Mach Learn (2015) 101:35–58
DOI 10.1007/s10994-014-5440-5

Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function

Elcin Kartal Koc · Hamparsum Bozdogan

Received: 1 December 2013 / Accepted: 20 March 2014 / Published online: 27 June 2014

Abstract This paper introduces information-theoretic measure of complexity (ICOMP) criterion for model selection in multivariate adaptive regression splines (MARS) to tradeoff efficiently between how well the model fits the data and the model complexity. As is well known, MARS is a popular nonparametric regression technique used to study the nonlinear relationship between a response variable and the set of predictors with the help of piecewise linear or cubic splines as basis functions. A critical aspect in determining the form of the nonparametric regression model during the MARS strategy is the evaluation of portfolio of submodels to select the best submodel with the appropriate number of knots over subset of predictors. In the usual regression modeling, when a large number of predictor variables are present in the model, and there is no precise information about the exact functional relationships among the variables, many model selection criteria still overfit the model. In this paper, to find the simplest model that balances the overfitting and underfitting for the model, ICOMP is proposed as a powerful model selection criterion for MARS modeling. Here, the model complexity is treated with respect to the interdependency of parameter estimates, as well as the number of free parameters in the model. We develop and study the performance of ICOMP along with several most popular model selection criteria such as Akaike's information criterion, Schwarz's Bayesian information criterion and generalized cross-validation in MARS modeling to select the best subset models. We provide two Monte Carlo simulation examples and a real benchmark example to demonstrate the utility and versatility of the proposed model selection approach to determine best functional form of the predictive model. Our numerical examples show that ICOMP provides a general model selection criterion with an insight to the interdependencies and/or correlational structure between parameter esti-

Editors: Vadim Strijov, Richard Weber, Gerhard-Wilhelm Weber, and Süreyya Ozogur Akyüz.

E. Kartal Koc (✉) · H. Bozdogan
Department of Statistics, Operations, and Management Science,
The University of Tennessee, Knoxville, TN 37996, USA
e-mail: kartalelcin@gmail.com

H. Bozdogan
e-mail: bozdogan@utk.edu

mates in the selected model. This new approach can also be applicable to many complex statistical modeling problems.

Keywords Model selection · Multivariate adaptive regression Splines (MARS) · Nonparametric regression · Information complexity

1 Introduction

In high dimensional data modeling, multivariate adaptive regression splines (MARS) is a popular nonparametric regression technique used to study the nonlinear relationship between a response variable and the set of predictor variables with the help of splines. MARS uses piecewise linear or cubic splines for local fit and applies an adaptive procedure to select the final model. MARS can be viewed as a generalization of stepwise linear regression or modification of the classification and regression trees (CART) to improve further CART's performance in the regression modeling (Friedman 1991).

In passing, we note that the underlying idea of MARS modeling appears to be similar to the *group method of data handling* (*GMDH*) which is a combinatorial heuristic, developed by Ivakhnenko dating back to 1966 (Ivakhnenko 1966), a Ukrainian cyberneticist, which constructs a mathematical model of a system in an evolutionary fashion. The algorithm is designed to model the functional relationship between the response and predictor variables which are learned directly from a self-organization of the data. It constructs high order regression type models beginning with a few basic quadratic equations and constructing a high-order polynomial of the Kolmogorov–Gabor type. The difference between MARS and GMDH is that, MARS uses piecewise linear or cubic splines instead of quadratic polynomials in several variables. For more on GMDH, we refer the readers to Ivakhnenko (1966) and Hild and Bozdogan (1995).

The popularity of MARS as a nonparametric modeling tool can be seen in its successful applications in many cross-disciplinary fields such as in medical research in detecting disease-risk relationship differences among gender subgroups (York et al. 2006), in studies of HIV reverse transcriptase inhibitors (Xu et al. 2004), in breast cancer diagnosis (Chou et al. 2004); in business in mining the customer credit (Lee et al. 2006), and in intrusion detection systems (Mukkamala et al. 2006); in molecular biology in chromatographic retention prediction of peptides (Put and Vander Heyden 2007), and many others, to mention a few.

In terms of most recent new algorithmic developments on the performance of MARS, in the literature, we see the uses of genetic algorithm for knot selection in Pitmann and McCulloch (2002). In the study of Weber et al. (2012), MARS algorithm is modified by introducing penalized residual sum of squares, and the problem is solved as Tikhonov regularization problem with conic quadratic optimization (Weber et al. 2012). In the studies of Yazici (2011) and Ozmen et al. (2011), the complexity of the method proposed in Weber et al. (2012) is reduced by bootsrapping and the capability of the method is enhanced to handle random input and output variables by robust method, respectively. Further, a time efficient forward selection procedure is proposed in Kartal Koc and Iyigun (2013) for MARS modeling.

A critical aspect in determining the form of the non-parametric regression model during the MARS strategy is the evaluation of submodels to select the best one with proper number of knots over the best subset of predictors. In the model fitting process, the function estimation is basically generated via a two-step procedure: *forward selection* and *backward elimination*. At each forward step, a candidate term (spline function) that most improves the overall 'goodness-of-fit' of the fitted model is added to the model. As discussed in Friedman (1991), at the end of this step there may be model terms that no longer sufficiently contributes to the model fit. Thus, by a backward step, the candidate term that least degrades the overall ’goodness-of-fit’ of the fitted model is eliminated from the model. In this respect, evaluation and selection of relevant subset of predictor variables with corresponding proper knots are the main concern of MARS to reduce the *curse of dimensionality*.

The problem of selecting the best spline functions, which are treated as the inputs, in MARS is solved by Friedman through a stepwise procedure using the modified generalized cross-validation (GCV) (not accounting for the selection bias) of Craven and Wahba (1979). Although Friedman avoids the overfitting problem in MARS by the modified GCV, in the literature, questions have been raised whether the modified GCV criterion is the ’best’ criterion for model selection in the MARS algorithm.

In the literature, initially Stevens (1991) appears to be the first to apply Akaike’s information criterion (AIC) (Akaike 1974), AIC’s modification (Akaike 1979), Schwarz Bayesian criterion (SBC) (Schwarz 1978), Amemiya’s prediction criterion (PC) (Amemiya 1980) including modified GCV in MARS for modeling the univariate and semi-multivariate time series systems. Although, these criteria are specifically designed for model selection and not just for the estimation of risk as stated in Barron and Xiao (1991), in regression modeling, when a large number of predictor variables are presented to the model, and there is no precise information about the exact relationships among the variables, such criteria still overfit the model. In addition, the complexity of a model increases as the number of independent and adjustable parameters (i.e., effective degrees of freedom of the model) increase. According to the qualitative principle of Occam’s Razor, we need to find the simplest model that judiciously balances overfitting and under-fitting of the model. To achieve this in MARS, our major objective is to introduce and develop for the first time Bozdogan’s information-theoretic measure of complexity (ICOMP) criterion (“I” for information and “COMP” for complexity) (Bozdogan 1988, 1990, 1994, 2000) within the MARS modeling framework .

In contrast to AIC-based information criteria, ICOMP approximates sum of two Kullback and Leibler (1951) distances that measures the lack of fit of the model and the model complexity in one criterion function using an entropic measure of the estimated covariance matrix of the model parameters. In this sense, the concept of model complexity here takes into account not only the number of free parameters in the model but also the interdependency of parameter estimates. Hence, a general model selection criterion with an insight to the correlational structure between parameter estimates in the selected model can be provided by ICOMP. Using ICOMP, also a better tradeoff between how well the model fits the data and the model complexity is achieved for MARS modeling. In addition, our objective also is to carry out a comprehensive Monte Carlo simulation study to compare the performance of the model selection criteria such as ICOMP, AIC, SBC, and GCV in MARS modeling which to our knowledge does not exist in the literature.

This paper is organized as follows. In Sect. 2, requisite background on MARS modeling and GCV criterion are given. Section 3 provides analytical derived forms of ICOMP based on the estimated inverse Fisher information matrix (IFIM) and the estimated posterior utility form of ICOMP along with the derived forms of AIC and SBC in MARS modeling. As an alternative to Tikhonov regularization, in this section we introduce a new and novel smoothed (or robust) covariance estimation procedures to resolve the problem of ill-conditioned model covariance matrices in MARS modeling and also use an eigenvalue stabilization method given in Thomaz (2004). In Sect. 4, performances of the model selection criteria in selecting the best subset of models are shown and studied via two Monte Carlo simulations, and on a real dataset to predict the body fat in obesity studies. Section 5 concludes the paper with a discussion and provides future directions in MARS modeling research.

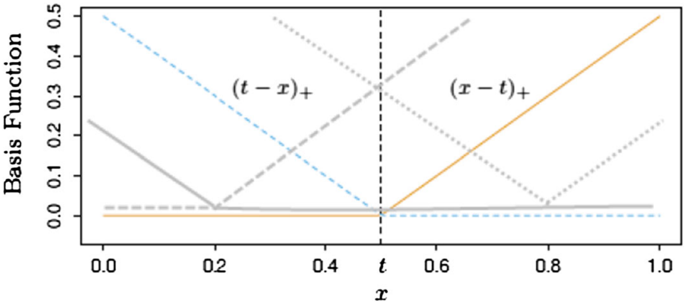


Fig. 1 The forms of BFs in MARS

2 Multivariate adaptive regression splines

MARS is developed by Friedman (1991) as a nonparametric regression technique to approximate a general type of model,

$$y = f(\mathbf{x}) + \varepsilon, \tag{1}$$

where, ε indicates the error term, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$ denotes the p number of predictor variables, and y is a response variable.

To approximate the nonlinear relationship between predictor variables, $\boldsymbol{x}$ and response variable, y, a flexible model estimate is provided using piecewise linear basis functions (BFs) of the form,

$$(x-t)_+ = \begin{cases} x-t, & if x > t \\ 0 & otherwise \end{cases} \text{ and } (t-x)_+ = \begin{cases} t-x, & if x < t \\ 0 & otherwise \end{cases}$$

where, the “+” means positive part.

As an example for univariate variable, x, the piecewise linear BFs (also called reflected pairs) for $t = 0.5$ are shown in Fig. 1, where t denotes the knot point (or breaking point).

The idea of MARS is to form reflected pairs for each predictor variable, x_j, $j \in \{1, \ldots, p\}$ with knots at each observed value, x_{ij}, $i \in \{1, \ldots, n\}$ of that variable, where n is the sample size. For the example given in Fig. 1, two other possible BFs with knots at $t = 0.2$ and $t = 0.8$ are displayed by shadow lines. The set of all possible reflected pairs with the corresponding knots, therefore, can be expressed by the set $\mathcal{S}$ in (2).

$$\mathcal{S} = \{(x_j - t)_+, (t - x_j)_+ | t \in \{x_{1j}, x_{2j}, \ldots, x_{nj}\},\ j \in \{1, \ldots, p\}\}. \tag{2}$$

The model building strategy of MARS is similar to the one developed in classical linear regression. However, instead of the original predictor variables, MARS uses the functions in set $\mathcal{S}$ or their products. The form of the MARS model defined to approximate the function in (1) is defined as

$$f(\mathbf{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(\mathbf{x}), \tag{3}$$

where, $B_m(\mathbf{x})$ represents a BF from set $\mathcal{S}$ or product of two or more such functions, and M is the number of BFs in the current model (Friedman 1991; Friedman and Silverman 1989).

For multiple variable cases, the expression $B_m(\mathbf{x})$ in (3) can also incorporate interactions between predictors. The interaction terms are created in MARS by multiplying an existing BF with a truncated linear function involving a new variable. Hence, product of two BFs

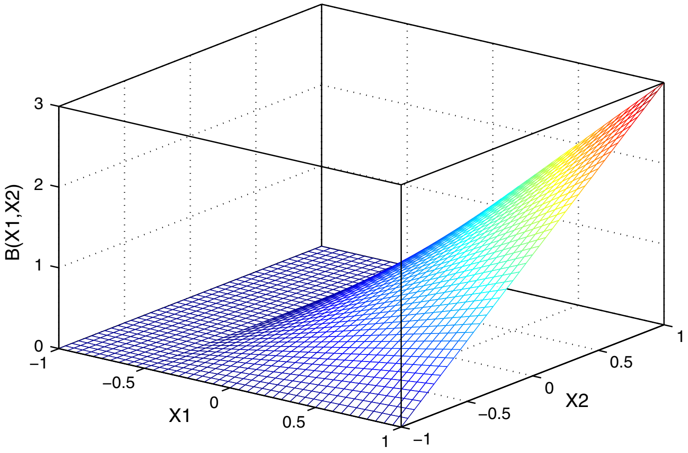


Fig. 2 Two-way interactions BFs

produces a result which is nonzero only over the space of predictors where both components are nonzero. In Fig. 2, the form of the function $B(x_{1,}x_2)$ resulted from the multiplication of two piecewise linear functions $(x_1 + 0.5)_+$ and $(x_2 + 1)_+$ is illustrated.

An example of MARS models built by piecewise linear and cubic splines for a two-dimensional noise-free function given by

$$y = f(x_1, x_2) = \sin(2\pi x_1)\cos(1.25\pi x_2) \tag{4}$$

are shown in Fig. 3a, b, respectively. The regression surface is build by using only nonzero components which locally obtained from the product of two BFs only when they are needed (Hastie et al. 2001).

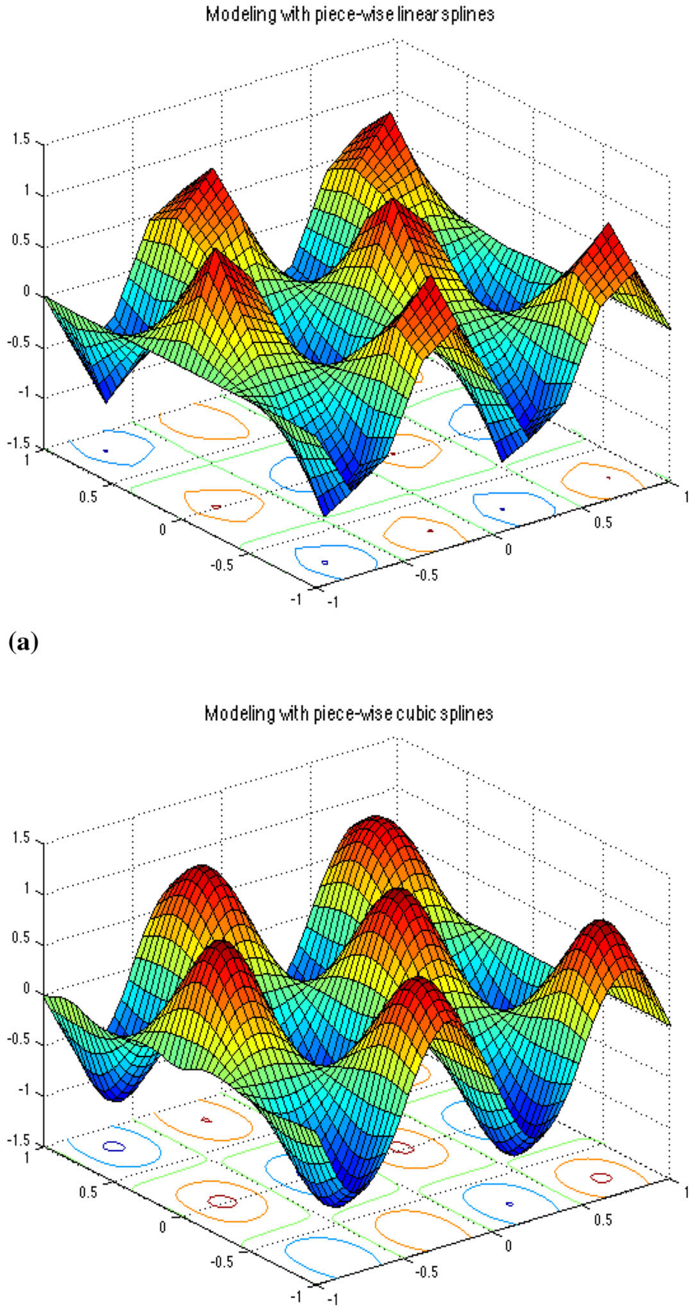
2.1 Traditional model selection in MARS with GCV

MARS builds a model by searching over all combinations of the variables and all values of each variable as the candidate knots through an adaptive procedure including a two-stage process: forward selection and backward elimination.

In the forward step, the algorithm starts with a model consisting of intercept term, β_0 and then the reflected pairs that give the maximum reduction in sum-of-squares residual error are added to the model iteratively until the maximum number of terms specified by the user is reached. Each new BF consists of a term already in the model multiplied with a new truncated linear function. At the end of this step, a large model typically overfitting the data is obtained. Figure 4a illustrates a simple example of how MARS would attempt to fit data during the forward step in a two dimension space using piecewise linear regression splines.

Following the forward-step, a backward elimination is implemented to refine the model fitting process. In this pruning step, the BFs contributing less to the model are eliminated step by step through modified GCV (Craven and Wahba 1979) until the best submodel is found. GCV depends on the idea of minimizing the average-squared residuals of the fit of the model given by

$$GCV(M) = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(y_i - \hat{f}_M(\mathbf{x}_i)\right)^2}{(1 - P(M)^*/n)^2}, \tag{5}$$



Modeling with piece-wise linear splines

(a)

Modeling with piece-wise cubic splines

(b)

Fig. 3 The plots of the piecewise linear and cubic types of MARS models. **a** Piecewise linear approximation, **b** Piecewise cubic approximation

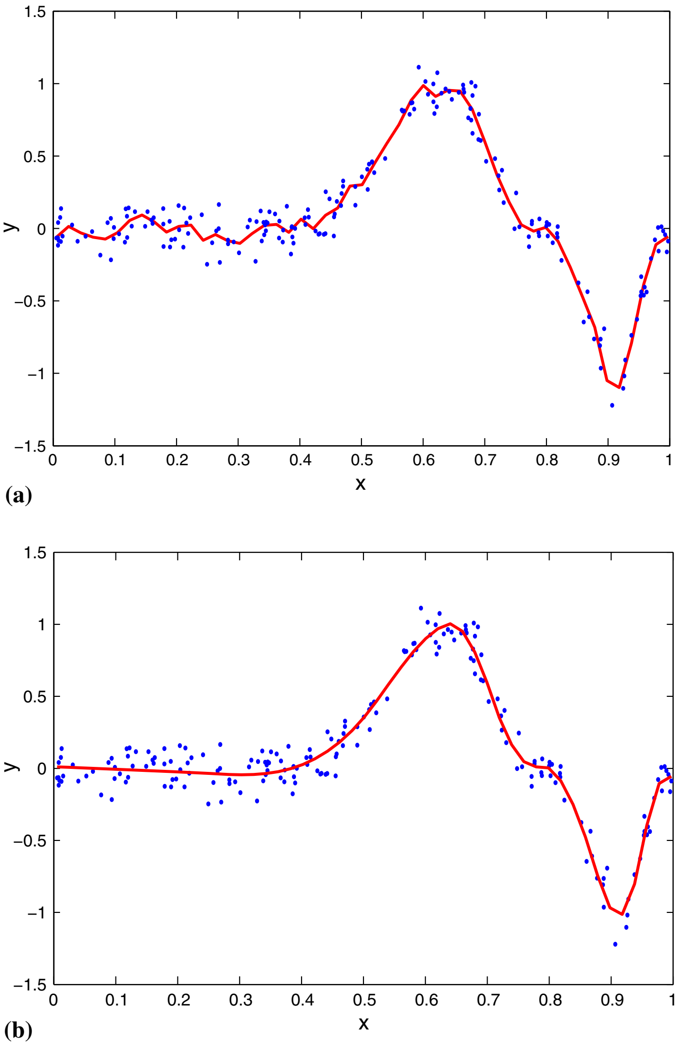


Fig. 4 Examples of MARS models fitted after forward and backward steps. **a** Forward-step model, **b** backward-step model

where, y_i is the ith observed response value; $\hat{f}_M(x_i)$ is the fitted response value obtained for the ith observed predictor vector, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = (1, .., n)$, n is the number of observations, and M represents the maximum number of BFs in the model.

In general, $P(M)$ is calculated by

$$P(M) = trace\left(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\right) + 1 \tag{6}$$

and it represents the cost penalty measure of a model, when there are M BFs in the model (Friedman 1991). In (6), $\mathbf{B}$ denotes the matrix of BFs with dimension $M \times n$.

Further, $P(M)$ in (6) represents the effective number of parameters which is a penalty measure for complexity. A modified form of $P(M)$ is used in the current MARS algorithm which is $P(M)^{*} = P(M) + dM$, where M is the number non-constant BFs in the MARS model. Note that, for an additive model, d is taken to be two, while it is taken to be three for an interaction model (Friedman 1991; Hastie et al. 2001). If the value of $P(M)$ is small, a large model including too many BFs is built. Otherwise, a smaller model is obtained. For a simple model with less lack-of-fit, the model with minimum GCV is chosen.

Figure 4b gives an example for a fitted model obtained after a backward step. As it is seen, the models obtained by the backward elimination step are smooth that keeps the fidelity of the data.

Friedman (1991) provides valuable insights into the use of GCV criterion for various types of MARS modeling. However, the criterion does not consider the complexity in terms of correlation within the model parameters.

3 ICOMP: a new information theoretic model selection criterion

In recent years, the statistical literature has placed more and more emphasis on information-based model selection and evaluation criteria. The necessity of introducing the concept of model evaluation has been recognized as one of the important technical areas, and the problem is posed on the choice of the best approximating model among a class of competing models by a suitable model evaluation criteria given a data set. Several of the popular model selection criteria have its underpinning to statistical information theory. They are based on the estimation of *Kullback-Leibler* information in high dimensions as a loss function (Kullback and Leibler 1951; Kullback 1968). The objective of information-based model selection criteria are to select a model that best incorporates the inference uncertainty (i.e., a measure of the lack-of-fit or badness-of-fit of the model) and parametric uncertainty (i.e., a measure of model parsimony and complexity).

Recently, based on Akaike's original AIC (Akaike 1973), many model-selection procedures which take the form of a penalized likelihood (a negative log likelihood plus a penalty term) have been proposed (Sclove 1987). For example, for AIC this form is given by

$$AIC(k) = -2\log L(\hat{\theta}_k) + 2k, \tag{7}$$

where, $L(\hat{\theta}_k)$ is the maximized likelihood function, $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter vector θ_k, and k is the number of independent parameters estimated. The first term in (7), $-2\log L(\hat{\theta}_k)$ is a measure of lack of fit, and $2k$ is the penalty term for the number of free parameters estimated in the model.

In AIC, a compromise takes place between the measure of lack-of-fit, and the number of parameters, which is considered as a measure of complexity that compensates for the bias in the lack-of-fit. The model with minimum AIC value is chosen as the best model to fit the data.

The use of AIC as a model selection criterion is popular because of its simplicity. However, it is well-known that in the context of complex modeling situations, AIC overfits the model order. In response to the over-fitting phenomenon in model selection, Schwarz (1978) introduced a Bayesian model selection criterion, abbreviated as SBC, assuming the data are generated from an exponential family of distributions. Independently, Rissanen (1978) introduced his minimum description length (MDL) criterion which takes the same form as SBC both defined as

$$MDL/SBC(k) = -2\log L(\hat{\theta}_k) + k\log(n). \tag{8}$$

Comparing with AIC, the SBC in (8) increases the penalty for adding additional terms to the model by a factor of $(1/2)\ln(n)$. In general, the model with minimum SBC or MDL is chosen as the best model to fit the data.

The development of ICOMP has been motivated in part by AIC, and in part by information complexity concepts and indices. In contrast to AIC, the new ICOMP procedure is based on the structural complexity of an element or set of random vectors via a generalization of the *information-based covariance complexity index* of Van Emden (1971).

A rationale for ICOMP as a model selection criterion is that it combines a badness-of-fit term (such as minus twice the maximum log likelihood) with a measure of complexity of a model differently than AIC, or its variants, by taking into account the interdependencies of the parameter estimates as well as the dependencies of the model residuals. The general form of ICOMP is based on the quantification of the concept of overall model complexity in terms of the *estimated inverse-Fisher information matrix* (IFIM). This approach results in an approximation to the sum of two Kullback-Leibler distances Kullback and Leibler (1951).

In contrast to AIC and SBC, ICOMP is designed to estimate a loss function given by Bozdogan (2004) as

$$Loss = Lack\ of\ Fit + Lack\ of\ Parsimony + Lack\ of\ Complexity \tag{9}$$

This is achieved by using the additivity property of information theory and the entropic developments in Rissanen (1976) in his final estimation criterion (FEC) in estimation and model identification problems, as well as AIC (Akaike 1973) and its analytical extensions in Bozdogan (1987). In the loss function in (9), by the third term, profusion of complexity, we mean the interdependencies or the correlations among the parameter estimates and the random error term of a model.

We define the general form of ICOMP as

$$ICOMP(k) = -2\log L(\hat{\theta}_k) + 2C\left(\hat{\Sigma}_{model}\right), \tag{10}$$

where $L(\hat{\theta}_k)$ is the maximized likelihood function, $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter vector θ_k, and C represents a real-valued complexity measure. In (10), $\hat{\Sigma}_{model} = \hat{C}ov(\hat{\theta})$ represents the estimated covariance matrix of the parameter vector of the model. This covariance matrix can be estimated in several ways, one of which uses celebrated Cramer-Rao lower bound (CRLB) matrix. The form of the estimated inverse Fisher information matrix (IFIM) of the model is obtained from

$$\hat{\mathcal{F}}^{-1} = \left\{-E\left(\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}\right)_{\hat{\theta}}\right\}^{-1}. \tag{11}$$

In (11), the expression in bracket is the matrix of second partial derivatives of the log-likelihood function of the fitted model evaluated at the maximum likelihood estimators. For more on IFIM, we refer the readers to Cramér (1946) and Rao (1945, 1947, 1948). By the estimated IFIM, an inherent measure of uncertainty or a precise measure of accuracy of the parameters which is estimated from the available data can be provided. The diagonal elements of IFIM contain the estimated variance of the estimated parameters, while the corresponding off-diagonals contain their covariances. Thus, ICOMP provides a universal criterion with IFIM which takes into account the entire parameter space of the model.

There are several forms and justifications of ICOMP based on (10) discussed in Bozdogan (1988, 1990, 2000, 2004, 2010) and Bozdogan and Bearse (1998). Here, we present only two of the general forms of ICOMP to be used in MARS modeling and show their derived analytical forms in the next section.

3.1 ICOMP based on estimated inverse Fisher information matrix (IFIM)

For a multivariate normal linear or nonlinear structural model, based on IFIM, ICOMP in (10) is defined as

$$ICOMP(IFIM) = -2\log L(\hat{\theta}_k) + 2C_1\left(\hat{\mathcal{F}}^{-1}(\hat{\theta}_k)\right), \tag{12}$$

where, C_1 denotes the *maximal entropic complexity* of the estimated IFIM, given by

$$C_1(\hat{\mathcal{F}}^{-1}) = \frac{s}{2}\log\left[\frac{tr(\hat{\mathcal{F}}^{-1})}{s}\right] - \frac{1}{2}\log|\hat{\mathcal{F}}^{-1}|. \tag{13}$$

In (13), s refers to the dimension or the rank of $\hat{\mathcal{F}}^{-1}$.

After some work, for a MARS model under the consideration that the random noise is normally distributed, the estimated IFIM is obtained as

$$\hat{C}ov(\hat{\beta}, \hat{\sigma}^2) = \hat{\mathcal{F}}^{-1} = \begin{bmatrix} \hat{\sigma}^2(B'B)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}, \tag{14}$$

where

$$\hat{\beta} = (B'B)^{-1}(B'y) \ and \ \hat{\sigma}^2 = \frac{(y - B\hat{\beta})'(y - B\hat{\beta})}{n}.$$

Using the definition in (12), ICOMP(IFIM) becomes

$$ICOMP(IFIM) = n\ln(2\pi) + n\ln(\hat{\sigma}^2) + n + 2C_1(\hat{\mathcal{F}}^{-1}(\hat{\theta}_M)), \tag{15}$$

where, the C_1 complexity is given by

$$C_1(\hat{\mathcal{F}}^{-1}(\hat{\theta}_M)) = (M+1)\log\left[\frac{tr\hat{\sigma}^2(B'B)^{-1} + \frac{2\hat{\sigma}^4}{n}}{M+1}\right] - \log|\hat{\sigma}^2(B'B)^{-1}| - \log\left(\frac{2\hat{\sigma}^4}{n}\right). \tag{16}$$

In (14), as the number of free parameters increases (i.e. as the size of B increases), the error variance $\hat{\sigma}^2$ gets smaller even though the complexity gets larger. Also, as $\hat{\sigma}^2$ increases, $(B'B)^{-1}$ decreases. Therefore, the use of $C_1(\hat{\mathcal{F}}^{-1}(\hat{\theta}_M))$ in information-theoretic model evaluation criteria achieves a trade-off between these two extremes and guards against the presence of multicollinearity. With ICOMP(IFIM), complexity is defined as the measure of the interaction or the dependency between its components. Hence, ICOMP(IFIM) provides a more judicious penalty term than AIC and SBC, and chooses simple models that provide more accurate and efficient parameter estimates over more complex models.

3.2 ICOMP as an estimate of posterior expected utility

By introducing some utility functions for both the lack of fit component of the model and the complexity of the parameters space of the model, a new class of ICOMP(IFIM) criteria are developed as a Bayesian criterion in maximizing a posterior expected utility (PEU) (Bozdogan and Haughton 1998; Bozdogan 2010). The idea of using two utility functions U_1 and U_2 that are multiplied to define a utility U whose posterior expectation is maximized to select a model is also considered by Poskitt (1987) and others. Poskitt defines $\log(U_1) = KL$, where

KL is Kullbrack-Liebler information, and it is considered as utility function by many authors. See Chaloner and Verdinelli (1995) for more details.

In this paper, a version of ICOMP(IFIM) derived from the multiplication of the utility U_1 by a utility U_2 equal to

$$U_2 = \exp\left[-\frac{k}{2}\log(n) - C_1(\hat{\mathcal{F}}^{-1})\right] \tag{17}$$

is used. With the choice of these utility functions, a more consistent ICOMP(IFIM) criterion whose formulation given in (18) is proposed and used in MARS modeling. This criterion provides a severe penalization for the overparametrization. Thus, simplest MARS models are chosen whenever there is nothing to be lost by doing so. This is very crucial for determining the nonlinear relationship between inputs and output variables with high multicollinearity.

$$ICOMP(IFIM)_{PEU} = -2\log L(\hat{\theta}_k) + k(1 + \log(n)) + 2C_1\left(\hat{\mathcal{F}}^{-1}(\hat{\theta}_k)\right), \tag{18}$$

3.3 Robust covariance estimation

In regression modeling, covariance matrices of the parameter estimates can often be ill-conditioned. That is, singularity occurs and that the condition number becomes very large. This is especially valid in the case of high multicollinearity between predictors, and that fact is usually indispensable in MARS modeling. To remedy the manifestation of the singular solutions, as an alternative to Tikhonov regularization (Taylan et al. 2010), we propose a new regularization of the covariance matrix of the parameter estimates in MARS modeling by adjusting the eigenvalues of the estimated covariance matrix, $\hat{\Sigma}$ given by

$$\Sigma^* = \hat{\Sigma} + \alpha I_p, \tag{19}$$

where, I_p is the p dimensional identity matrix. This is often called the "naive" ridge regularization.

Usually, the ridge parameter, α, is chosen to be very small. For different ridge parameters, many smoothed (or robust) covariance estimators have been developed as a way to data-adaptively improve ill-conditioned and/or singular covariance matrix in MARS. Several of these smoothed covariance estimators perturb the diagonals, and hence, the eigenvalues enough to achieve well-conditioned covariance matrix. In this study, we propose to use Maximum Likelihood/ Empirical Bayes (MLE/EB) given in (20), Stipulated Ridge (SRE) (Shurygin 1983) given in (21), and Thomaz Stabilization method (Thomaz 2004) given in (23).

MLE/EB:

$$\hat{\Sigma}_{MLE/EB} = \hat{\Sigma} + \frac{p-1}{ntr(\hat{\Sigma})} I_p, \tag{20}$$

SRE:

$$\hat{\Sigma}_{SRE} = \hat{\Sigma} + p(p-1)(2ntr(\hat{\Sigma})^{-1} I_p \tag{21}$$

Thomaz Stabilization:

$$\hat{\Sigma}_{Thomaz} = V\Lambda^* V, \tag{22}$$

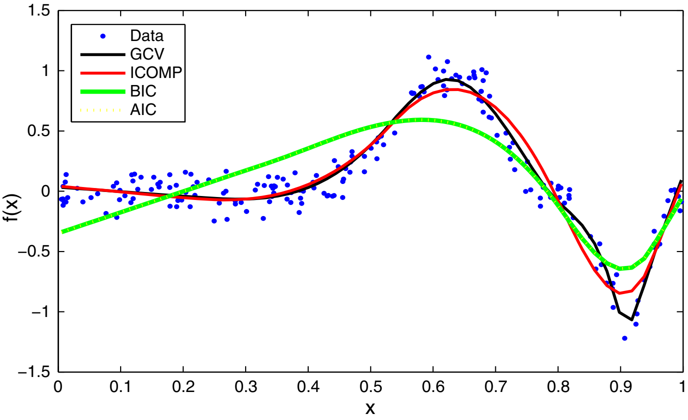


Fig. 5 MARS fits obtained by different model selection criteria

where,

$$\Lambda^* = \begin{bmatrix} \max(\lambda_1, \bar{\lambda}) & 0 & \cdots & 0 \\ 0 & \max(\lambda_2, \bar{\lambda}) & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \max(\lambda_p, \bar{\lambda}) \end{bmatrix} \tag{23}$$

where λ_i is the ith eigenvalue, $\bar{\lambda}$ is the arithmetic mean of the eigenvalues, and V is the matrix of eigenvalues.

4 Numerical examples

In this section, ICOMP(IFIM)PEU criterion given in (18) is implemented for MARS modeling, and its performance on best subset of variable selection is compared with that of GCV, AIC and SBC criteria using Monte Carlo simulations. We also show our results on a real dataset to predict body fat in obesity studies. As mentioned before, model selection criterion has an important effect on the selection of proper knots and influential variables over the response variable. In Fig. 5, we illustrate how different models can be fitted by using different model selection criteria. Since the selected number and locations of knots are different over the same variables, different forms of MARS models may be obtained for the same underlying dataset. Because of this fact, MARS modeling is studied under the model selection framework. It is note that, the model selection criteria studied in this paper are implemented to MARS algorithm using ARESLab (Jekabsons 2011) toolbox written entirely in MATLAB$^{(R)}$ (2010) environment. This toolbox uses the main functionality of MARS technique described by Friedman (1991).

To carry out a subset selection of variables, two Monte Carlo simulation protocols are implemented. The first protocol includes a nonlinear functional form between predictors and response, while the second simulation protocol refers collinear variable structure. MARS models are built for 100 different datasets generated using the same function in each protocol. To provide some insight regarding the importance of variables as predictors over the dependent variable, and to see whether the true model can be selected or not, the final MARS model fit is analyzed by ANOVA decomposition form given in (24).

$$\hat{f}(\mathbf{x}) = \beta_0 + \sum_{k_m=1} \beta_m B_m(x_i) + \sum_{k_m=2} \beta_m B_m(x_i, x_j) + \sum_{k_m=3} \beta_m B_m(x_i, x_j, x_k) + \cdots \qquad (24)$$

MARS refits the model after removing all terms involving the variable to be assessed and calculates the reduction in goodness of fit. All variables are then ranked according to their impact on goodness of fit. By the ANOVA decomposition in (24), it is possible to identify which variables enter to the model, whether they are purely additive, or are involved in interactions with other variables. In (24), the first term in ANOVA decomposition represents only the main effects, while the second and third terms reflect two-way and three-way interactions, respectively. The other terms denotes four-way interaction terms or etc. For each MARS model, the resulting ANOVA decomposition is examined to see whether the correct subset of models can be selected or not. Furthermore, the prediction and accuracy performances of the final models selected by each criterion are evaluated using the measures such as mean squared error (MSE) including residual sum of squares and number of terms in the model and multiple coefficient of determination (R^2).

4.1 Monte Carlo simulation: example 1

In over first Monte Carlo simulation study, the performance of ICOMP(IFIM)PEU criteria is demonstrated on a simulated dataset using a nonlinear function given in Friedman (1991). We start by creating datasets using a ten-dimensional function with Gaussian noise. The data consists of a 10-dimensional unit hypercube ($x_i = rand(0, 1),\ i = 1, \ldots, 10$).

$$y = 10sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + 0.5\varepsilon, \qquad (25)$$

where, $\varepsilon \sim N(0, 1)$, the standard normal distribution.

Note that, while the first three variables are nonlinear in function in (25), the next are linear to the output, and the last 5 variables have no effect on the response y. Therefore, true model includes the predictors x_1, x_2, x_3, x_4 and x_5.

For this simulation protocol, the maximal number of basis functions (BFs) is set to 21 including the intercept term, and maximum interaction level is limited to 2. That is, only pairwise products of BFs are allowed. The model is piecewise linear type. MARS algorithm is applied under these specifications using GCV, AIC, SBC and ICOMP(IFIM)PEU. An example of MARS model obtained through GCV criterion for a data generated with 200 observations is illustrated in Table 1.

The final MARS model in Table 1 includes 15 BFs including the main effects of $\{x_1, x_2, x_3, x_4, x_5\}$ and interaction terms x_1x_2 and x_1x_3. The ANOVA decomposition of the corresponding MARS model is given in Table 2. As described in the paper of Friedman (1991), the first column lists the function number. Each function is a sum of overall basis functions involving only the predictor in the last column. The second column gives the standard deviation of the functions. This gives us an indication of the (relative) importance of the corresponding function to the overall model and can be interpreted in a manner similar to a standardized regression coefficient in a linear regression model. The third column lists the GCV score for a model obtained by removing the entire basis functions corresponding to that particular ANOVA function. Hence, whether this ANOVA function is making an important contribution to the model, or whether it just slightly helps to improve the global GCV score can

Table 1 Mars equation

$$y = 8.43 + 10.4BF1 + 18.5BF2 - 20.3BF3 + 17.4BF4 + 4.13BF5 - 5.39BF6 - 99.8BF7 - 28.1BF8 + 18.4BF9 + 10.5BF10 - 20BF11 + 18.5BF12 + 9.55BF13 + 1.83BF14 + 25.6BF15,$$

where,

$$BF1 = \max(0, x_4 - 0.0504),\ BF2 = \max(0, x_2 - 0.569),\ BF3 = \max(0, 0.569 - x_2),$$

$$BF4 = \max(0, x_3 - 0.502),\ BF5 = \max(0, x_5 - 0.582),\ BF6 = \max(0, 0.582 - x_5),$$

$$BF7 = BF2 \times \max(0, x_1 - 0.515),\ BF8 = \max(0, 0.938 - x_1) \times \max(0, x_2 - 0.878),$$

$$BF9 = \max(0, 0.938 - x_1) \times \max(0, 0.878 - x_2),\ BF10 = \max(0, x_1 - 0.749),$$

$$BF11 = \max(0, 0.749 - x_1)\ BF12 = \max(0, x_3 - 0.927),\ BF13 = \max(0, 0.927 - x_3),$$

$$BF14 = \max(0, 0.938 - x_1) \times \max(0, x_3 - 0.125),$$

$$BF15 = \max(0, 0.938 - x_1) \times \max(0, 0.125 - x_3).$$

Table 2 ANOVA decomposition of MARS model selected with GCV criterion

Func.	Std	GCV	#Basis	#Params	Variables
1	5.445	99.198	2	5.0	1
2	5.852	59.167	2	5.0	2
3	1.388	57.831	3	7.5	3
4	2.969	45.321	1	2.5	4
5	1.442	4.304	2	5.0	5
6	4.967	47.939	3	7.5	1 2
7	0.367	1.108	2	5.0	1 3

be judged. The fourth column gives the number of BFs comprising the ANOVA function while the fifth column provides an estimate of the additional number of linear degrees-of-freedom. The last column gives the particular predictor variables associated with the ANOVA function.

In this example, the MARS model in Table 1 selects $\{x_1, x_2, x_3, x_4, x_5, x_1x_2, x_1x_3\}$ as the best subset model using GCV criterion. The first two ANOVA functions (corresponding x_1 and x_2) give the largest contribution to the model, as well as the effect of interaction between x_1 and x_2. The last ANOVA function which corresponds to the BFs developed for the predictors x_1 and x_3 gives very small contribution. Hence, the term may be removed from the model.

MARS models are built using ICOMP(IFIM)PEU, AIC, SBC and GCV for 100 replication of the above simulation protocol with different sample sizes. The performances of the model selection criteria in selecting the best subset of predictors are analyzed in terms of percentage hits over 100 trials through ANOVA tables as in Table 2. In Table 3, the percent number of hits are given for only three types of models (in column 2). The first one refers to the models including exactly the true predictors $\{x_0, x_1, x_2, x_3\}$. The second one determines the models in which the relative importance of the true predictors is more than 90 %, and the last one describes the models for which at least one of the true predictors is not selected by the criteria. Note that, the percent contribution is calculated over the standard deviations of the functions given in ANOVA decomposition shown in Table 2. Depending on the results in Table 3, the following conclusions can be drawn:

- For small sample size, $n = 50$, GCV criterion selects the true model with the highest frequency and less number of BFs. However, the increase in the rates at which GCV selects exactly the true model does not improve dramatically as the sample size increases.

Table 3 % model hits out of 100 simulations

Criteria	Models selected	Main effect and interaction terms			
		Sample size			
		50	100	200	500
GCV	1	45	36	76	100
	2	86	97	100	100
	3	13	3	0	0
	BFs	10.9	14.4	15.4	16.0
AIC	1	8	32	83	99
	2	79	96	100	100
	3	14	0	0	0
	BFs	14.8	15.9	16.1	16.3
SBC	1	11	32	73	99
	2	82	97	99	100
	3	11	0	1	0
	BFs	13.9	15.2	14.9	15.6
ICOMP(IFIM)PEU	1	32	42	83	100
	2	76	96	100	100
	3	20	0	0	0
	BFs	11.8	13.9	15.3	15.7

For $n \geq 100$, all criteria be able select the models including the all true predictors with about 100 %. In other words, the percent hit of the models for which at least one of the true predictors is not selected by the criteria is zero. The highest percent hit ratio of models with exact true predictors is achieved by ICOMP(IFIM)PEU for $n \geq 100$.

- When the relative importance of true predictors is examined, it is observed that the percent hit of models dominated by the true predictors and interaction terms is at least 96 % for all criteria, and for $n \geq 100$.
- For large sample size $n = 500$, all criteria are able to select the true model including predictors and interaction terms with 100 % for GCV and ICOMP(IFIM)PEU, and with 99 % for AIC and BIC.
- For $n \geq 100$, the simplest models, the ones with minimum average number of BFs, are selected by ICOMP(IFIM)PEU and SBC criteria.

In order to evaluate and compare the prediction (generalization) ability of MARS models, models are analyzed for 100 simulated train and test datasets genetared with n = 100 and n = 20 observations, respectively through MSE and R^2 measures. The corresponding results are shown in Table 4. Looking at Table 4, although the average values of the performance measures are close to each others for all criteria, the models selected by AIC show better performance both in training and testing datasets. However, the highest numbers of BFs in the final models belongs to AIC models. The simplest models including less number of BFs are selected by ICOMP(IFIM)PEU and GCV criterion.

Table 4 Comparison of MARS models with respect to model selection criteria for n = 100

Criteria	avgMSE		avgR^2		BFs
	Train	Test	Train	Test	
GCV	0.382	0.782	0.988	0.973	14.5
AIC	0.349*	0.737*	0.989*	0.975*	16.0
SBC	0.367	0.766	0.988	0.974	15.0
ICOMP(IFIM)PEU	0.389	0.806	0.987	0.972	13.9

* indicates better performance

4.2 Monte Carlo simulation: example 2

In this Monte Carlo simulation example, a different simulation protocol (Bozdogan 2004) is used in which highly collinear input variables are generated. The first five variables are simulated using the following protocol.

$$\begin{aligned} x_1 &= 10 + \varepsilon_1 \\ x_2 &= 10 + 0.3\varepsilon_1 + \alpha\varepsilon_2 \\ x_3 &= 10 + 0.3\varepsilon_1 + 0.5604\alpha\varepsilon_2 + 0.8282\alpha\varepsilon_3 \\ x_4 &= -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5 \\ x_5 &= -5 + 0.5x_1 + x_2 + 0.5\varepsilon_5 \end{aligned}$$

where, $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5 \sim N(0, \sigma^2 = 1)$. The parameter $\alpha = \sqrt{1 - 0.3^2}$ controls the degree of collinearity in the predictors. Then, the response variable y is generated from:

$$y = -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5\varepsilon, \tag{26}$$

where, ε is independent and identically distributed (i.i.d.) according to $N(0, \sigma^2 = 1)$ for $i = 1, 2, \ldots, n$.

Further, five redundant variables $x_6, \ldots, x_{10}$ are generated using the uniform random numbers given by $x_6 = 6 \times rand(0, 1), \ldots, x_{10} = 10 \times rand(0, 1)$ and a MARS model of y on $X = \{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ is fitted for different sample sizes n = 50, 100, 200, 500 and 1,000, where $x_0 = 1$ constant column of $(n \times 1)$ vector of ones.

In this simulation study, it is expected that model selection criteria would pick the set of variables $\{x_0, x_1, x_2, x_3\}$ to be the best subset through the MARS algorithm. Using the ANOVA decomposition analysis, model selection performances of criteria are analyzed by examining the predictors selected in the final model and their corresponding relative importance within the model.

In Table 5, the percentages of simulations in which the criteria select three types of models in 100 trials are given. The first row for each criterion refers to models including exactly the true predictors $\{x_0, x_1, x_2, x_3\}$. The second one determines the models in which the relative importance of the true predictors is more than 90 %, and the last one describes the models to which at least one of the true predictors is not selected via the criteria. Based on the results presented in Table 5, the following conclusions can be drawn:

- For small sample sizes, ICOMP(IFIM)PEU performs better in selecting the true set of predictors than others, although it misses the true subset in 44 trials. However, AIC only selects the models that do not include the true model with 16 % of the simulation. That

Table 5 % model hits out of 100 simulations

Criteria	Models selected	Sample size				
		50	100	200	500	1,000
GCV	1	18	18	14	15	12
	2	25	35	55	78	96
	3	33	7	0	0	0
	BFs	5.7	7.0	7.1	7.9	8.6
AIC	1	0	0	0	0	0
	2	0	3	6	38	69
	3	16	1	0	0	0
	BFs	13.6	12.9	12.9	12.8	12.9
SBC	1	3	21	30	45	50
	2	7	40	69	95	98
	3	18	0	0	0	0
	BFs	9.9	7.2	9.4	6.1	6.8
ICOMP(IFIM)PEU	1	23	55	66	84	85
	2	26	65	86	99	100
	3	44	7	0	0	0
	BFs	5.0	5.2	5.3	5.1	5.1

may be due to the fact that the models built by ICOMP(IFIM)PEU includes less number of BFs than the models of AIC and other criteria.

- As the sample size increases, the percent hit of models including exactly the true predictors is improved for ICOMP(IFIM)PEU, rising dramatically to 84 % for n=500. However, other criteria, especially GCV and AIC, cannot show such an improvement in selecting exactly the true model. In this respect, GCV performs much better than AIC and SBC for small sample sizes. However, as the sample size increase, it tends to pick models with extra predictor variables as it is always the case for AIC. AIC cannot select the exact true model for all sample sizes. The performance of SBC on selecting exactly the true model is improved slightly, rising to 50 % for n = 1,000.
- GCV and AIC show higher tendency to pick the models including extra variables other than true predictors. This result can be validated by examining the percent hits in picking the models in which true predictors have more than 90 % contribution. While the exact true model can be selected with 12 % by GCV, the rate of selecting models which are mainly dominated by the true predictors become 96 % for n = 1,000. In overall simulations, SBC selects models including 8.6 BFs, on the average. This may, however, indicate that there is no substantial tendency for SBC to overfit. It is difficult to draw the same conclusion for AIC. All models selected by AIC include the true model (e.g. $\{x_0, x_1, x_2, x_3\}$). For 31 trials, the contributions of true predictors are even less than 90 %, which may indicate over parameterization. This conclusion can also be supported by the excessive number of BFs in the final models.
- ICOMP(IFIM)PEU also performs very well in picking the models in which true predictors have more than 90 % contribution. Even if some extra variables are selected by ICOMP(IFIM)PEU, their contributions to the model is less than 10 % for at least 86 % of hits for $n \geq 200$.

Table 6 Comparison of MARS models selected by the corresponding model selection criteria for n=200

Criteria	avgMSE		avgR^2		BFs
	Train	Test	Train	Test	
GCV	0.251	0.347	0.894	0.838	7.4
AIC	0.241*	0.371	0.901*	0.832	12.2
SBC	0.253	0.340	0.893	0.837	6.7
ICOMP(IFIM)PEU	0.262	0.312*	0.888	0.844*	5.5

* indicates better performance

Table 7 List of variables for body fat data

y = Percent body fat from Siri's equation	
x_0 = Constant	x_7 = Hip circumference
x_1 = Age (years)	x_8 = Thigh circumference
x_2 = Weight (lbs)	x_9 = Knee circumference
x_3 = Height (in)	x_{10} = Ankle circumference
x_4 = Neck circumference	x_{11} = Extended biceps circum.
x_5 = Chest circumference	x_{12} = Forearm circumference
x_6 = Abdomen 2 circumference	x_{13} = Wrist circumference

The estimation and prediction performances of the MARS models are also evaluated for each criterion with respect to MSE and R^2. Table 6 shows that the models selected by AIC criterion for training datasets have better performance than the other models with respect to all performance measures. However, the prediction performance of the corresponding models for new datasets is not as good as ICOMP(IFIM)PEU models although including less number of BFs in the models. This may be because of the overfitting problem caused by the excessive number of BFs in AIC models.

4.3 A real benchmark example: prediction of body fat in obesity studies

The practical utility and the importance of the new class of model selection criteria proposed for MARS modeling is applied on a real body fat dataset to determine a predictive model in obesity studies. This data is composed of body measurement observations from n = 252 men. There are 13 regressors, listed in Table 7. A method for accurately computing the percent body fat from simple body measurements without requiring underwater weighing is highly desirable (Bozdogan and Howe 2012). Before the construction of models, variables are normalized to make them comparable, and the MARS algorithm is run for 30 maximum numbers of BFs with no interaction terms.

The equation of the MARS model built by ICOMP(IFIM)PEU criterion is given in Table 8. It is noted that, during the implementation of ICOMP criterion for body fat data, we encountered with the problem of ill-conditioned covariance matrices. The higher condition numbers obtained for the covariance matrices may refer to the existence of high correlation between the BFs selected. This situation may affect the model selection performance of the criteria. In order to make the ICOMP criterion reliable, the methods proposed in Sect. 3 are applied to the model covariance matrices. Since the method that can decrease the condition

Table 8 MARS model with ICOMP(IFIM)PEU

$$y = 12.8 - 5.93BF1 + 2.16BF2 - 1.55BF3 + 11.2BF4 - 6.68BF5 + 5.28BF6$$

where,

$$BF1 = \max(0, x_2 + 0.406),\ BF2 = \max(0, 0.397 - x_{13}),\ BF3 = \max(0, 1.13 - x_{11})$$

$$BF4 = \max(0, x_6 + 0.905),\ BF5 = \max(0, -0.74 - x_7),\ BF6 = \max(0, -0.945 - x_{10})$$

Table 9 ANOVA decomposition of MARS model selected with ICOMP(IFIM)PEU criterion

Func.	Std	ICOMP(IFIM)PEU	#Basis	#Params	Variables
1	4.747	1508.6	1	2.5	2
2	10.119	1656.1	1	2.5	6
3	1.544	1442.5	1	2.5	7
4	0.808	1432.1	1	2.5	10
5	1.357	1435.1	1	2.5	11
6	1.494	1439.1	1	2.5	13

Table 10 MARS model with GCV

$$y = 23.3 - 3.18BF1 + 2.23BF2 - 3.56BF3 - 1.78BF4 + 12.4BF5 - 18.9BF6+,4.89BF7+ 8.76BF8 - 13.5BF9 + 18.3BF10 - 8.71BF11 - 1.3BF12 - 10.6BF13 + 4.21BF14 + 13.2BF15.$$

where,

$$BF1 = \max(0, x_2 + 0.406),\ BF2 = \max(0, 0.397 - x_{13}),\ BF3 = \max(0, x_{11} - 1.13),$$

$$BF4 = \max(0, 1.13 - x_{11}),\ BF5 = \max(0, -0.905 - x_6),\ BF6 = \max(0, -0.74 - x_7),$$

$$BF7 = \max(0, -0.945 - x_{10}),\ BF8 = \max(0, x_6 - 0.0597),\ BF9 = \max(0, 0.0597 - x_6),$$

$$BF10 = \max(0, -0.202 - x_4),\ BF11 = \max(0, -0.861 - x_4),\ BF12 = \max(0, -0.15 - x_1),$$

$$BF13 = \max(0, 0.00326 - x_4),\ BF14 = \max(0, 0.0412 - x_7),\ BF15 = \max(0, -1.13 - x_7).$$

numbers is Thomaz regularization given in (23), this regularization method is applied for body fat dataset.

Minimizing ICOMP(IFIM)PEU leads to the best model as the one including weight, abdomen 2, hip, ankle, extended biceps and wrist circumferences. ANOVA decomposition table of the corresponding MARS model given in Table 9 shows that weight and abdomen 2 circumferences are the important predictors with the highest contributions.

The MARS model fitted using GCV criterion is given in Table 10. The final model includes 16 BFs with constant term. According to the ANOVA decomposition in Table 11, although the model includes 8 different variables, the highest contribution is supplied by weight and abdomen 2 circumferences according to GCV.

Once the MARS algorithm is run with AIC, the fitted model in Table 12 is obtained. The model includes 21 BFs including constant term, and the variables selected into the model are similar to the ones selected by GCV. However, the model selected by AIC is larger than the model selected by GCV. The main contribution is again supplied by two variables: weight and abdomen 2 circumferences (see Table 13).

Finally, MARS algorithm is applied to body fat data using SBC. The fitted model in Table 14 includes only 8 BFs including 6 variables: weight, abdomen 2, hip, ankle, extended

Table 11 ANOVA decomposition of MARS model selected with GCV criterion

Func.	Std	GCV	#Basis	#Params	Variables
1	0.644	19.39	1	2.5	1
2	2.551	32.35	1	2.5	2
3	1.284	20.01	3	7.5	4
4	9.590	140.31	3	7.5	6
5	1.515	20.81	3	7.5	7
6	0.748	19.28	1	2.5	10
7	1.500	28.58	2	5.0	11
8	1.540	24.28	1	2.5	13

Table 12 MARS model with AIC

$$
\begin{aligned}
y = 3.53 &- 2.25BF1 - 1.13BF2 + 2.12BF3 - 3.38BF4 - 1.83BF5 + 8.32BF6 + 4.38BF7 \\
&- 13.1BF8 - 6.82BF9 + 5.51BF10 + 4.01BF11 - 5.17BF12 + 11.3BF13 + 8.02BF14 \\
&- 9.09BF15 + 2.66BF16 - 5.81BF17 - 11.7BF18 + 4.88BF19 + 13.2BF20.
\end{aligned}
$$

where,

$BF1 = \max(0, x_2 + 0.406)$, $BF2 = \max(0, x_{13} - 0.397)$, $BF3 = \max(0, 0.397 - x_{13})$,

$BF4 = \max(0, x_{11} - 1.13)$, $BF5 = \max(0, 1.13 - x_{11})$, $BF6 = \max(0, x_6 + 0.905)$,

$BF7 = \max(0, -0.905 - x_6)$, $BF8 = \max(0, x_7 + 0.74)$, $BF9 = \max(0, -0.74 - x_7)$,

$BF10 = \max(0, -0.945 - x_{10})$, $BF11 = \max(0, 0.961 - x_1)$, $BF12 = \max(0, 0.0597 - x_6)$,

$BF13 = \max(0, x_4 + 0.202)$, $BF14 = \max(0, -0.202 - x_4)$, $BF15 = \max(0, -0.861 - x_4)$,

$BF16 = \max(0, x_1 + 0.15)$, $BF17 = \max(0, -0.15 - x_1)$, $BF18 = \max(0, x_4 - 0.00326)$,

$BF19 = \max(0, 0.0412 - x_7)$, $BF20 = \max(0, x_7 + 1.13)$.

Table 13 ANOVA decomposition of MARS model selected with AIC criterion

Func.	Std	AIC	#Basis	#Params	Variables
1	0.978	1448.8	3	7.5	1
2	1.802	1441.9	1	2.5	2
3	1.372	1439.3	4	10.0	4
4	9.302	1588.3	3	7.5	6
5	1.599	1435.3	4	10.0	7
6	0.843	1449.1	1	2.5	10
7	1.519	1467.5	2	5.0	11
8	1.762	1455.8	2	5.0	13

biceps and wrist circumferences. Again, the weight and abdomen 2 circumferences are the most significant variables in the model (see Table 15).

Overall, the weight and abdomen 2 circumferences are selected as the important predictors with the highest contributions by all criteria. As always the case, the model selected with AIC includes more BFs than the models selected by other model selection criteria. ICOMP(IFIM)PEU and SBC can capture the contribution of important variables with less number BFs.

Table 14 MARS model with SBC

$y = 10.9 - 3.73BF1 + 2.4BF2 - 1.61BF3 + 10.3BF4 - 11.4BF5 - 2.91BF6 + 1.98BF7 + 9.55BF8$

where,

$BF1 = \max(0, x_2 + 0.406)$, $BF2 = \max(0, 0.397 - x_{13})$, $BF3 = \max(0, 1.13 - x_{11})$,

$BF4 = \max(0, x_6 + 0.905)$, $BF5 = \max(0, x_7 + 0.74)$, $BF6 = \max(0, 0.0597 - x_6)$,

$BF7 = \max(0, -0.202 - x_4)$, $BF8 = \max(0, x_7 + 1.13)$.

Table 15 ANOVA decomposition of MARS model selected with SBC criterion

Func.	Std	SBC	#Basis	#Params	Variables
1	2.990	1422.8	1	2.5	2
2	0.952	1415.6	1	2.5	4
3	10.481	1642.2	2	5.0	6
4	1.590	1432.9	2	5.0	7
5	1.410	1423.9	1	2.5	11
6	1.659	1428.9	1	2.5	13

Table 16 Estimation and prediction performance of MARS models with the corresponding model selection criteria

Criteria	avgMSE		avgR^2		BFs
	Train	Test	Train	Test	
GCV	15.660	26.187	0.783	0.590	13
AIC	14.624*	27.986	0.805*	0.579	21.8
SBC	15.469	26.314	0.786	0.590	13.7
ICOMP(IFIM)PEU	17.052	23.126*	0.759	0.629*	7.8*

* indicates better performance

Table 16 gives the performances of MARS models obtained as a result of 10-fold cross validation study. The models selected with different criteria show close performances both for training and test datasets. Although AIC models perform better for training data, their prediction performance is not as good as the others. This is due to the fact that the model selected by AIC includes excessive number of BFs which causes an overfitting problem. On the other hand, the simplest models are selected by ICOMP(IFIM)PEU criterion, and its corresponding models have better prediction performances for new datasets.

5 Conclusion and discussion

In multivariate adaptive regression splines (MARS), comparison of the submodels during the backward elimination step plays a crucial role in the estimation of nonlinear relationship between predictors and output. By minimizing a model selection criterion, both the accuracy and the complexity of models can be controlled in each step of backward iterations. Most of the criteria in the literature such as AIC, BIC and GCV consider the complexity as the number of free parameters within a model, and determine the model dimension with an additional penalized term, a cost function of number of free parameters in the model. In this paper, however, a new information-based model selection criterion, ICOMP, is proposed to be used in MARS which also handles the interdependency of parameter estimates and the model complexity. ICOMP selects the best number of breaking points, and corresponding basis functions in MARS by taking into account the interaction or dependency between the components as well as the lack of model fit and model parsimony.

In this paper, the model selection performances of ICOMP is evaluated and compared with AIC, SBC, and GCV using two Monte Carlo simulation protocols and on a real body fat dataset. Following the results of first simulation protocol, over 100 simulation datasets, it is observed that the capability of ICOMP(IFIM)PEU for selecting models with exact true predictors is higher than the other criteria. For small sample size, GCV selects the true model with the highest frequency of all the criteria. However, the increase in the rates at which GCV selects exactly the true model does not improve dramatically as the sample size increases. For large sample sizes, all criteria are able to select the true model including main effects and interaction between true predictors with about 100 %.

With the second Monte Carlo simulation protocol, the model selection performances of criteria are evaluated for the datasets including high multicollinear structure. The results show that ICOMP(IFIM)PEU performs better in selecting the true set of predictors than the others for small sample sizes. As the sample size increases, the rates at which ICOMP(IFIM)PEU selected exactly the true model improved dramatically. AIC can never select the exact true model for all sample sizes. As well as GCV, AIC shows higher tendencies to pick the models including extra variables besides the true predictors. This conclusion can also be supported by the excessive number of BFs in the final models.

Overall, ICOMP(IFIM)PEU criterion can be used as a powerful criteria for the submodel selection of MARS algorithm due to its better performances on the selection of true models with less number of BFs and high generalization capability.

The existing forward selection and backward elimination procedures of MARS are computationally expensive and does not quarantee globally optimal solution. As it is observed in the simulation studies, MARS selects many redundant terms into the model. Although this can be prevented by a model selection criterion to some extent, it is not always possible to select the correct model due to the stepwise nature of MARS. In this respect, we will try to develop a data adaptive "open architecture" for model building via the intelligent Genetic Algorithm (GA) as our optimizer along with ICOMP criterion. In a future study, we shall develop and score the misspecified form of ICOMP criteria given in Bozdogan (2004).

Acknowledgments This research for the first author has been supported as a Visiting Post-doctoral Scholar by the Turkish Scientific and Technological Research Council of Turkey at the University of Tennessee under Professor Hamparsum Bozdogan.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrox, F. Csaki (Eds.) *Second International Symposium on Information Theory* (pp. 267–281). Academiai Kiado, Budapest.

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

Akaike, H. (1979). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, *30*, 9–14.

Amemiya, T. (1980). Selection of regressors. *International Economic Review*, *21*, 331–354.

Barron, A. R., & Xiao, X. (1991). Discussion: Multivariate adaptive regression splines. *Annals of Statistics*, *19*, 67–82.

Bozdogan, H. (1987). Model selection and akaike's information criterion: The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Bozdogan, H. (1988). Icomp: A new model-selection criteria. In H. Bock (Ed.), *Classification and related methods of data analysis*. Amsterdam, North-Holland: Elsevier Science Publishers.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communication in Statistics, Theory and Methods*, *19*, 221–278.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity, vol. 2. In H. Bozdogan (Ed.) *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* (pp. 69–113). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, *44*, 62–91.

Bozdogan, H. (2004). Intelligent statistical data mining with information complexity and genetic algorithms. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery*. Boca Raton, FL: Chapman and Hall/CRC.

Bozdogan, H. (2010). A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation. *Istanbul University Journal of the School and Business Administration*, *39*, 370–398.

Bozdogan, H., & Bearse, P. (1998). Subset selection in vector autoregressive models using the genetic algorithm with information complexity as the fitness function. *Systems Analysis Modeling and Simulation*, *31*, 61–91.

Bozdogan, H., & Haughton, D. (1998). Informational complexity criteria for regression models. *Computational Statistics and Data Analysis*, *28*, 51–76.

Bozdogan, H., & Howe, J. A. (2012). Misspecified multivariate regression models using the genetic algorithm and information complexity as the fitness function. *European Journal of Pure and Applied Mathematics*, *5*, 211–249.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design a review. *Statistical Science*, *10*, 273–304.

Chou, S. M., Lee, T. S., Shao, Y. E., & Chen, I. F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *27*(1), 133–142.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numerische Mathematik*, *31*, 377–403.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*, 1–61.

Friedman, J. H., & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modelling. *Technometrics*, *31*, 3–21.

Hastie, T. J., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning, data mining, inference and prediction*. New York: Springer.

Hild, C., & Bozdogan, H. (1995). The use of information-based model evaluation criteria in the GMDH algorithm. *Systems Analysis Modeling Simulation*, *20*, 29–50.

Ivakhnenko, A. G. (1966). Group method of data handling: A rival of the method of stochastic approximation. *Soviet Automatic Control*, *13*, 43–71.

Jekabsons, G. (2011). ARESLab: Adaptive regression splines toolbox for matlab/Octave. http://www.cs.rtu.lv/jekabsons/.

Kartal Koc, E., Iyigun, C. (2013). Restructuring forward step of mars algorithm using a new knot selection procedure based on a mapping approach. *Journal of Global Optimization*. doi:10.1007/s10898-013-0107-5.

Kullback, A., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover.

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113–1130.

Mukkamala, S., Sung, A. H., Abraham, A., & Ramos, V. (2006). Intrusion detection systems using adaptive regression spines. *Enterprise information systems VI* (pp. 211–218). Berlin: Springer.

Ozmen, A., Weber, G. W., & Batmaz, I. (2011). RCMARS: Robustification of CMARS with different scenarios under polyhedral uncertainty set. *Communications in Nonlinear Science and Numerical Simulation (CNSNS): Nonlinear, Fractional and Complex*, *16*, 4780–4787.

Pitmann, J., & McCulloch, C. (2002). Adaptive splines and genetic algorithms. *Journal of Computational and Graphical Statistics*, *11*(3), 615–638.

Poskitt, D. (1987). Precision, complexity and Bayesian model determination. *Journal of the Royal Statistical Society, Series B (Methodological)*, *49*(2), 199–208.

Put, R., & Vander Heyden, Y. (2007). The evaluation of two-step multivariate adaptive regression splines for chromatographic retention prediction of peptides. *Proteomics*, *7*(10), 1664–1677.

Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, *37*, 81–91.

Rao, C. (1947). Minimum variance and the estimation of several parameters. *Proceedings of the Cambridge Philosophical Society*, *43*, 280.

Rao, C. (1948). Sufficient statistics and minimum variance estimates. *Proceedings of the Cambridge Philosophical Society*, *45*, 213.

Rissanen, J. (1976). Minmax entropy estimation of models for vector process. In R. K. Mehra & D. G. Lainiotis (Eds.), *System identification* (pp. 97–119). New York: Academic Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.

Shurygin, A. (1983). The linear combination of the simplest discriminator and fisher's one. In Nauka (ed.) *Applied statistics*. Moscow, Russia.

Stevens, J. (1991). *An investigation of multivariate adaptive regression splines for modeling and analysis of univariate and semi-multivariate time series systems*. Ph.D. thesis, Naval Postgraduate School.

Taylan, P., Weber, G. W., & Ozkurt-Yerlikaya, F. (2010). A new approach to multivariate adaptive regression splines by using tikhonov regularization and continuous optimization. *TOP*, *18*(2), 377–395.

Thomaz, C. (2004). *Maximum entropy covariance estimate for statistical pattern recognition*. Ph.D. thesis, University of London and for the Diploma of the Imperial College (D.I.C.).

Van Emden, M. (1971). An analysis of complexity. In *Mathematical centre tracts*, vol. 35. Amsterdam: Mathematisch Centrum.

Weber, G. W., Batmaz, I., Köksal, G., Taylan, P., & Yerlikaya-Özkurt, F. (2012). CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. *Inverse Problems in Science and Engineering*, *20*, 371–400.

Xu, Q. S., Daszykowski, M., Walczak, B., Daeyaert, F., De Jonge, M., Heeres, J., et al. (2004). Multivariate adaptive regression splines—Studies of HIV reverse transcriptase inhibitors. *Chemometrics and Intelligent Laboratory Systems*, *72*(1), 27–34.

Yazici, C. (2011). *A computational approach to nonparametric regression: Bootstrapping the cmars method*. Master's thesis, Middle East Technical University, Ankara, Turkey.

York, T. P., Eaves, L. J., & van den Oord, E. J. (2006). Multivariate adaptive regression splines: A powerful method for detecting disease-risk relationship differences among subgroups. *Statistics in Medicine*, *25*(8), 1355–1367.