Research Article

An Extension of the Quadratic Error Function for Learning Imprecise Data in Multivariate Nonlinear Regression

Castro Gbêmmali Hounmenou ^{1,2}, Kossi Essona Gneyou,³ and Romain Glélé Kakaï¹

¹Laboratoire de Biomathématiques et d'Estimations Forestières, Université d'Abomey-Calavi, Cotonou, Benin

²Institut de Mathématiques et de Sciences Physiques, Porto-Novo, Benin

³Laboratoire de Modélisations Mathématiques et Applications, Université de Lomé, Lomé, Togo

Correspondence should be addressed to Castro Gbêmmali Hounmenou; castro.hounmenou@imsp-uac.org

Received 8 February 2020; Accepted 18 June 2020; Published 29 July 2020

Academic Editor: Alessandro De Gregorio

Copyright © 2020 Castro Gbêmmali Hounmenou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multivariate noises in the learning process are most of the time supposed to follow a standard multivariate normal distribution. This hypothesis does not often hold in many real-world situations. In this paper, we consider an approach based on multivariate skew-normal distribution. It allows for a multiple continuous variation from normality to nonnormality. We give an extension of the generalized least squares error function in a context of multivariate nonlinear regression to learn imprecise data. The simulation study and application case on real datasets conducted and based on multilayer perceptron neural networks (MLP) with bivariate continuous response and asymmetric revealed a significant gain in precision using the new quadratic error function for these types of data rather than using a classical generalized least squares error function having any covariance matrix.

1. Introduction

Let X and Y be, respectively, $\mathbb{R}^p$ and $\mathbb{R}^q$ valued random vectors defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $(X_t, Y_t)_{t \in \mathbb{N}}$ be independent identically distributed (i.i.d) replications of (X, Y) . One seeks to build the best estimator $\hat{F}$ of the regression function $F: x \mapsto \mathbb{E}[Y | X = x]$ in the sense of a loss function $\mathbb{L}$ where the bias $F(x) - \mathbb{E}[\hat{F}(x)]$ and the term of variance $\mathbb{E}[(\hat{F}(x) - \mathbb{E}[\hat{F}(x)])^2]$ must be conjointly minimized and given the bias-variance balance to quantify the accuracy of the estimator. Let $\mathcal{S} = \{F_\theta(X), \theta \in \Theta, X \in \mathbb{R}^p\}$ be not necessarily a nonlinear parametric regression model family with $Y = F_\theta(X) + \varepsilon$, $F_\theta: \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a continuous function of θ for fixed X and measurable for all θ fixed and Θ is a compact (i.e., closed and bounded) subset of the set of possible parameters of the regression model family. We assume that $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ is an observed data set coming from a true model $(X_t, Y_t)_{t \in \mathbb{N}}$ for which the true regression function is F_{θ^0} , for a θ^0 (possibly not unique) in the interior of Θ .

The nonlinear regression is often applicable in modeling empirical data for forecasting purposes, where relationships that underlie the measurements can be strongly nonlinear, nonunivocal, noisy with dynamic nature [1–6]. Nonlinear regression techniques such as projection-pursuit regression, support vector machine, and multilayer perceptron are applied in various fields [7–10]. For instance in agronomy, crop yield can be predicted, where $Y = \{\text{agricultural yields, insurance policy}\}$ and $X = \{\text{agronomic, edaphic, climatic and economic parameters}\}$; in climatology, estimation of long wave atmospheric radiation with $Y = \{\text{heating rates profile, radiation fluxes}\}$ and $X = \{\text{temperature, moisture, O}_3, \text{CO}_2, \text{cloud parameters profiles, surface fluxes, etc.}\}$, and in finance, estimation of financial characteristics with $Y = \{\text{future stock index, exchange rate, option price, etc.}\}$ and $X = \{\text{financial data up to a certain time}\}$.

Estimation of parameters is done using learning techniques. It consists to estimate the true parameter θ^0 based on the observations D_n . This can be done by minimizing the mean square error (MSE) function [11]: $R(\theta) = \mathbb{E}[\|Y -$

$F_\theta(x)\|_{L^2}^2$. The minimization of this quadratic error function leads to a biased θ^0 under the assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_q)$, when D_n is imprecise. To solve this problem of the misspecification of the quadratic error function, Badran and Thiria [1] proposed another function by supposing $\varepsilon \sim \mathcal{N}(0, \Sigma(x))$ using Bayesian approach. This hypothesis is arbitrary if not false due to the nature of some data in some fields of application [12]. In this work, we propose an extension of Badran and Thiria [1] results using new hypothesis on ε . One of these hypotheses is that ε follows the multivariate skew-normal distribution. This distribution refers to a parametric class of probability density function (p.d.f.) that extends the normal distribution using an additional shape parameter that regulates the skewness, allowing for a multiple continuous variation from normality to nonnormality. In addition, we compare the newly established error function with that of Badran and Thiria [1] in case of any covariance matrix, especially when the matrix is diagonal and its diagonal elements are constant on the one hand and nonidentical on the other hand. This comparison is made on simulated and real data in order to evaluate the possible performance of the proposed error function.

In Section 2 of the paper, we recall some existing concepts. In Section 3, we establish the main results. Section 4 presents simulation study and application case on real dataset. The conclusion is considered in Section 5. The proofs of the main results are presented in the Appendix.

2. Preliminary Definitions

2.1. Imprecise Data and Multivariate Skew-Normal Distribution

Definition 1. Imprecise data are data inconsistent in the way they are collected and/or registered (rounding, grouping, etc.). It is data with asymmetry or data showing a slight or heavy tail of the distribution.

Definition 2. A random multivariate variable $\varepsilon = (\varepsilon_1, \dots, \varepsilon_q)^\top$ of dimension $q \in \mathbb{N}^*$ is said to follow a multivariate skew-normal distribution, denoted by $\varepsilon \sim \mathcal{SN}(0, \bar{\Omega}, \alpha)$, if it is continuous and its density function is given by (Azzalini and Dalla Valle [13]):

$$h_q(\varepsilon) = 2\phi_q(\varepsilon; \bar{\Omega})\Phi(\alpha^\top \varepsilon), \varepsilon \in \mathbb{R}^q, \quad (1)$$

where $\phi_q(\cdot, \cdot)$ is the density function of a q -dimensional multivariate normal distribution with standardized marginals and correlation matrix $\bar{\Omega}$, $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution, and α is a q -dimensional vector that regulates departure from symmetry.

Consider the variance-covariance matrix $\Sigma = D_{1/\sigma}^{-1} \bar{\Omega} D_{1/\sigma}^{-1}$ with $D_{1/\sigma} = \text{diag}(1/\sigma_1, \dots, 1/\sigma_q)$, the diagonal matrix of dimension q of the inverse of standard deviation $1/\sigma$. The expression of the probability density function of ε is then

$$h_q(\varepsilon) = 2(2\pi)^{-q/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} \varepsilon^\top \Sigma^{-1} \varepsilon\right) \int_{-\infty}^{\alpha^\top \varepsilon} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right) du, \quad (2)$$

where $|\Sigma|$ denotes the determinant of matrix Σ . For more details, see Azzalini and Dalla Valle [13].

2.2. Bayes Approach. Suppose that $x^\top = (x_1, \dots, x_n)$ is a vector of n observations whose probability density function $p(x, \theta)$ depends on the values of k -dimensional parameter $\theta = (\theta_1, \dots, \theta_k)^\top$. Suppose also that θ itself has a probability density function $\pi(\theta)$. Then, denoting by $p(x)$ and $p(x|\theta)$, the marginal probability density function of x , and the conditional probability density function of x given θ , respectively, one has

$$p(x, \theta) = p(x|\theta)\pi(\theta) = p(\theta|x)p(x), \quad (3)$$

where given the observed data x , the conditional probability density function of θ when $p(x) \neq 0$ is

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)}. \quad (4)$$

By considering $p(x) = \mathbb{E}[p(x|\theta)] = c^{-1}$, we may write (4) alternatively as

$$p(\theta|x) = cp(x|\theta)\pi(\theta). \quad (5)$$

Therefore, Bayes formula in (5) is often written as $p(\theta|x) \propto p(x|\theta)\pi(\theta)$, which denotes that $p(\theta|x)$ is proportional to $p(x|\theta)\pi(\theta)$.

Definition 3. The prior distribution of θ , $\pi(\theta)$ in (5) is the probability distribution of θ without knowledge of x . In contrast, the posterior distribution of θ given x , $p(\theta|x)$ in (5) is the probability distribution of θ given x .

A Bayesian statistic model is performed from a parametric statistical model $p(x|\theta)$ and a prior distribution of parameter $\pi(\theta)$.

2.3. Learning. The learning consists to estimate θ of the regression function $F_\theta: = \mathbb{E}(Y|X=x)$ that minimizes the loss function $R(\theta) = \mathbb{E}(\mathbb{L}(\theta, x, y))$. A simple expression of this function is the generalized least squares (GLS) error function given by

$$R(\theta) = \iint (y - F_\theta(x))^\top \Sigma(x)^{-1} (y - F_\theta(x)) d\mathbb{P}(x, y), \quad (6)$$

where $d\mathbb{P}(x, y) = p(x, y)dx dy$ with $p(x, y)$, the joint probability density function of (X, Y) , and $\Sigma^{-1}(x)$ is the inverse of the covariance matrix $\Sigma(x)$ of the conditional random variable $Y|X=x$, which is a symmetric positive definite matrix of order q .

3. Main Results

3.1. Existence of the Minimum of the GLS Error Function. The following result provides the solution of the minimization problem of (6).

Theorem 1. Let $\Sigma(x)$ be the unknown covariance matrix of the conditional random variable $Y|X=x$, $F_\theta(x) \in \mathbb{R}^q$, $q \in \mathbb{N}$

be the output vector of the regression function, and $\mathbb{E}(Y | X = x)$ be the conditional mean vector of the observations $y \in \mathbb{R}^q$.

For any value of $(x, \Sigma(x))$, $\text{argmin}R(\theta)$ with respect to θ exists and is given by

$$\text{argmin}_{\theta} \left[\int [\mathbb{E}(Y | X = x) - F_{\theta}(x)]^{\top} \Sigma(x)^{-1} [\mathbb{E}(Y | X = x) - F_{\theta}(x)] p(x) dx \right]. \quad (7)$$

The proof of Theorem 1 is given in the Appendix.

However, θ from (7) cannot be analytically obtained because $\Sigma^{-1}(x)$ and the marginal probability density function $p(x)$ are unknown. We consider the case where $\Sigma^{-1}(x)$ is known and does not depend on x . Thus, the solution of the minimization problem of (6) leads to the following theorem.

Theorem 2. Let $\mathbf{U} = (u_1, \dots, u_q)$ be orthogonal matrix of the eigen vectors of $\Sigma(x)$ and σ_k^2 the eigenvalue associated to u_k ($1 \leq k \leq q$). Suppose $\Sigma(x)^{-1} = \mathbf{U} \mathbf{D}^{-1} \mathbf{U}^{\top}$ does not depend on x with $\mathbf{D}^{-1} = \text{diag}(1/\sigma_1^2, \dots, 1/\sigma_q^2)$, the $q \times q$ inverse diagonal matrix. Thus, we have from (6) the existence of $\text{argmin}_{\theta} mR(\theta)$ given by

$$\text{argmin}_{\theta} \left[\sum_{k=1}^q \left(\frac{u_k^{\top} [\mathbb{E}(Y | X = x) - F_{\theta}(x)]}{\sigma_k} \right)^2 p(x) dx \right]. \quad (8)$$

The proof of Theorem 2 is given in the Appendix.

Remark 1. Under the conditions of Theorem 1 or 2, the minimum of problem of (6) with respect to θ is reached with $F_{\theta}(x) \approx \mathbb{E}(Y | X = x)$.

The argument of minimizing the theoretical cost function R given by (6) with respect to θ is difficult to get because $p(x)$ and $\Sigma(x)$ are unknown. However, we can obtain a finite set of independent observations $D_n = \{(y_t, x_t), t = 1, \dots, n\}$, $n \in \mathbb{N}$, which allows to define the discrete approximation of the theoretical cost function.

Therefore, $F_{\theta}(x) \approx \mathbb{E}(Y | X = x)$ is obtained by applying the optimization procedure on the empirical cost function $R_n(\theta, x, y)$ given by

$$\text{argmin}_{\theta} R_n(\theta, x, y) = \text{argmin}_{\theta} \left[\sum_{t=1}^n (y_t - F_{\theta}(x_t))^{\top} \Sigma(x)^{-1} (y_t - F_{\theta}(x_t)) - \ln |\Sigma(x)^{-1}| - 2 \ln \int_{-\infty}^{\alpha^{\top} (y_t - F_{\theta}(x_t))} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} u^2\right) du \right], \quad (11)$$

with $\Sigma(x) = D_{1/\sigma}^{-1} \bar{\Omega} D_{1/\sigma}^{-1}$.

The proof of Theorem 3 is given in the Appendix.

In (11), it appears as an extension of the empirical expression (9). The additional terms $\ln \Phi(\alpha^{\top} (y_t - F_{\theta}(x_t)))$ and $\ln |\Sigma(x)^{-1}|$ in the expression (9) contribute in the process to take into account of the imprecise data. They can

$$R_n(\theta, x, y) = \sum_{t=1}^n R_t(\theta), \quad (9)$$

with

$$R_t(\theta) = (y_t - F_{\theta}(x_t))^{\top} \Sigma(x_t)^{-1} (y_t - F_{\theta}(x_t)), \quad (10)$$

which is a discrete approximation of the theoretical cost function R given by (7).

In the next subsection, we use Bayesian approach to take into account uncertainties appearing during the regression process by considering some hypotheses linked to the nature of ε . In this case, the maximum likelihood model provides an extension of the empirical cost function (9).

3.2. Extension of GLS considering the Output Noise. Results considered below are established under the following assumptions related to the random i.i.d. noises $\varepsilon_t = Y_t - F_{\theta}(X_t)$.

- A1. $\forall t \in [1, n]$ and $\forall 1 \leq j \leq p; \forall 1 \leq i \leq q$, $\text{Cov}(X_{t_j}, \varepsilon_{t_i}) \neq 0$
- A2. $\varepsilon_t | X_t = x_t \sim \mathcal{SN}(0, \bar{\Omega}, \alpha)$
- A3. x_t is not spoiled and does not depend on θ

Theorem 3. Assume that the conditions A1–A3 hold. Let $R_n(\theta, x, y)$ be the empirical cost function, $p(\theta | D_n)$ be the posterior distribution of the model parameters ($\theta \in \Theta$) known as finite set of independent observations $D_n = \{(y_t, x_t), t = 1, \dots, n\}$, $n \in \mathbb{N}^*$ and $\pi(\theta)$, the prior distribution of θ which is assumed to have a uniform distribution, and $p(D_n | \theta)$ be the likelihood of D_n knowing θ . Thus,

be known or not; if not, they have to be considered and estimated during the minimization process.

Corollary 1. Under the conditions of Theorem 3, if the vector α is equal to the vector null, we have the same results with Badran and Thiria [1], which is

$$\operatorname{argmin}_{\theta} R_n(\theta, x, y) = \operatorname{argmin}_{\theta} \left[\sum_{t=1}^n (y_t - F_{\theta}(x_t))^{\top} \sum (x)^{-1} (y_t - F_{\theta}(x_t)) - \ln \left| \sum (x_t)^{-1} \right| \right]. \quad (12)$$

The simulation study presented below is based on multilayer perceptron neural networks (MLP) with one hidden layer. The output variable is continuous and bivariate. This output variable has a deterministic part (its true value) plus an additional random part which is asymmetric. The same MLP architecture is used to predict the bivariate response from simulated data at different sample sizes. Three different empirical cost functions have been used in the learning process of the imprecise data generated for prediction purposes. These are (i) the classical generalized least squares function (CGLS) (see (10)), (ii) the extension of the generalized least squares function proposed by [1] (EGLS_B) (see (12)), and (iii) the new extension of the generalized least squares function that we proposed (NEGLS) (see (11)).

The coefficient of determination, R^2 , and the mean absolute percentage error, $MAPE$, allowed comparing the performance of the estimation of the nonlinear regression function based on these three empirical cost functions.

4. Simulation Study and Real Data Applications

In this section, we consider the regressor F_{θ} as a multilayer perceptron neural network (MLP).

A MLP is a set of units (neurons) organized in successive layers (first layer = input layer, last layer = output layer, and the middle layer(s) = hidden layer(s)), which are interconnected with the weights (θ). The connections are always directed from lower layers to upper layers and the neurons of the same layer are not interconnected.

The MLP model (F_{θ}) considered here has four input neurons noted X ($X \in \mathbb{R}^4$), three hidden units which represent 75% of input neurons, and two output neurons denoted Y , $Y \in \mathbb{R}^2$ (see Figure 1).

Consider

$w^{\top} = (w_{11}, \dots, w_{14}, \dots, w_{31}, \dots, w_{34}; w_{10}, \dots, w_{30}) \in \mathbb{R}^{15}$ as a vector of parameters between the input and hidden layers, $\beta^{\top} = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23}; \beta_{10}, \beta_{20}) \in \mathbb{R}^8$ as a vector of parameters between the hidden and output layers and $\Theta_m \subset \mathbb{R}^{23}$ as a compact subset of possible parameters of the regression model family $\mathcal{S} = \{F_{\theta}(X), \theta \in \Theta_m, X \in \mathbb{R}^4\}$. g and f (real value functions) are considered in the model as the output and hidden-unit activation function, respectively.

4.1. Simulation Plan. Let $X \in \mathbb{R}^4$ be a vector of input variables and $Y \in \mathbb{R}^2$, a bivariate output variable. We have $Y = G(X) + \varepsilon$ where $G: \mathbb{R}^4 \rightarrow \mathbb{R}^2$ and $\varepsilon \in \mathbb{R}^2$ is a vector of residual variables. To guarantee the nonlinearity of G , we chose

$$G(X_1, X_2, X_3, X_4) = (c_2 X_2 \exp(c_4 X_4), c_1 \log(X_1) + c_3 X_3). \quad (13)$$

The coefficient used in the simulation study was pre-specified as $c_1 = 3.06$; $c_2 = 0.15$; $c_3 = 5.47$; $c_4 = -0.07$. Concerning the vector $X = (X_1, X_2, X_3, X_4)^{\top}$, its components, respectively, follow normal, log-normal, binomial negative, and Poisson laws, with $X_1 \sim \mathcal{N}(\mu = 3.86, \sigma^2 = 0.35)$, $X_2 \sim \log - \mathcal{N}(\mu = 5.12, \sigma^2 = 1.16)$, $X_3 \sim \mathcal{BN}(\nu = 0.36; p = 12.18)$, and $X_4 \sim P(\lambda = 2.4)$. The set of distributions of the different X components allowed having X distribution of probability unknown as in the previous theorems. The additional random part at $G(X)$ is $\varepsilon = (\varepsilon_1, \varepsilon_2)$ with $\varepsilon \sim \mathcal{SN}(0, \Sigma(x), \alpha)$, where $\alpha = (-4, 4)^{\top}$ corresponding to a case of bivariate skew-normal distribution. Consider any covariance matrix $\Sigma(x)$ assumed to be known. Thus, we arbitrarily chose $\Sigma(x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ so that the errors $\varepsilon = (\varepsilon_1, \varepsilon_2)$ are centered, reduced, and uncorrelated, and $\Sigma(x) = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$ so that the errors $\varepsilon = (\varepsilon_1, \varepsilon_2)$ are centered, reduced, and correlated (see Figure 2).

The conditions A1–A3 and the previous information made it possible to randomly generate a population of size $N = 10,000$.

Simulation's algorithm is presented in the following steps:

Step 1. Generate the input variables X_1 to X_4 and ε from their respective distribution for $N = 10,000$.

Step 2. Calculate $G(X)$ using (13) from values of coefficients c_1 to c_4 and generated input variables X_1 to X_4 .

Step 3. Generate outcome variables Y for $N = 10,000$, from each ε distribution adds to $G(X)$.

Step 4. Consider a multivariate Bootstrap sample of size n ($n = 50, 150, 1000$) from the populations generated at previous steps.

Step 5. For each Bootstrapping sample, we run MLP model on 70% of the data (training dataset) and the remaining 30% (test dataset) are intended for predictive performance analysis.

The characteristics considered for the model are as follows:

- (1) Activation functions ($f(z) = (e^z - e^{-z}) / (e^z + e^{-z})$ and $g(z) = z$);
- (2) Number of hidden neurons = 3;
- (3) Learning rate = 40%;
- (4) The initial value of $\theta \sim \mathcal{U}[-3, 3]$;
- (5) Learning algorithm = standard backpropagation, and here each of the three empirical cost functions (CGLS, EGLS_B, NEGLS) is considered.

Figure 1: Multilayer neural network with one hidden layer: MLP 4-1-2.

FIGURE 2: Simulation of (a) uncorrelated and (b) correlated bivariate normal skew errors for $n = 1000$.

Step 6. For each combination of n and ε distribution, we repeat the previous step 1000 times, and the mean of the performance criteria is computed.

The R software (v. 3.3.6) [14] is used for data generation and model implementation.

4.2. Performance Criteria. Two performance criteria are used on test data: coefficient of determination, R^2 , and mean absolute percentage error, $MAPE$ [15, 16].

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - F_{\theta}(x_t))^2}{\sum_{t=1}^n (y_t - \bar{y})^2}, \quad (14)$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{F_{\theta}(x_t) - y_t}{y_t} \right| \times 100.$$

A R^2 close to “1” indicates good performance. $MAPE$ is a measure of the accuracy of the forecast. A value close to “0” indicates good accuracy.

4.3. Findings Related to Simulation Study. Comparing the performance of nonlinear regression models based on multilayer perceptual neural networks using three different quadratic error functions in the learning process reveals that the larger the sample size, the more accurate the model (low value of $MAPE$ and high values of R^2). On the other hand, when the errors are centered, reduced, and uncorrelated, the error functions GLS and $EGLS_B$ give the same results but $NEGLS$ gives a better result than them (Table 1). For centered, reduced, and correlated errors, the three error functions each used in the learning process lead to different results. The best performance regardless of sample size is obtained with the quadratic error function $NEGLS$, followed by $EGLS_B$ and finally GLS (Table 1).

4.4. Real Data Applications: AIS Dataset. In this subsection, an application of the new quadratic error function and the two other to real-world data is shown. Data refer to 13 biological characteristics of 102 male and 100 female athletes

TABLE 1: Comparison of statistical performances.

Error function	Constrained on the errors	$n = 50$		$n = 150$		$n = 1000$	
		R^2	$MAPE$	R^2	$MAPE$	R^2	$MAPE$
GLS	Errors are centered, reduced, and uncorrelated	0.599	3.586	0.658	3.365	0.857	1.011
		0.819	4.646	0.452	2.831	0.783	0.845
$EGLS_B$		0.599	3.586	0.658	3.365	0.857	1.011
		0.819	4.646	0.452	2.831	0.783	0.845
$NEGLS$		0.643	2.434	0.787	2.105	0.933	0.415
		0.902	2.917	0.929	1.783	0.945	0.326
GLS		0.705	2.733	0.772	1.758	0.821	0.681
		0.493	1.734	0.590	0.718	0.630	0.429
$EGLS_B$		0.798	2.069	0.832	1.334	0.899	0.511
		0.588	1.240	0.667	0.526	0.755	0.297
$NEGLS$	0.805	1.540	0.901	1.258	0.912	0.426	
	0.693	1.011	0.704	0.333	0.919	0.134	

GLS =classical generalized least squares; $EGLS$ =extension of generalized least squares proposed by I; $NEGLS$ =new extension of generalized least squares proposed by us; R^2 = coefficient of determination; $MAEP$ = mean absolute percentage error.

TABLE 2: Comparison of statistical performances of quadratic error functions with imprecise real dataset.

Error function	Constrained on the errors	$n = 202$	
		R^2	$MAPE$
GLS	Errors are centered and correlated	0.504	4.11
		0.481	2.540
$EGLS_B$		0.799	3.586
		0.819	1.646
$NEGLS$		0.871	2.117
		0.899	0.718

recorded at the Australian Institute of Sport, courtesy of Richard Telford and Ross Cunningham. The data have been presented and examined by [17] and are available in the package “sn” of R software under the name “ais.” The set is repeatedly used in the literature about skewed distributions. In our application, we only use a subset of these characteristics used by Azzalini [18] for fitting linear models with skew-elliptical error term and in addition to some biological characteristics. Output bivariate variable considered is body mass index (BMI) and lean body mass (LBM) with $\alpha = (4.999810, -1.767682)^\top$ and $\Sigma(x) = \begin{pmatrix} 7.034139 & 20.02137 \\ 20.02137 & 117.90713 \end{pmatrix}$. The input variables are red cell count (RCC), white cell count (WCC), hematocrit (Hc), hemoglobin (Hg), sum of skin folds (SSF), and body fat percentage (Bfat). These data were fitted to nonlinear regression models based on multilayer perceptron neural networks as in the simulation study.

4.5. Findings Related to Real Data Applications. Analysis of the performance of the MLP model on real data using the quadratic error functions $NEGLS$, $EGLS_B$, and GLS in a context where the errors are centered and correlated (Table 2) shows that $NEGLS$ gives the best precision and low bias compared with the other two functions (low value of $MPAE$ and high values of R^2).

5. Conclusion

In the case of imprecise data prediction, nonlinear regression models are more appropriate. In the context of parametric models with multivariate response, parameter estimation based on the traditional generalized least squares error function (GLS) leads to biases and the models obtained are less efficient. Results obtained from our simulation study reveal the importance of developing a new quadratic error function to learn these types of data rather than using a traditional GLS error function. Our proposed error function provides more flexibility to take into account the nature of noise in the imprecise data during the learning process. The next step will be to develop a learning algorithm and a package to learn imprecise data for prediction purposes of nonlinear system.

Appendix

A. Proof of Theorem 1

Let

$$R(\theta) = \iint (y - F_\theta(x))^\top \Sigma(x)^{-1} (y - F_\theta(x)) d\mathbb{P}(x, y) \quad (A.1)$$

be the least squares error function (cost function) in the case of a nonlinear regression model (F_θ). Replace $d\mathbb{P}(x, y)$ by $p(x, y)dx dy$ in equation (A.1), and we have

$$R(\theta) = \iint (y - F_\theta(x))^\top \Sigma(x)^{-1} (y - F_\theta(x)) p(x, y) dx dy. \quad (A.2)$$

Using the Bayes rules, $p(x, y) = p(y|x)p(x)$, in equation (A.2) gives

$$R(\theta) = \iint \left(\int (y - F_\theta(x))^\top \Sigma(x)^{-1} (y - F_\theta(x)) p(y|x) dy \right) p(x) dx. \quad (A.3)$$

We can now rewrite (A.2) as

$$R(\theta) = \int A(\theta)p(x)dx, \quad (\text{A.4})$$

with

$$A(\theta) = \int [(y - \mathbb{E}(Y|X=x)) + (\mathbb{E}(Y|X=x) - F_\theta(x))]^\top \Sigma(x)^{-1} [(y - \mathbb{E}(Y|X=x)) + (\mathbb{E}(Y|X=x) - F_\theta(x))] p(y|x) dy. \quad (\text{A.5})$$

Using the distributivity rule with respect to $\Sigma^{-1}(x)$ and doing some additional simplifications, equation (A.5) becomes

$$\begin{aligned} A(\theta) = & \int (y - \mathbb{E}(Y|X=x))^\top \Sigma(x)^{-1} (y - \mathbb{E}(y|X=x)) p(y|x) dy \\ & + 2(\mathbb{E}(Y|X=x) - F_\theta(x))^\top \Sigma(x)^{-1} \int (y - \mathbb{E}(y|X=x)) p(y|x) dy \\ & + (\mathbb{E}(Y|X=x) - F_\theta(x))^\top \Sigma(x)^{-1} (\mathbb{E}(Y|X=x) - F_\theta(x)) \int p(y|x) dy. \end{aligned} \quad (\text{A.6})$$

In equation (A.6),

$$\int (y - \mathbb{E}(Y|X=x))^\top \Sigma(x)^{-1} (y - \mathbb{E}(y|X=x)) p(y|x) dy \quad (\text{A.7})$$

does not depend on θ . $\mathbb{E}(Y|X=x) - F_\theta(x)$ and $\Sigma^{-1}(x)$ does not depend on y , and then we have

$$\begin{aligned} \int (y - \mathbb{E}(Y|X=x)) p(y|x) dy &= 0, \\ \int p(y|x) dy &= 1. \end{aligned} \quad (\text{A.8})$$

The minimization of $A(\theta)$ with respect to θ depends only on $(\mathbb{E}(Y|X=x) - F_\theta(x))^\top \Sigma(x)^{-1} (\mathbb{E}(Y|X=x) - F_\theta(x))$. We therefore have

$$\operatorname{argmin}_\theta R(\theta) = \operatorname{argmin}_\theta \left(\int (\mathbb{E}(Y|X=x) - F_\theta(x))^\top \Sigma(x)^{-1} (\mathbb{E}(Y|X=x) - F_\theta(x)) p(x) dx \right). \quad (\text{A.9})$$

B. Proof of Theorem 2

Replacing $\Sigma(x)^{-1} = \mathbf{U}D^{-1}\mathbf{U}^\top$ in equation (A.2) gives

$$R(\theta) = \iint (\mathbf{U}^\top y - \mathbf{U}^\top F_\theta(x))^\top D^{-1} (\mathbf{U}^\top y - \mathbf{U}^\top F_\theta(x)) p(x, y) dx dy. \quad (\text{B.1})$$

And the discrete form of $R(\theta)$ can be written as follows:

$$R(\theta) = \sum_{k=1}^q \iint \left(\frac{u_k y - u_k F_\theta(x)}{\sigma_k} \right)^2 p(x, y) dx dy = \sum_{k=1}^q \iint \left(\frac{u_k y - u_k F_\theta(x)}{\sigma_k} \right)^2 p(y|x) dy p(x) dx. \quad (\text{B.2})$$

Let us consider

$$B(\theta) = \int \left(\frac{u_k y - u_k F_\theta(x)}{\sigma_k} \right)^2 p(y|x) dy. \quad (\text{B.3})$$

With some development at the internal of B by adding $1/\sigma_k^2 [\mathbb{E}(u_k y | X = x)]^2 - 1/\sigma_k^2 [\mathbb{E}(u_k y | X = x)]^2$, it becomes

$$B(\theta) = \left(\frac{\mathbb{E}(u_k y | X = x) - u_k F_\theta(x)}{\sigma_k} \right)^2 + 1, \quad (\text{B.4})$$

$$R(\theta) = \sum_{k=1}^q \left(\frac{u_k [\mathbb{E}(y | X = x) - F_\theta(x)]}{\sigma_k} \right)^2 p(x) dx + \sum_{k=1}^q \int p(x) dx. \quad (\text{B.5})$$

Therefore,

$$\text{argmin}_\theta R(\theta) = \text{argmin}_\theta \sum_{k=1}^q \left(\frac{u_k^\top [\mathbb{E}(Y | X = x) - F_\theta(x)]}{\sigma_k} \right)^2 p(x) dx. \quad (\text{B.6})$$

C. Proof of Theorem 3

Consider the following expression:

$$\text{argmax}_\theta (p(\theta | D_n)) = \text{argmin}_\theta [-2 \ln(p(\theta | D_n))]. \quad (\text{C.1})$$

Using the Bayes formula in (4), we have

$$\text{argmin}_\theta [-2 \ln(p(\theta | D_n))] = \text{argmin}_\theta [-2 \ln(p(D_n | \theta)) - 2 \ln(\pi(\theta)) + 2 \ln(p(D_n))]. \quad (\text{C.2})$$

From this equation, $p(D_n)$ is constant with respect to θ (see (5)). $\ln(\pi(\theta))$ is a regularization term and corresponds to an additional constraint on the θ distribution. By considering $\theta \sim \mathcal{U}[a, b]$, a and b are not depending on θ . Then, we can remove $p(D_n)$ and $\pi(\theta)$ in the minimization process of $\ln(p(\theta | D_n))$. By considering the independent observations, we get

$$\ln(p(D_n | \theta)) = \sum_{t=1}^n \ln(p((x_t, y_t) | \theta)). \quad (\text{C.3})$$

Its decomposition gives

$$\ln(p(D_n | \theta)) = \sum_{t=1}^n \ln(p(y_t | x_t, \theta)) + \sum_{t=1}^n \ln(p(x_t | \theta)). \quad (\text{C.4})$$

The first term of (C.4) is the probability that y_t comes from x_t when the conditional probability is generated by the nonlinear regression model $F_\theta(x_t)$. In the second term, considering that the hypothesis A3 is satisfied, it can be removed in the minimization process.

Recall that ε is the stochastic part of Y ($\varepsilon_t = Y_t - Y_v$) and depends on X , with Y_v being the deterministic part of Y . It satisfies Definition 1. Assuming that hypotheses A1 and A2 hold, we have from (2) of Definition 2:

$$p(\varepsilon | x) = 2(2\pi)^{-q/2} |\sum(x)|^{-1/2} \exp\left(-\frac{1}{2}(y - y_v)^\top \sum(x)^{-1}(y - y_v)\right) \Phi(\alpha^\top (y - y_v)), \quad (\text{C.5})$$

with

$$\sum(x) = D_{1/s}^{-1} \bar{\Omega} D_{1/s}^{-1},$$

$$\Phi(\alpha^\top (y - y_v)) = \int_{-\infty}^{\alpha^\top (y - y_v)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du. \quad (\text{C.6})$$

Assume an acceptable nonlinear regression model such that $y_{vt} = F_\theta(x_t) \approx \mathbb{E}(Y | X = x_t)$; then, we can simplify the term $\ln(p(y_t | x_t, \theta))$ by replacing y_{vt} with $F_\theta(x_t)$ and have

$$\begin{aligned} -2 \ln(p(y_t | x_t, \theta)) &= (y_t - F_\theta(x_t))^\top \sum (x)^{-1} (y_t - F_\theta(x_t)) - \ln \left| \sum (x_t)^{-1} \right| \\ &\quad - 2 \ln \Phi(\alpha^\top (y_t - F_\theta(x_t))) - 2 \ln 2 + q \ln 2 \pi. \end{aligned} \quad (\text{C.7})$$

Hence, we can define the empirical risk from (C.7) as

$$\begin{aligned} \operatorname{argmin}_\theta [R_N(\theta, x, y)] &= \operatorname{argmin}_\theta [-2 \ln(p(\theta | D_n))] \\ &= \operatorname{argmin}_\theta \left[\sum_{t=1}^n R_t(\theta) \right], \end{aligned} \quad (\text{C.8})$$

with

$$\begin{aligned} R_t(\theta) &= \left[(y_t - F_\theta(x_t))^\top \sum (x)^{-1} (y_t - F_\theta(x_t)) \right] \\ &\quad - \ln \left| \sum (x_t)^{-1} \right| - 2 \ln \Phi(\alpha^\top (y_t - F_\theta(x_t))). \end{aligned} \quad (\text{C.9})$$

Data Availability

The Australian Institute of Sport (AIS) data used to support the results of this study have been posted on <https://rdrr.io/cran/MoEClust/man/ais.html> or <https://www.rdocumentation.org/packages/sn/versions/1.5-4/topics/ais>. In addition, these data were made public in the book by Cook and Weisberg [17].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] F. Badran and S. Thiria, "Les perceptrons multicouches: de la régression non-linéaire aux problèmes inverses," *Journal de Physique IV (Proceedings)*, vol. 12, no. 1, pp. 157–188, 2002.
- [2] M. Cottrell, M. Olteanu, F. Rossi, J. Rynkiewicz, and N. Vialaneix, "Neural networks for complex data," *Künstliche Intelligenz*, vol. 26, pp. 1–8, 2012.
- [3] C. López-Martín, "Predictive accuracy comparison between neural networks and statistical regression for development effort of software projects," *Applied Soft Computing*, vol. 27, pp. 434–449, 2015.
- [4] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481–1497, 1990.
- [5] M. Kordos and W. Duch, "Variable step search algorithm for feedforward networks," *Neurocomputing*, vol. 71, no. 13–15, pp. 2470–2480, 2008.
- [6] J. Gaudart, B. Giusiano, and L. Huiart, "Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data," *Computational Statistics & Data Analysis*, vol. 44, no. 4, pp. 547–570, 2004.
- [7] H. Y. Hen and J. N. Hwang, *Handbook of Neural Network Signal Processing*, CRC Press, London, UK, 2001.
- [8] K. Gurney, *An Introduction to Neural Networks*, CRC Press, London, UK, 2018.
- [9] D. Graupe, *Principles of Artificial Neural*, World Scientific Publishing Company, Singapore, 3rd edition, 2013.
- [10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer Series in Statistics, Berlin, Germany, 2011.
- [11] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Vol. 15, SIAM, Philadelphia, PA, USA, 1995.
- [12] J. Rynkiewicz, "General bound of overfitting for MLP regression models," *Neurocomputing*, vol. 90, pp. 106–110, 2012.
- [13] A. Azzalini and A. Dalla Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, no. 1, pp. 715–726, 1996.
- [14] R Core Team, *R 3.3.6: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019, <https://www.R-project.org/>.
- [15] H. Elarabi and N. F. Taha, "Effect of different factors of neural network on soil profile of khartoum state," *American Journal of Earth Sciences*, vol. 1, no. 3, pp. 62–66, 2014.
- [16] H. A. Kazem and J. H. Yousif, "Comparison of prediction methods of photovoltaic power system production using a measured dataset," *Energy Conversion and Management*, vol. 148, no. 1, pp. 1070–1081, 2017.
- [17] R. D. Cook and S. Weisberg, *An Introduction to Regression Graphics*, John Wiley & Sons, New York, NY, USA, 1994.
- [18] A. Azzalini, *The R Package "sn": The Skew-Normal and Related Distributions Such as the Skew-T (version 1.5-2)*, 2018, <http://azzalini.stat.unipd.it/SN>.