - e(\hat{\boldsymbol{h}}_n)\right] \\ & \le E\left[W(\hat{\boldsymbol{h}}_n)\right] - 1_v\sigma^2 - r(\hat{\boldsymbol{h}}_n) \\ & \le E\left[W(\boldsymbol{h}_n^*)\right] - 1_v\sigma^2 - r(\hat{\boldsymbol{h}}_n) \\ & \le E\left[\mathrm{WT}(\boldsymbol{h}_n^*)\right]\left[1 + e(\boldsymbol{h}_n^*)\right] + \left[r(\boldsymbol{h}_n^*) - r(\hat{\boldsymbol{h}}_n)\right]. \end{aligned}$$

Further,

$$\begin{aligned} & \frac{E\left[\mathrm{WT}(\hat{\boldsymbol{h}}_n)\right]}{E\left[\mathrm{WT}(\boldsymbol{h}_n^*)\right]} \\ & \le \left[\frac{1 + e(\boldsymbol{h}_n^*)}{1 - e(\hat{\boldsymbol{h}}_n)}\right] + \left\{\frac{1}{E\left[\mathrm{WT}(\boldsymbol{h}_n^*)\right]}\right\}\left[\frac{r(\boldsymbol{h}_n^*) - r(\hat{\boldsymbol{h}}_n)}{1 - e(\hat{\boldsymbol{h}}_n)}\right] \\ & \le \left[\frac{1 + e(\boldsymbol{h}_n^*)}{1 - e(\hat{\boldsymbol{h}}_n)}\right] + \frac{|1_v - 1|(|\mu_{1v}(\boldsymbol{h}_n^*)| + |\mu_{1v}(\hat{\boldsymbol{h}}_n)|)}{|\min\left\{\phi_W[\mu_{1v}(\boldsymbol{h}_n^*)], \phi_W[\mu_{1v}(\hat{\boldsymbol{h}}_n)]\right\}\mu_{2v}(\boldsymbol{h}_n^*)|}\left[\frac{2}{1 - e(\hat{\boldsymbol{h}}_n)}\right]. \end{aligned}$$

As $\mu_{1v}(\boldsymbol{h}_n)$ and $\mu_{1v}^2(\boldsymbol{h}_n)/\mu_{2v}(\boldsymbol{h}_n)$ tend to 0, $e(\boldsymbol{h}_n)$, $r(\boldsymbol{h}_n)$ and $p[\mu_{1v}(\boldsymbol{h}_n)]/\mu_{2v}(\boldsymbol{h}_n)$ tend to 0. Since $(1_v - 1)\mu_{1v}(\boldsymbol{h}_n)/\mu_{2v}(\boldsymbol{h}_n^*)$ tends to 0,

$$\frac{E\left[\mathrm{WT}(\hat{\boldsymbol{h}}_n)\right]}{E\left[\mathrm{WT}(\boldsymbol{h}_n^*)\right]} \to 1.$$

6.3 Appendix C: Theorem 1 and Corollary 1 applied to weighted smoothing

Denote $\boldsymbol{Z} = \boldsymbol{V}^{1/2}\boldsymbol{B}$ and let $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{Q}$ by singular value decomposition, where $\boldsymbol{U}$ and $\boldsymbol{Q}$ are orthogonal matrices and $\boldsymbol{D}$ is an $n \times p_n$ matrix whose diagonal entries are the square roots of the eigenvalues $h_1, \ldots, h_{\min(n,p_n)}$ of $\boldsymbol{Z}^t\boldsymbol{Z} = \boldsymbol{B}^t\boldsymbol{V}\boldsymbol{B}$ and with all other entries equal to 0.

As given in Sect. 6.1., $E\left[\mathrm{WT}(\boldsymbol{\lambda})\right] = b^2(\boldsymbol{\lambda}) + \sigma^2\mu_{2v}(\boldsymbol{\lambda})$, where

$$\psi(\boldsymbol{V})b^2(\boldsymbol{\lambda}) = \boldsymbol{f}^t\left[\boldsymbol{I} - \boldsymbol{H}_v(\boldsymbol{\lambda})\right]^t\boldsymbol{V}\left[\boldsymbol{I} - \boldsymbol{H}_v(\boldsymbol{\lambda})\right]\boldsymbol{f},$$

and $\psi(\boldsymbol{V})\mu_{2v}(\boldsymbol{\lambda}) = \mathrm{Tr}\left[\boldsymbol{H}_v^t(\boldsymbol{\lambda})\boldsymbol{V}\boldsymbol{H}_v(\boldsymbol{\lambda})\boldsymbol{V}^{-1}\right]$.

Let s_{ij} be the (i, j)th element of the matrix $(\boldsymbol{D}^t\boldsymbol{D} + \sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t)^{-1}$. Then,

$$\begin{aligned}
\max_{i,j} |s_{ij}| &\leq \left\| \left(\boldsymbol{D}^t\boldsymbol{D} + \sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t \right)^{-1} \right\|_2 \\
&= \sigma_{p_n}^{-1} \left(\boldsymbol{D}^t\boldsymbol{D} + \sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t \right) \\
&\leq \left[\sigma_{p_n}(\boldsymbol{D}^t\boldsymbol{D}) + \sigma_{p_n} \left(\sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t \right) \right]^{-1} \\
&\leq \frac{1}{\sum_{j=1}^{k} \lambda_j h^*_{j,p_n}},
\end{aligned}$$

and

$$\begin{aligned}
s_{ii} &\geq \sigma_{p_n} \left[\left(\boldsymbol{D}^t\boldsymbol{D} + \sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t \right)^{-1} \right] \\
&= \sigma_1^{-1} \left(\boldsymbol{D}^t\boldsymbol{D} + \sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t \right) \\
&\geq \left[\sigma_1(\boldsymbol{D}^t\boldsymbol{D}) + \sigma_1 \left(\sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t \right) \right]^{-1} \\
&\geq \frac{1}{h_1 + \sum_{j=1}^{k} \lambda_j h^*_{j,1}},
\end{aligned}$$

where $\|\boldsymbol{M}_1\|_2$ is the matrix 2-norm of the matrix $\boldsymbol{M}_1$ and $\sigma_n(\boldsymbol{M}_1) \leq \sigma_{n-1}(\boldsymbol{M}_1) \leq \cdots \leq \sigma_2(\boldsymbol{M}_1) \leq \sigma_1(\boldsymbol{M}_1)$ are the singular values (or eigenvalues) of the $n \times n$ symmetric matrix $\boldsymbol{M}_1$. Then,

$$\begin{aligned}
\psi(\boldsymbol{V})\mu_{1v}(\boldsymbol{\lambda}) &= \mathrm{Tr} \left[\boldsymbol{U}\boldsymbol{D} \left(\boldsymbol{D}^t\boldsymbol{D} + \sum_{j=1}^{k} \lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t \right)^{-1} \boldsymbol{D}^t\boldsymbol{U}^t \right] \\
&= \sum_{i=1}^{\min(n,p_n)} h_i s_{ii}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{\sum_{i=1}^{\min(n,p_n)} h_i}{\sum_{j=1}^{k} \lambda_j h^*_{j,p_n}} \\
&\leq c_0 n^{q_2} \sum_{i=1}^{\min(n,p_n)} i^{-q_1} \\
&\approx c_0 c_1 n^{q_2},
\end{aligned}$$

and

$$\begin{aligned}
\psi(\boldsymbol{V})\mu_{2v}(\boldsymbol{\lambda}) &= \mathrm{Tr}\left\{\left[\boldsymbol{D}\left(\boldsymbol{D}^t\boldsymbol{D} + \sum_{j=1}^{k}\lambda_j \boldsymbol{Q}\boldsymbol{P}_j\boldsymbol{Q}^t\right)^{-1}\boldsymbol{D}^t\right]^2\right\} \\
&= \sum_{i=1}^{\min(n,p_n)}\sum_{j=1}^{\min(n,p_n)} h_i h_j s_{ij}^2 \\
&\geq \sum_{i=1}^{\min(n,p_n)} h_i^2 s_{ii}^2 \\
&\geq \sum_{i=1}^{\min(n,p_n)} \left(\frac{h_i}{h_1 + \sum_{j=1}^{k}\lambda_j h^*_{j,1}}\right)^2 \\
&\approx c_2 \sum_{i=1}^{\min(n,p_n)} i^{-2q_1} \\
&\approx c_2 c_3,
\end{aligned}$$

where c_0, c_1, c_2 and c_3 are some constants. Thus, both $\mu_{1v}(\boldsymbol{\lambda})$ and $0 \leq \mu_{1v}^2(\boldsymbol{\lambda})/\mu_{2v}(\boldsymbol{\lambda}) \leq c_0^2 c_1^2 n^{2q_2}/[c_2 c_3 \psi(\boldsymbol{V})]$ tend to 0 as n tends to infinity. Further, both $e(\boldsymbol{\lambda})$ and $r(\boldsymbol{\lambda})$ tend to 0. Therefore, $W(\boldsymbol{\lambda})$ is nearly an unbiased estimator of $WP(\boldsymbol{\lambda})$ implied by Theorem 1. By differentiating $\psi(\boldsymbol{V})b^2(\boldsymbol{\lambda})$ and $\psi(\boldsymbol{V})\mu_{2v}(\boldsymbol{\lambda})$,

$$\left[\frac{\psi(\boldsymbol{V})\partial b^2(\boldsymbol{\lambda})}{\partial\lambda_j}\right]_{\boldsymbol{\lambda}=0} = -2\boldsymbol{f}^t\left[\boldsymbol{I} - \boldsymbol{H}_v(0)\right]^t \boldsymbol{V}\left[\frac{\partial \boldsymbol{H}_v(\boldsymbol{\lambda})}{\partial\lambda_j}\right]_{\boldsymbol{\lambda}=0}\boldsymbol{f} = 0,$$

and $\psi(\boldsymbol{V})\partial\mu_{2v}(\boldsymbol{\lambda})/\partial\lambda_j < 0$. Thus, $E[\mathrm{WT}(\boldsymbol{\lambda})]$ has strictly negative gradient at $\boldsymbol{\lambda} = 0$ and

$$\lim_{\boldsymbol{\lambda}\to\infty} E\left[\mathrm{WT}(\boldsymbol{\lambda})\right] = [1/\psi(\boldsymbol{V})]\sum_{i=1}^{n}\tilde{f}_i^2 > 0,$$

where $\boldsymbol{V}^{1/2}\boldsymbol{f} = \tilde{\boldsymbol{f}} = (\tilde{f}_1, \ldots, \tilde{f}_n)^t$. These implies that $E[\mathrm{WT}(\boldsymbol{\lambda})]$ has at least one minimizer $\boldsymbol{\lambda}_n^* > 0$. Finally, if $1_v - 1 = O(n^{-q})$ and $q \geq 1/2$, Corollary 1 holds.

Acknowledgments The author would like to thank two referees and the editor for helpful suggestions that led to a substantial improvement in the paper. This research is partly supported by Taiwan NSC Grant (Project: NSC 95-2118-M-029-005).

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics, 22*, 203–217.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

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

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

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

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. New York: Dekker.

Golub, G. H., Van Loan, C. F. (1993). *Matrix computations*. Baltimore and London: The Johns Hopkins University Press.

Green, P. J., Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman and Hall.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin: Springer.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32*, 1–49.

Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics, 15*, 958–975.

Opsomer, J., Wang, Y., Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science, 16*, 134–153.

O'Sullivan, F., Yandell, B., Raynor, W. (1986). Automatic smoothing of regression function in generalized linear models. *Journal of the American Statistical Association, 81*, 96–103.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics, 12*, 1215–1230.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika, 68*, 45–54.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association, 93*, 341–348.

Wei, W. H. (2005). The smoothing parameter, confidence interval and robustness for smoothing splines. *Journal of Nonparametric Statistics, 17*, 613–642.