Nonparametric Statistics
Vol. 16(1–2), February–April 2004, pp. 187–195

Taylor & Francis
Taylor & Francis Group

RESIDUAL INFORMATION CRITERION FOR SINGLE-INDEX MODEL SELECTIONS

P. A. NAIK* and CHIH-LING TSAI†

Graduate School of Management, University of California, Davis, California 95616, USA

(Received 27 September 2002; Revised 28 April 2003; In final form 30 May 2003)

We develop a residual information criterion (RIC) for single-index models using the residual log-likelihood approach. The proposed criterion selects both the smoothing parameter and explanatory variables. Thus, it is a general selection criterion that provides a unified approach to model selection across both parametric and nonparametric functions. Monte Carlo studies demonstrate that RIC performs satisfactorily except when the sample size is small and the signal-to-noise ratio is weak. An application of RIC is illustrated for marketing a new medical technology.

Keywords: Local polynomial regression; Residual likelihood; Sliced inverse regression; Variable and smoothing estimator selections

1 INTRODUCTION

In data analysis, linear regression models have been widely used to study the relationship between the response variable, y, and a vector of explanatory variables, x. In practice, however, the functional form that links the response variable to the set of explanatory variables may not be known. In such situations, the single-index model (see Horowitz, 1998) provides an approach to incorporate unknown, potentially nonlinear, relationship between the response variable and the index variable $x'\beta$. Specifically, a link function g relates the expected response $E(y)$ to the index variable $x'\beta$ in the single-index model: $E(y) = g(x'\beta)$. The linear regression model is its special case when g is an identity function; nonparametric regression model also is a special case when x contains one variable with $\beta = 1$.

In linear as well as nonparametric regression models, various model selection criteria have been proposed and studied over the last three decades. In the context of variable selections, the selection criterion can be either efficient (*e.g.*, Akaike information criterion (AIC), Akaike, 1973) or consistent (Bayesian information criterion (BIC), Schwarz, 1978). There is no general agreement on whether efficiency or consistency is preferred (see Burnham and Anderson, 2002, Ch. 6; McQuarrie and Tsai, 1998, Ch. 2 for detailed discussions). Recently, Naik and Tsai (2001) extended the applicability of Hurvich and Tsai's (1989) efficient criterion AIC_C from linear regression models to single-index models. However, a consistent criterion for single-index models is not available. More recently, Shi and Tsai (2002) applied residual likelihood

* Corresponding author. E-mail: cltsai@ucdavis.edu
† E-mail: panaik@ucdavis.edu

ISSN 1048-5252 print; ISSN 1029-0311 online
DOI: 10.1080/10485250310001624800

approach to obtain a consistent criterion *i.e.*, the residual information criterion (RIC) for linear regression models. Complementing Naik and Tsai (2001) and extending Shi and Tsai (2002), we derive RIC for single-index model and investigate its performance. The resulting criterion simultaneously chooses relevant variables and the smoothing parameter for unknown link functions.

The rest of this paper is organized as follows. In Section 2, we introduce the residual likelihood function and derive RIC for single-index models. Section 3 presents Monte Carlo results that show RIC performs well in most situations, except when the sample size is small or the Signal-to-noise ratio (SNR) is weak. In this case, the use of RIC may result in underfitting. Section 4 applies RIC to test a new concept of medical technology for market introduction. Finally, in Section 5, we conclude by suggesting possible avenues for further research.

2 DERIVATION OF RIC

2.1 Model Structures

Consider the collection of candidate models

$$Y = g(X\beta) + e, \tag{2.1}$$

where $Y = (y_1, \ldots, y_n)'$, $X = (x_1, \ldots, x_n)'$ is an $n \times p$ matrix of random regressor values, x_i and β are $p \times 1$ vectors, $g(X\beta)$ is an $n \times 1$ vector with ith component $g(x_i'\beta)$ $(i = 1, \ldots, n)$, e for given $X = x$ is distributed as $N(0, \sigma^2 I_n)$, and σ is an unknown scalar. Furthermore, we assume that g is a differentiable function and $\|\beta\| = 1$ for identification (see Carroll *et al.*, 1997). The log-likelihood function for model (2.1), omitting irrelevant terms, is

$$L\{(\beta, \sigma^2); Y\} = -\frac{n}{2}\log(\sigma^2) - \frac{\{Y - g(X\beta)\}'\{Y - g(X\beta)\}}{2\sigma^2}. \tag{2.2}$$

For the given g, the least squares estimator of β, $\hat{\beta}$, is a $\sqrt{n}$-consistent estimator. Thus, $g(X\hat{\beta}) = g(X\beta) + V(\hat{\beta} - \beta) + o_p(1/\sqrt{n}) = g(X\beta) + H_p e + o_p(1/\sqrt{n})$, where $H_p = V(V'V)^{-1}V'$, $V = \partial g(X\beta)/\partial\beta = \dot{g}(X\beta)X$, and $\dot{g}$ is the derivative of g. Thus, the residual for the given g is $\hat{e} = Y - g(X\hat{\beta}) = (I - H_p)e + o_p(1/\sqrt{n})$. We next apply the residual likelihood approach (Verbyla, 1993), which is the same as the marginal likelihood (McCullagh and Nelder, 1989, Ch. 7) or the restricted likelihood (Diggle *et al.*, 1994, Ch. 4), to obtain the residual log-likelihood for candidate models:

$$L\{\sigma^2; \hat{e}\} = -\frac{(n-p)}{2}\log(\sigma^2) - \frac{\log|V'V|}{2} - \frac{e'(I - H_p)e}{2\sigma^2} + o_p\left(\frac{1}{n}\right). \tag{2.3}$$

If the link function g is an identity function, then ignoring the remainder term from Eq. (2.3) yields the residual log-likelihood function of the linear regression model given by Verbyla (1990). To derive RIC, we consider it as the residual log-likelihood function for the single-index model.

Suppose that data Y are generated from a model, which constitutes the nearest representation of the true situation. Adopting Linhart and Zucchini's (1986) terminology, we call such a model the operating model:

$$Y = g_0(X_0\beta_0) + \varepsilon,$$

where $X_0 = (x_{10}, \ldots, x_{n0})'$ is an $n \times p_0$ matrix of random regressor values, x_{i0} and β_0 are $p_0 \times 1$ vectors, $g_0(X_0\beta_0)$ is an unknown $n \times 1$ vector with ith component $g_0(x_{i0}'\beta_0)$ $(i = 1, \ldots, n)$, ε for given $X_0 = x_0$ is distributed as $N(0, \sigma_0^2 I_n)$, and σ_0 is an unknown scalar. In addition, we assume that g_0 is a differentiable function and $\|\beta_0\| = 1$ for identification. Then the residual log-likelihood for the operating model is $L_0\{(g_0, \sigma_0^2); \hat{\varepsilon}\}$, which has the same form as Eq. (2.3) except that $p, g, \hat{e}, \sigma^2, V$, and H_p in Eq. (2.3) are replaced by $p_0, g_0, \hat{\varepsilon}, \sigma_0^2, V_0$, and $H_{p0} = V_0(V_0'V_0)^{-1}V_0'$, respectively, where $\hat{\varepsilon} = Y - g_0(X_0\hat{\beta}_0)$, $\hat{\beta}_0$ is the estimator of β_0, and $V_0 = \partial g_0(X_0\beta_0)/\partial\beta_0$. Using the residual log-likelihood function, we next obtain a new model selection criterion.

2.2 RIC Criterion for Single-Index Models

A useful measure of the discrepancy between the residual log-likelihood functions of candidate and operating models is the Kullback-Leibler information. Omitting terms that are not functions of the candidate model (Linhart and Zucchini, 1986), we obtain the twice of the Kullback-Leibler information metric:

$$\begin{aligned}
\delta &= E_0[-2L\{\sigma^2; \hat{e}\}] \\
&= (n-p)\log(\sigma^2) + \log|V'V| + E_0\frac{e'(I-H_p)e}{\sigma^2} \\
&= (n-p)\log(\sigma^2) + \log|V'V| + (n-p)\frac{\sigma_0^2}{\sigma^2} \\
&\quad + \{g_0(X_0\beta_0) - g(X\beta)\}'(I-H_p)\frac{g_0(X_0\beta_0) - g(X\beta)}{\sigma^2}, \qquad (2.4)
\end{aligned}$$

where E_0 denotes the expectation under the operating model.

To assess the quality of the above Kullback-Leibler information in the light of data, we need to find parameter estimators of (β, g, σ^2). For the unknown g, Billinger (1983) showed that the ordinary least squares (OLS) estimator is a $\sqrt{n}$-consistent estimator of β up to a constant of proportionality. Later, Duan and Li (1991) developed the sliced inverse regression (SIR) estimator, and showed that the SIR estimator is $\sqrt{n}$-consistent estimator and it usually has a smaller variance than the OLS estimator. In addition, it is easy to compute because it does not require iterative computation even though the link function g is not known. Hence, in this paper, we apply SIR rather than OLS. (For other estimators of single-index models, see Horowitz, 1998, Carroll *et al.*, 1997, and Hristache *et al.*, 2001.) Based on the SIR $\hat{\beta}$, we first construct the index $t = X\hat{\beta}$, and then apply nonparametric regression smoother to estimate $g(t)$ (see Fan and Gijbels, 1996; Simonoff, 1996). Finally, the estimator of σ^2 is the residual sum of squares divided by the degrees of freedom.

Replacing (β, g, σ^2) in Eq. (2.4) with $(\hat{\beta}, \hat{g}, \hat{\sigma}^2)$, respectively, we obtain

$$\begin{aligned}
\hat{\delta} &= (n-p)\log(\hat{\sigma}^2) + \log|\hat{V}'\hat{V}| + (n-p)\frac{\sigma_0^2}{\hat{\sigma}^2} \\
&\quad + \{g_0(X_0\beta_0) - \hat{g}(X\hat{\beta})\}'(I-\hat{H}_p)\frac{g_0(X_0\beta_0) - \hat{g}(X\hat{\beta})}{\hat{\sigma}^2},
\end{aligned}$$

where $\hat{V}$ is V evaluated at $\beta = \hat{\beta}$ and $g = \hat{g}$, $\hat{H}_p = \hat{V}(\hat{V}'\hat{V})^{-1}\hat{V}'$, $\hat{\sigma}^2 = \{Y - \hat{g}(X\hat{\beta})\}'\{Y - \hat{g}(X\hat{\beta})\}/(n-\hat{m})$, $\hat{m} = tr\{(I-\hat{H}_p)(I-\hat{H}_{np})\}$, and $\hat{H}_{np}Y = \hat{g}(X\hat{\beta})$. To assess the quality of

a candidate model with respect to data, we compute the expectation of $\hat{\delta}$:

$$\begin{aligned}\Delta &= E_0\{\hat{\delta}\} \\ &= E_0\{(n-p)\log(\hat{\sigma}^2)\} + E_0(\log|\hat{V}'\hat{V}|) + (n-p)\sigma_0^2 E_0\left(\frac{1}{\hat{\sigma}^2}\right) \\ &\quad + E_0\left[\{g_0(X_0\beta_0) - \hat{g}(X\hat{\beta})\}'(I-\hat{H}_p)\frac{g_0(X_0\beta_0)-\hat{g}(X\hat{\beta})}{\hat{\sigma}^2}\right]. \end{aligned} \tag{2.5}$$

Given the collection of competing candidate models, we can then select the model that results in the smallest Δ.

In order to compute Δ, we adapt Naik and Tsai's (2001) four assumptions which are: (1) the parametric component of a candidate model includes the parametric component of the operating model; that is the columns of X can be rearranged so that $X_0\beta_0 = X\beta^*$, where $\beta^* = (\beta_0', \beta_1')'$, and β_1 is a $(p-p_0)\times 1$ vector of zeros; (2) there exists a smoother matrix $\tilde{H}_{np}$ so that $\tilde{g}(X\beta^*) = \tilde{H}_{np}Y$. That is, $\tilde{g}$ is the projection of Y through the hat matrix $\tilde{H}_{np}$; (3) $E_0\{\tilde{g}(X\beta^*)\} = g_0(X\beta^*)$; (4) $\hat{g}(X\hat{\beta}) - \tilde{g}(X\beta^*) = \tilde{V}(\hat{\beta}-\beta^*) + o_p(1/\sqrt{n}) = \tilde{H}_p(Y-\tilde{g}(X\beta^*)) + o_p(1/\sqrt{n})$, where $\tilde{H}_p = \tilde{V}(\tilde{V}'\tilde{V})^{-1}\tilde{V}'$, $\tilde{V} = \partial\tilde{g}(X\beta)/\partial\beta|_{\beta=\beta^*} = \tilde{g}_\cdot(X\beta^*)X$, and $\tilde{g}_\cdot$ is the derivative of $\tilde{g}$. The first assumptions was made in the derivation of AIC for parametric model (see Akaike, 1973; Linhart and Zucchini, 1986, p. 245). The second and the third assumptions indicate that the estimator is a linear function of Y and an unbiased estimator, respectively. Assumption 2 is often considered in the context of nonparametric regressions. Assumption 3 was made by Cleveland and Delvin (1988) even though it may not hold in practice. Assumption 4 implies that the estimator of the link function can be replaced by its first order Taylor approximation. As in Hurvich *et al*. (1998) and Naik and Tsai (2001), the above assumptions are made only to facilitate the derivation of a selection criterion whose performance is satisfactory in finite samples (see Section 3).

Under Assumptions (1)–(4), we have $g_0(X\beta^*) - \tilde{g}(X\beta^*) = -\tilde{H}_{np}\varepsilon$ and $\tilde{g}(X\beta^*) - \hat{g}(X\hat{\beta}) \approx -\tilde{H}_p\{\varepsilon + g_0(X\beta^*) - \tilde{g}(X\beta^*)\} = -(\tilde{H}_p - \tilde{H}_p\tilde{H}_{np})\varepsilon$. Hence, $g_0(X\beta^*) - \hat{g}(X\hat{\beta}) \approx -(\tilde{H}_p + \tilde{H}_{np} - \tilde{H}_p\tilde{H}_{np})\varepsilon$, and $Y - \hat{g}(X\hat{\beta}) \approx (I - \tilde{H}_p - \tilde{H}_{np} + \tilde{H}_p\tilde{H}_{np})\varepsilon$. Furthermore, we approximate $\hat{H}_p$ and $\hat{H}_{np}$ with $\tilde{H}_p$ and $\tilde{H}_{np}$, respectively. Then, Δ in Eq. (2.5) can be approximated by

$$\begin{aligned}\tilde{\Delta} &= E_0\{(n-p)\log(\hat{\sigma}^2)\} + E_0(\log|\hat{V}'\hat{V}|) \\ &\quad + (n-p)\tilde{m}\sigma_0^2 E_0\left\{\frac{1}{\varepsilon'(I-\tilde{H}_p-\tilde{H}_{np}+\tilde{H}_p\tilde{H}_{np})'(I-\tilde{H}_p-\tilde{H}_{np}+\tilde{H}_p\tilde{H}_{np})\varepsilon}\right\} \\ &\quad + \tilde{m}E_0\left\{\frac{\varepsilon'(\tilde{H}_p+\tilde{H}_{np}\tilde{H}_p\tilde{H}_{np})'(I-\tilde{H}_p)(\tilde{H}_p+\tilde{H}_{np}-\tilde{H}_p\tilde{H}_{np})\varepsilon}{\varepsilon'(I-\tilde{H}_p-\tilde{H}_{np}+\tilde{H}_p\tilde{H}_{np})'(I-\tilde{H}_p-\tilde{H}_{np}+\tilde{H}_p\tilde{H}_{np})\varepsilon}\right\}, \end{aligned}$$

where $\tilde{m} = \mathrm{tr}\{(I-\tilde{H}_p)(I-\tilde{H}_{np})\}$. Approximating $\log|\hat{V}'\hat{V}|$ with $p\log(n)$, and upon algebraic simplification, we get

$$\begin{aligned}\tilde{\Delta} &\approx E_0\{(n-p)\log(\hat{\sigma}^2)\} + p\log(n) + (n-p)\tilde{m}\sigma_0^2 E_0\left\{\frac{1}{\varepsilon'(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\varepsilon}\right\} \\ &\quad + \tilde{m}E_0\left\{\frac{\varepsilon'\tilde{H}_{np}'(I-\tilde{H}_p)\tilde{H}_{np}\varepsilon}{\varepsilon'(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\varepsilon}\right\}. \end{aligned} \tag{2.6}$$

Next, applying the method proposed by Cleveland and Delvin's (1988) and Hastie and Tibshirani (1990), the distributions of $\varepsilon'(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\varepsilon$ and $\{\varepsilon'\tilde{H}'_{np}(I-\tilde{H}_p)\tilde{H}_{np}\varepsilon\}/\{\varepsilon'(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\varepsilon\}$ can be approximated by $(\delta_2/\delta_1)\chi^2_{\delta_1^2/\delta_2}$ and $(\nu_1/\delta_1)F_{\nu_1^2/\nu_2,\delta_1^2/\delta_2}$, respectively, where $\delta_1=\mathrm{tr}\{(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\}$, $\delta_2=\mathrm{tr}[\{(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\}^2]$, $\nu_1=\mathrm{tr}\{\tilde{H}'_{np}(I-\tilde{H}_p)\tilde{H}_{np}\}$ and $\nu_2=\mathrm{tr}[\{\tilde{H}'_{np}(I-\tilde{H}_p)\tilde{H}_{np}\}^2]$. Hence,

$$E_0\left\{\frac{1}{\varepsilon'(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\varepsilon}\right\}\approx\left(\frac{\delta_1}{\delta_2}\right)\left(\frac{\delta_1^2}{\delta_2}-2\right),$$

and

$$E_0\left\{\frac{\varepsilon'\tilde{H}'_{np}(I-\tilde{H}_p)\tilde{H}_{np}\varepsilon}{\varepsilon'(I-\tilde{H}_{np})'(I-\tilde{H}_p)(I-\tilde{H}_{np})\varepsilon}\right\}\approx\left\{\frac{\nu_1(\delta_1/\delta_2)}{(\delta_1^2/\delta_2-2)}\right\}.$$

Substituting the above expressions in Eq. (2.6), we propose an estimator of Δ:

$$\mathrm{RIC}^*=(n-p)\log(\hat{\sigma}^2)+p\log(n)+\hat{m}\left\{\frac{(\hat{\delta}_1/\hat{\delta}_2)(n-p+\hat{\nu}_1)}{\hat{\delta}_1^2/\hat{\delta}_2-2}\right\},$$

where $\hat{m}$, $\hat{\delta}_1$, $\hat{\delta}_2$ and $\hat{\nu}_1$ are $\tilde{m}$, δ_1, δ_2 and ν_1 evaluated at $\beta=\hat{\beta}$ and $g=\hat{g}$. We further simplify RIC* so that it is easy to compute. Specifically, we adapt Hurvich *et al.*'s (1998) approach to approximate $\hat{\delta}_2$ by $\hat{\delta}_1$, and then replace $\hat{\delta}_1$ and $\hat{\nu}_1$ by $\hat{m}$ and $\mathrm{tr}\{(I-\hat{H}_p)\hat{H}_{np}\}$, respectively. Consequently, we obtain the RIC:

$$\mathrm{RIC}=(n-p)\log(\hat{\sigma}^2)+p\log(n)+\frac{\hat{m}[\mathrm{tr}\{(I-\hat{H}_p)\hat{H}_{np}\}+(n-p)]}{\hat{m}-2}. \tag{2.7}$$

The accuracy of approximation of RIC to RIC* was examined in Monte Carlo simulations (not reported here), and was found to be excellent.

The RIC criterion unifies model selection across both parametric and nonparametric functions. For example, in parametric regression models, $\tilde{H}_{np}=0$ and $\mathrm{tr}(\tilde{H}_p)=p$. Hence, after subtracting the constant $n+2$, Eq. (2.7) results in

$$\mathrm{RIC}=(n-p)\log(\hat{\sigma}^2)+p\log(n)-p+\frac{4}{(n-p-2)},$$

which is the RIC of Shi and Tsai (2002).

3 SIMULATIONS

In this section, we use Monte Carlo simulations to investigate the performance of RIC as a function of sample size, SNR, and shape of the link function. For the sake of brevity, we report the results for the following settings even though we conducted extensive simulation

TABLE I Frequency of Number of Variables Selected by RIC in 1000 Repetitions for the Operating Model $\exp(-X_0'\beta_0)$ with $p_0 = 5$.

		Variables, p									
n	*SNR*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
30	3	226	58	45	86	379	65	30	24	14	73
	5	109	24	23	74	652	67	23	9	8	11
	8	33	12	14	39	827	44	18	4	3	6
50	3	136	38	34	76	548	76	24	13	17	38
	5	28	13	16	46	803	48	23	14	6	3
	8	6	6	5	19	888	56	11	5	3	1

studies. Here, we consider the sample sizes $n = 30$ and 50. True link functions are $g_0(X_0\beta_0) = \exp(-X_0\beta_0)$ and $\sin(X_0\beta_0)$, where $\beta_0 = (1/\sqrt{55})(1, 2, 3, 4, 5)'$, X_0 is an $n \times 5$ matrix, and the ith row of X_0, $(x_{i1}, \ldots, x_{i5})$, contains five independent standard normal random variables. The explanatory variables of the candidate single-index models are stored in the $n \times 10$ matrix X containing independent standard normal random variables in a nested fashion. In other words, columns 1 to p, $p = 1, \ldots, 10$, define the matrix of explanatory variables for the candidate single-index model with p regressors. The true single-index model contains the explanatory variables corresponding to the first five columns of X. We take $\varepsilon \sim N(0, \sigma_0^2)$, and $\text{SNR} = R_y/\sigma_0^2 = 3, 5$ and 8, where R_y is the range of $g_0(X_0\beta_0)$. We perform 1000 replications for each of the settings described above. In each realization, we apply sliced inverse regression (SIR) and local polynomial regression to estimate β and g, respectively.

Table I presents the frequency of model selection by RIC when the true link function is $\exp(-X_0\beta_0)$ which exhibits a decreasing trend. When the SNR is weak (SNR = 3), the correct model is not easily discernible, and RIC tends to underfit severely. This phenomenon holds even when $n = 50$, albeit not so severely. As SNR gets larger, the model estimation improves, and so the extent of underfitting decreases. In addition, RICs large penalty function prevents overfitting, especially when the SNR increases. This finding can be seen more clearly in Table II where the true link function is a non-monotonic function, $\sin(X_0\beta_0)$. Indeed, RIC performs quite well when SNR = 8. In conclusion, the performance of RIC improves as the sample size increases or the SNR gets strong. A similar finding has been noticed in linear regression model selection (see Shi and Tsai, 2002).

In addition to variable selections, we compute the average of the normed SIR estimates $\hat{\beta}_{\text{SIR}}$ and their standard deviations. We find that $\hat{\beta}_{\text{SIR}}$ provides a good estimate of β. As n or SNR increases, both the accuracy and precision of $\hat{\beta}_{\text{SIR}}$ increase. Since this finding is consistent with Naik and Tsai (2001), we do not present our results here. Instead, we compute the average

TABLE II Frequency of Number of Variables Selected by RIC in 1000 Repetitions for the Operating Model $\sin(X_0'\beta_0)$ with $p_0 = 5$.

		Variables, p									
n	*SNR*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
30	3	349	35	39	64	497	15	0	0	1	0
	5	96	7	3	22	869	3	0	0	0	0
	8	22	3	0	8	960	6	1	0	0	0
50	3	94	8	9	31	842	15	0	1	0	0
	5	5	0	0	0	991	3	1	0	0	0
	8	1	0	0	0	996	3	0	0	0	0

TABLE III Bandwidth Estimates When the Correct Model is Chosen.

	$exp(-X_0'\beta_0)$			$sin(X_0'\beta_0)$		
SNR	*3*	*5*	*8*	*3*	*5*	*8*
$n = 30$	1.715	1.107	0.654	1.776	1.268	0.854
$n = 50$	1.552	0.865	0.452	1.405	0.763	0.443

TABLE IV The Average ASE in 1000 Repetitions.

	$exp(-X_0'\beta_0)$			$sin(X_0'\beta_0)$		
SNR	*3*	*5*	*8*	*3*	*5*	*8*
$n = 30$	3.281	1.966	1.343	0.098	0.053	0.030
$n = 50$	2.687	1.765	1.143	0.065	0.030	0.017

smoothing parameter estimate $\hat{h}$ when the correct model is chosen. Table III shows that $\hat{h}$ becomes larger when n or SNR gets smaller. Thus, RIC tends to oversmooth as sample size or SNR decreases.

To further illustrate the performance of RIC on the smoothing parameter selection, we use the same simulation settings given above except that we only consider the candidate model with $p = 5$. Let h_{RIC} be the smoothing parameter chosen by RIC, and the average squared errors (ASE), $\text{ASE} = \sum_{i=1}^{n}\{\hat{g}_{h_{\text{RIC}}}(x_i'\hat{\beta}_{\text{SIR}}) - g_0(x_i'\beta_0)\}^2/n$. Table IV shows that the ASE decreases as the SNR or the sample size increases. This implies that the performance of RIC in smoothing parameter selection improves as n or SNR gets large. Thus, RIC can be used to select both the smoothing parameter and relevant regressors when the sample size is large and/or the signal is strong.

4 EMPIRICAL EXAMPLE

We illustrate the application of RIC to test a new concept of medical technology for market introduction. A professional market research company conducted 'concept tests' to gauge whether the market would adopt a new medical technology. This market research study enables the company to determine the important attributes of the new technology (see Dolan, 1993, p. 83 for details). In this proprietary study, 46 respondents indicated their purchase intention on a 5-point scale, where 5 represents 'very likely to adopt', and 1 denotes 'not at all likely.' Using 10-point scales, where 10 indicates 'strongly agree' and 1 means 'strongly disagree,' the respondents stated their extent of agreement on the importance of nineteen attributes of the new concept for medical technology.

The goal of concept testing is to determine which key attributes drive respondents' purchase intention. To this end, typically, linear regression is applied to concept-test data, where purchase intention serves as the dependent variable and the set of attributes constitutes the independent variables. When we apply linear regression to the above proprietary data, we find that adjusted R^2 is 8.95%, and that only one of the 19 attributes (namely, micronised particles) is significant using the t-test at the 5% significance level. When we apply SIR to estimate the single-index model, purchase intention $= g$(set of 19 attributes), we do not pre-specify functional forms for $g(\cdot)$, thus relaxing the linearity assumption inherent in linear regression models. Using Chen and Li's (1998) t-test, we then discover two more important attributes (namely, very

large molecules, and variation of biologicals) in addition to the micronised particles. Thus, SIR detects the effects of key attributes that the linear regression could not; this is because SIR estimator is more efficient than the OLS (Duan and Li, 1991).

When we apply the RIC criterion for this single-index model, we attain the minimum RIC value of 86.47, the bandwidth is 1.4, and the resulting adjusted R^2 equals 29.07%. More importantly, we find that the retained set of important variables includes two more attributes (namely, nano-particles and early drug development) in addition to the above three attributes. In other words, RIC retains 5 variables as the key attributes to predict purchase intention. Thus, by augmenting the set of important variables, RIC not only assists the new product team to identify specific attributes that drive purchase intention, but also prevents the premature elimination of important attributes that would otherwise result from the use of t-tests for linear or inverse regressions.

5 CONCLUSION

In this paper, we applied residual log-likelihood approach to obtain the selection criterion RIC for jointly selecting nonparametric smoothing parameter and relevant variables in single-index models. We can extend this approach to three research areas. The first is to derive RIC for partially linear single-index models (Carroll *et al*., 1997) and multiple-index models (Ichimura and Lee, 1991). The second is to obtain RIC for single-index models with general covariances that include the weighted or autocorrelated structures. The third is to study the efficacy of RIC with alternative parametric parameter estimators, *e.g.*, the OLS estimator, Brillinger (1983), the weighted average derivative estimator, Powell *et al*. (1989), and the maximum quasi-likelihood estimator, Carroll *et al*. (1997) as well as nonparametric smoothing parameter estimators (*e.g.*, cubic spline smooth estimator, quadratic local polynomial estimator, and kernel estimator). We believe that such efforts would yield better methods for nonparametric and semiparametric data analysis.

Acknowledgement

We are grateful to the referee for helpful comments. Chih-Ling Tsai's research was supported in part by National Institutes of Health grant DA-01-0433.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Csaki, F. (Eds.), *2nd International Symposium on Information Theory*. Akademia Kiado, Budapest, pp. 267–281.

Brillinger, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In: Bickel, P. J., Doksum, K. A. and Hodges, J. L. (Eds.), *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday*. Wadsworth, Belmont, pp. 97–114.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*, 2nd ed. Springer, New York.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Am. Statist. Assoc.*, **92**, 477–489.

Chen, C. H. and Li, K. C. (1998). Can sir be a popular as multiple linear regression? *Statistica Sinica*, **8**, 289–316.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.*, **83**, 596–610.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press, New York.

Dolan, R. J. (1993). *Managing the New Product Development Process*. Addison Wesley, New York.

Duan, N. and Li, K. C. (1991). Slicing regression: A link free regression method. *Ann. Statist.*, **19**, 505–530.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, New York.

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

Horowitz, J. L. (1998). *Semiparametric Methods in Econometrics*. Springer, New York.
Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. of Statist.*, **29**, 595–623.
Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.
Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc. B*, **60**, 271–293.
Ichimura, H. and Lee, L. F. (1991). Semiparametric least squares estimation of multiple index models: Single equation estimation. In: Barnett, W. A., Powell, J. and Tauchen, G. (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press, Cambridge, pp. 3–49.
Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley, New York.
McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, New York.
McQuarrie, A. D. R. and Tsai, C. L. (1998). *Regression and Time Series Model Selection*. World Scientific Publishing, Singapore.
Naik, P. A. and Tsai, C. L. (2001). Single-index model selections. *Biometrika*, **88**, 821–832.
Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403–1430.
Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Shi, P. and Tsai, C. L. (2002). Regression model selection–a residual likelihood approach. *J. R. Statist. Soc.*, **64**, 237–252.
Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.
Verbyla, A. P. (1993). Modelling variance heterogeneity: Residual maximum likelihood and diagnostics. *J. R. Statist. Soc.*, **55**, 493–508.
Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Aust. J. Statist.*, **32**, 227–230.