Electronic Journal of Statistics
Vol. 7 (2013) 742–762
ISSN: 1935-7524
DOI: 10.1214/13-EJS788

Generalized predictive information criteria for the analysis of feature events

Mike K. P. So

Department of Information Systems,
Business Statistics and Operations Management,
The Hong Kong University of Science and Technology,
Kowloon, Hong Kong
e-mail: immkpso@ust.hk

and

Tomohiro Ando

Graduate School of Business Administration,
Keio University,
Kanagawa 223-8526, Japan
e-mail: andoh@kbs.keio.ac.jp

Abstract: This paper develops two weighted measures for model selection by generalizing the Kullback-Leibler divergence measure. The concept of a model selection process that takes into account the special features of the underlying model is introduced using weighted measures. New information criteria are defined using the bias correction of an expected weighted loglikelihood estimator. Using weight functions that match the features of interest in the underlying statistical models, the new information criteria are applied to simulated studies of spline regression and copula model selection. Real data applications are also given for predicting the incidence of disease and for quantile modeling of environmental data.

Keywords and phrases: Feature matching, information criteria, model selection, weighted Kullback-Leibler measure.

Received August 2012.

1. Introduction

An information theoretic approach (ref. [2]) for model selection expresses a statistical model in the form of a probability distribution. The model is evaluated using an estimate of the Kullback-Leibler information (ref. [16]) as an overall measure of the divergence of the fitted model from the true model, which is generating the data. According to [2], if the specified model contains the true model and the model is estimated by the maximum likelihood method, then Akaike's information criterion (AIC) can be used to evaluate the constructed models. Previous studies have developed information criteria, as estimators of the expected log-likelihood, under model mis-specification (e.g. [22]), and for evaluating the mis-specified models constructed by various types of estimation procedures (e.g. [14, 20]). [21, 3, 4] extend the information criteria approach to the Bayesian paradigm. These studies focus on the overall fitness of the constructed model to the true underlying structure. This paper proposes methods, which will allow model selection to emphasize some features of the true model.

When selecting a model, we usually have well-defined purposes to guide our selection of the class of models. For example, we may want to identify distributions that can assess the probability of the occurrence of rare events. In a bivariate setting, we may select a good copula model to capture the tail dependence of two random variables. In medical disease prognosis, it is important to develop statistical models that can accurately predict the incidence of a disease. A common characteristic of the above examples is that prior to model selection some key features of the true model have been identified. From both the practical and the statistical points of view, it is desirable for model selection schemes to incorporate relevant features into the selection procedure. A feature capturing concept was proposed by [23] who emphasized feature matching in time series modeling. They suggested doing model estimation by treating the multiple-step conditional mean forecasts and autocorrelation functions as features. In this paper, we develop two information criteria which embed some pre-specified relevant features that the true model should contain. We first define a weighted Kullback-Leibler (KL) measure as follows.

Definition. Suppose G and F represent the true underlying distribution and the fitted distribution of a random variable Y, respectively. Define a positive real-valued function $w(y)$ which is bounded and does not depend on the sample size. Denote the expectations with respect to G and F by $E_G[\cdot]$ and $E_F[\cdot]$, respectively. Under the regularity condition

$$E_G[w(Y)] \geq E_F[w(Y)], \tag{1.1}$$

a weighted KL measure is defined as

$$K_w(G,F) = E_G\left[w(Y)\log\frac{g(Y)}{f(Y)}\right], \tag{1.2}$$

where $g(y)$ and $f(y)$ are probability densities or probability mass functions of G and F, respectively.

The weighted KL measure is a measure satisfying (i) $K_w(G,F) \geq 0$; and (ii) $K_w(G,F) = 0$ if and only if $G = F$. A proof is given in the Appendix. Note that an alternative weighted KL measure satisfying (i) and (ii) above is defined as

$$\widetilde{K}_w(G,F) = E_G\left[w(Y)\log\frac{g(Y)}{f(Y)}\right] - (E_G[w(Y)] - E_F[w(Y)]). \tag{1.3}$$

The weighted KL measures in Equations (1.2) and (1.3) reduce to the traditional KL measure, denoted by $K(G,F)$, when $w(y) = 1$. The main reason for the construction of the weighted KL measures is to put more/less weight on the feature region of G so that any discrepancy between G and F in the feature region will be emphasized/de-emphasized via $E_G\left[w(Y)\log\frac{g(Y)}{f(Y)}\right]$. The constraint

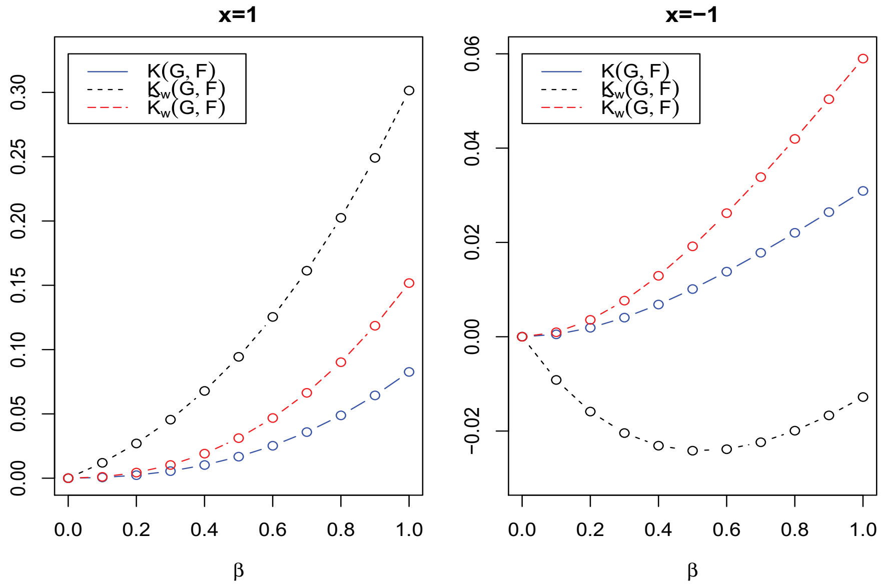

FIG 1. *The plot of $K(G, F)$, $K_w(G, F)$ and $\widetilde{K}_w(G, F)$ against β, which represents the "difference" between the true logistic model G and the constant probability model F.*

in Equation (1.1) is called the feature condition and is a core condition in constructing the weighted KL measure in Equation (1.2). We give examples below to show how the feature condition is linked to the characteristics of the true model that we want the selected model to match.

Example 1 Consider a discrete probability distribution $G(y)$, where "$y = 0$" represents a non-disease group and "$y = 1$" represents a disease group. Assume that $P_G(Y = 1|X = x) = \exp(-2+\beta x)/[1+\exp(-2+\beta x)]$, $\beta > 0$. The distribution $F(x)$ can represent a screening test detecting the disease and the fitted model is $P_F(Y = 1) = \exp(-2)/[1 + \exp(-2)] \approx 0.12$. If the main focus of the model selection is the sensitivity of the screening test, we can specify $w(0) = 1$ and $w(1) > 1$. In this case, $E_G[w(Y)] = 1+(w(1)-1)P_G(Y = 1|X = x)$ and $E_F[w(Y)] = 1+(w(1)-1)P_F(Y = 1)$. So $E_G[w(Y)] - E_F[w(Y)] = (w(1) - 1)[P_G(Y = 1|X = x) - P_F(Y = 1)]$ is greater than or equal to 0 if $P_G(Y = 1|X = x) \geq P_F(Y = 1)$ or $x \geq 0$. We plot $K(G, F)$, $K_w(G, F)$ and $\widetilde{K}_w(G, F)$ versus β for $w(1) = 2$ in Figure 1. We find that both $K_w(G, F)$ and $\widetilde{K}_w(G, F)$ increase faster than $K(G, F)$ with β if $x \geq 0$. This is understandable because when $x \geq 0$, the feature condition in Equation (1.1) is satisfied. Even though when $x < 0$ where Equation (1.1) does not hold and $K_w(G, F)$ may not be positive, $\widetilde{K}_w(G, F)$ is still greater than $K(G, F)$ for all β, indicating that $\widetilde{K}_w(G, F)$ can have higher discriminating power than $K(G, F)$ in practice.

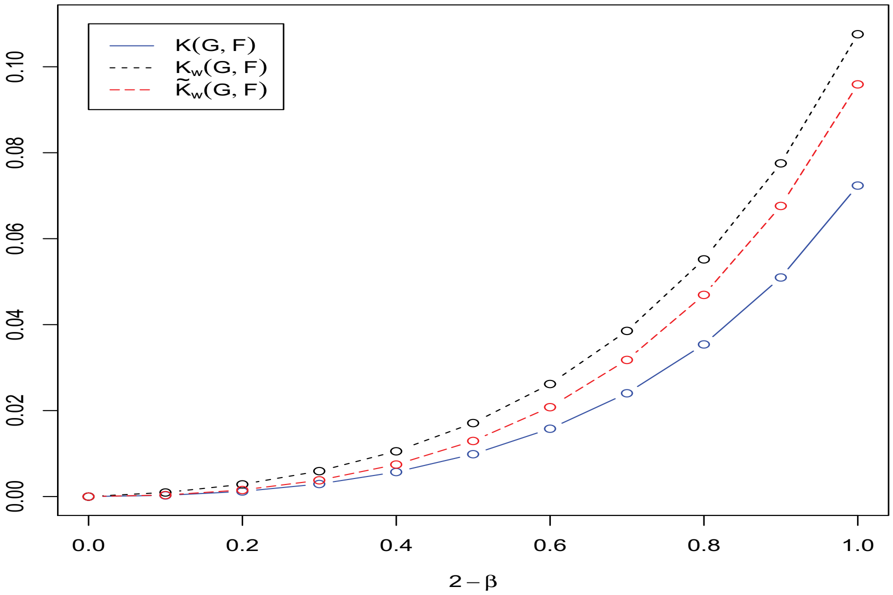

Fig 2. *The plot of $K(G,F)$, $K_w(G,F)$ and $\widetilde{K}_w(G,F)$ against $2-\beta$, which represents the 'difference" between the true exponential power distribution G and the standard normal distribution F.*

Example 2 Suppose the true model $G(y)$ is an exponential power distribution with mean 0 and variance 1. Its density is $g(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)}e^{-(|y|/\alpha)^\beta}$ with $\alpha = \sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}}$, and the fitted model is the standard normal. If the feature of interest is the tail part of the true model, we can set $w(y)$= 2 if $|y| > 3$ and $w(y) = 1$ otherwise. In this case, $E_G[w(Y)] = P_G(|Y| > 3) + 1$ and $E_F[w(Y)] = P_F(|Y| > 3) + 1$ which satisfies the feature condition as $P_G(|X| > 3) \geq P_F(|X| > 3)$. The heavy-tailed feature of the distribution is emphasized with the choice of $w(y)$ which puts higher weight on the "tail discrepancy". Figure 2 is the plot of $K(G,F)$, $K_w(G,F)$, and $\widetilde{K}_w(G,F)$ and shows that both $K_w(G,F)$ and $\widetilde{K}_w(G,F)$ are greater than $K(G,F)$ when β decreases from 2, or the exponential power distribution becomes more heavy-tailed. We expect that our weighted KL measures will give a larger discrepancy between the true and the fitted model if the fitted model cannot capture the target feature. In practice, this property of $K_w(G,F)$ and $\widetilde{K}_w(G,F)$ can help to differentiate potential models with respect to the target feature, as the divergence between the true model and the fitted "wrong" model will be amplified by suitably defining the weights to match the feature.

This paper has three objectives. First, two weighted KL measures, that allow certain characteristics of the underlying model to be incorporated into the model selection are proposed. Second, we develop two new information criteria based on the weighted KL measures to determine the best model among candidate choices. Third, we demonstrate the advantage of having the weight function $w(y)$ in the model selection. The focus is on spline regression and copula model selection. The remainder of the paper is organized as follows. In Section 2, we develop two information criteria for model selection using the weighted KL measures. Section 3 investigates the performance of the proposed information criteria on spline regression and copula model selection, using Monte Carlo simulations. In Section 4, we apply the developed approach to real data. Section 5 provides a discussion, including possibilities for future work.

2. Main result

Let $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ be a set of n observations from the true model $G(y)$. For $\alpha = 1, \ldots, n$, we express the fitted model density function of y_α as $f(y_\alpha; \theta)$, where θ is the model parameter. One can estimate the parameter θ by maximizing the penalized weighted log-likelihood:

$$\ell_w(\theta, y) - \lambda p(\theta), \tag{2.1}$$

where

$$\ell_w(\theta, y) = \frac{1}{n} \left[\sum_{\alpha=1}^{n} w(y_\alpha) \log f(y_\alpha; \theta) \right], \tag{2.2}$$

the weights $w(y_\alpha)$ can be specified by decision makers to target for specific features in $G(y)$, λ is a regularization parameter and $p(\theta)$ is a penalty function. As shown in Section 3.1, the penalty term can improve the performance in spline regression. After obtaining the parameter estimate $\widehat{\theta}$ by maximizing Equation (2.1), we produce a model by replacing the parameter θ with the sample estimate $\widehat{\theta}$. The problem is how to choose the best model among the candidates. This section constructs two information criteria for evaluating the fitted model for possible model mis-specification from a predictive point of view.

Based on the definition of the weighted measure $K_w(G, F)$, we assess the goodness of the fitted model based on the expected weighted log-likelihood given by

$$\int \ell_w(\widehat{\theta}, z) dG(z) = \int \left[\sum_{\alpha=1}^{n} w(z_\alpha) \log f(z_\alpha; \widehat{\theta}) \right] dG(z_1, \ldots, z_n), \tag{2.3}$$

where $z = (z_1, \ldots, z_n)^{\mathrm{T}}$ are replicates of the observations y_α's. Note that if we specify the weight function as $w(z_\alpha) = 1$, $\alpha = 1, 2, \ldots,$ Equation (2.3) will reduce to the expected log-likelihood extensively analyzed by [14]. We note here that the expected weighted log-likelihood depends on the unknown true distribution $G(y)$ and on the observed data $y_1, \ldots, y_n$. A natural estimator of the expected weighted log-likelihood is the sample based weighted log-likelihood

$\ell_w(\widehat{\theta}, y)$, which is formally obtained by replacing the unknown true distribution with the empirical distribution, putting probability mass $1/n$ on each observation. The weighted log-likelihood generally induces a positive bias as an estimator of the expected weighted log-likelihood, because the same data are used both to estimate the parameters of the model and to evaluate the expected log-likelihood. We therefore consider the bias correction of the log-likelihood. The bias is defined by

$$\mathrm{bias}(G) = \int \left[\ell_w(\widehat{\theta}, y) - \int \ell_w(\widehat{\theta}, z) dG(z) \right] G(y). \tag{2.4}$$

Then an estimator of the expected weighted log-likelihood is given by

$$\ell_w(\widehat{\theta}, y) - \widehat{\mathrm{bias}}(G), \tag{2.5}$$

where $\widehat{\mathrm{bias}}(G)$ is an estimator of the bias of the empirical weighted log-likelihood in estimating the expected weighted log-likelihood. The following theorem provides an expression of the bias term.

Theorem 2.1. *Suppose that the specified model does not necessarily contain the true model generating the data. Assume that $\theta \in \Theta$, where Θ is a compact set of $\Re^q$, $q < n$, for the expansion in Equation (5.1) to be valid. Then a bias term is expressed as*

$$\mathrm{bias}(G) = \frac{1}{n} \mathrm{tr} \left[\int T^{(1)}(z; G) \left. \frac{\partial \ell_w(\theta, z)}{\partial \theta^{\mathrm{T}}} \right|_{T(G)} dG(z) \right] + o(n^{-1}),$$

where $T^{(1)}(z; G)$ is the first order derivative of the functional

$$T^{(1)}(y; G) = R(G)^{-1} \left. \frac{\partial \left\{ w(z) \log f(\theta, z) - \lambda p(\theta) \right\}}{\partial \theta} \right|_{\theta = T(G)}, \tag{2.6}$$

with

$$R(G) = - \int \frac{\partial^2 \left\{ \ell_w(\theta, z) - \lambda p(\theta) \right\}}{\partial \theta \partial \theta^{\mathrm{T}}} dG(z).$$

The Appendix gives the proof of Theorem 2.1 which relies on the innovative idea of [14] who introduced the statistical functional framework to information theoretic approach. By replacing the unknown distribution G with the empirical distribution $\widehat{G}$, we can calculate the bias term. After correcting the bias of the log-likelihood, we propose a new information criterion

$$\mathrm{IC} = -2 \sum_{\alpha=1}^{n} w(y_\alpha) \log f(y_\alpha; \widehat{\theta}) + 2 \mathrm{tr} \{ R(\widehat{G})^{-1} K(\widehat{G}) \}, \tag{2.7}$$

with

$$\begin{aligned} R(\widehat{G}) &= -\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial^2 \left\{w(y_\alpha)\log f(y_\alpha;\theta) - \lambda p(\theta)\right\}}{\partial\theta\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\widehat{\theta}}, \\ K(\widehat{G}) &= \frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial \left\{w(y_\alpha)\log f(y_\alpha;\theta) - \lambda p(\theta)\right\}}{\partial\theta} \cdot \frac{\partial \left\{w(y_\alpha)\log f(y_\alpha;\theta)\right\}}{\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\widehat{\theta}}. \end{aligned} \tag{2.8}$$

We select the model that minimizes the IC score in Equation (2.7) with respect to the models of interest defined by $f(y;\widehat{\theta})$ and λ.

Corollary 1. *If $\widehat{\theta}$ is the maximum penalized likelihood estimator, i.e. it is obtained by maximizing*

$$\frac{1}{n}\left[\sum_{\alpha=1}^{n}\log f(y_\alpha;\theta) - \lambda p(\theta)\right],$$

then the bias term in Equation (2.7) will become

$$\begin{aligned} R(\widehat{G}) &= -\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial^2 \{\log f(y_\alpha;\theta) - \lambda p(\theta)\}}{\partial\theta\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\widehat{\theta}}, \\ K(\widehat{G}) &= \frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial \{\log f(y_\alpha;\theta) - \lambda p(\theta)\}}{\partial\theta} \cdot \frac{\partial \left\{w(y_\alpha)\log f(y_\alpha;\theta)\right\}}{\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\widehat{\theta}}. \end{aligned} \tag{2.9}$$

Note that the model estimated by the maximum likelihood can be handled by setting $\lambda = 0$. See Section 3.2 for a consideration of the copula model selection. A number of recent statistical packages (e.g., R software) implement the maximum likelihood method to estimate various types of statistical models, including regression, classification, time series and other models. Thus, the range of models to which our criteria can be applied is very wide.

Corollary 2. *If the weighted KL measure $\widetilde{K}_w(G,F)$ is used, an information criterion is defined by*

$$\begin{aligned} \widetilde{\mathrm{IC}} = -2\sum_{\alpha=1}^{n} w(y_\alpha)\log f(y_\alpha;\widehat{\theta}) + 2n\int w(x) f(x|\widehat{\theta})dx \\ +2\mathrm{tr}\{R(\widehat{G})^{-1}K(\widehat{G})\}, \end{aligned} \tag{2.10}$$

where the bias term $R(\widehat{G})^{-1}K(\widehat{G})$ follows Equations (2.8) and (2.9) when we use a maximum penalized weighted likelihood estimator and a maximum penalized likelihood estimator, respectively.

Remark. [1] also used weighted likelihood estimates, but for robust model selection using AIC rather than for developing new information criteria. Comparing with [6] who proposed focused information criterion to focus on some parameters in competitive models, we do model selection with respect to feature events which are matched by weight functions. If we set $w(y)$ to be a constant, the proposed criteria IC and $\widetilde{\mathrm{IC}}$ reduce to GIC (ref. [14]), which further reduces to AIC if the model $f(y;\theta)$ is correctly specified, giving $R(\widehat{G}) = K(\widehat{G})$. Therefore, the proposed criterion can be regarded as an extension of the standard information theoretic approach. The developed information criteria require an i.i.d. framework. It is our conjecture that similar results, when the sample size is large, can be proved under the time series context such as an AR(p) model. See, for instance, one of the examples in [15].

3. Simulation study

3.1. Penalized B-spline regression

Monte Carlo experiments are conducted to compare the performance of our method with the standard weighted likelihood approach and GIC. Suppose we have n independent observations $\{(y_\alpha, x_\alpha); \alpha = 1, 2, \dots, n\}$, where y_α are response variables and x_α are explanatory variables. In generalized linear models y_α are assumed to follow the exponential family of distributions with densities $f(y_\alpha|x_\alpha;\xi_\alpha,\phi) = \exp\left[\{y_\alpha\xi_\alpha - u(\xi_\alpha)\}/\phi + v(y_\alpha,\phi)\right]$, where $u(\cdot)$ and $v(\cdot,\cdot)$ are functions specific to each distribution, and ϕ is an unknown scale parameter. The conditional expectation $E(y_\alpha|x_\alpha) = \mu_\alpha = u'(\xi_\alpha)$ is linked to the predictor $\eta_\alpha = h(\mu_\alpha)$ with the link function $h(\cdot)$.

Using the basis expansion approach given in [8, 13], the unknown predictor η_α is approximated by a linear combination of basis functions $\eta_\alpha = \sum_{j=1}^{m} \beta_j b_j(x_\alpha) = \beta^{\mathrm{T}} b(x_\alpha)$, where $\beta = (\beta_1, \dots, \beta_m)^{\mathrm{T}}$ is the m-dimensional coefficient vector and $b(x) = (b_1(x), \dots, b_m(x))^{\mathrm{T}}$ is the m-dimensional basis function vector. With the basis expansion predictor, the probability density function of the generalized linear model is given by

$$f(y_\alpha|x_\alpha;\theta) = \exp\left(\left[y_\alpha r\left\{\beta^{\mathrm{T}} b(x_\alpha)\right\} - s\left\{\beta^{\mathrm{T}} b(x_\alpha)\right\}\right]/\phi + v(y_\alpha,\phi)\right), \quad (3.1)$$

where $\theta = (\beta^{\mathrm{T}}, \phi)^{\mathrm{T}}$, $r(\cdot) = u'^{-1} \circ h^{-1}(\cdot)$, $s(\cdot) = u \circ u'^{-1} \circ h^{-1}(\cdot)$ and $\circ$ is the convolution operator. The penalty function in Equation (2.1) is defined by

$$p(\theta) = \sum_{j=2}^{m} (\Delta^2 \beta_j)^2 = \beta^{\mathrm{T}} D \beta, \quad (3.2)$$

where Δ is the difference operator, i.e. $\Delta\beta_j = \beta_j - \beta_{j-1}$ and D is a matrix representation of the difference operator. Putting these equations with the weight function $w(\cdot)$, which is specified below, the unknown parameter θ is estimated by maximizing Equation (2.1).

The remaining problem is how to choose the smoothing parameter λ. Substituting the density and penalty functions given by Equations (3.1) and (3.2) into Equation (2.7), we derive a tailor-made version of model selection criteria for evaluating the generalized linear models using basis expansion predictor:

$$\begin{aligned}\mathrm{IC} \quad &= \quad -2\sum_{\alpha=1}^{n} w(y_\alpha)\left[y_\alpha r\left\{\widehat{\beta}^{\mathrm{T}} b(x_\alpha)\right\}/\widehat{\phi} - s\left\{\widehat{\beta}^{\mathrm{T}} b(x_\alpha)\right\}/\widehat{\phi} + v(y_\alpha,\widehat{\phi})\right] \\ &\quad + 2\mathrm{tr}\left\{R^{-1}(\widehat{\theta})K(\widehat{\theta})\right\}, \end{aligned} \tag{3.3}$$

where $K(\widehat{\theta})$ and $R(\widehat{\theta})$ are the $(m+1)\times(m+1)$ matrices given by

$$\begin{aligned} K(\widehat{\theta}) &= \frac{1}{n}\begin{pmatrix} B^{\mathrm{T}}W\Lambda/\widehat{\phi} - \lambda D\widehat{\beta}e^{\mathrm{T}} \\ p^{\mathrm{T}} \end{pmatrix}\left(\Lambda WB/\widehat{\phi}, p\right), \\ R(\widehat{\theta}) &= \frac{1}{n}\begin{pmatrix} B^{\mathrm{T}}W\Gamma B/\widehat{\phi} + n\lambda D & B^{\mathrm{T}}W\Lambda e/\widehat{\phi}^2 \\ e^{\mathrm{T}}\Lambda WB/\widehat{\phi}^2 & -q^{\mathrm{T}}e \end{pmatrix}, \end{aligned}$$

respectively. Here $B = (b(x_1),\ldots,b(x_n))^{\mathrm{T}}$, $W = \mathrm{diag}\{w(y_1),\ldots,w(y_n)\}$, $e = (1,\ldots,1)^{\mathrm{T}}$, Λ and Γ are $n\times n$ diagonal matrices and p and q are n-dimensional vectors with the α-th diagonal elements and the α-th elements given by

$$\begin{aligned} \Lambda_{\alpha\alpha} &= \frac{y_\alpha - \widehat{\mu}_\alpha}{u''(\widehat{\xi_\alpha})h'(\widehat{\mu}_\alpha)}, \\ \Gamma_{\alpha\alpha} &= \frac{(y_\alpha - \widehat{\mu}_\alpha)\{u'''(\widehat{\xi_\alpha})h'(\widehat{\mu}_\alpha) + u''(\widehat{\xi_\alpha})^2 h''(\widehat{\mu}_\alpha)\}}{\{u''(\widehat{\xi_\alpha})h'(\widehat{\mu}_\alpha)\}^3} + \frac{1}{u''(\widehat{\xi_\alpha})h'(\widehat{\mu}_\alpha)^2}, \\ p_\alpha &= w(y_\alpha)\left[-\frac{y_\alpha r\{\widehat{\beta}^{\mathrm{T}}b(x_\alpha)\} - s\{\widehat{\beta}^{\mathrm{T}}b(x_\alpha)\}}{\widehat{\phi}^2} + \frac{\partial}{\partial\phi}v(y_\alpha,\phi)\bigg|_{\phi=\widehat{\phi}}\right],\ q_\alpha = \frac{\partial p_\alpha}{\partial\phi}\bigg|_{\phi=\widehat{\phi}}. \end{aligned}$$

We choose the smoothing parameter λ as the minimizer of the criteria.

To illustrate the use of Equation (3.3), we consider P-spline Gaussian regression modeling in the simulation. Data sets $\{(x_\alpha, y_\alpha); \alpha = 1,\ldots,n\}$ are repeatedly generated from the true regression model $y_\alpha = m(x_\alpha) + \varepsilon_\alpha$ for $x_\alpha = (2\alpha-1)/(2n)$. The errors ε_α are assumed to be independently distributed according to the normal distribution with mean 0 and variance σ^2. Two true functions are studied in the simulation: $m(x) = \sin(5\pi x^{2.5})$ and $m(x) = 0.005(1-48x+218x^3+145x^4) + \cos(4\pi(x-1)^3)$. We estimate the unknown function $m(x)$ by using P-spline Gaussian regression model:

$$f(y_\alpha|x_\alpha;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left\{y_\alpha - \beta^{\mathrm{T}}b(x_\alpha)\right\}^2\right], \qquad \alpha = 1,\ldots,n, \tag{3.4}$$

with $\theta = (\beta^{\mathrm{T}},\sigma^2)^{\mathrm{T}}$. Following the suggestion of Eilers & Marx (1996), we set a large number of basis functions $m = 20$ and optimize the value of the smoothing parameter λ in each Monte Carlo experiment. The candidates for the smoothing parameter are chosen on a geometrical grid of 20 knots between $\log_{10}(\lambda) = 0$ and $\log_{10}(\lambda) = -9$. Here we specify the weight function as $w_\alpha = 1/3$ if $(x_\alpha < 0.5)$

Table 1

Comparison of the average weighted mean squared errors based on various criteria and using 100 simulated datasets. Figures in the second line give estimated standard deviations. MWL refers to the standard maximum weighted likelihood approach

$m(x) = \sin(5\pi x^{2.5})$						
σ^2	n	Our method	MWL	MWL	MWL	GIC
			$(m = 6)$	$(m = 9)$	$(m = 12)$	
0.2	50	0.0137	0.1019	0.1001	0.0446	0.0139
		0.0050	0.0028	0.0034	0.0040	0.0052
0.2	100	0.0073	0.1000	0.0971	0.0408	0.0077
		0.0025	0.0021	0.0024	0.0024	0.0026
0.5	50	0.0859	0.1289	0.1402	0.0944	0.0868
		0.0411	0.0237	0.0354	0.0371	0.0415
0.5	100	0.0401	0.1107	0.1136	0.0639	0.0407
		0.0151	0.0091	0.0108	0.0143	0.0153

$m(x) = 0.005(1 - 48x + 218x^3 + 145x^4) + \cos(4\pi(x-1)^3)$						
σ^2	n	Our method	MWL	MWL	MWL	GIC
			$(m = 6)$	$(m = 9)$	$(m = 12)$	
0.2	50	0.0154	0.3789	0.1871	0.0302	0.0158
		0.0061	0.0036	0.0047	0.0055	0.0063
0.2	100	0.0086	0.3737	0.1769	0.0263	0.0091
		0.0030	0.0022	0.0026	0.0029	0.0030
0.5	50	0.0806	0.4075	0.2280	0.0821	0.0810
		0.0377	0.0226	0.0286	0.0322	0.0379
0.5	100	0.0404	0.3857	0.1966	0.0509	0.0410
		0.0229	0.0098	0.0159	0.0195	0.0230

and $w_\alpha = 5/3$ if $(x_\alpha \geq 0.5)$. This weight setting is considered when the relative importance of prediction is different in a particular range of x. In this case, we have more emphasis on the predictive performance on "large" x. Taking $u(\xi_\alpha) = \xi_\alpha^2/2$, $\phi = \sigma^2$, $v(y_\alpha, \psi) = -y_\alpha^2/(2\sigma^2) - \log(2\pi\sigma^2)/2$ and $h(\mu_\alpha) = \mu_\alpha$ in the criteria of Equation (3.3), we obtain a tailor-made version of an information criterion for evaluating the estimated P-spline Gaussian regression model.

Table 1 compares the averaged "weighted" mean squared error

$$\mathrm{WMSE} = \frac{1}{n}\sum_{\alpha=1}^{n} w_\alpha \left\{m(x_\alpha) - \widehat{y}(x_\alpha)\right\}^2$$

between the true and estimated functions $\widehat{y}(x_\alpha) = \hat{\beta}^{\mathrm{T}} b(x_\alpha)$. Again, this WMSE is set to align with the objective of emphasizing the prediction of y when $x \geq 0.5$. This performance measure is used because we want to be able to emphasize specific features that are selected by users according to their research focus. We also estimate the regression model by GIC and the standard weighted maximum likelihood approach, which is implemented by setting the smoothing parameter $\lambda = 0$ in Equation (2.1). As the parameter estimate for θ can not be obtained by the weighted likelihood method when we set a large number of basis functions such as $m = 15$, we prepare several values of the number of basis functions $m = \{6, 9, 12\}$ and then construct the regression model. In addition, by setting $w(y_\alpha) = 1$, the model $f(y_\alpha|x_\alpha;\theta)$ in Equation (3.4) is also estimated by the usual maximum penalized likelihood approach. This approach is employed to test the importance of the weight function. An optimal value of the smoothing parameter λ is determined by GIC (ref. [14]). The simulation results are obtained from 100 repeated Monte Carlo trials. It may be seen from the simulation results in Table 1 that our method is superior to the competitors; it gives the smallest value of the average WMSE in all combinations of n and σ^2.

Finally, we note that our criteria can also be applied to P-spline logistic regression modeling. It is easy to derive the information criteria for evaluating the estimated P-spline logistic regression model by taking $u(\widehat{\xi}_\alpha) = \log\{1 + \exp(\widehat{\xi}_\alpha)\}$, $v(y_\alpha, \phi) = 0$, $h(\widehat{\mu}_\alpha) = \log\{\widehat{\mu}_\alpha/(1 - \widehat{\mu}_\alpha)\}$, and $\phi = 1$ in Equation (3.3). In a Poisson model, we shall take $u(\widehat{\xi}_\alpha) = \exp(\widehat{\xi}_\alpha)$, $v(y_\alpha, \phi) = -\log(y_\alpha!)$, $h(\widehat{\mu}_\alpha) = \log(\widehat{\mu}_\alpha)$ and $\phi = 1$ in Equation (3.3).

3.2. Copula model selection

Recently, there has been active research on copula model selection; see for example [5, 11, 19, 9]. We apply our IC in this section to find the best copula models to explain certain dependence structure of the data. Let $y_\alpha = (y_{\alpha 1}, \ldots, y_{\alpha p})^{\mathrm{T}}$, $\alpha = 1, \ldots, n$, be p-dimensional independent observations. With $y_{\alpha j}$, $j = 1, \ldots, p$, distributed as Uniform[0,1], the distribution of $y_\alpha = (y_{\alpha 1}, \ldots, y_{\alpha p})^{\mathrm{T}}$ is specified by $F(y_\alpha; \theta) = C(y_{\alpha 1}, \ldots, y_{\alpha p}; \theta)$, where $C(\cdot)$ is a copula function and θ is the copula parameter. The main objective is to select an appropriate $C(\cdot)$. To perform the copula model selection, we use the IC in Equation (2.7) with θ estimated by MLE and the bias term in Equation (2.9) is adopted. As discussed in the previous sections, a novelty of our approach is that we can choose the weight function with respect to a special feature of the distribution of the random variables of interest to do the model selection. Suppose $p = 2$ and we want to select an appropriate copula function for $y_{\alpha 1}$ and $y_{\alpha 2}$, where a feature of attention is their lower tail dependence structure. In this case, we use $w(y_\alpha) = \lambda$ if $y_{\alpha 1}, y_{\alpha 2} \leq 0.1$ and $w(y_\alpha) = 1$ otherwise, where $\lambda > 0$. The rationale of this choice of $w(y_\alpha)$ is to emphasize or de-emphasize the influence of the co-occurrence of extreme observations, i.e. when both $y_{\alpha 1} < 0.1$ and $y_{\alpha 2} < 0.1$. To study the effect of our approach, we simulate $n = 100$ observations from three copula functions with different lower tail dependence behavior. They are the t copula with the copula density $c(u_1, u_2) = t_\rho^{(2)}(T^{-1}(u_1), T^{-1}(u_2))/[t^{(1)}(T^{-1}(u_1))t^{(1)}(T^{-1}(u_2))]$, the Clayton copula, $C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$, and the Gumbel copula, $C(u_1, u_2) = \exp\{-[(-\log u_1)^\theta + (-\log u_2)^\theta]^{1/\theta}\}$, where $t^{(1)}$ is the univariate t density with degrees of freedom ν, $t_\rho^{(2)}$ is the bivariate t density with correlation ρ and degrees of freedom ν, and $T^{(-1)}$ is the inverse of the univariate t distribution. We choose $\rho = 0.59$ and $\nu = 8$ in the t-copula, and $\theta = 0.67$ and 1.67 for the Clayton and Gumbel copulas, respectively to match their Kendall's tau. By construction, both t and Clayton copulas have positive lower tail dependence, whereas the Gumbel copula has zero tail dependence. We select the best model among the fitted t, Clayton and Gumbel copula models using the IC.

Table 2

Simulation results of 100 replications of size $n = 100$ generated by the t, Clayton and Gumbel copulas. The table shows the median ratios of the KL measure, the L^1-norm, the L^2-norm and the Hellinger measure for different λ over that for the benchmark method ($\lambda = 1$)

λ	0.2	0.4	0.6	0.8	1	2	4	6	8
					t copula				
KL	1.41	1.30	1.14	1.00	1.00	0.70	0.73	0.74	0.79
L^1-norm	1.16	1.12	1.03	1.00	1.00	0.70	0.76	0.76	0.89
L^2-norm	1.58	1.56	1.51	0.95	1.00	0.81	0.91	1.02	1.23
Hellinger	1.56	1.48	1.26	1.00	1.00	0.74	0.80	0.77	0.85
					Clayton copula				
KL	1.73	1.55	1.32	1.06	1.00	0.97	0.95	0.95	0.88
L^1-norm	1.25	1.17	1.08	1.02	1.00	0.94	0.90	0.90	0.89
L^2-norm	1.45	1.46	1.31	1.44	1.00	0.84	0.89	0.84	0.78
Hellinger	1.52	1.44	1.12	1.06	1.00	0.92	0.83	0.83	0.79
					Gumbel copula				
KL	0.74	0.93	0.97	0.98	1.00	1.40	13.58	27.32	31.72
L^1-norm	0.86	0.88	0.94	0.97	1.00	1.16	3.48	5.28	5.80
L^2-norm	0.77	0.78	0.94	0.90	1.00	1.39	21.22	61.59	75.82
Hellinger	0.76	0.79	0.88	0.94	1.00	1.36	12.50	28.92	35.32

As the target feature is the lower tail dependence, we want to see how fitted models match the true model in $\mathcal{A} = \{y_\alpha : y_{\alpha 1}, y_{\alpha 2} \leq 0.1\}$. Statistically, we assess the performance by comparing the true tail distribution $g(y|\mathcal{A}) = g(y)I(y \in \mathcal{A})/P_G(y \in \mathcal{A})$ and the fitted tail distribution $f(y|\mathcal{A}) = f(y)I(y \in \mathcal{A})/P_F(y \in \mathcal{A})$, where f and F are from a fitted model, and P_G and P_F are probabilities evaluated under G and F, respectively. Four performance measures are adopted, namely, the KL measure, $\int_{\mathcal{A}} g(y|\mathcal{A}) \log \frac{g(y|\mathcal{A})}{f(y|\mathcal{A})} dy$; the L^1-norm, $\int_{\mathcal{A}} |g(y|\mathcal{A}) - f(y|\mathcal{A})| dy$; the L^2-norm, $\int_{\mathcal{A}} [g(y|\mathcal{A}) - f(y|\mathcal{A})]^2 dy$; and the Hellinger measure, $\int_{\mathcal{A}} [\sqrt{g(y|\mathcal{A})} - \sqrt{f(y|\mathcal{A})}]^2 dy$. The smaller the performance measures, the better the fitted models.

The model selection results using the IC in Equation (2.7) with different λ are produced for 100 replications of each of the three copula models. Table 2 shows the ratio of the median of the KL measure based on 100 replications, for each λ, over the median KL measure of the benchmark ($\lambda = 1$), and the respective ratio of the median L^1-norm, L^2-norm and the Hellinger measure. The smaller-than-one ratio of the median performance measure means that the model selection based on the IC is better than that of the benchmark method. For the t-copula data generating process, our method outperforms (most of the median ratios are less than 1) the benchmark when $\lambda > 1$. For Clayton, we observe the same pattern that for $\lambda > 1$, all median ratios are less than one. For Gumbel, the outperformance appears in $\lambda < 1$, where the best-performed cases are $\lambda = 0.2$ and 0.4. The above findings suggest that when λ is appropriately chosen to focus on one feature of the copula models, our IC based on the weighted KL measure $K_w(G, F)$ is superior to that based on the usual KL measure. For Clayton and t copulas, which have lower tail dependence, it is reasonable to choose $\lambda > 1$ to emphasize the effect of extreme observations, whereas for

Gumbel, which has zero lower tail dependence, $\lambda < 1$ is appropriate because we want to de-emphasize the effect of extreme observations to match with the zero or weak low tail dependence of the true model. In practice, whether we want to emphasize or de-emphasize the extreme observations can depend on some prior belief regarding the tail dependence property of the true model.

4. Real data analysis

4.1. Predicting the incidence of disease

A motivation for introducing the KL measures is that in medical research, the proportion of disease cases, e.g., heart-disease and cancer, is much small relative to the proportion of non-disease cases. If we can design a suitable weight function that weighs more heavily on the disease cases, we expect that the accuracy rate (i.e., the chance of correctly predicting a cancer patient to have cancer) can be improved. We consider an analysis of South African heart disease data (ref. [18]) studied in [10], pp. 122 to illustrate our method using a logistic regression model. The data set consists of 160 heart-disease cases ($y = 1$) and a sample of 302 controls ($y = 0$). Using a set of four predictors, tobacco (cumulative tobacco x_1), ldl (low density lipoprotein cholesterol x_2), famhist (family history of heart disease x_3), and age (age at onset x_4), we model the conditional probability $\Pr(y = 1|x) = \pi(x)$ by

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \sum_{j=1}^{4} \beta_j x_j,$$

where $x = (x_1, \ldots, x_4)^{\mathrm{T}}$. Then the weighted log-likelihood function for (x_α, y_α), $\alpha = 1, \ldots, 462$, in terms of $\beta = (\beta_0, \ldots, \beta_4)^{\mathrm{T}}$ is

$$\ell_w(\beta, y) \quad = \quad \frac{1}{462}\sum_{\alpha=1}^{462} w(y_\alpha)\{y_\alpha \log \pi(x_\alpha) + (1-y_\alpha)\log(1-\pi(x_\alpha))\}.$$

Because our concern is to increase the chance of correctly predicting a heart-disease patient to have heart disease, we set the weight values as $w(1) = 4$ and $w(0) = 1$. The unknown parameter vector $\beta = (\beta_0, \ldots, \beta_4)^{\mathrm{T}}$ is estimated by maximizing the weighted log-likelihood $\ell_w(\beta, y)$. The corresponding IC criterion in Equation (2.7) is then

$$\mathrm{IC} = -2\sum_{\alpha=1}^{462} w(y_\alpha)\left[y_\alpha \widehat{\beta}^{\mathrm{T}} x_\alpha - \log\left\{1 + \exp(\widehat{\beta}^{\mathrm{T}} x_\alpha)\right\}\right] + 2\mathrm{tr}\left\{R^{-1}(\widehat{G})K(\widehat{G})\right\},$$

where $\widehat{\beta}$ is the parameter estimate, $K(\widehat{G})$ and $R(\widehat{G})$ are 5×5 matrices given by $K(\widehat{G}) = X^{\mathrm{T}}\Lambda^2 X/462$ and $R(\widehat{G}) = X^{\mathrm{T}}\Gamma X/462$, with $X = (x_1, \ldots, x_{462})^{\mathrm{T}}$, Λ and Γ being 462×462 diagonal matrices

$$\begin{aligned}
\Lambda &= \mathrm{diag}\left[w(y_1)\{y_1 - \widehat{\pi}(x_1)\}, \ldots, w(y_{462})\{y_{462} - \widehat{\pi}(x_{462})\}\right], \\
\Gamma &= \mathrm{diag}\left[w(y_1)\widehat{\pi}(x_1)(1 - \widehat{\pi}(x_1)), \ldots, w(y_n)\widehat{\pi}(x_{462})(1 - \widehat{\pi}(x_{462}))\right],
\end{aligned}$$

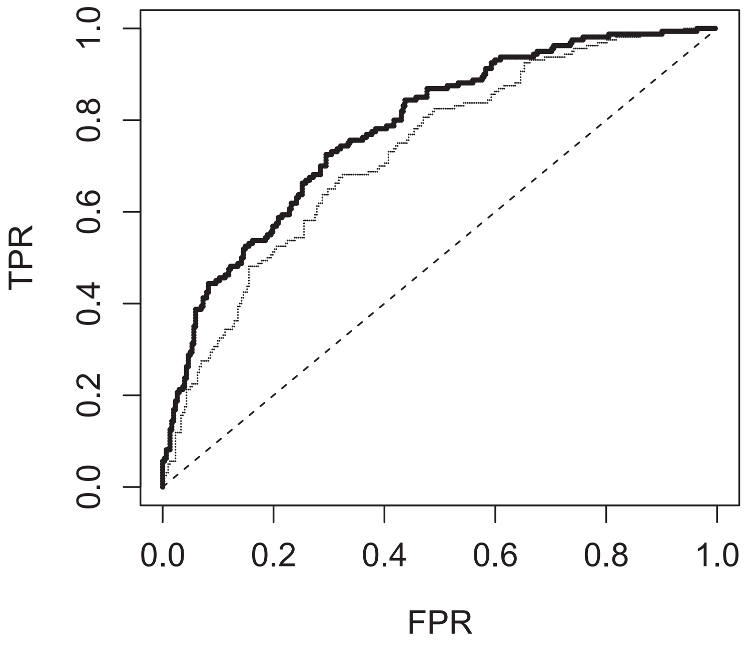

Fig 3. *The ROC curves obtained using our IC score (thick line) and from the GIC (thin line).*

where $\widehat{\pi}(x_\alpha)$ is the estimated conditional probability. As a result, we obtain the minimum IC score as -1.052 with the model including all the four predictors. The thick line in Figure 3 shows the receiver operating characteristic (ROC) curve, obtained by plotting the fraction of true positives out of the positives (TPR = true positive rate) versus the fraction of false positives out of the negatives (FPR = false positive rate).

To compare the performance with the un-weighted method ($w(y_\alpha) = 1$ for $\alpha = 1, \ldots, 462$), we use the GIC in [14] to do the model selection again. The selected model contains only two predictors x_1 and x_2 with the ROC curve given by the thin line in Figure 3. As this ROC curve lies below the curve obtained from our approach, the former model obtained by the IC score is superior. We observe an improved predictive power of the logistic regression model when it has more weights on the disease cases and when the model is selected using our IC score.

4.2. Quantile modeling of environmental data

In contrast to classical linear regression models, where a conditional expectation of the response variable is in focus, the quantile regression tries to estimate the τ-th conditional quantile of y given $x = (x_1, \ldots, x_p)^{\mathrm{T}}$ as

$$q_\tau(y|x) = \beta_0(\tau) + \sum_{j=1}^{p} \beta_j(\tau) x_j,$$

where $\beta_0(\tau), \ldots, \beta_p(\tau)$ are coefficients dependent on the quantile τ. When we set $\tau = 0.5$, the model reduces to the conditional median regression, which is more robust to outliers than the conditional mean regression. The unknown parameters are estimated by maximizing

$$\ell_w(\beta, y) = -\frac{1}{n}\sum_{\alpha=1}^{n} w(y_\alpha)\rho_\tau(y_\alpha - \beta(\tau)^{\mathrm{T}} x_\alpha)$$

with $\rho_\tau(u) = u(\tau - I(u < 0))$, the usual loss function for standard quantile regression modeling and $\beta(\tau) = (\beta_0(\tau), \ldots, \beta_p(\tau))^T$. To evaluate estimated quantile regression models, we use the IC score in Equation (2.7). The criterion is then

$$\mathrm{IC} = -2\sum_{\alpha=1}^{n} w(y_\alpha)\rho_\tau(y_\alpha - \hat{\beta}(\tau)^{\mathrm{T}} x_\alpha) + 2\mathrm{tr}\left\{R^{-1}(\widehat{G})K(\widehat{G})\right\}, \tag{4.1}$$

with $X = (x_1, \ldots, x_n)^{\mathrm{T}}$, $K(\widehat{G}) = \frac{1}{n}\tau(1-\tau)X^{\mathrm{T}}WX$, and $R(\widehat{G}) = \frac{1}{n}X^{\mathrm{T}}MX$. Here $W = \mathrm{diag}\{w(y_1), \ldots, w(y_n)\}$, $M = \mathrm{diag}\{g(\xi_1(\tau)), \ldots, g(\xi_n(\tau))\}$ is the n-dimensional diagonal matrix with the α-th element of M being the τ-th quantile of the density function $g(\xi_\alpha(\tau))$, $\xi_\alpha(\tau) = G^{-1}(\tau|x_\alpha)$, and $P(y_\alpha \leq y|x_\alpha) = G(y|x_\alpha)$. Although our formulation in Equation (4.1) allows heterogeneous weights $w(y_\alpha)$ for observations y_α, the IC score with the equal weight $w(y_t) = 1$ alone, is new in the literature. In the study of precipitation data discussed below, we set the weight function as $w(y_\alpha) = 1 + I(y_\alpha \geq a_u)$ for upper quantiles and $w(y_\alpha) = 1 + I(y_\alpha \leq a_l)$ for lower quantiles, where a_u and a_l are some thresholds. This setting is in line with the environmental issues like drought, climate change, flooding, etc, in which the precipitation is either very high or very low.

First, we apply our method to monthly precipitation data collected in one of the meteorological stations of the Hong Kong Observatory. The data period is from September, 1997 to December, 2010. The quantile autoregression model in [12] is considered:

$$q_\tau(y_\alpha) = \beta_0(\tau) + \sum_{j=1}^{p} \beta_j(\tau) y_{\alpha-j},$$

where y_α is the observed precipitation value at time α. Setting the value of τ as 5% and 95%, we construct two quantile functions. To emphasize the effect of extreme precipitations, we use the weighted functions $w(y_\alpha) = 1 + I(y_\alpha \geq 800)$ for the 95% quantile and $w(y_\alpha) = 1 + I(y_\alpha \leq 20)$ for the 5% quantile. To determine an optimal lag value of p, the IC score is used. We fit the candidate models of lag $p = \{0, 1, \ldots, 20\}$ and then determine the best lag that minimizes the IC score. When we construct the IC score in Equation (4.1), the values of the τ-th quantile of the density function $g(\xi_\alpha(\tau))$ in the matrix M are estimated by the adaptive kernel method in [17]. As a result, we find that the optimal lags for the two quantile functions are $p = 12$. Figure 4(a) shows the estimation result.

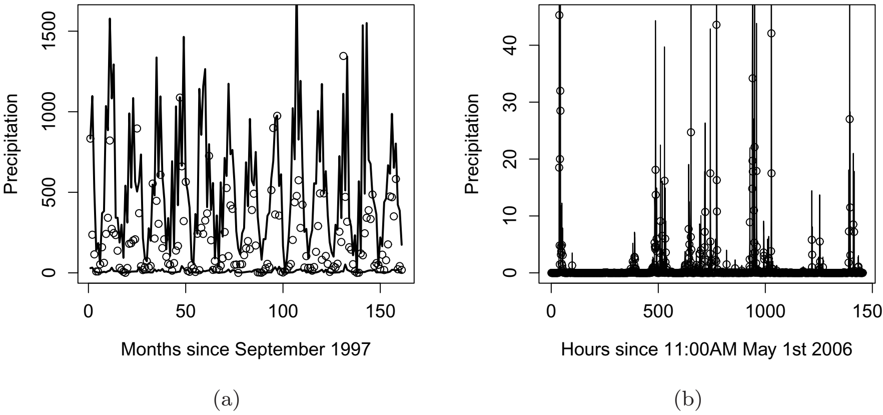

Fig 4. *Precipitation data. The solid lines are the fitted quantile functions and circles are realized values.*

The solid lines are the fitted quantile functions $q_\tau(y_\alpha) = \widehat{\beta}_0(\tau)+\sum_{j=1}^{p} \widehat{\beta}_j(\tau)y_{\alpha-j}$ and circles are realized values.

Next we apply our method to hourly precipitation data, from 1:00 May 1, 2006 to 23:00 June 30, 2006. The same modeling procedure described above is used. Here our focus is $\tau = 95\%$ percentile which has implications for flooding, and we use the weight function $w(y_\alpha) = 1 + I(y_\alpha \geq 10)$ for the 95% quantile. The optimal lag is $p = 1$. Again this result seems to be very natural because the most recent observation contains useful information for forecasting. Figure 4(b) shows the estimation result.

5. Discussion

The concept of model selection with respect to some model features is introduced in the information-theoretic paradigm. The standard KL measure is generalized to have a weight function designed for the specified features. The IC score based on the weighted KL measure, $K_w(G, F)$, is defined in Theorem 2.1 by deriving the bias correction of the expected weighted loglikelihood estimator, $\int \ell_w(\widehat{\theta}, z)dG(z)$. The scope of applications of our IC score is very wide. In real-life applications, there are many cases that allow us to use the weighted KL measure. Examples are a logistic probability model (Example 1) and an exponential power distribution (Example 2), described in Section 1. We show the usefulness of our statistical modeling procedures through simulation studies, including penalized B-spline regression (Section 3.1) and copula model selection (Section 3.2). We also illustrate our information criteria by applying it to real data, predicting the incidence of disease (Section 4.1) and quantile modeling of monthly precipitation rate (Section 4.2). We believe that our statistical modeling would contribute to the advancement of empirical studies.

The bias correction term can be computed using bootstrap methods (ref. [7]) if their analytical forms are not tractable. As the bootstrap analogues of the expected weighted log-likelihood of Equation (2.3) and the sample based weighted log-likelihood are $\eta_w^{(b)} = n^{-1}\sum_{\alpha=1}^{n} w(y_\alpha)\log f(y_\alpha;\widehat{\theta}^*)$ and $\hat{\eta}_w^{(b)} = n^{-1}\sum_{\alpha=1}^{n} w(y_\alpha^*)\log f(y_\alpha^*;\widehat{\theta}^*)$, the bootstrap bias estimator, an estimator of Equation (2.4), is given by $E_{y^*}(\eta_w^{(b)} - \hat{\eta}_w^{(b)})$. Here $y^* = (y_1^*,\ldots,y_n^*)^{\mathrm{T}}$ is the empirical distribution based on bootstrap samples that has the probability n^{-1} at each data point y_α^* $(\alpha = 1,\ldots,n)$, and $\widehat{\theta}^*$ is the parameter estimate based on the bootstrap sample y^*, i.e., the maximizer of the weighted penalized likelihood: $\ell_w(\theta, y^*) - \lambda p(\theta)$. This approach provides a direct computational way of assessing the constructed model. The numerical approach to constructing information criteria has been examined by [14, 3].

We have given examples demonstrating how $w(y)$ is designed to match any model feature of interest. The superior performance is demonstrated in various regression-type problems and in copula model selection. Determining how to align $w(y)$ with the purpose of the statistical modeling will be an interesting topic for further research.

Appendix

Proof of the weighted KL measure

Assume without loss of generality that G and F are continuous. From Konishi & Kitagawa (2008, pp. 30), we have

$$\log\frac{f(y)}{g(y)} \le \frac{f(y)}{g(y)} - 1,$$

which implies that

$$\begin{aligned} -K_w(G,F) &= \int w(y)\log\frac{f(y)}{g(y)}g(y)dy \\ &\le \int w(y)\left(\frac{f(y)}{g(y)} - 1\right)g(y)dy \\ &= \int w(y)f(y)dy - \int w(y)g(y)dx \\ &= E_F[w(Y)] - E_G[w(Y)] \le 0, \end{aligned}$$

and so (i) is satisfied. It is obvious that $K_w(G,G) = 0$. When $K_w(G,F) = 0$, the above shows that both $E_F[w(Y)] - E_G[w(Y)] = 0$ and $\int w(y)\log\frac{f(y)}{g(y)}g(y)dy = \int w(y)\big(\frac{f(y)}{g(y)} - 1\big)g(y)dy$. Therefore, (ii) is satisfied. The above inequalities also indicate that $\widetilde{K}_w(G,F) \ge 0$ for all fitted model F.

Proof of Theorem 2.1

We denote the observed time series $y_1, \dots, y_n$ and its replicates as y and z, respectively. Note that the true estimate $\theta = T(G)$, which maximizes the penalized weighted log-likelihood, can be expressed as a functional:

$$\int \frac{\partial\{\ell_w(\theta, z) - \lambda p(\theta)\}}{\partial\theta}\bigg|_{\theta=T(G)} dG(z) = 0.$$

Then the stochastic expansion of each of the elements of $\widehat{\theta} = T(\widehat{G})$ around $T(G)$ is expressed as

$$\begin{aligned}\widehat{\theta}_i = T_i(G) + \frac{1}{n}\sum_{\alpha=1}^{n} T_i^{(1)}(y_\alpha; G) \\ +\frac{1}{2n^2}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} T_i^{(2)}(y_\alpha, y_\beta; G) + o_p(n^{-1}),\end{aligned} \tag{5.1}$$

where $T_i^{(1)}(y_\alpha; G)$ and $T_i^{(2)}(y_\alpha, y_\beta; G)$ are the first and second order derivatives of the functional $T(\cdot)$. Putting Equation (5.1) into a Talyor expansion of $\ell_w(\widehat{\theta}, z)$ around $\theta = T(G)$ gives

$$\begin{aligned}
& E_{G(z)}[\ell_w(\widehat{\theta}, z)] \\
\approx & \int \ell_w(T(G), z) dG(z) + \sum_{i=1}^{p}(\widehat{\theta}_i - T_i(G)) \int \frac{\partial \ell_w(\theta, z)}{\partial\theta_i}\bigg|_{\theta=T(G)} dG(z) \\
& + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\widehat{\theta}_i - T_i(G))(\widehat{\theta}_j - T_j(G)) \int \frac{\partial^2 \ell_w(\widehat{\theta}, z)}{\partial\theta_i\partial\theta_j}\bigg|_{\theta=T(G)} dG(z) \\
= & \int \ell_w(T(G), z) dG(z) + \frac{1}{n}\sum_{i=1}^{p}\sum_{\alpha=1}^{n} T_i^{(1)}(y_\alpha; G) \int \frac{\partial \ell_w(\theta, z)}{\partial\theta_i}\bigg|_{\theta=T(G)} dG(z) \\
& + \frac{1}{2n^2}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\Bigg[\sum_{i=1}^{p} T_i^{(2)}(y_\alpha, y_\beta; G) \int \frac{\partial \ell_w(\theta, z)}{\partial\theta_i}\bigg|_{\theta=T(G)} dG(z) \\
& + \sum_{i=1}^{p}\sum_{j=1}^{p} T_i^{(1)}(y_\alpha; G) T_j^{(1)}(y_\beta; G) \int \frac{\partial^2 \ell_w(\theta, z)}{\partial\theta_i\partial\theta_j}\bigg|_{\theta=T(G)} dG(z)\Bigg] + o_p(n^{-1}).
\end{aligned}$$

Also, we have

$$\begin{aligned}
\ell_w(\widehat{\theta}, y) \approx & \ \ell_w(T(G), y) + \sum_{i=1}^{p}(\widehat{\theta}_i - T_i(G)) \frac{\partial \ell_w(\theta, y)}{\partial\theta_i}\bigg|_{\theta=T(G)} \\
& + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\widehat{\theta}_i - T_i(G))(\widehat{\theta}_j - T_j(G)) \frac{\partial^2 \ell_w(\theta, y)}{\partial\theta_i\partial\theta_j}\bigg|_{\theta=T(G)}
\end{aligned}$$

$$
\begin{aligned}
&= \ell_w(T(G),y) + \frac{1}{n}\sum_{i=1}^{p}\sum_{\alpha=1}^{n} T_i^{(1)}(y_\alpha;G)\frac{\partial \ell_w(\theta,y)}{\partial\theta_i}\bigg|_{\theta=T(G)} \\
&\quad + \frac{1}{2n^2}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\Bigg[\sum_{i=1}^{p} T_i^{(2)}(y_\alpha,y_\beta;G)\frac{\partial \ell_w(\theta,y)}{\partial\theta_i}\bigg|_{\theta=T(G)} \\
&\quad + \sum_{i=1}^{p}\sum_{j=1}^{p} T_i^{(1)}(y_\alpha;G)T_j^{(1)}(y_\beta;G)\frac{\partial^2 \ell_w(\theta,y)}{\partial\theta_i\partial\theta_j}\bigg|_{\theta=T(G)}\Bigg] + o_p(n^{-1}).
\end{aligned}
$$

Taking the expectations yields

$$
\begin{aligned}
&\int \ell_w(T(G),y)dG(y) \\
&= \int \ell_w(T(G),z)dG(z) + \frac{1}{n}\left[b^{\mathrm{T}}a - \frac{1}{2}\mathrm{tr}[\Sigma(G)J(G)]\right] + o(n^{-1}), \\
&\int \left[\int \ell_w(T(G),z)dG(z)\right]dG(y) \\
&= \int \ell_w(T(G),z)dG(z) + \frac{1}{n}\left[b^{\mathrm{T}}a - \frac{1}{2}\mathrm{tr}[\Sigma(G)J(G)]\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{p}\int T_i^{(1)}(z;G)\frac{\partial \ell_w(T(G),z)}{\partial\theta_i}\bigg|_{\theta=T(G)} dG(z) + o(n^{-1}),
\end{aligned}
$$

where $a = (a_1,\ldots,a_p)^{\mathrm{T}}$ and $b = (b_1,\ldots,b_p)^{\mathrm{T}}$ are given as

$$
a_i = \int \frac{\partial \ell_w(\theta,z)}{\partial\theta_i}\bigg|_{\theta=T(G)} dG(z) \quad \text{and} \quad b = \int [\widehat{\theta} - T(G)]dG(z) + o(n^{-1}),
$$

respectively. The $p \times p$ matrix $\Sigma = (\sigma_{ij})$ is the estimator of the variance covariance matrix of $\sqrt{n}(\widehat{\theta} - T(G))$, and

$$
J(G) = -\int \frac{\partial^2 \ell_w(\theta,z)}{\partial\theta\partial\theta^{\mathrm{T}}}\bigg|_{\theta=T(G)} dG(z).
$$

Finally, we have

$$
\begin{aligned}
&\int \left[\ell_w(\widehat{\theta},y) - \int \ell_w(\widehat{\theta},z)dG(z)\right] G(y) \\
&= \frac{1}{n}\sum_{i=1}^{p}\int T_i^{(1)}(z;G)\frac{\partial \ell_w(\theta,z)}{\partial\theta_i}\bigg|_{\theta=T(G)} dG(z) + o(n^{-1}) \\
&= \frac{1}{n}\mathrm{tr}\left[\int T^{(1)}(z;G)\frac{\partial \ell_w(\widehat{\theta},z)}{\partial\theta^{\mathrm{T}}}\bigg|_{T(G)} dG(z)\right] + o(n^{-1}).
\end{aligned}
$$

The function $T^{(1)}(z;G)$ for the penalized weighted likelihood estimator is given as

$$T^{(1)}(z;G) = R(G)^{-1}\frac{\partial\{w(z)\log f(\theta,z) - \lambda p(\theta)\}}{\partial\theta}\bigg|_{\theta=T(G)}.$$

Then, replacing the expectation with the empirical distribution, the bias term is obtained.

Acknowledgements

The authors thank the editor, the associate editor and a referee for their valuable suggestions and comments. This work is supported by an HKUST Grant (RPC11BM019) and Zengin Foundation for Studies on Economics and Finance, Japan.

References

[1] Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statistics & Probability Letters* **56** 289–300. MR1892990

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd Inter. Symp. on Information Theory* 267–281. Budapest: Akademiai Kiado. MR0483125

[3] Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* **94** 443–458. MR2380571

[4] Ando, T. (2012). Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences* **31** 13–38. MR2976700

[5] Chen, X and Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics* **135** 125–154. MR2328398

[6] Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* **98** 900–916. MR2041482

[7] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall. MR1270903

[8] Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statistical Science* **11** 89–121. MR1435485

[9] Gronneberg, S. (2010). The copula information criterion and its implications for the maximum pseudo-likelihood estimators. *Dependence Modeling* 113–138. MR2856971

[10] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer. MR2722294

[11] Huard, D., Evin, G. and Favre, A. C. (2006). Bayesian copula selection. *Computational Statistics & Data Analysis* **51** 809–822. MR2297490

[12] Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association* **101** 980–990. MR2324109

[13] Konishi, S., Ando, T. and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91** 27–43. MR2050458

[14] Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika* **83** 875–890. MR1440051

[15] Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York: Springer. MR2367855

[16] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22** 79–86. MR0039968

[17] Portnoy, S. and Koenker, R. (1989). Adaptive L estimation of linear Models. *Annals of Statistics* **17** 362–381. MR0981456

[18] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities. *South African Medical Journal* **64** 430–436.

[19] Silva, R. S. and Lopes, H. F. (2008). Copula, marginal distributions and model selection: a Bayesian note. *Statistics and Computing* **18** 313–320. MR2413387

[20] Sin, C. Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics* **71** 207–225. MR1381082

[21] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *Journal of the Royal Statistical Society, Series B* **64** 583–639. MR1979380

[22] Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models (in Japanese). *Mathematical Science* **153** 12–18.

[23] Xia, Y. and Tong, H. (2011). Feature matching in time series modeling. *Statistical Science* **26** 21–46. MR2849904