TI 2008-046/4
Tinbergen Institute Discussion Paper

MDL Mean Function Selection in Semiparametric Kernel Regression Models

Jan G. De Gooijer[1]
Ao Yuan[2]

[1] *University of Amsterdam, and Tinbergen Institute;*
[2] *National Human Genome Center, Howard University.*

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

MDL Mean Function Selection in Semiparametric Kernel Regression Models *

Jan G. De Gooijer[1] and Ao Yuan[2]

[1] Department of Quantitative Economics and Tinbergen Institute, University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: j.g.degooijer@uva.nl

[2] Statistical Genetics and Bioinformatics Unit
National Human Genome Center, Howard University
Washington DC, USA
e-mail: ayuan@howard.edu

Abstract

We study the problem of selecting the optimal functional form among a set of non-nested nonlinear mean functions for a semiparametric kernel based regression model. To this end we consider Rissanen's minimum description length (MDL) principle. We prove the consistency of the proposed MDL criterion. Its performance is examined via simulated data sets of univariate and bivariate nonlinear regression models.

Key words: Kernel density estimator; Maximum likelihood estimator; Minimum description length; Nonlinear regression; Semiparametric model.

AMS 2000 subject classification: 62G20, 62G05.

*Forthcoming in: *Communications in Statistics: Theory and Methods, 2008*

1 Introduction

Consider a general (non)linear regression problem with observations

$$Y_i]|(\mathbf{X}_i, \boldsymbol{\theta}^*) = g(\boldsymbol{\theta}^*, \mathbf{X}_i) + \varepsilon_i, \quad (i = 1, \ldots, n), \tag{1}$$

where $\boldsymbol{\theta}^*$ is the "true" value of the model parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset I\!R^k$, and $g(\boldsymbol{\theta}, \boldsymbol{x}) = E(Y_i|(\mathbf{X}_i, \boldsymbol{\theta}))$ is a specified conditional mean function of the k-dimensional parameter vector $\boldsymbol{\theta}$ and the covariates $\mathbf{X}$, and where it is assumed that the $(Y_i, \mathbf{X}_i, \varepsilon_i)$'s are *i.i.d.* realizations from a common random source $(Y, \mathbf{X}, \varepsilon)$. In the usual parametric regression model, the ε_i's are assumed to have some known distribution $f(\cdot)$. Then the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ has desirable optimal properties. when $f(\cdot)$ is the true residual distribution. But, in practice, $f(\cdot)$ is unknown. Its choice is often based on convenience, and usually restricted to a limited few. In an attempt to derive a general regression method without imposing too strong subjective model assumptions, Yuan & De Gooijer (2007) (hereafter YDG) studied a semiparametric regression model, in which $g(\cdot, \cdot)$ is specified parametrically, and $f(\cdot)$ is modeled by a kernel density estimator; see Section 2 for a brief introduction. They proved that, under fairly general regularity conditions, the MLE $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is consistent, is asymptotically normal with rate $\sqrt{n}$, and efficient. YDG showed that the nonparametric pseudo-likelihood ratio test statistic has the Wilks property. Further, using this test statistic, they presented simulation results for selecting a subset of parameters in a set of nested models.

One important topic, not considered by these authors, concerns the selection of the optimal functional form of $g(\cdot, \cdot)$, among a set of non-nested or separate models. A good $g(\cdot, \cdot)$ should fit the data well and be as simple (smooth) as possible. Goodness-of-fit implies that the corresponding semiparametric log-likelihood function, evaluated at $\hat{\boldsymbol{\theta}}_n$, is relatively large. But typically, this will favor a sophisticated function $g(\cdot, \cdot)$. So there is a trade-off between goodness-of-fit and model complexity. Different $g_i(\cdot, \cdot)$'s, taken from a finite set of available functions, may have the same parametric dimensions but different parametrizations, or different parametric dimensions and the corresponding model may not necessarily be nested. One way to solve this model-selection problem is through using Rissanen's (1986, 1987, 1996) "minimum description length" (MDL) principle. Within the information theoretic view of model selection, the MDL principle rests on somewhat the same foundations as does the idea underlying the commonly used information criteria AIC and BIC. However, in contrast to these latter two model-selection criteria, it allows comparisons between non-nested regression models. The MDL principle has been successfully applied to a wide variety of model-selection problems in the fields of computer

science, electrical engineering, and database mining; see, e.g., Grünwald *et al.* (2005). Good tutorial introductions are provided by Bryant & Cordero-Braña (2000), Hansen & Yu (2001), and Lanterman (2001).

In this paper, we propose an MDL model-selection criterion for semiparametric kernel based regression models (Section 3). In addition, we study the consistency of the proposed MDL criterion. Its performance is examined via simulated data sets of univariate and bivariate nonlinear regression models (Section 4).

2 Semiparametric kernel regression

For a fixed $g(\cdot,\cdot)$, given $\mathbf{x}$ and $\boldsymbol{\theta}$, assume that the ε_i's are *i.i.d.* with common unknown density $f(\cdot)$. Hence, $Y_i|(\mathbf{X}_i,\boldsymbol{\theta}) \sim f(\cdot - g(\boldsymbol{\theta},\mathbf{x}))$. There are several potential ways to estimate $f(\cdot)$ nonparametrically. One direct approach is the Nadaraya-Watson estimator, given by $f_n(\varepsilon) = (nh_n)^{-1}\sum_{j=1}^{n} K((\varepsilon - Y_j + g(\boldsymbol{\theta},\mathbf{X}_j)/h_n)$ where $K(\cdot)$ is a probability density (kernel), and h_n a positive sequence (bandwidth) with $h_n \to 0$ as $n \to \infty$. The main idea of YDG is to plug in the estimator $f_n(\cdot)$ for the "true" density of the $\varepsilon_i = Y_i - g(\boldsymbol{\theta},\boldsymbol{X}_i)$'s. From the construction of $f_n(\cdot)$, this involves terms of $K(\cdot)$ evaluated at the data points $Y_i - Y_j - g(\boldsymbol{\theta},\mathbf{X}_i) + g(\boldsymbol{\theta},\mathbf{X}_j)$ $(i,j = 1,\ldots,n)$, which for some specifications of $g(\cdot,\cdot)$, will cause the cancellation of some parameters in the difference $-g(\boldsymbol{\theta},\mathbf{X}_i)+g(\boldsymbol{\theta},\mathbf{X}_j)$, and thus gives rise to an identifiability problem; see YDG for a simple method to overcome this problem. In the rest of this paper we assume, without loss of generality, that $g(\cdot,\cdot)$ is identifiable. Thus, instead of modeling the distribution of the ε_i's, the idea is to model the distribution of the $Z_i = Y_i - g(\boldsymbol{\theta},\mathbf{X}_i)$'s. Let $f(\cdot)$ and $f_n(\cdot)$ denote the true density of Z_i and its estimate respectively. Since all Z_j's are used in the construction of $f_n(\cdot)$ at each data point Z_i, the nonparametric likelihood specification will contain some unwanted values of $h_n^{-1}K(0)$. So, using the delete-one version of $f_n(\cdot)$, the likelihood function of $\mathbf{Y} = (Y_1,\ldots,Y_n)$ given $\mathbf{X} = (\mathbf{X}_1,\ldots,\mathbf{X}_n)$ is given by

$$\ell_n(\mathbf{Y}|\boldsymbol{\theta},\mathbf{X}) = \prod_{i=1}^{n} f_{(n,i)}(Z_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{(n,i)}(Y_i - g(\boldsymbol{\theta},\mathbf{X}_i)) \tag{2}$$

where $f_{(n,i)}(Z_i|\boldsymbol{\theta}) = \{(n-1)h_n\}^{-1}\sum_{j\neq i} K((Z_i - Z_j)/h_n)$. Maximizing (2) over $\boldsymbol{\theta}$ yields the MLE $\hat{\boldsymbol{\theta}}_n$ as the inferred regression relationship in $E(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})$.

3 MDL selection of $g(\cdot,\cdot)$

Initiated by Kolmogorov's theory of algorithmic or descriptive complexity, Rissanen (1978) developed the MDL principle of model-selection. Loosely defined: choose the model that gives the shortest description of data. The conversion of this principle into an explicit criterion resulted in a number of different versions with different interpretations. Here, we adopt Rissanen's (1996) formulation of the MDL criterion based on the expected Fisher information.

A precise formulation requires a bit of notation. Let $\mathcal{G} = \{g(\cdot,\cdot)\}$ be a finite set of candidate mean functions under consideration, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_j : j = 1,\ldots,h.\}$ be the collection of parametrizations of interest. The $\boldsymbol{\theta}_j$'s may or may not be nested within each other, or $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ both in $\boldsymbol{\Theta}$ may have the same dimension but different parametrization. Next consider a fixed density $f(\cdot|\boldsymbol{\theta}_j)$, with parameter $\boldsymbol{\theta}_j$ running through a subset $\Gamma_j \subset I\!R^k$, to emphasize the index of the parameter, we denote the MLE of $\boldsymbol{\theta}_j$ under model $f(\cdot|\cdot)$ by $\hat{\boldsymbol{\theta}}_j$ (instead of by $\hat{\boldsymbol{\theta}}_n$ to emphasize the dependence on the sample size), $I(\boldsymbol{\theta}_j)$ the Fisher information for $\boldsymbol{\theta}_j$ under $f(\cdot|\cdot)$, $|I(\boldsymbol{\theta}_j)|$ its determinant, and k_j the dimension of $\boldsymbol{\theta}_j$. Then the MDL criterion (Rissanen, 1996) chooses $\boldsymbol{\theta}_j$ so as to minimize

$$-\sum_{i=1}^{n} \log f(Y_i|\hat{\boldsymbol{\theta}}_j) + \frac{k_j}{2}\log\frac{n}{2\pi} + \log\int_{\Gamma_j}\sqrt{|I(\boldsymbol{\theta}_j)|}d\boldsymbol{\theta}_j, \quad (j = 1,\ldots,h). \tag{3}$$

The second and third term are often referred to as a complexity penalty. Note that Schwarz's BIC seeks a $\boldsymbol{\theta}_j$ which minimizes

$$BIC_j = -\sum_{i=1}^{n} \log f(Y_i|\hat{\boldsymbol{\theta}}_j) + \frac{k_j}{2}\log n, \quad (j = 1,\ldots,h). \tag{4}$$

When $f(\cdot|\cdot)$ is known, both the MDL and BIC criteria have reasonable explanations, though the results may not be the same. But when $f(\cdot|\cdot)$ depends on a functional form $g(\cdot,\cdot)$, BIC does not take this extra complexity into account, while in MDL, this extra bit of uncertainty is reflected in $I(\boldsymbol{\theta}_j)$ which, since it depends on $g(\cdot,\cdot)$, will be denoted by $I_g(\boldsymbol{\theta}_j)$. For the semiparametric regression model $I_g(\boldsymbol{\theta}_j)$ is given by

$$I_g(\boldsymbol{\theta}_j) = E\left(\frac{f^{(1)}(Z)}{f(Z)}\right)^2 E_{\boldsymbol{\theta}_j}\left(g^{[1]}(\boldsymbol{\theta}_j,\mathbf{X})\right)\left(g^{[1]}(\boldsymbol{\theta}_j,\mathbf{X})\right)', \tag{5}$$

where $g^{[1]}(\cdot,\cdot)$ is the first partial derivative vector $\partial g(\boldsymbol{\theta}_j,\mathbf{X})/\partial\boldsymbol{\theta}_j$.

Given the fact that in the semiparametric kernel based regression model $f(\cdot|\cdot)$, $\hat{\boldsymbol{\theta}}_j$, and $I(\boldsymbol{\theta}_j)$ also depend on $g(\cdot,\cdot)$ we label them with the subscript g when necessary. Now, our objective is to choose the optimal $g(\cdot,\cdot)$. Since $f_g(\cdot|\cdot)$ is unknown, a typical way is to apply the MDL

model-selection criterion on the pseudo log-likelihood, which is obtained by replacing $f_g(\cdot|\cdot)$, as given in (2), by $f_{n,g,i}(\cdot|\cdot)$. Thus we choose $g(\cdot,\cdot) \in \mathcal{G}$ to minimize the criterion

$$MDL(g) = -\sum_{i=1}^{n} \log f_{n,g,i}(Z_i|\hat{\boldsymbol{\theta}}_{g,j}) + \frac{k_j}{2}\log\frac{n}{2\pi} + \log\int_{\Gamma_j}\sqrt{|\hat{I}_g(\boldsymbol{\theta}_j)|}d\boldsymbol{\theta}_j, \tag{6}$$

where $\boldsymbol{\theta}_j$ is the (only) corresponding parameter(s) on which $g(\cdot,\cdot)$ is defined. Here $\hat{I}_g(\boldsymbol{\theta}_j)$ is an estimator of (5). From YDG (2007, Remark 6) this estimator is given by

$$\hat{I}_g(\boldsymbol{\theta}_j) = \Big(\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{f^{(1)}_{n,i}(z_i)}{f_{n,i}(z_i)}\Big]^2\Big)\frac{1}{n}\sum_{i=1}^{n}\Big(g^{[1]}(\boldsymbol{\theta}_j,\mathbf{x}_i)\Big)\Big(g^{[1]}(\boldsymbol{\theta}_j,\mathbf{x}_i)\Big)', \tag{7}$$

and where the notation $f^{(1)}_{n,i}(z_i)$ denotes the first derivative of $f(\cdot|\cdot)$ with respect to z. Given two mean functions $g_1(\cdot,\cdot)$ and $g_2(\cdot,\cdot)$ in $\mathcal{G}$, function $g_2(\cdot,\cdot)$ is preferred over $g_1(\cdot,\cdot)$ if $MDL(g_2) < MDL(g_1)$.

Clearly, (6) takes into account the dimensionality of $\boldsymbol{\theta}_j$ and the complexity of $g_j(\cdot,\cdot)$ simultaneously. Note that each $g(\cdot,\cdot)$ can only be defined on one of the $\boldsymbol{\theta}_j$'s, but each $\boldsymbol{\theta}_j$ may have more than one $g(\cdot,\cdot)$'s defined on it. For example, let $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i})'$, $\boldsymbol{\theta} = (\theta_1,\theta_2,\theta_3)$, $\boldsymbol{\theta}_1 = (\theta_1,\theta_2)$ and $\boldsymbol{\theta}_2 = (\theta_2,\theta_3)$. We can define $g_1(\boldsymbol{\theta},\mathbf{X}) = g_1(\boldsymbol{\theta}_1,\mathbf{X}_i) = \theta_1 X_{1i} + \theta_2 X_{2i}$, $g_2(\boldsymbol{\theta},\mathbf{X}_i) = g_2(\boldsymbol{\theta}_1,\mathbf{X}_i) = \theta_1 X_{1i}^2 + \theta_2 X_{2i}^2$, and $g_3(\boldsymbol{\theta},\mathbf{X}_i) = g_3(\boldsymbol{\theta}_2,\mathbf{X}_i) = \theta_2 X_{2i} + \theta_3 X_{3i}$. Here, for the same $\boldsymbol{\theta}_1$, we have $g_1(\cdot,\cdot)$ and $g_2(\cdot,\cdot)$ defined on it. In other words, if $g^*(\cdot,\cdot)$ is the true data generating mechanism, we are not sure if $g^*(\cdot,\cdot)$ will minimize (6), but (6) is still a reasonable criterion to use.

To study the consistency of $\hat{I}_g(\boldsymbol{\theta}_j)$, we impose the following conditions:

(A1) $h_n \to 0$ and $\sum_{n=1}^{\infty}\exp(-\varepsilon n h_n^4) < \infty$, for all $\varepsilon > 0$.

(A2) $K^{(i)}(\cdot)$ is bounded $(i = 0, 1, 2)$.

(A3) $\int K(u)du = 1$, $\int K^{(1)}(u)du = 0$, and $\int uK^{(1)}(u)du = -1$.

(A4) $f^{(1)}(\cdot)$ is uniformly continuous.

(A5) $g^{[1]}(\cdot,\cdot)$ is continuous.

(A6) $0 < \inf_{\boldsymbol{\theta}\in\Gamma}|I_g(\boldsymbol{\theta})| \le \sup_{\boldsymbol{\theta}\in\Gamma}|I_g(\boldsymbol{\theta})| < \infty$.

(A7) $\mathbf{X}$ has compact support.

(A8) $\inf_z f(z) > 0$.

Conditions (A1)-(A7) are practical and easy to satisfy. Note that if we use a Gaussian kernel, then (A3) is satisfied. Condition (A8) is used by a number of authors (Hall 1986; Joe 1989; Hall & Morton 1993). The following theorem, omitting the subscript j for simplicity, asserts the consistency of $\hat{I}_g(\boldsymbol{\theta})$. A proof is given in the Appendix.

Theorem. *Suppose (A1)-(A8) hold. Then for any compact* Γ*, we have*

$$\log \int_{\Gamma} \sqrt{|\hat{I}_g(\boldsymbol{\theta})|} d\boldsymbol{\theta} \to \log \int_{\Gamma} \sqrt{|I_g(\boldsymbol{\theta})|} d\boldsymbol{\theta}, \quad a.s.$$

In the $MDL(g)$ criterion (6), the integration in the last term can be well approximated by Monte Carlo methods (see, e.g., Robert & Casella, 1999). One suitable method is importance sampling[1], as described below. Denote by $\hat{I}_{n,g}(\boldsymbol{\theta}_j) = H_n J_{n,g}(\boldsymbol{\theta}_j)$ where $H_n = n^{-1} \sum_{i=1}^{n} [f_{n,i}^{(1)}(z_i)/ f_{n,i}(z_i)]^2$, and $J_{n,g}(\boldsymbol{\theta}_j) = n^{-1} \sum_{i=1}^{n} (g^{[1]}(\boldsymbol{\theta}_j, \mathbf{x}_i)(g^{[1]}(\boldsymbol{\theta}_j, \mathbf{x}_i))'$. Then $\int_{\Gamma_j} \sqrt{I_{n,g}(\boldsymbol{\theta}_j)} d\boldsymbol{\theta}_j = \sqrt{H_n} \int_{\Gamma_j} \sqrt{|J_{n,g}(\boldsymbol{\theta}_j)|} d\boldsymbol{\theta}_j$. To compute the integration on the right-hand side, specify the support set Γ_j of $\boldsymbol{\theta}_j$. If there is no information for a particular form of Γ_j just take $\mathbb{R}^k$ for convenience. Let $\chi_{\Gamma_j}(\cdot)$ be the indicator function on the set Γ_j. Select an arbitrary density for $\boldsymbol{\theta}_j$ only for sampling purpose. For instance, the k_j-variate standard normal density $\phi(\boldsymbol{\theta}_j)$. Given a large integer value m ($m = 10,000$ in the simulations), then, for $u = 1, \ldots, m$ do the following: i) independently sample $\boldsymbol{\theta}_{j,u} \sim \phi(\boldsymbol{\theta}_j)$; ii) compute $\sqrt{|J_{n,g}(\boldsymbol{\theta}_{j,u})| \chi_{\Gamma_j}(\boldsymbol{\theta}_{j,u})}/\phi(\boldsymbol{\theta}_{j,u})$. Then, by the SLLN, we have (as $m \to \infty$)

$$\frac{1}{m} \sum_{u=1}^{m} \frac{\sqrt{|J_{n,g}(\boldsymbol{\theta}_{j,u})|}}{\phi(\boldsymbol{\theta}_{j,u})} \chi_{\Gamma_j}(\boldsymbol{\theta}_{j,u}) \stackrel{a.s.}{\to} E_{\boldsymbol{\theta}_j} \left(\frac{\sqrt{|J_{n,g}(\boldsymbol{\theta}_j)|}}{\phi(\boldsymbol{\theta}_j)} \chi_{\Gamma_j}(\boldsymbol{\theta}_j) \right) = \int_{\Gamma_j} \sqrt{|J_{n,g}(\boldsymbol{\theta}_j)|} d\boldsymbol{\theta}_j.$$

4 Numerical Studies

This section examines the performance of the MDL model-selection criterion (6) for non-nested nonlinear regression models via two sets of simulation experiments. For comparison, we also compute (6) without the complexity penalty term $\log \int_{\Gamma_j} \sqrt{|\hat{I}_g(\boldsymbol{\theta}_j)|} d\boldsymbol{\theta}_j$.

4.1 Univariate regression: 2 parameters

Consider the nonlinear regression model (1). We simulate data from the following four univariate regression models

$$\begin{aligned}
g_1(\cdot,\cdot) &= \frac{\theta_1(1) X_{1i}}{\theta_1(2) + X_{1i}}, && \theta_1(1) = 2,\ \theta_1(2) = 2, \\
g_2(\cdot,\cdot) &= \theta_2(1) \exp(\theta_2(2) X_{1i}), && \theta_2(1) = -0.5,\ \theta_2(2) = -2, \\
g_3(\cdot,\cdot) &= \frac{\theta_3(1) X_{1i}}{X_{1i} + \theta_3(2) X_{1i}^2}, && \theta_3(1) = 0.5,\ \theta_3(2) = 0.8, \\
g_4(\cdot,\cdot) &= \theta_4(1) X_{1i} + \theta_4(2) X_{1i}^2, && \theta_4(1) = 1,\ \theta_4(2) = -0.5.
\end{aligned}$$

[1] In the simulations the so-called VEGAS algorithm of G.P. Lepage was used for this purpose.

Model function $g_1(\cdot,\cdot)$ is the so-called Michaelis-Menten equation. It was fitted to empirical data by Bates and Watts (1988, Appendix A.1.3). Model functions $g_2(\cdot,\cdot) - -g_4(\cdot,\cdot)$ are adapted versions of models listed in Appendix 7 of Bates & Watts (1988). Hence, all above models are of interest to applied research. Note that in all cases the mean functions have no constant term. So, we don't have the identifiability problem. Figure 1.a) shows a plot of the four mean functions.

We draw 1000 random samples of sizes $n = 100$ and 200 from each model. We sample Z_i's from a standard gamma distribution with density $f(z) = z\exp(-z)$, $z > 0$. The covariate X_{1i} has a Uniform (-1,3) distribution. Throughout the simulations we use the biweight kernel $K(u) = \frac{15}{16}(1-u^2)\chi(|u| \leq 1)$. In all cases the so-called *rule of thumb* bandwidth selector of Deheuvels (1977) was adopted; see YDG for some discussion on this choice. Table 1 presents a ranking of the total number of models selected.

MDL identifies the "true" model specification in all cases, for both sample sizes. However, whether or not the complexity integral term is included makes quite a difference. Indeed, for each of the true models $g_1(\cdot,\cdot)$, $g_2(\cdot,\cdot)$, and $g_3(\cdot,\cdot)$, the total number of models ranking first is much lower for the MDL criterion than for the MDL criterion without penalty term (printed in parentheses). An exception is $g_2(\cdot,\cdot)$ with substantial larger number of correctly identified models for the MDL criterion than for the MDL criterion without penalty term. Note that for $n = 200$ the MDL criterion without penalty term incorrectly chooses $g_1(\cdot,\cdot)$ 636 times while it picks the true model $g_2(\cdot,\cdot)$ in only 348 out of 1000 cases. In contrast, MDL picks the true model in 867 cases and $g_1(\cdot,\cdot)$ in 114 cases. These observations suggest that, in finite samples, the complexity integral term in the MDL criterion plays a crucial role to identify the true model. Regarding the overall ranking, given the true models $g_1(\cdot,\cdot)$, $g_2(\cdot,\cdot)$, and $g_4(\cdot,\cdot)$, all competing model specifications are most likely to end up at ranks 2–4. Except for $g_3(\cdot,\cdot)$ which, apart from ranking first, also ranks second when $n = 100$ and $n = 200$.

4.2 Bivariate regression: 2–4 parameters

A large number of nonlinear regression models have been proposed to describe the equilibrium moisture content, M_e, of many biological and agricultural materials as a function of relative humidity, RH, and the solid material temperature, T_s, i.e. $M_{e,i} = g(\boldsymbol{\theta}, T_{s,i}, RH_i)$ $(i = 1,\ldots,n)$. Recently, Ribeiro *et al.* (2005), using experimental data for *Bixa orellana* seeds, a tropical shrub

which grows quickly in Brazil, India, and East Africa, evaluated the following five models[2]:

$$g_1(\cdot,\cdot) = \Big(\frac{\ln(1-RH_i)}{-\theta_1(1)T_{s,i}}\Big)^{1/\theta_1(2)}, \qquad \theta_1(1) = 1.05\times 10^{-4}, \theta_1(2) = 1.68,$$

$$g_2(\cdot,\cdot) = \Big(\frac{\ln(1-RH_i)}{-\theta_2(1)(T_{s,i}+\theta_2(3))}\Big)^{1/\theta_2(2)}, \qquad \theta_2(1) = 1.05\times 10^{-4}, \theta_2(2) = 1.71,$$
$$\theta_2(3) = 3.38,$$

$$g_3(\cdot,\cdot) = \frac{-1}{\theta_3(2)}\ln\Big(\frac{(T_{s,i}+\theta_3(3))\ln(RH_i)}{-\theta_3(1)}\Big), \qquad \theta_3(1) = 124.57, \theta_3(2) = 0.08,$$
$$\theta_3(3) = -10.69,$$

$$g_4(\cdot,\cdot) = \frac{-1}{\theta_4(3)T_{s,i}^{\theta_4(4)}}\ln\Big(\frac{\ln(RH_i)}{-\theta_4(1)T_{s,i}^{\theta_4(2)}}\Big), \qquad \theta_4(1) = 0.992, \theta_4(2) = 0.486,$$
$$\theta_4(3) = 0.004, \theta_4(4) = 0.816,$$

$$g_5(\cdot,\cdot) = \Big(\frac{-\exp(\theta_5(1)T_{s,i}+\theta_5(3))}{\ln(RH_i)}\Big)^{1/\theta_5(2)}, \qquad \theta_5(1) = -4.19\times 10^{-2}, \theta_5(2) = 2.575,$$
$$\theta_5(3) = 9.$$

Figure 1.b) shows a plot of the five mean functions at temperature $T_s = 25°$C. Note that, as opposed to Figure 1.a), these functions show very similar patterns.

We draw 1000 random samples of size $n = 100$ from each model. We sample Z_i's from a standard normal distribution. The covariate $T_{s,i}$ has a Uniform (20,50) distribution, and the covariate RH_i has a Uniform (0,1) distribution. The two covariates are independent. Table 2 presents a ranking of the total number of models selected. Not surprisingly, the total number of correctly chosen models is always very high when the data generating mechanism is the true one. Clearly, there is less distinction between the MDL criterion with or without the complexity integral term. This feature was also noticed for sample sizes $n > 100$. Interestingly, in all four cases $g_5(\cdot,\cdot)$ is not the true data generating mechanism, model $g_5(\cdot,\cdot)$ ends up last in ranking. Hence, there is less support for Ribeiro *et al.*'s (2005) conclusion, albeit using different statistical methodology, to select $g_5(\cdot,\cdot)$ as the best mean function.

References

Bahadur, R.R. & Zabell, S.L., 1979. Large deviations of the sample mean in general vector spaces. Ann. Probab. 7, 587-621.

[2]Model functions $g_2(\cdot,\cdot)$, $g_3(\cdot,\cdot)$, and $g_5(\cdot,\cdot)$ have been adopted as standard equations by the American Society of Agricultural Engineers for describing water sorption isotherms.

Bates, D.M. & Watts, D.G., 1988. Nonlinear Regression Analysis & Its Applications. Wiley, New York.

Bryant, P.G. & Cordero-Braña, O.I., 2000. Model selection using the minimum description length principle. Amer. Statist. 54, 257-268.

Deheuvels, P., 1977. Estimation non paramétrique de la densité par histogrammes généralisés". Rev. Statist. Appl. 35, 5-42.

Dvoretzky, A., Kiefer, J. & Wolfowitz, J., 1956. Asymptotic minimax character of the sample distribution function and of the classic multinomial estimator. Ann. Math. Statist. 27, 642-669.

Grünwald, P.D., Myung, I.J. & Pitt, M.A. (Eds.), 2005. Advances in Minimum Description Length: Theory and Applications, Cambridge, MA: MIT Press.

Hall, P., 1986. On powerful distributional tests based on sample spacings. J. Multivariate Statist. 19, 201-225.

Hall, P. & Morton, S.C., 1993. On the estimation of entropy. Ann. Inst. Statist. Math. 45, 69-88.

Hansen, M. & Yu, B., 2001. Model selection and the principle of minimum description length. J. Amer. Statist. Assoc. 96, 746-774.

Joe, H., 1989. Estimation of entropy and other functionals of a multivariate density. Ann. Inst. Statist. Math. 41, 683-697.

Kotz, S. & Johnson, N.L., 1982. Encyclopedia of Statistical Sciences, Vol. 6. Wiley, New York.

Lanterman, A.D., 2001. Schwarz, Wallace, and Rissanen: Intertwinning themes in theories of model selection. Int. Statist. Rev. 69, 185-212.

Ribeiro, J.A., Oliveira, D.T., Passos, M.L. & Barrozo, M.A.S., 2005. The use of nonlinearity measures to discriminate the equilibrium moisture equations for *Bixa orellana* seeds. J. Food Engineering 66, 63-68.

Rissanen, J., 1978. Modelling by shortest data description. Automatica 14, 465-471.

Rissanen, J., 1986. Stochastic complexity and modeling. Ann. Statist. 14, 1080-1100.

Rissanen, J., 1987. Stochastic complexity (with discussion). J. Roy. Statist. Soc. Ser. B 49, 223-265.

Rissanen, J., 1996. Fisher information and stochastic complexity. IEEE Trans. Inform. Theory 42, 40-47.

Robert, C.P. & Casella, G., 1999. Monte Carlo Statistical Methods. Springer, New York.

Yuan, A. & De Gooijer, J.G., 2007. Semiparametric regression with kernel error model. Scand.

J. Statist. 34, 841-869.

Appendix: Proof of Theorem

Proof: We first prove

$$\sup_z |f_n^{(r)}(z) - f^{(r)}(z)| = o(1) \quad a.s. \quad (r = 0, 1). \tag{8}$$

Here we only prove the case $r = 1$. The case $r = 0$ is similar and simpler. We have

$$\sup_z |f_n^{(1)}(z) - f^{(1)}(z)| \le \sup_z |f_n^{(1)}(z) - Ef_n^{(1)}(z)| + \sup_z |Ef_n^{(1)}(z) - f^{(1)}(z)| := V_{n,1} + V_{n,2}.$$

Let $F_n(\cdot)$ and $F(\cdot)$ respectively be the empirical and true distributions of the Z_i's. Note that (A2) implies that $K^{(j)}(\cdot)$ has bounded variation $0 < \tau_j < \infty$ $(j = 0, 1)$, so we have

$$V_{n,1} = \frac{1}{h_n^2} \sup_z | \int K^{(1)}(\frac{z-y}{h_n}) dF_n(y) - \int K^{(1)}(\frac{z-y}{h_n}) dF(y)|$$

$$\le \frac{1}{h_n^2} \sup_z \int |F_n(y) - F(y)||dK^{(1)}(\frac{z-y}{h_n})|$$

$$\le \frac{1}{h_n^2} \sup_z \int |F_n(z - h_n u) - F(z - h_n u)||dK^{(1)}(u)| \le \frac{\tau_1}{h_n^2} \sup_y |F_n(y) - F(y)|.$$

By the result on large deviation in Dvoretzky *et al.* (1956), there are positive constants C and $0 < \alpha \le 2$, such that

$$P\left(\frac{\tau_1}{h_n^2} \sup_y |F_n(y) - F(y)| > \epsilon\right) \le C \exp(-\alpha \epsilon^2 \tau_1^{-2} n h_n^4), \qquad \forall\ \epsilon > 0.$$

This together with (A1), and the Borel-Cantelli lemma we have $V_{n,1} \to 0$ (a.s).

Using (A3), for some $0 \le \beta_n \le 1$, we have

$$Ef_n^{(1)}(z) = \frac{1}{h_n^2} \int_{-\infty}^{+\infty} K^{(1)}(\frac{z-y}{h_n}) f(y) dy = -\frac{1}{h_n} \int_{+\infty}^{-\infty} K^{(1)}(u) f(z - h_n u) du$$

$$= \frac{1}{h_n} \int_{-\infty}^{+\infty} K^{(1)}(u)[f(z) - h_n u f^{(1)}(z - \beta_n h_n u)] du = -\int u K^{(1)}(u) f^{(1)}(z - \beta_n h_n u) du.$$

For any $\epsilon > 0$, by (A4), there is a $\delta > 0$ such that $\sup_{|\beta_n h_n u| \le \delta} \sup_z |f^{(1)}(z - \beta_n h_n u) - f^{(1)}(z)| \le \epsilon/2$. Also, by (A4), there is $0 < C < \infty$ such that $\sup_z |f^{(1)}(z - \beta_n h_n u) - f^{(1)}(z)| \le C$. By (A3), there is an n_0 such that for $n > n_0$, $\int_{|\beta_n h_n u| > \delta} |u K^{(1)}(u)| du < \epsilon/2$. So by (A3) again, we have

$$\sup_z |Ef_n^{(1)}(z) - f^{(1)}(z)| = \sup_z \left| \int u K^{(1)}(u)[f^{(1)}(z - \beta_n h_n u) - f^{(1)}(z)] du \right|$$

$$\le \int_{|\beta_n h_n u| \le \delta} |u K^{(1)}(u)| \sup_z |f^{(1)}(z - \beta_n h_n u) - f^{(1)}(z)| du$$

$$+\int_{|\beta_n h_n u|>\delta} |uK^{(1)}(u)| \sup_z |f^{(1)}(z-\beta_n h_n u) - f^{(1)}(z)| du \le \frac{\epsilon}{2}\int |uK^{(1)}(u)|du + C\frac{\epsilon}{2}.$$

Since $\epsilon > 0$ is arbitrary, we have $V_{n,2} \to 0$. This completes the proof of (8) for the case $r = 1$.

Now, by (8), we have

$$\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{f_{n,i}^{(1)}(z_i)}{f_{n,i}(z_i)}\Big]^2 = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{f^{(1)}(z_i)+o(1)}{f(z_i)+o(1)}\Big]^2 = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{f^{(1)}(z_i)}{f(z_i)}\Big]^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{o(1)[f(z_i)-f^{(1)}(z_i)]^2}{f^2(z_i)[f(z_i)+o(1)]^2}.$$

For large n, the second term on the right-hand side above is asymptotically equivalent to $o(1)C$, by the SLLN and (A8), for some $0 < C < \infty$. Thus,

$$\hat{I}_g(\boldsymbol{\theta}) = \Big(\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{f^{(1)}(z_i)}{f(z_i)}\Big]^2\Big)\frac{1}{n}\sum_{i=1}^{n}\Big(g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i)\Big)\Big(g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i)\Big)' + o(1) \quad a.s.$$

Now we only need to prove

$$\frac{1}{n}\sum_{i=1}^{n}\Big(g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i)\Big)\Big(g^{[1]}(\boldsymbol{\theta},\boldsymbol{x}_i)\Big)' \to E_{\boldsymbol{\theta}}\Big(g^{[1]}(\boldsymbol{\theta},\mathbf{X})\Big)\Big(g^{[1]}(\boldsymbol{\theta},\mathbf{X})\Big)', \ (a.s.) \text{ uniformly for } \boldsymbol{\theta}\in\Gamma. \quad (9)$$

To show (9) for any corresponding components of the matrices we assume, without loss of generality, that $g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i)$ is one-dimensional. Let S be the support of $\boldsymbol{X}$. By (A5), (A7) and the compactness of Γ, $g^{[1]}(\cdot,\cdot)$ is uniformly continuous on $\Gamma \times S$. Thus, for any $\epsilon > 0$, there are finite number of points $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_J$ in Γ such that for any $\boldsymbol{\theta}\in\Gamma$, there is a $\boldsymbol{\theta}_j \in \Gamma$ satisfying

$$|g^{[1]}(\boldsymbol{\theta},\mathbf{x}) - g^{[1]}(\boldsymbol{\theta}_j,\mathbf{x})| < \epsilon/3, \quad \forall \mathbf{x}\in S \quad \text{and} \quad |Eg^{[1]}(\boldsymbol{\theta},\mathbf{X}) - Eg^{[1]}(\boldsymbol{\theta}_j,\mathbf{X})| < \epsilon/3.$$

The index j is dependent on $\boldsymbol{\theta}$, and we just write it as j in the following. Now we have

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Gamma}\Big|\frac{1}{n}\sum_{i=1}^{n}g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i) - Eg^{[1]}(\boldsymbol{\theta},\mathbf{X})\Big| &\le \frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\theta}\in\Gamma}\Big|g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i) - g^{[1]}(\boldsymbol{\theta}_j,\mathbf{x}_i)\Big| \\ &+\Big|\frac{1}{n}\sum_{i=1}^{n}g^{[1]}(\boldsymbol{\theta}_j,\mathbf{x}_i) - Eg^{[1]}(\boldsymbol{\theta}_j,\mathbf{X})\Big| + \Big|Eg^{[1]}(\boldsymbol{\theta}_j,\mathbf{X}) - Eg^{[1]}(\boldsymbol{\theta},\mathbf{X})\Big| \\ &\le 2\epsilon/3 + \Big|\frac{1}{n}\sum_{i=1}^{n}g^{[1]}(\boldsymbol{\theta}_j,\mathbf{x}_i) - Eg^{[1]}(\boldsymbol{\theta}_j,\mathbf{X})\Big|.\end{aligned}$$

By the large deviation inequality (according to a result initiated by Cramér-Chernoff, extended by Bahadur & Zabell, 1979; concisely stated in Kotz & Johnson, 1982, pp. 32-33), and stated in the form of an inequality in YDG, 2007), there are constants $0 < C_0 < \infty$ and $0 < C(\epsilon)$ such that $P(|\frac{1}{n}\sum_{i=1}^{n}g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i) - g^{[1]}(\boldsymbol{\theta}_j,\mathbf{x}_i)| \ge \epsilon/3) \le C_0\exp(-nC(\epsilon))$, thus by the Borel-Cantelli lemma the last term in the above is bounded by $\epsilon/3$ (a.s.) and we have

$$\lim_n \sup_{\boldsymbol{\theta}\in\Gamma}\Big|\frac{1}{n}\sum_{i=1}^{n}g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i) - Eg^{[1]}(\boldsymbol{\theta},\mathbf{X})\Big| \le \epsilon. \quad a.s.$$

Since $\epsilon > 0$ is arbitrary, we have $\frac{1}{n}\sum_{i=1}^{n}g^{[1]}(\boldsymbol{\theta},\mathbf{x}_i) \to Eg^{[1]}(\boldsymbol{\theta},\mathbf{X})$ (a.s.) uniformly for $\boldsymbol{\theta}\in\Gamma$ and (9) is proved $\square$

Table 1: Total number of true models selected out of 1000 replications by the MDL criterion and by the MDL criterion without complexity integral term (values in parentheses). Univariate regression with 2 parameters. Given a true model the value, say Z, of entry (rank$_i$,$g_i(\cdot,\cdot)$), with $g_i(\cdot,\cdot)$ $(i = 1,\ldots,4)$ in top row and rank$_i$ in first column, means that model $g_i(\cdot,\cdot)$ ended up at rank$_i$ Z times out of 1000 replications.

	Model $g(\cdot,\cdot)$ selected							
Rank	$g_1(\cdot,\cdot)$	$g_2(\cdot,\cdot)$	$g_3(\cdot,\cdot)$	$g_4(\cdot,\cdot)$	$g_1(\cdot,\cdot)$	$g_2(\cdot,\cdot)$	$g_3(\cdot,\cdot)$	$g_4(\cdot,\cdot)$
	True model $g_1(\cdot,\cdot)$							
	$n = 100$				$n = 200$			
1	525 (906)	377 (58)	76 (28)	22 (7)	888 (968)	69 (9)	41 (23)	2 (0)
2	355 (80)	493 (693)	108 (157)	44 (70)	93 (29)	801 (779)	78 (152)	28 (40)
3	119 (14)	122 (229)	541 (439)	318 (318)	19 (3)	127 (205)	587 (525)	267 (267)
4	1 (0)	8 (19)	375 (376)	616 (605)	0 (0)	3 (7)	294 (300)	703 (693)
	True model $g_2(\cdot,\cdot)$							
	$n = 100$				$n = 200$			
1	3 (284)	980 (690)	13 (18)	4 (8)	114 (636)	867 (348)	19 (16)	0 (0)
2	768 (630)	20 (292)	110 (42)	102 (36)	837 (355)	124 (620)	37 (25)	2 (0)
3	197 (77)	0 (17)	380 (409)	423 (497)	49 (9)	9 (32)	649 (657)	293 (302)
4	32 (9)	0 (1)	497 (531)	471 (459)	0 (0)	0 (0)	295 (302	705 (698)
	True model $g_3(\cdot,\cdot)$							
	$n = 100$				$n = 200$			
1	6 (98)	381 (244)	590 (630)	23 (28)	6 (103)	408 (261)	584 (633)	2 (3)
2	155 (249)	248 (261)	362 (289)	235 (201)	164 (259)	321 (326)	391 (310)	124 (105)
3	340 (333)	197 (237)	44 (76)	419 (354)	466 (350)	95 (215)	24 (55)	415 (380)
4	499 (320)	174 (258)	4 (5)	323 (417)	364 (288)	176 (198)	1 (2)	459 (512)
	True model $g_4(\cdot,\cdot)$							
	$n = 100$				$n = 200$			
1	10 (74)	433 (316)	49 (36)	508 (574)	15 (57)	423 (358)	34 (27)	528 (558)
2	87 (258)	407 (414)	256 (196)	250 (132)	165 (282)	447 (438)	168 (133)	220 (147)
3	321 (353)	131 (173)	411 (317)	137 (157)	330 (344)	82 (126)	440 (354)	148 (176)
4	582 (315)	29 (97)	284 (451)	105 (137)	490 (317)	48 (78)	358 (486)	104 (119)

Table 2: Total number of true models selected out of 1000 replications by the MDL criterion and by the MDL criterion without complexity integral term (values in parentheses). Bivariate regression with 2–4 parameters and $n = 100$. Given a true model the value, say Z, of entry (rank$_i$,$g_i(\cdot,\cdot)$), with $g_i(\cdot,\cdot)$ ($i = 1,\ldots,5$) in top row and rank$_i$ in first column, means that model $g_i(\cdot,\cdot)$ ended up at rank$_i$ Z times out of 1000 replications.

	Model $g(\cdot,\cdot)$ selected				
Rank	$g_1(\cdot,\cdot)$	$g_2(\cdot,\cdot)$	$g_3(\cdot,\cdot)$	$g_4(\cdot,\cdot)$	$g_5(\cdot,\cdot)$
	True model $g_1(\cdot,\cdot)$				
1	757 (805)	243 (193)	0 (0)	0 (0)	0 (0)
2	243 (195)	757 (805)	0 (0)	0 (0)	0 (0)
3	0 (0)	0 (2)	3 (0)	997 (998)	0 (0)
4	0 (0)	0 (0)	997 (1000)	3 (0)	0 (0)
5	0 (0)	0 (0)	0 (0)	0 (0)	1000 (1000)
	True model $g_2(\cdot,\cdot)$				
1	86 (107)	914 (891)	0 (0)	0 (2)	0 (0)
2	912 (881)	86 (109)	0 (0)	2 (10)	0 (0)
3	2 (12)	0 (0)	12 (0)	986 (988)	0 (0)
4	0 (0)	0 (0)	988 (1000)	12 (0)	0 (0)
5	0 (0)	0 (0)	0 (0)	0 (0)	1000 (1000)
	True model $g_3(\cdot,\cdot)$				
1	0 (0)	0 (0)	1000 (907)	0 (93)	0 (0)
2	0 (0)	0 (0)	0 (93)	1000 (907)	0 (0)
3	610 (648)	390 (352)	0 (0)	0 (0)	0 (0)
4	389 (351)	610 (648)	0 (0)	0 (0)	1 (1)
5	1 (1)	0 (0)	0 (0)	0 (0)	999 (999)
	True model $g_4(\cdot,\cdot)$				
1	0 (0)	0 (0)	0 (0)	1000 (1000)	0 (0)
2	80 (94)	920 (906)	0 (0)	0 (0)	0 (0)
3	920 (906)	80 (94)	0 (0)	0 (0)	0 (0)
4	0 (0)	0 (0)	1000 (997)	0 (0)	0 (0)
5	0 (0)	0 (0)	0 (3)	0 (0)	1000 (1000)
	True model $g_5(\cdot,\cdot)$				
1	0 (0)	0 (0)	0 (0)	0 (0)	1000 (1000)
2	36 (45)	63 (80)	830 (302)	70 (573)	0 (0)
3	50 (34)	51 (40)	125 (595)	775 (331)	0 (0)
4	596 (604)	259 (243)	38 (78)	107 (75)	0 (0)
5	318 (317)	627 (637)	7 (25)	48 (21)	0 (0)

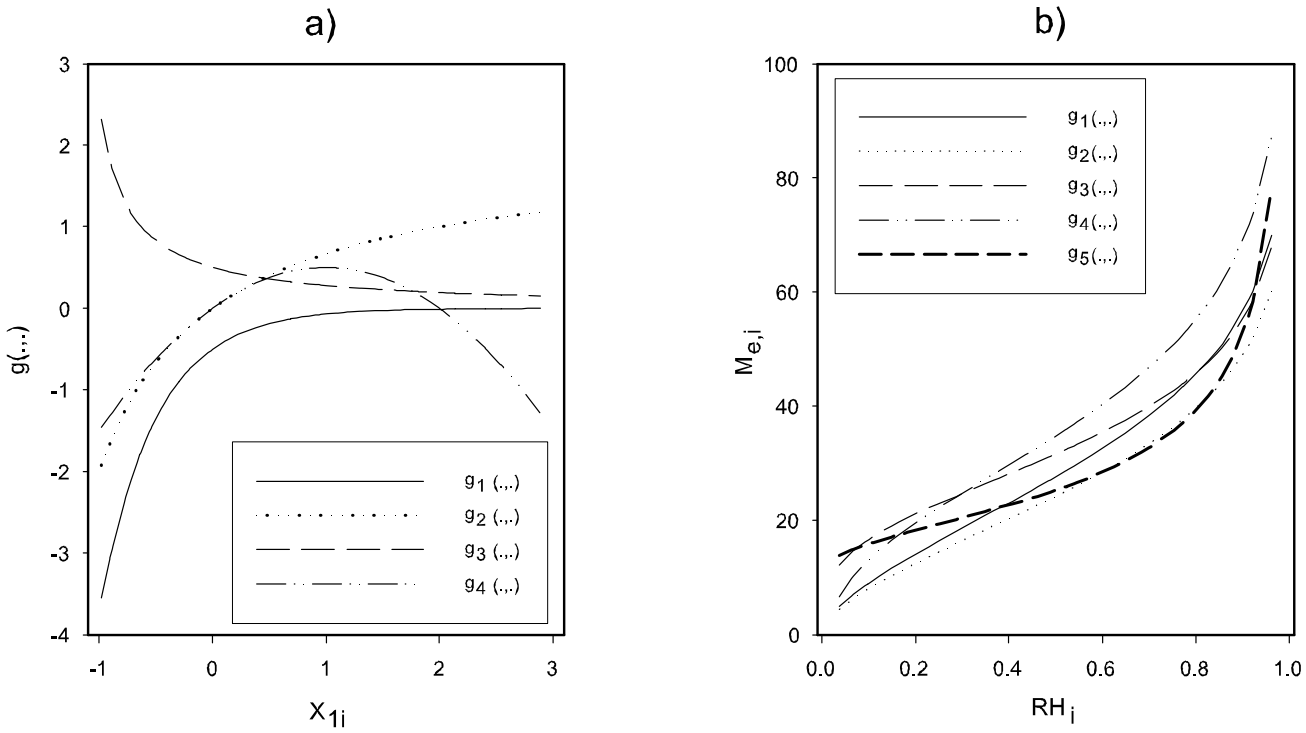

Figure 1: *True regression functions: a) Univariate case and b) bivariate case.*