Advances in Data Analysis and Classification manuscript No.
(will be inserted by the editor)

An Isotonic Trivariate Statistical Regression Method

Simone Fiori

Received: November 13, 2012. Resubmitted in revised form: February 11, 2013. Accepted: April 03, 2013.
Published as: **S. Fiori**, "An Isotonic Trivariate Statistical Regression Method", *Advances in Data Analysis and Classification* (Springer), Vol. 7, No. 2, pp. 209 - 235, 2013

Abstract The present research work outlines the main ideas behind statistical regression by a 2-independent-variates and 1-dependent-variate model based on the invariance of measures in probabilistic spaces. The principle of probabilistic measure invariance, applied under the assumption that the model be isotonic, leads to a system of differential equations. Such differential system is reformulated in terms of an integral equation that affords an iterative numerical solution. Numerical tests performed on the devised statistical regression procedure illustrate its features.

Keywords Statistical regression · Dominant independent variates · Isotonic regression · Integral equation

Mathematics Subject Classification (2000) 62G08 · 68P99 · 65R20

1 Introduction

Regression is a data mining technique that predicts a target value such as profit, sales, house values or temperature on the basis of designated predictors. For example, a regression model could be used to predict the value of a house (target) based on location, number of rooms and lot size (predictors). A regression task begins with a data set in which the target values as well as the predictors values are known. For example, a regression model that predicts house values could be developed on the basis of observed data for a large number of properties over a sufficiently large period of time. A regression algorithm estimates the value of the target as a function of the predictors for each case in the data set. These relationships between predictors and target

are summarized in a model, which can then be applied to a different data set in which the target values are unknown. Regression has several applications in trend analysis, business planning, marketing, financial forecasting, drug response prediction as well as environmental modeling. Multivariate nonlinear regression refers to nonlinear regression with two or more predictors.

Statistical multivariate regression provides a useful tool to build up a model of a phenomenon under observation. The qualification *statistical* refers to the distinguishing feature of such class of regression methods that do not make direct use of the data set to infer a model underlying the data but that makes use of their statistical features. Statistical regression is applied in a variety of research fields, such as integrated electronics [1,16], environmental research [14,18], fluids transport [28], atmospheric research [15], artificial intelligence [17], geotechnical engineering [30], toxicology research [5], biomass engineering [19], analysis of food quality [8], analysis of water quality in lakes [25], chemical engineering [23] as well as demography [2]. The largest part of available statistical regression techniques concerns bivariate regression, although multivariate statistical regression is of great interest in applications (see, e.g., [5,21]). In the present context, the statistical features are summarized by the joint probability density functions of the target and of the predictors. As intended here, *isotonic trivariate statistical regression* is based on two main assumptions:

- The physical phenomenon under observation relates a set of independent variables with a single dependent variable. Among the independent variables, only *two of them* play a prominent role in the description of the phenomenon under observation, while the remaining independent variables are regarded as nuisance parameters. Such assumption explains the qualification *trivariate*.
- The statistical regression model is *monotonically increasing or decreasing* (or, equivalently, it is of *dose-response* type). The hypothesis of monotonicity in data modeling occurs frequently in applied fields such as data regression and data mining [26] and explains the qualification *isotonic*. Isotonic statistical regression is also referred to, in the scientific literature, as *regression under order restrictions* [3,22].

The present section explains the statistical regression problem in an analytic fashion. It is assumed that $n+3$ variates of interest in a regression problem are related by the functional relationship:

$$y = \Phi(x_1, x_2, \nu_1, \ldots, \nu_n), \tag{1}$$

where $y \in \mathcal{Y}$ represents the *dependent variate* or *target*, $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^2$ represent the *dominant independent variates* or *predictors* and $(\nu_1, \ldots, \nu_n) \in \mathcal{N} \subset \mathbb{R}^n$ represent the *nuisance variates*, according to a physical phenomenon described by the function Φ. The nuisance variates may account for measurement errors of the variables y, x_1, x_2. The joint statistical features of the target and of the predictors are described by the joint probability density function $p_{y,x_1,x_2,\nu_1,\ldots,\nu_n}(y, x_1, x_2, \nu_1, \ldots, \nu_n)$, where, by a slight abuse of notation, the variates have been confused with their realizations. The marginal joint

probability density function of the dependent variate and of the dominant independent variates is described by:

$$p_{y,x_1,x_2}(y,x_1,x_2) \stackrel{\text{def}}{=} \int_{\mathcal{N}} p_{y,x_1,x_2,\nu_1,\ldots,\nu_n}(y,x_1,x_2,\nu_1,\ldots,\nu_n)\,\mathrm{d}\nu_1\cdots\mathrm{d}\nu_n. \quad (2)$$

Then, the other required marginal probability density functions may be obtained by integrating with respect to the variables y, x_1, x_2.

The discrimination between the dominant variates and the nuisance variates may be effected on the basis of a correlation analysis between each predictors and the target and by selecting the two predictors that show a significantly larger correlation with the target compared to the remaining predictors.

Trivariate isotonic statistical regression is about determining a model of the relationship between the two dominant independent variates $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ and the dependent variate $y \in \mathcal{Y}$, described by a function $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$, on the basis of a *probabilistic measure invariance principle*. The condition that the model be isotonic is formalized by:

$$\frac{\partial f}{\partial x_1} \neq 0 \text{ and } \frac{\partial f}{\partial x_2} \neq 0, \forall (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2. \quad (3)$$

Except for the condition that the model be isotonic, the shape of the model f is unrestricted. The present research extends the previous work on bivariate isotonic regression [12,13] and was deemed necessary as the connection between the two problems and methods is nontrivial. It is important to underline that the present approach, unlike different approaches known from the scientific literature (see, e.g., [9], for a general approach and [29] for a recent discussion on advantages/disadvantages of optimization-based approaches), does not rely on a formulation based on optimization but it is based uniquely on the principle of probabilistic measure invariance of statistics. In the case of nonlinear regression, optimization methods have been used to determine the parameters which *best fit* the data, by minimizing, in the most common cases, a least squares expression. As a matter of fact, in general, there is no closed-form expression for the best-fitting parameters, therefore, usually numerical optimization algorithms are applied to determine the best values of the parameters. There may be many local minima of the function to be optimized and even the global minimum may produce a biased estimate. According to the survey [27], it is unlikely that the optimal parameters which minimize some least squares formulation are the only reasonable parameters, because there are several sources of uncertainty that can contribute to difficulty in identifying optimal parameter values in nonlinear problems. In fact, the data may be affected by significant uncertainties (such as missing values, measurement errors and systematic biases), the nonlinear inverse problems may involve discontinuities which result in multiple values for the optimal parameters due to complexities in the underlying physics, and the model form can also influence the parameter settings. The statistical regression method suggested within the present contribution stems from a different starting point. In fact, it is not based on any parametric model and does not rely on the raw data, but it is

based on an unrestricted model (which can take any possible shape, except for the fundamental requirement of monotonicity with respect to the descriptors) and on cumulative joint statistical features of the data.

The present manuscript is organized as follows. Section 2 of the present paper discusses the formalization of the statistical regression problem at hand on the basis of a probabilistic measure invariance principle. Section 3 proposes a reformulation of the differential constraints on the model arising from the probabilistic invariance principle as an integral equation and some possible boundary conditions to complete the formulation. Section 4 discusses the numerical implementation of the iterative procedure to solve such integral equation and illustrates the behavior of the devised numerical implementation by numerical tests. Section 5 discusses in details the implementation of the regression method and illustrates its numerical features either on synthetic and real-world data sets. Section 6 concludes the paper.

2 Isotonic trivariate statistical regression by a probabilistic measure invariance principle

Consider the non-linear system with two input variables and two output variables:

$$(y, z) = \varphi(x_1, x_2), \tag{4}$$

where $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^2$, $(y, z) \in \mathcal{Y} \times \mathcal{Z} \subset \mathbb{R}^2$ and $\varphi : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y} \times \mathcal{Z}$. The non-linear system is supposed to be invertible in the domain of interest and its inverse is denoted by $\varphi^{-1} : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}_1 \times \mathcal{X}_2$.

Assume that (x_1, x_2) are random variables with joint probability density function denoted by $p_{x_1,x_2}(x_1, x_2)$. Such random variables transform into two random variables (y, z) that are distributed according to the probability density function $p_{y,z}(y, z)$. Define the Jacobian of the system φ as:

$$J_\varphi(x_1, x_2) \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} \\ \frac{\partial z}{\partial x_1} & \frac{\partial z}{\partial x_2} \end{bmatrix}. \tag{5}$$

The principle of invariance of probabilistic measures prescribes that the joint probability density function of the output variates be related with the joint probability density function of the input variates by the relationship:

$$|\det J_\varphi(x_1, x_2)|(p_{y,z} \circ \varphi)(x_1, x_2) = p_{x_1,x_2}(x_1, x_2), \tag{6}$$

where symbol $|\cdot|$ denotes absolute value. As the non-linear function φ was supposed to be invertible, the matrix-function $J(x_1, x_2)$ is non-singular in $\mathcal{X}_1 \times \mathcal{X}_2$, and hence the scalar-function $\det J_\varphi(x_1, x_2)$ is everywhere non-zero in $\mathcal{X}_1 \times \mathcal{X}_2$. Consequently, the sign of the quantity $\det J_\varphi$ stays constant over the domain of regression. The main idea behind isotonic statistical regression based on the above considerations is that, whenever the probability density functions p_{x_1,x_2} and $p_{y,z}$ are *known* and a non-linear model that links the four

variates x_1, x_2, y, z as in equation (4) is sought for, the equation (6) provides (differential) constraints to be satisfied by the model.

In the case of interest in the present work, the non-linear model includes three variates x_1, x_2, y. To fix the ideas, assume that the regression model $y = f(x_1, x_2)$ is chosen such that $\frac{\partial f}{\partial x_1} > 0$ and $\frac{\partial f}{\partial x_2} > 0$. In order to exploit the relationship (6) between four variates to model a relationship between the three variates of interest in the present context, construct the non-linear system:

$$(y, z) = \varphi(x_1, x_2) \stackrel{\text{def}}{=} (f(x_1, x_2), x_2). \tag{7}$$

In this case, the det-Jacobian reads $\det J_\varphi = \frac{\partial f}{\partial x_1} > 0$, thus the equation (6) gives:

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = \frac{p_{x_1,x_2}(x_1, x_2)}{p_{y,x_2}(f(x_1, x_2), x_2)}. \tag{8}$$

Further, construct the non-linear system:

$$(y, z) = \varphi(x_1, x_2) \stackrel{\text{def}}{=} (f(x_1, x_2), x_1). \tag{9}$$

In this case, the det-Jacobian reads $\det J_\varphi = -\frac{\partial f}{\partial x_2} < 0$, thus the equation (6) gives:

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = \frac{p_{x_1,x_2}(x_1, x_2)}{p_{y,x_1}(f(x_1, x_2), x_1)}. \tag{10}$$

The above two conditions must be satisfied at the same time, therefore the non-linear 2-to-1 regression model f may be calculated by solving the system of two partial differential equations (8) and (10) equipped with suitable boundary conditions. The required marginal probability density functions are obtained as:

$$p_{y,x_1}(y, x_1) = \int_{\mathcal{X}_2} p_{y,x_1,x_2}(y, x_1, x_2)\,\mathrm{d}x_2, \tag{11}$$

$$p_{y,x_2}(y, x_2) = \int_{\mathcal{X}_1} p_{y,x_1,x_2}(y, x_1, x_2)\,\mathrm{d}x_1, \tag{12}$$

$$p_{x_1,x_2}(x_1, x_2) = \int_{\mathcal{Y}} p_{y,x_1,x_2}(y, x_1, x_2)\,\mathrm{d}y \tag{13}$$

and are known, while the system of partial differential equations (8) and (10) needs to be solved for the unknown model f.

It is worth recalling that the relationship between the independent variables and the dependent variable is captured by the conditional probability density function $p_{y|x_1,x_2}(y, x_1, x_2)$, which is linked to the marginal probability density function of the three variates by:

$$p_{y,x_1,x_2}(y, x_1, x_2) = p_{y|x_1,x_2}(y, x_1, x_2) p_{x_1,x_2}(x_1, x_2). \tag{14}$$

The choice of the type of predicted dependency (monotonically increasing or monotonically decreasing) relies ultimately on the understanding of the physical phenomenon behind the data. A help about this matter may come

from a correlation analysis of the dependency of the variate-pair (x_1, y) and of the dependency of the variate-pair (x_2, y). From the statistical viewpoint, the hypothesis about the kind of dependency based on the understanding of the phenomenon underlying the data may be analyzed from an inferential viewpoint. In this respect, the hypothesis test described in [11], which assumes Gaussian errors, and hypothesis test described in [10], which is purely non-parametric, are of use. If the hypothesis test confirms that the sought-for relationship is (non strictly) monotonic, it is still necessary to deal with the possibility that a special case occurs, namely, that the model be *constant*. Specific tests were developed in order to detect such occurrence, as described in [6, 7]. In addition, if the relationship is isotonic in a strict sense, then a further analysis may be carried out in order to check if the restrictions underlying the regression method are fulfilled: Among others, the existence of the first-order derivative at any point, being the derivative strictly positive. The work [4] allows to check for the strict monotonicity behavior provided that the regression is continuously differentiable (to certain degree) and the standardized errors are independent, identically distributed.

3 Statistical regression: Solution of the differential system

The present section deals with the analytic treatment of the system of partial differential equations (8) and (10). Such system of partial differential equations is an instance of the differential system:

$$\begin{cases} \frac{\partial y}{\partial x_1} = F_1(y, x_1, x_2), \\ \frac{\partial y}{\partial x_2} = F_2(y, x_1, x_2), \end{cases} \tag{15}$$

where:

$$F_1(y, x_1, x_2) \stackrel{\text{def}}{=} \frac{p_{x_1,x_2}(x_1, x_2)}{p_{y,x_2}(y, x_2)},\ F_2(y, x_1, x_2) \stackrel{\text{def}}{=} \frac{p_{x_1,x_2}(x_1, x_2)}{p_{y,x_1}(y, x_1)}. \tag{16}$$

The differential system (15) needs to be equipped with a set of suitable boundary conditions as, for instance, $y(\tilde{x}_1, \tilde{x}_2) = \tilde{y}$, provided that $(\tilde{x}_1, \tilde{x}_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ and $\tilde{y} \in \mathcal{Y}$. The functions $F_1 : \mathcal{Y} \times \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ and $F_2 : \mathcal{Y} \times \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$ are known and the solution of the system of differential equations is written as $y = f(x_1, x_2)$.

The functions F_1 and F_2 are defined only on the *supports* of the probability density functions p_{y,x_2} and p_{y,x_1}. Namely, define:

$$\mathrm{Supp}(p_{y,x_1}) \stackrel{\text{def}}{=} \{(y, x_1) \in \mathcal{Y} \times \mathcal{X}_1 | p_{y,x_1}(y, x_1) \neq 0\}, \tag{17}$$

$$\mathrm{Supp}(p_{y,x_2}) \stackrel{\text{def}}{=} \{(y, x_2) \in \mathcal{Y} \times \mathcal{X}_2 | p_{y,x_2}(y, x_2) \neq 0\}, \tag{18}$$

then the function F_1 is well-defined over the set $\mathcal{X}_1 \times \mathrm{Supp}(p_{y,x_2})$, while the function F_2 is well-defined over $\mathcal{X}_2 \times \mathrm{Supp}(p_{y,x_1})$.

The approach discussed in the present paper to solve the problem (15) is based on the reformulation of the differential system of equations (8) and (10) as a non-linear integral equation.

In what follows, it is assumed that the sets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}$ are bounded intervals. Moreover, it pays to define $\underline{y} \stackrel{\text{def}}{=} \min\{\mathcal{Y}\}$.

3.1 Reformulation of the system (15) as a non-linear integral equation

Define a *kernel function* $q : \mathcal{Y} \to \mathbb{R}^+$. As $y = f(x_1, x_2)$, the following identity holds:

$$q(y)\,\mathrm{d}y = q(f(x_1, x_2))\left[\frac{\partial f(x_1, x_2)}{\partial x_1}\,\mathrm{d}x_1 + \frac{\partial f(x_1, x_2)}{\partial x_2}\,\mathrm{d}x_2\right]. \tag{19}$$

Define the gradient of the function f, in rectangular coordinates, as ∇f, and a smooth path γ that joins the boundary point of coordinates $(\tilde{x}_1, \tilde{x}_2)$ with the point of current coordinates (x_1, x_2). Then, the following integral equality holds:

$$\int_{\underline{y}}^{y} q(u)\,\mathrm{d}u = c + \int_{\gamma} q(f(\mathbf{r}))\nabla f(\mathbf{r}) \cdot\, \mathrm{d}\mathbf{r}, \tag{20}$$

where $c \in \mathbb{R}$ is a constant that depends on the boundary condition, $\mathbf{r} = (\xi_1, \xi_2)$ denotes the coordinates over the path and $\cdot$ denotes inner product. (The integral on the right-hand side is a *line integral* over the path γ.) If the path γ is parameterized by functions $\xi_1 = \xi_1(u)$ and $\xi_2 = \xi_2(u)$, with $u \in [0,\ 1]$, such that $\xi_1(0) = \tilde{x}_1$, $\xi_2(0) = \tilde{x}_2$ and $\xi_1(1) = x_1$, $\xi_2(1) = x_2$, by the equations (15), the equation (20) takes on the form of a functional equation:

$$\begin{gathered}\textstyle\int_{\underline{y}}^{f(x_1,x_2)} q(u)\,\mathrm{d}u = \\ c + \textstyle\int_0^1 q(f(\xi_1,\xi_2))\left[F_1(f(\xi_1,\xi_2),\xi_1,\xi_2)\frac{\mathrm{d}\xi_1}{\mathrm{d}u} + F_2(f(\xi_1,\xi_2),\xi_1,\xi_2)\frac{\mathrm{d}\xi_2}{\mathrm{d}u}\right]\,\mathrm{d}u.\end{gathered} \tag{21}$$

The kernel function may be chosen in different ways. In the following, a few possible choices are explored and discussed.

The first choice of kernel function discussed in the present paper is based on the following auxiliary quantities. Define:

$$p_y(y) \stackrel{\text{def}}{=} \int_{\mathcal{X}_1 \times \mathcal{X}_2} p_{y,x_1,x_2}(y, x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2, \tag{22}$$

$$P_y(y) \stackrel{\text{def}}{=} \int_{\underline{y}}^{y} p_y(u)\,\mathrm{d}u, \tag{23}$$

and denote by $P_y^{-1} : [0,\ 1] \to \mathcal{Y}$ the inverse of the function $P_y : \mathcal{Y} \to [0,\ 1]$. Define the conditional probability density functions:

$$p_{x_1|y}(x_1, y) \stackrel{\text{def}}{=} \frac{p_{y,x_1}(y, x_1)}{p_y(y)}, \tag{24}$$

$$p_{x_2|y}(x_2, y) \stackrel{\text{def}}{=} \frac{p_{y,x_2}(y, x_2)}{p_y(y)}. \tag{25}$$

The integrands on the right-hand side of the integral equation (21) obey the system of equations (8) and (10) and thus the following identities hold:

$$p_y(f(x_1,x_2))\frac{\partial f(x_1,x_2)}{\partial x_1} = \frac{p_{x_1,x_2}(x_1,x_2)}{p_{x_2|y}(x_2,f(x_1,x_2))}, \tag{26}$$

$$p_y(f(x_1,x_2))\frac{\partial f(x_1,x_2)}{\partial x_2} = \frac{p_{x_1,x_2}(x_1,x_2)}{p_{x_1|y}(x_1,f(x_1,x_2))}. \tag{27}$$

Taking $q = p_y$, the integral equation (21) may be rewritten compactly as:

$$P_y(f(x_1,x_2)) = c + \int_0^1 \left[\frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{x_2|y}(\xi_2,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_1}{\mathrm{d}u} + \frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{x_1|y}(\xi_1,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_2}{\mathrm{d}u}\right]\mathrm{d}u. \tag{28}$$

Such formulation coincides with the formulation obtained in [12,13] in the case that a *single (dominant) independent variate is present* and hence constitutes a generalization of the 1-to-1 regression equation studied in the previous contributions [12,13]. The above integral equation may be rewritten as the functional equation in the unknown f:

$$f(x_1,x_2) = P_y^{-1}\left[c + \int_0^1 \left(\frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{x_2|y}(\xi_2,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_1}{\mathrm{d}u} + \frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{x_1|y}(\xi_1,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_2}{\mathrm{d}u}\right)\mathrm{d}u\right]. \tag{29}$$

In the case that the independent variables x_1, x_2 are statistically independent of each other, the identity $p_{x_1,x_2} = p_{x_1}p_{x_2}$ holds, hence the functional equation (29) simplifies into:

$$f(x_1,x_2) = P_y^{-1}\left[c + \int_0^1 \left(\frac{p_{x_1}(\xi_1)p_{x_2}(\xi_2)}{p_{x_2|y}(\xi_2,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_1}{\mathrm{d}u} + \frac{p_{x_1}(\xi_1)p_{x_2}(\xi_2)}{p_{x_1|y}(\xi_1,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_2}{\mathrm{d}u}\right)\mathrm{d}u\right]. \tag{30}$$

The general integral equation in the case of mixed monotonically increasing/decreasing behavior with respect to the independent variables reads:

$$f(x_1,x_2) = P_y^{-1}\left[c + \int_0^1 \left(\pm\frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{x_2|y}(\xi_2,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_1}{\mathrm{d}u} \pm \frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{x_1|y}(\xi_1,f(\xi_1,\xi_2))}\frac{\mathrm{d}\xi_2}{\mathrm{d}u}\right)\mathrm{d}u\right]. \tag{31}$$

The main disadvantage of the above formulation is that it requires the computation of conditional probabilities and that the division by the conditional probabilities leads to computational burden and numerical difficulties. Moreover, the cumulative density function (23) results invertible only if $p_y(y) \neq 0$ for all $y \in \mathcal{Y}$.

Another possible choice of the kernel function that was explored is $q = 1$. Such choice leads to the integral equation:

$$f(x_1,x_2) = \underline{y} + c + \int_0^1 \left(\pm\frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{y,x_2}(f(\xi_1,\xi_2),\xi_2)}\frac{\mathrm{d}\xi_1}{\mathrm{d}u} \pm \frac{p_{x_1,x_2}(\xi_1,\xi_2)}{p_{y,x_1}(f(\xi_1,\xi_2),\xi_1)}\frac{\mathrm{d}\xi_2}{\mathrm{d}u}\right)\mathrm{d}u, \tag{32}$$

where the signs can be chosen according to the character of the model. Such equation looks simpler than the equation (31). However, an important problem about the numerical integration of such equation arises. In fact, it is clear that the integrands on the right-hand side of above equation may be evaluated correctly only over the support of the probability density functions $p_{y,x1}$ and p_{y,x_2}. Such supports do not possess a regular/simple shape, in general, which makes it difficult to implement numerically the line integral over a curve γ which should belong entirely to the supports. An example of curly supports is illustrated in the Figure 1. The exemplary supports were obtained with

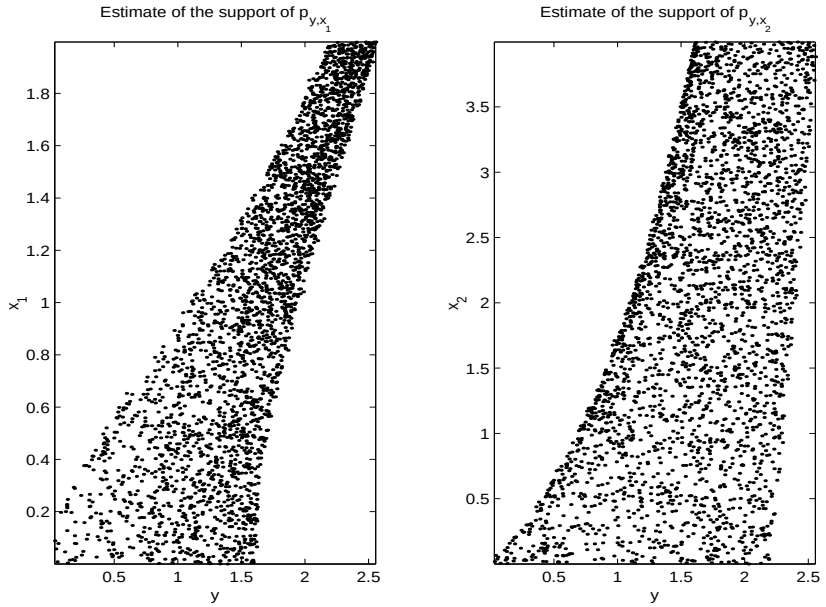

Fig. 1 An example of curly supports (numerically estimated) of the probability density functions p_{y,x_1} and p_{y,x_2}.

$f(x_1, x_2) = \log(2x_1^2 + x_2 + 1)$, x_1 randomly drawn from a uniform distribution in $[0,\ 2]$, x_2 randomly drawn from a uniform distribution in $[0,\ 4]$ and $y = f(x_1, x_2) + \nu_1$, where ν_1 is a Gaussian noise with zero mean and variance 10^{-4}. A total of $3,000$ samples were generated to get an approximate picture of the supports $\mathrm{Supp}(p_{y,x_1})$ and $\mathrm{Supp}(p_{y,x_2})$.

The present formulation that relies on recasting the system of first-order partial differential equations (15) into a non-linear functional equation enjoys the following features:

- The formulation (32) does not need complicated calculations to evaluate the right-hand side.
- The resulting model is certainly monotonic by construction.

The differential (and the integral) formulation needs suitable boundary conditions to become complete.

3.2 Boundary conditions

In principle, consistent boundary conditions may be chosen freely on the basis of any prior knowledge on the model function f. The fact that there is not any *intrinsic* way o establishing a boundary condition implies that the model f may be estimated *up to an additive constant.*

A special boundary condition is devised as follows. Define $\underline{x}_1 \stackrel{\text{def}}{=} \min\{\mathcal{X}_1\}$ and $\underline{x}_2 \stackrel{\text{def}}{=} \min\{\mathcal{X}_2\}$. As the function $y = f(x_1, x_2)$ is monotonically increasing with respect to both independent variables, the identity $f(\underline{x}_1, \underline{x}_2) = \underline{y}$ holds. The choice $\tilde{x}_1 = \underline{x}_1$ and $\tilde{x}_2 = \underline{x}_2$ leads to the condition $c = 0$.

Another special boundary condition, that retraces the center-of-mass-to-center-of-mass-mapping condition utilized in [12] is as follows. Define:

$$x_1^{\mathrm{m}} \stackrel{\text{def}}{=} \int_{\mathcal{X}_1 \times \mathcal{X}_2} p_{x_1,x_2}(x_1, x_2) x_1 \, \mathrm{d}x_1 \, \mathrm{d}x_2, \tag{33}$$

$$x_2^{\mathrm{m}} \stackrel{\text{def}}{=} \int_{\mathcal{X}_1 \times \mathcal{X}_2} p_{x_1,x_2}(x_1, x_2) x_2 \, \mathrm{d}x_1 \, \mathrm{d}x_2, \tag{34}$$

$$y^{\mathrm{m}} \stackrel{\text{def}}{=} \int_{\mathcal{Y}} p_y(y) y \, \mathrm{d}y. \tag{35}$$

A center-of-mass-to-center-of-mass-mapping condition casts as $f(x_1^{\mathrm{m}}, x_2^{\mathrm{m}}) = y^{\mathrm{m}}$. The center-of-mass-to-center-of-mass-mapping condition leads to a value of the displacement $c = y^{\mathrm{m}} - \underline{y}$.

4 Numerical solution of the differential system (15)

The present section discusses the problem of numerically representing the quantities of interest and of numerically solving the general system (15) of two partial differential equations with no regard to the statistical regression problem in particular. The next section will deal with the specific problem of statistical regression.

It is assumed that the set $\mathcal{X}_1$ is a bounded interval $[\underline{x}_1,\ \overline{x}_1]$ and that the set $\mathcal{X}_2$ is a bounded interval $[\underline{x}_2,\ \overline{x}_2]$. The interval $\mathcal{X}_1$ is subdivided into B_1 subintervals of equal width $h_1 \stackrel{\text{def}}{=} (\overline{x}_1 - \underline{x}_1)/B_1$ and the interval $\mathcal{X}_1$ is subdivided into B_2 subintervals of equal width $h_2 \stackrel{\text{def}}{=} (\overline{x}_2 - \underline{x}_2)/B_2$.

The solution $y = f(x_1, x_2)$ of the differential system (15) is represented by a $(B_2+1)\times(B_1+1)$ matrix $\mathbf{y}$ of components $y_{i,j}$, with $i = 1, \ldots, B_2+1$ and $j = 1, \ldots, B_1+1$. Note that the index i represents the rows of the matrix $\mathbf{y}$, hence it is associated with the variable x_2, while the index j represents the columns of the matrix $\mathbf{y}$, hence it is associated with the variable x_1. The quantity $y_{i,j}$ represents an approximation of the true value $y(\underline{x}_1 + (j-1)h_1, \underline{x}_2 + (i-1)h_2)$.

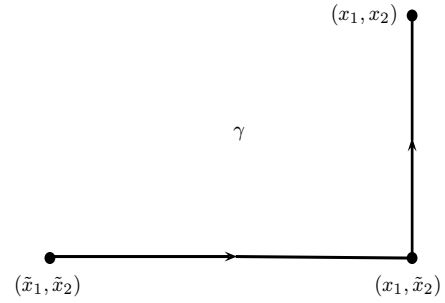

Fig. 2 Integration line γ to solve the functional equation (21).

The structure of the functional equation (21) suggests a possible algorithmic solution based on a fixed-point iteration. In the present section it is assumed that the support of the function p_{y,x_1} coincides with the rectangle $\mathcal{Y} \times \mathcal{X}_1$ and that the support of the function p_{y,x_2} coincides with the rectangle $\mathcal{Y} \times \mathcal{X}_2$. If such assumption holds true, the integration path γ may be chosen coincident with the union of the horizontal segment that connects the point $(\tilde{x}_1, \tilde{x}_2)$ with the point $(x_1, \tilde{x}_2)$ and of the vertical segment that connects the point $(x_1, \tilde{x}_2)$ with the point (x_1, x_2), as illustrated in the Figure 2, where $\tilde{x}_1 \in \mathcal{X}_1$ and $\tilde{x}_2 \in \mathcal{X}_2$ may be chosen arbitrarily.

The functional equation (21), for a unitary kernel function, for an integration path γ as in Figure 2 with $\tilde{x}_1 = \underline{x}_1$ and $\tilde{x}_2 = \underline{x}_2$, and with boundary condition set as $f(\underline{x}_1, \underline{x}_2) = \underline{y}$, reads:

$$f(x_1, x_2) = \underline{y} + \int_{\underline{x}_1}^{x_1} F_1(f(u, \underline{x}_2), u, \underline{x}_2)\,\mathrm{d}u + \int_{\underline{x}_2}^{x_2} F_2(f(x_1, u), x_1, u)\,\mathrm{d}u. \quad (36)$$

The above functional equation may be solved numerically on a computation platform – upon discretization of the domains $\mathcal{X}_1 \times \mathcal{X}_2$ and $\mathcal{Y}$ – by means of any numerical quadrature rule and by putting into effect an appropriate iterative numerical scheme. The quadrature rule of choice is the *left Riemann sum*:

$$\int_a^b \varphi(u)\,\mathrm{d}u \approx \varphi(a)(b-a),$$

for a, b finite and $\varphi : [a, b] \to \mathbb{R}$ integrable. Although an approximate quadrature of the above integral might be obtained by other known methods, e.g., the midpoint rule, the trapezoid rule or the Cavalieri-Simpson rule:

$$\int_a^b \varphi(u)\,\mathrm{d}u \approx \varphi\left(\frac{a+b}{2}\right)(b-a),\ \int_a^b \varphi(u)\,\mathrm{d}u \approx \frac{\varphi(a)+\varphi(b)}{2}(b-a),\ \cdots,$$

it should be recognized that the use of a 'left Riemann sum' facilitates the handling of the boundary conditions and *ensures that the monotonicity of the*

model is preserved. In such context, the functional equation (36) gives rise to the set of equations:

$$y_{i,j} = \underline{y} + h_1 \sum_{s=1}^{j-1} F_1(y_{1,s}, \underline{x}_1 + (s-1)h_1, \underline{x}_2) + h_2 \sum_{r=1}^{i-1} F_2(y_{r,j}, \underline{x}_1 + (j-1)h_1, \underline{x}_2 + (r-1)h_2), \quad (37)$$

for $i = 2, \ldots, B_2 + 1$ and $j = 2, \ldots, B_1 + 1$. Further conditions are:

$$y_{1,j} = \underline{y} + h_1 \sum_{s=1}^{j-1} F_1(y_{1,s}, \underline{x}_1 + (s-1)h_1, \underline{x}_2), \text{ for } j > 1, \quad (38)$$

$$y_{i,1} = \underline{y} + h_2 \sum_{r=1}^{i-1} F_2(y_{r,1}, \underline{x}_1, \underline{x}_2 + (r-1)h_2), \text{ for } i > 1, \quad (39)$$

$$y_{1,1} = \underline{y}. \quad (40)$$

It is convenient to rewrite such system of non-linear equations by the conventional representation:

$$\mathbf{y} = T(\mathbf{y}), \quad (41)$$

where the matrix $\mathbf{y}$ represents the whole set of unknowns and the symbol T represents a non-linear operator. Then, a fixed-point iterative scheme may generate a sequence $\mathbf{y}^{(k)}$, with $k = 0, 1, 2, \ldots$, of increasingly-refined approximations of the true solution to the system (15), computed by the recurrence rule:

$$\mathbf{y}^{(k+1)} = T(\mathbf{y}^{(k)}), \quad (42)$$

where the quantity $\mathbf{y}^{(0)}$ denotes a suitably-chosen initial guess. Due to the non-linear structure of the fixed-point recurrence rule (42), no theoretical results are available at present about its convergence features, which are rather analyzed from a numerical perspective in Section 5.

A pseudocode to implement a possible instance of the above numerical scheme is reported in the Algorithm 1. Note that in the MATLAB© framework, the two inner cycles of the Algorithm 1 may be implemented efficiently by making use of the function '`cumsum`'. The purpose of the pseudocode explained in the present section is to illustrate and clarify the basic notions behind the implementation of the discussed numerical procedures. More refined and better optimized versions could be implemented, indeed.

4.1 Numerical tests

The numerical procedures devised in the present section will be tested on two test-problems:

Algorithm 1 Pseudocode to implement the numerical solution of the differential system (15) of partial differential equations reformulated as the non-linear integral equation (21).

▷ Input domain boundaries $\underline{x}_1, \overline{x}_1, \underline{x}_2, \overline{x}_2$, numbers of subdivisions B_1, B_2, functions F_1, F_2 and boundary value $\underline{y}$
Compute widths $h_1 = \frac{\overline{x}_1 - \underline{x}_1}{B_1}$ and $h_2 = \frac{\overline{x}_2 - \underline{x}_2}{B_2}$
Set $y_{i,j}^{(0)} = \underline{y}$ for every $i = 1, \ldots, B_2 + 1$ and $j = 1, \ldots, B_1 + 1$
for $k = 0, \overline{1}, 2, \ldots$ **do**
  for $i = 2$ to $B_2 + 1$ **do**
    Compute $y_{i,1}^{(k+1)} = \underline{y} + h_2 \sum_{r=1}^{i-1} F_2(y_{r,1}^{(k)}, \underline{x}_1, \underline{x}_2 + (r-1)h_2)$
    for $j = 2$ to $B_1 + 1$ **do**
      if $i = 1$ **then**
        Compute $y_{1,j}^{(k+1)} = \underline{y} + h_1 \sum_{s=1}^{j-1} F_1(y_{1,s}^{(k)}, \underline{x}_1 + (s-1)h_1, \underline{x}_2)$
      end if
      Compute $y_{i,j}^{(k+1)} = \underline{y} + h_1 \sum_{s=1}^{j-1} F_1(y_{1,s}^{(k)}, \underline{x}_1 + (s-1)h_1, \underline{x}_2) + h_2 \sum_{r=1}^{i-1} F_2(y_{r,j}^{(k)}, \underline{x}_1 + (j-1)h_1, \underline{x}_2 + (r-1)h_2)$
    end for
  end for
end for
▷ Output result y

- **Test problem 1**: The problem is defined by the functions $F_1(y, x_1, x_2) = y - 2x_1 - x_2 + 1$ and $F_2(y, x_1, x_2) = y - 2x_1 - x_2$, by the intervals $\mathcal{X}_1 = [0,\ 1]$ and $\mathcal{X}_1 = [0,\ 2]$ and by the boundary condition $f(0, 0) = 1$. The exact solution of this problem is $f(x_1, x_2) = 2x_1 + x_2 + 1$. The partitions cardinality are $B_1 = 10$ and $B_2 = 15$.
- **Test problem 2**: The problem is defined by the functions $F_1(y, x_1, x_2) = 4x_1 e^{-y}$ and $F_2(y, x_1, x_2) = e^{-y}$, by the intervals $\mathcal{X}_1 = [0,\ 1]$ and $\mathcal{X}_1 = [0,\ 2]$ and by the boundary condition $f(0, 0) = 0$. The exact solution of this problem is $f(x_1, x_2) = \log(2x_1^2 + x_2 + 1)$. The partitions cardinality are $B_1 = 15$ and $B_2 = 20$.

The pairs (x_1, x_2) of independent variates are generated uniformly in $\mathcal{X}_1 \times \mathcal{X}_2$ and the variate y is generated by the rule $y = f(x_1, x_2) + \nu_1$, with ν_1 being a zero-mean Gaussian noise. A total of $3,000$ samples were made available for the numerical simulations.

The following numerical results illustrate the behavior of the Algorithm 1 on the two test problems. The following discrepancy measures are made use of in order to evaluate the progress of the iterative method:

$$\text{Point-wise absolute error at } (i,j) = \left| \frac{y_{i,j}^{(k+1)} - y_{i,j}^{\star}}{y_{i,j}^{\star}} \right|, \tag{43}$$

$$\text{Mean relative error} = \frac{\|\mathbf{y}^{(k+1)} - \mathbf{y}^{(k)}\|_{\mathrm{F}}}{(B_1+1)(B_2+1)}, \tag{44}$$

$$\text{Mean absolute error} = \frac{\|(\mathbf{y}^{(k+1)} - \mathbf{y}^{\star}) \oslash \mathbf{y}^{\star}\|_{\mathrm{F}}}{(B_1+1)(B_2+1)}, \tag{45}$$

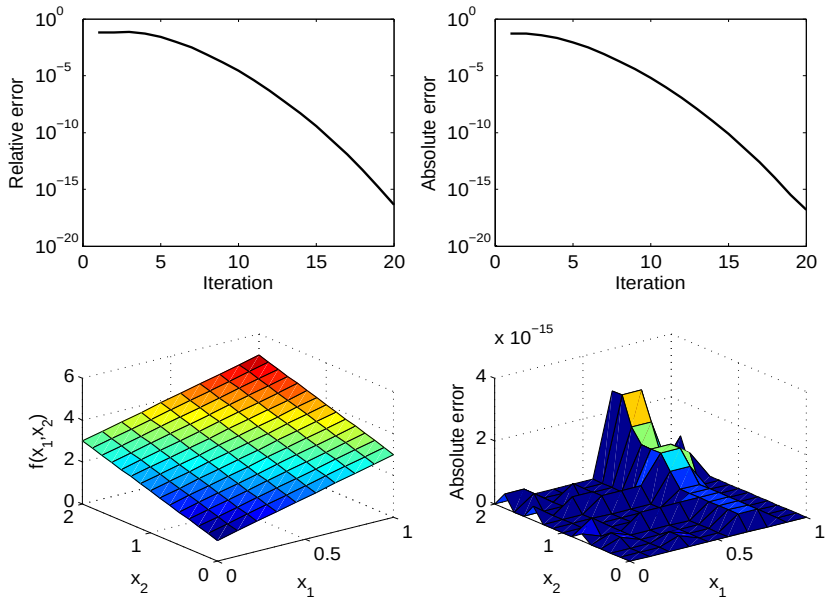

Fig. 3 Numerical behavior of the Algorithm 1 on the Test problem 1. Top-left panel: Mean relative error. Top-right panel: Mean absolute error. Bottom-left panel: Estimated solution superimposed with the actual solution (lighter colors surface). Bottom-right panel: Point-wise absolute error.

where $\mathbf{y}^\star$ denotes the matrix of values corresponding to the actual solution f, symbol $\oslash$ denotes entry-wise division and symbol $\|\cdot\|_{\mathrm{F}}$ denotes Frobenius norm. Note that the above errors refer to the actual value of the model, hence they take into account the additive noise ν_1 too.

The Figure 3 illustrates the numerical behaviour of the Algorithm 1 on the Test problem 1, run over 20 iterations, in terms of relative error and absolute error. The variance of the noise ν_1 was 10^{-4}. The data-model to estimate is linear in both independent variables and the estimation result is excellent over the whole domain $\mathcal{X}_1 \times \mathcal{X}_2$.

The Figure 4 illustrates the numerical behavior of the Algorithm 1 on the Test problem 2, run over 15 iterations, in terms of relative error and absolute error. The variance of the noise ν_1 was 10^{-4}. The data-model to estimate is non-linear in both independent variables and the estimation result looks acceptable over the whole domain $\mathcal{X}_1 \times \mathcal{X}_2$, although it seems better in the central part of the domain than near the border of the domain.

The above numerical experiments were conducted by choosing a number of samples and subdivision numbers that might mimic a real-world situation. The number of subdivisions may be increased and, up to a certain extent, this would increase the precision of the estimation (at the expense of an increased computation burden). From the above numerical results it may be concluded

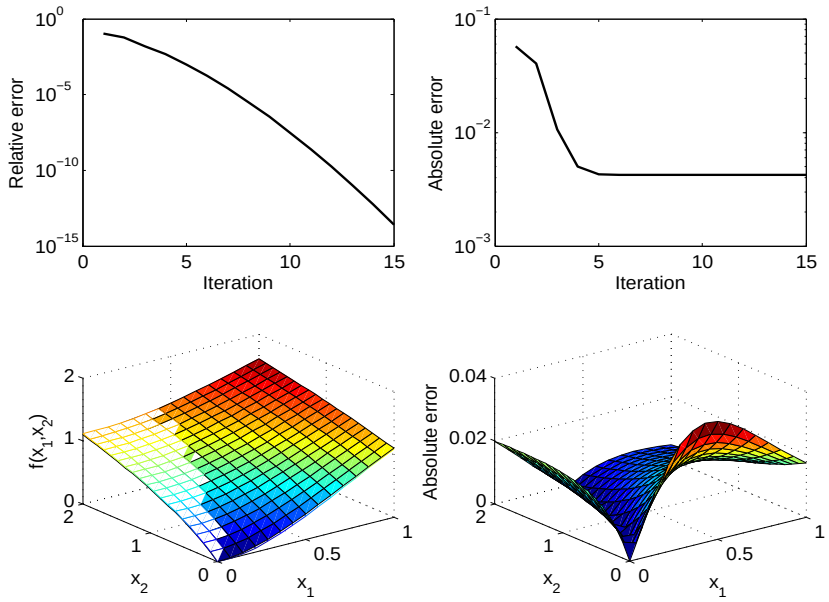

Fig. 4 Numerical behavior of the Algorithm 1 on the Test problem 2. Top-left panel: Mean relative error. Top-right panel: Mean absolute error. Bottom-left panel: Estimated solution superimposed with the actual solution (lighter colors surface). Bottom-right panel: Pointwise absolute error.

that the Algorithm 1 works properly and that it converges in a few iterations to a solution close to the actual one.

5 Numerical implementation of statistical regression

The present section discusses the problem of numerically coping with the statistical regression problem in particular. It basically covers two topics: How to estimate numerically the required probability density functions and how to modify the Algorithm 1 to the case where the functions F_1 and F_2 are given by numerical approximations of the relationships (16) and a different boundary condition is used.

In order to improve the numerical accuracy of the devised statistical regression procedure, the center-of-mass-to-center-of-mass-mapping boundary condition will be made use of. The available data are grouped into N sample-pairs $({}^s x_1, {}^s x_2)$, $({}^s y, {}^s x_1)$ and $({}^s y, {}^s x_2)$. The empirical average values of the samples are computed as:

$$x_1^{\mathrm{m}} = \frac{\sum_s {}^s x_1}{N},\ x_2^{\mathrm{m}} = \frac{\sum_s {}^s x_2}{N},\ y^{\mathrm{m}} = \frac{\sum_s {}^s y}{N}. \tag{46}$$

The use of the center-of-mass-to-center-of-mass-mapping boundary condition causes a *constant bias* of the result that could be mitigated by a least-squares fitting of the obtained model.

5.1 Estimation of the required probability density functions

The first implementation issue concerns the estimation of the joint probability density functions involved in the statistical regression theory devised in Section 2. The probability density functions to estimate are p_{x_1,x_2}, p_{y,x_1} and p_{y,x_2}. The goal of density estimation is to take a finite sample of data and to make inferences about the underlying probability density function everywhere, including where no data are observed. Two popular methods for multivariate density estimation are [24]:

- **Histogram-based estimates**: Such method is based on counting the number of samples that fall within each subdivision that the domain of the involved variates is partitioned into. This is the method utilized in the previous contribution [12,13]. This method is non-parametric and computationally inexpensive. However, when joint probability density functions estimates are sought for, the obtainable estimates are sensible only in presence of a large number of samples because the total number of partitions grows quickly with the number of partitions of the support of each variate. The histogram-based method provides a purely numerical representation of probability density function and does not possess, as is, any interpolation ability to predict the values of the distribution besides the values present in the data sets.
- **Mixture of kernels**: Such method is based on the approximation of a probability density function as a superposition of properly scaled and centered kernels (for example, Gaussian 'bells'). In kernel density estimation, the contribution of each data point is smoothed out from a single point into a region of space surrounding it. Aggregating the individually smoothed contributions gives an overall picture of the structure of the data and of its density function. Such method is computationally more expensive than other non-parametric methods but it is profitable even when the number of samples is limited. The mixture-of-kernel method provides a continuous functional representation of the estimate of probability density functions, hence it inherently exhibits interpolation as well as extrapolation abilities.

The present research is based on the assumption that the available data are enough to extract their probability density function by a histogram method, which also warrants a limited computational burden. Therefore, the histogram-based estimation method only is made use of within the present contribution.

The histogram-based estimate of the joint probability density function of two variates $z_1 \in \mathcal{Z}_1$ and $z_2 \in \mathcal{Z}_2$, with $\mathcal{Z}_1$ is a bounded interval $[\underline{z}_1,\ \overline{z}_1]$ and $\mathcal{Z}_2$ is a bounded interval $[\underline{z}_2,\ \overline{z}_2]$, may be implemented as follows. Let S_1 and S_2 be the number of subdivisions (or *bins*) of the intervals $\mathcal{Z}_1$ and $\mathcal{Z}_2$, respectively. The widths w_1 and w_2 of the subdivisions are related to the number of partitions by:

$$S_1 = \left\lfloor \frac{\overline{z}_1 - \underline{z}_1}{w_1} \right\rfloor,\ S_2 = \left\lfloor \frac{\overline{z}_2 - \underline{z}_2}{w_2} \right\rfloor, \tag{47}$$

where symbol $\lfloor\cdot\rfloor$ returns the nearest integer toward $-\infty$. Denote by $H_{i,j}^{z_1,z_2}$ the histogram count of the (i,j)th bin and by $({}^sz_1,{}^sz_2) \in \mathcal{Z}_1 \times \mathcal{Z}_2$ the sth sample-pair. The indexes-pair of the subdivision that the sample-pair $({}^sz_1,{}^sz_2)$ falls in is given by:

$$j_s = \left\lfloor \frac{{}^sz_1 - \underline{z}_1}{w_1} \right\rfloor + 1,\ i_s = \left\lfloor \frac{{}^sz_2 - \underline{z}_2}{w_2} \right\rfloor + 1. \tag{48}$$

In order to *force* the support of the estimated probability density function to coincide with the whole domain $\mathcal{Z}_1 \times \mathcal{Z}_2$, any zero entry of the matrix H may be set to 1. Such setting distorts the shape of the actual probability density function p_{z_1,z_2} but, if the number of samples is sufficiently large, the caused distortion is negligible.

In order to make the estimated histogram smoother, it may be convolved by a smoothing kernel $\mathbf{K}$, as, for instance:

$$\mathbf{K} = \mathbf{K}_3 \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{49}$$

Then the smoothed-out histogram estimate is computed by $\tilde{\mathbf{H}}^{z_1,z_2} = \mathbf{H}^{z_1,z_2} * \mathbf{K}$.

The obtained probability density function estimate is represented by a $(S_2+1)\times(S_1+1)$ matrix $\hat{\mathbf{p}}^{z_1,z_2} = A\tilde{\mathbf{H}}^{z_1,z_2}$, where the normalization constant $A > 0$ ensures that the probability density function integrates to the unity, namely, the constant A is chosen in such a way that:

$$A\sum_{i,j} \tilde{H}_{i,j}^{z_1,z_2} w_1 w_2 = 1. \tag{50}$$

The number of partitions of the interval $\mathcal{X}_1$ is denoted by S_1 and its width is denoted by w_1, the number of partitions of the interval $\mathcal{X}_2$ is denoted by S_2 and its width is denoted by w_2, and the number of partitions of the interval $\mathcal{Y}$ is denoted by S_y and its width is denoted by w_y. The width of partitions may be chosen on the basis of the rules adapted from [24]:

$$w_1 = \frac{7}{2}\hat{\sigma}_1 N^{-\frac{1}{4}},\ w_2 = \frac{7}{2}\hat{\sigma}_2 N^{-\frac{1}{4}},\ w_y = \frac{7}{2}\hat{\sigma}_y N^{-\frac{1}{4}}, \tag{51}$$

where $\hat{\sigma}_1$ denotes the empirical standard deviation of the variate x_1, $\hat{\sigma}_2$ denotes the empirical standard deviation of the variate x_2 and $\hat{\sigma}_y$ denotes the empirical standard deviation of the variate y and N denotes the cardinality of the sets $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{Y}$.

5.2 Algorithmic formulation of statistical regression

On the basis of the histogram estimation method recalled in the previous subsection, matrices $\hat{\mathbf{p}}^{x_1,x_2}$ of size $(S_2+1)\times(S_1+1)$, computed on the basis of the sample-pairs $({}^s x_1, {}^s x_2)$, $\hat{\mathbf{p}}^{y,x_1}$ of size $(S_1+1)\times(S_y+1)$, computed on the basis of the sample-pairs $({}^s y, {}^s x_1)$ and $\hat{\mathbf{p}}^{y,x_2}$ of size $(S_2+1)\times(S_y+1)$, computed on the basis of the sample-pairs $({}^s y, {}^s x_2)$ are available, which represent the joint probability density functions required by the statistical regression procedure to compute the functions F_1 and F_2 to integrate numerically. The number of available samples is denoted again by N.

Although the domain of the independent variates is $\mathcal{X}_1 \times \mathcal{X}_2$, it is convenient to define a sub-domain $[\underline{\underline{x}}_1,\ \overline{\overline{x}}_1] \times [\underline{\underline{x}}_2,\ \overline{\overline{x}}_2] \subset \mathcal{X}_1 \times \mathcal{X}_2$ where regression will actually be performed. Such a sub-domain gets partitioned into $B_1 \times B_2$ partitions of widths:

$$h_1 \stackrel{\text{def}}{=} \frac{1}{B_1}\left(\overline{\overline{x}}_1 - \underline{\underline{x}}_1\right),\ h_2 \stackrel{\text{def}}{=} \frac{1}{B_2}\left(\overline{\overline{x}}_2 - \underline{\underline{x}}_2\right). \tag{52}$$

The above sub-domain is supposed to contain the point $(x_1^{\mathrm{m}}, x_2^{\mathrm{m}})$.

Any numerical value of a variable possesses a different representation in the matrix that represents a joint probability density function and in the matrices that will represent the functions $F_1(y, x_1, x_2)$ and $F_2(y, x_1, x_2)$, due to the different sizes of the domains and of the widths of the subdivisions. Consider such relationships in details:

- For the variable x_1, denote by c its representation in the model-domain and by n_1 its representation in the histogram domain. Given a model-domain index $c \in \{1, \ldots, B_1+1\}$, its corresponding variable value is $x_1 = \underline{\underline{x}}_1 + (c-1)h_1$. The corresponding index in the histogram domain is $n_1 = \lfloor (x_1 - \underline{\underline{x}}_1)/w_1 \rfloor + 1$. The overall index-to-index mapping reads, thus:

$$n_1(c) \stackrel{\text{def}}{=} \left\lfloor \frac{\underline{\underline{x}}_1 + (c-1)h_1 - \underline{x}_1}{w_1} \right\rfloor + 1; \tag{53}$$

- For the variable x_2, denote by r its representation in the model-domain and by n_2 its representation in the histogram domain. Given a model-domain index $r \in \{1, \ldots, B_2+1\}$, its corresponding variable value is $x_2 = \underline{\underline{x}}_2 + (r-1)h_2$. The corresponding index in the histogram domain is $n_2 = \lfloor (x_2 - \underline{x}_2)/w_2 \rfloor + 1$. The overall index-to-index mapping reads, thus:

$$n_2(r) \stackrel{\text{def}}{=} \left\lfloor \frac{\underline{\underline{x}}_2 + (r-1)h_2 - \underline{x}_2}{w_2} \right\rfloor + 1; \tag{54}$$

- For the variable y, given a value in $\mathcal{Y}$, the corresponding index in the histogram domain is:

$$n_y(y) \stackrel{\text{def}}{=} \left\lfloor \frac{y - \underline{y}}{w_y} \right\rfloor + 1. \tag{55}$$

The values of the regression model over the sub-domain of interest are represented by a $(B_2+1)\times(B_1+1)$ matrix $\mathbf{y}$. The functions F_1 and F_2 may be represented numerically by matrices of sizes $(B_2+1)\times(B_1+1)$, whose entries are calculated as follows:

– The entry of position (r,c) of the matrix $\mathbf{F}_1$ is calculated as:

$$(F_1)_{r,c} = \pm\frac{\hat{p}^{x_1,x_2}_{n_2(r),n_1(c)}}{\hat{p}^{y,x_1}_{n_1(c),n_y(y_{r,c})}}; \tag{56}$$

– The entry of position (r,c) of the matrix $\mathbf{F}_2$ is calculated as:

$$(F_2)_{r,c} = \pm\frac{\hat{p}^{x_1,x_2}_{n_2(r),n_1(c)}}{\hat{p}^{y,x_1}_{n_2(r),n_y(y_{r,c})}}. \tag{57}$$

The indexes in the model-domain corresponding to the average values of the independent variates are compute as:

$$r^{\mathrm{m}} \stackrel{\mathrm{def}}{=} \left\lfloor \frac{x_2^{\mathrm{m}} - \underline{x}_2}{w_2} \right\rfloor + 1,\ c^{\mathrm{m}} \stackrel{\mathrm{def}}{=} \left\lfloor \frac{x_1^{\mathrm{m}} - \underline{x}_1}{w_1} \right\rfloor + 1, \tag{58}$$

and the center-of-mass-to-center-of-mass-mapping boundary condition is expressed by $y_{r^{\mathrm{m}},c^{\mathrm{m}}} = y^{\mathrm{m}}$.

The numerical implementation of the regression method consists in the modification of the Algorithm 1 to integrate the functions F_1 and F_2 on lines that connect any point of indexes (r,c) in the model domain to the 'boundary point' of indexes $(r^{\mathrm{m}},c^{\mathrm{m}})$ (which actually locates allegedly at the center of the domain). (A number of numerical issues that were taken care of in the actual computer implementation were omitted from the present description for brevity.)

5.3 Numerical tests

In the present section, numerical tests about the estimation ability of the devised statistical regression algorithm are presented and discussed, with reference to the same *Test problem 1* and *Test problem 2* of subsection 4.1. The numerical evaluation of the behavior of the devised statistical regression algorithm is made through the error evaluation functions described in the subsection 4.1. It is worth noting here that the expression (45) of the mean absolute error utilizes a weight $\frac{1}{\sqrt{(B_1+1))(B_2+1)}}$ which is the same for each entry of the matrix $(\mathbf{y}^{(k+1)} - \mathbf{y}^{\star}) \oslash \mathbf{y}^{\star}$. A different way of measuring the absolute estimation error would be to weight each entry of the matrix $(\mathbf{y}^{(k+1)} - \mathbf{y}^{\star}) \oslash \mathbf{y}^{\star}$ by the corresponding entry of the estimated probability density function $\hat{\mathbf{p}}^{x_1,x_2}$. Such weighting scheme gives rise to the

$$\text{Weighted absolute error} = \|\overline{\mathbf{p}}^{x_1,x_2} \otimes (\mathbf{y}^{(k+1)} - \mathbf{y}^{\star}) \oslash \mathbf{y}^{\star}\|_{\mathrm{F}}, \tag{59}$$

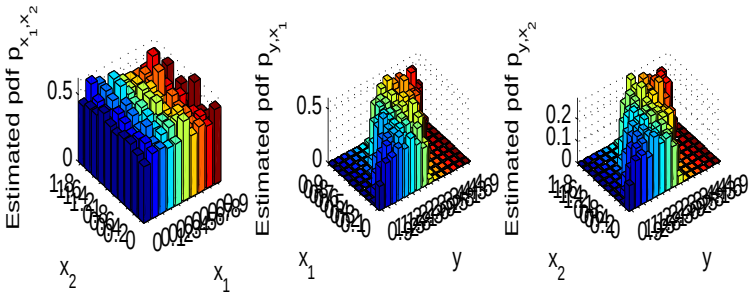

Fig. 5 Result of numerical estimation of the joint probability density functions p_{x_1,x_2}, p_{y,x_1} and p_{y,x_2}, from left to right, on the Test problem 1. Bivariate probability estimates are obtained by a histogram-based method.

where symbol $\otimes$ denotes entry-wise multiplication and matrix $\overline{\mathbf{p}}^{x_1,x_2}$ denotes the probability-matrix $\hat{\mathbf{p}}^{x_1,x_2}$ expressed in the model domain. Analogously, the point-wise absolute error (43) may be modified by weighting each error by the corresponding probability value:

$$\text{Weighted point-wise absolute error at } (r,c) = \hat{p}^{x_1,x_2}_{n_2(r),n_1(c)} \left| \frac{y^{(k+1)}_{r,c} - y^{\star}_{r,c}}{y^{\star}_{r,c}} \right|, \tag{60}$$

The first numerical experiment is about applying the devised statistical regression method to the Test problem 1. In order to obtain a good estimate of the probability density functions by the histogram method, a considerably high number of samples was selected, namely $N = 10,000$. The number of subdivisions of the axes for probability density function estimation were automatically selected as $S_1 = 10$, $S_2 = 10$ and $S_y = 15$. The number of subdivisions of the axes for model estimation were selected as $B_1 = 60$ and $B_2 = 70$. The variance of the noise ν_1 was 10^{-2}. In order to correctly evaluate the discrepancy between the estimated model and the actual model, the constant additive estimation bias due to the boundary condition was gotten rid of. The sub-domain of interest for model estimation was selected as $[0.2034,\ 0.8644] \times [0.3768,\ 1.7390]$. The Figure 5 illustrates the result of numerical estimation of the necessary joint probability density functions. No smoothing kernel was made use of. The Figure 6 illustrates the numerical behavior of the statistical regression algorithm, run over 10 iterations, in terms of mean absolute error and of weighted absolute error. In this experiment, the data-model to estimate is linear in both independent variables and the estimation result is good over the chosen sub-domain. The statistical regression algorithm does converge in a few iterations (less than 5, in this example). The weighted absolute error, in particular, evidences how the modeling procedure is less accurate at the border of the support of the probability density functions. The Figure 7 compares the estimated model with the actual model, on the two sections $f(x_1, x_2^{\mathrm{m}})$ and $f(x_1^{\mathrm{m}}, x_2)$. The figure also shows some data-points superimposed to the

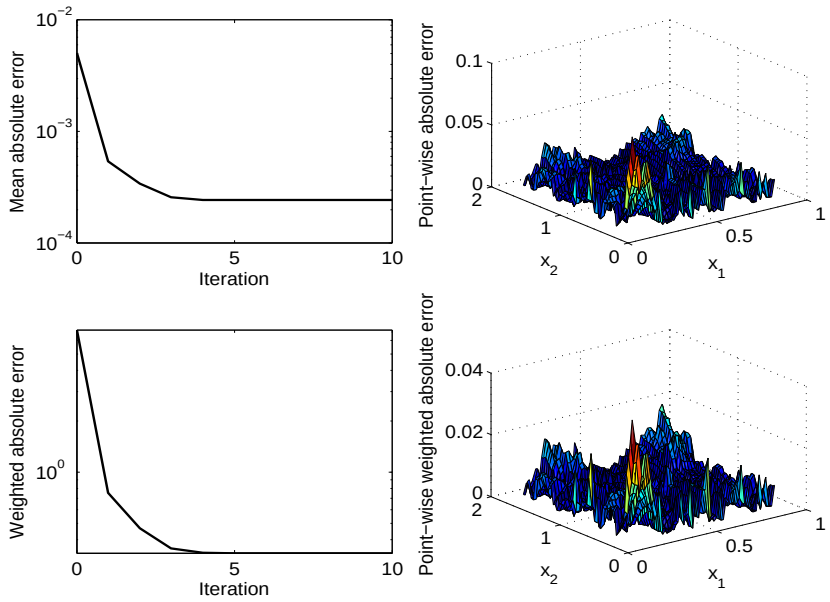

Fig. 6 Numerical behavior of the statistical regression algorithm on the Test problem 1. Top-left panel: Mean absolute error during iteration of the algorithm. Top-right panel: Point-wise absolute error after 10 iterations. Bottom-left panel: Weighted mean absolute error during iteration of the algorithm. Bottom-right panel: Weighted point-wise absolute error after 10 iterations.

estimated model. The estimated model looks in excellent agreement with the actual model in the selected sub-domain.

The second numerical experiment is about applying the devised statistical regression method to the Test problem 2. In order to obtain a good estimate of the probability density functions, $N = 10,000$ samples were generated. The number of subdivisions of the axes for probability density function estimation were automatically selected as $S_1 = 10$, $S_2 = 10$ and $S_y = 15$. The number of subdivisions of the axes for model estimation were selected again as $B_1 = 60$ and $B_2 = 70$. The variance of the noise ν_1 was 25×10^{-4}. The sub-domain of interest for model estimation was selected again as $[0.2034,\ 0.8644] \times [0.3768,\ 1.7390]$. The Figure 8 illustrates the result of numerical estimation of the necessary joint probability density functions. No smoothing kernel was made use of. The Figure 9 illustrates the numerical behaviour of the statistical regression algorithm, run over 10 iterations, in terms of absolute error. In this experiment, the data-model to estimate is non-linear in both independent variables and the estimation result is good over the chosen sub-domain. The statistical regression algorithm does converge in a few iterations (less than 5, in this example). The Figure 10 compares, in particular, the estimated model with the actual model, on the two sections $f(x_1, x_2^{\mathrm{m}})$ and $f(x_1^{\mathrm{m}}, x_2)$. The figure also shows some data-points superimposed to the esti-

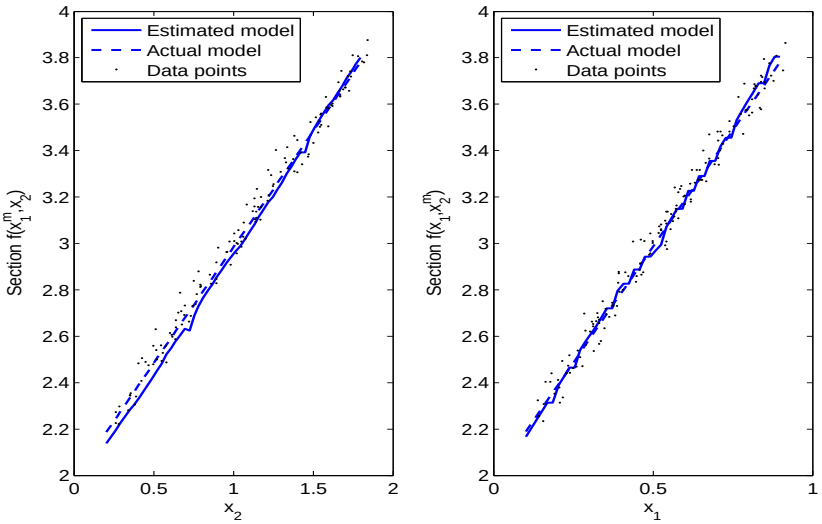

Fig. 7 Numerical behaviour of the statistical regression algorithm on the Test problem 1. Comparison of the estimated model with the actual model on the two sections $f(x_1, x_2^{\mathrm{m}})$ and $f(x_1^{\mathrm{m}}, x_2)$.

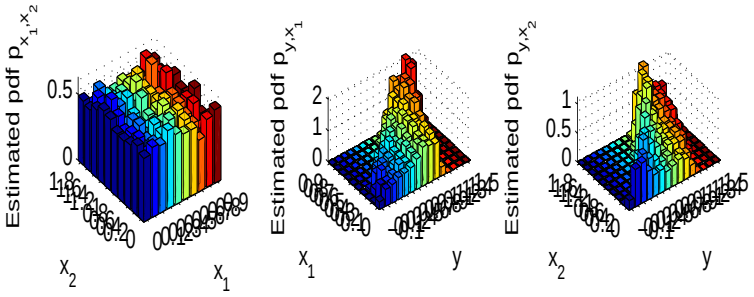

Fig. 8 Result of numerical estimation of the joint probability density functions p_{x_1,x_2}, p_{y,x_1} and p_{y,x_2}, from left to right, on the Test problem 2. Bivariate probability estimates are obtained by a histogram-based method.

mated model. The estimated model looks in good agreement with the actual model in the selected sub-domain.

The third numerical experiment concerns the empirical convergence analysis of the devised statistical regression algorithm on the data set of Test problem 1. In particular, the experiment aims at examining the convergence capability of the fixed-point algorithm (42) in relation to the number of available data-points N. The number N varies, with a logarithmic law, from $1,000$ to $100,000$ and, for each value of N, as much as 50 independent trials were conducted (by generating a new data-set over each trial). The number of successful trials over the whole set of trials and the average mean absolute error (calculated over the successful trials) are shown in the Figure 11. From the

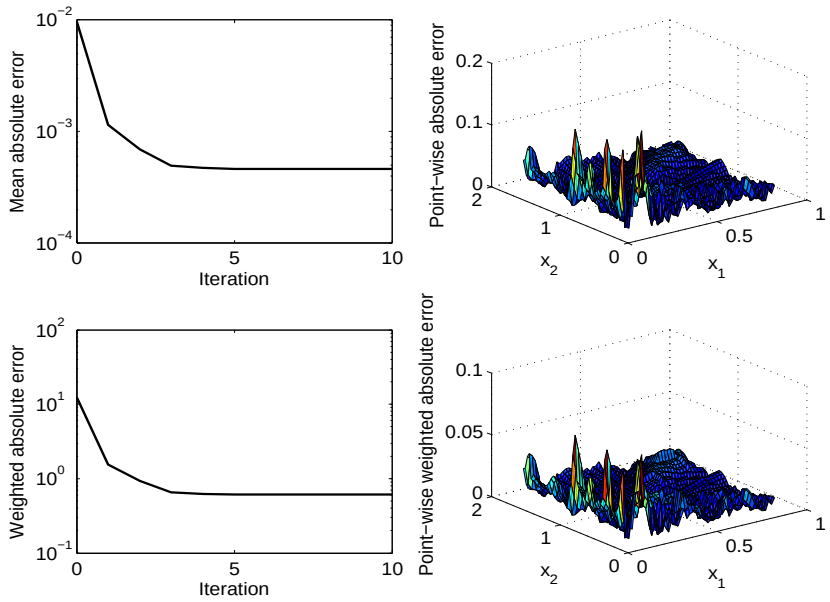

Fig. 9 Numerical behavior of the statistical regression algorithm on the Test problem 2. Top-left panel: Mean absolute error during iteration of the algorithm. Top-right panel: Point-wise absolute error after 10 iterations. Bottom-left panel: Weighted mean absolute error during iteration of the algorithm. Bottom-right panel: Weighted point-wise absolute error after 10 iterations.

above empirical results, it is concluded that the fixed-point algorithm (42) converges steadily if the number of available data-points is sufficiently large to allow a correct estimation of the bivariate probability density functions.

The fourth numerical experiment is about applying the devised statistical regression method to the statistical regression problem posed in the contribution [5], which concerns the formation of *acrylamide*[1] during the process of cooking French fries. Human exposure to acrylamide through consumption of French fries and other foods has been recognized as a potential health concern. Documented studies have found that increased dietary intake of acrylamide is associated with increased risks of postmenopausal endometrial and ovarian cancer, particularly among nonsmokers, that consumption of French fries during preschool years is associated with a slightly increased risk of breast cancer later in life, and that dietary acrylamide is significantly associated with increased risk of oral cavity cancer in female nonsmokers [5]. It has been observed that acrylamide formation can be greatly influenced by food processing conditions: At high temperatures during food processing, foods rich in carbo-

[1] Acrylamide [20] is a chemical compound with chemical formula C_3H_5NO. It is a white odorless crystalline solid, soluble in water, ethanol, ether and chloroform. Acrylamide decomposes in the presence of acids, bases, oxidizing agents, iron and iron salts. It decomposes non-thermally to form ammonia, while its thermal decomposition produces carbon monoxide, carbon dioxide and oxides of nitrogen.

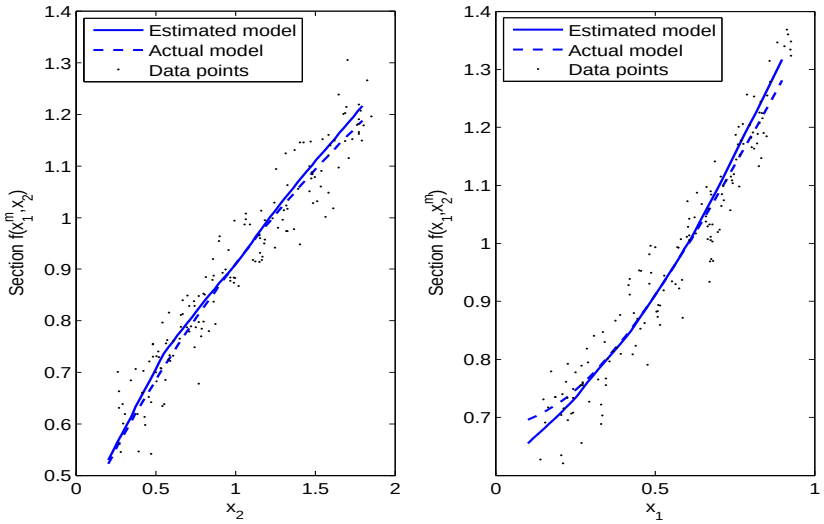

Fig. 10 Numerical behavior of the statistical regression algorithm on the Test problem 2. Comparison of the estimated model with the actual model on the two sections $f(x_1, x_2^{\mathrm{m}})$ and $f(x_1^{\mathrm{m}}, x_2)$.

hydrate can go through a series of reactions, to form acrylamide. In terms of food processing factors affecting acrylamide levels in French fries, frying time and temperature are the most important key parameters influencing acrylamide formation. The study presented in [5] is based on a model that uses cooking time and temperature as independent variables to predict the concentrations of acrylamide in French fries. Such data set is used in the present experiment as a case-study for the statistical regression method devised in the Section 2. In this context, the predictor variable x_1 denotes the cooking temperature expressed in °C, the predictor variable x_2 denotes the cooking time in seconds, while the target variable y denotes the concentration of acrylamide, expressed in μg/kg. The Figure 12 shows the result of modeling in terms of two cross-sections $f(x_1, x_2^{\mathrm{m}})$ and $f(x_1^{\mathrm{m}}, x_2)$. The obtained model seems to fit the data satisfactorily. The obtained model is able to predict the sudden rise of the concentration of acrylamide at around $x_1 = 180$°C, as predicted in [5]. It is also important to underline that, due to the intrinsic limitation of the proposed statistical regression method, which only allows to model monotonic behaviors, the obtained model is unable to predict a decrease of acrylamide concentration for $180\text{°C} < x_1 < 190\text{°C}$, which was observed in [5] (allegedly, acrylamide is eliminated through the evaporation process when temperature rises over 190°C).

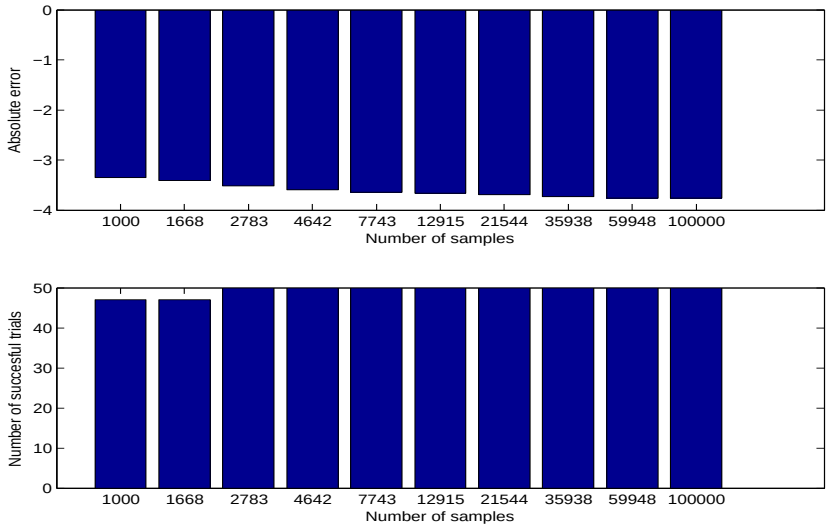

Fig. 11 Convergence of the statistical regression algorithm on the Test problem 1. Top panel: Mean absolute error, averaged over the number of successful trials. Bottom panel: Number of successful trials over a total of 50 trials versus changing cardinality N of the data sets.

6 Conclusions

The present research work outlined the main ideas behind isotonic statistical trivariate regression based on the invariance of measures in probabilistic spaces. The principle of probabilistic measure invariance leads to a system of two differential constraints to be obeyed to by the sought-after model, along with the assumed constrain of monotonicity of the model with respect to both independent variables. Such differential system is reformulated in terms of a single integral equation that affords an iterative numerical solution. Implementation issues such as the numerical representation and calculation of the required quantities, the estimation of required joint probability density functions from the available data and the numerical solution of the obtained integral equation were discussed. Numerical tests performed on the devised statistical regression procedure illustrated its features and confirmed that the model estimation result is good over the chosen sub-domains and that the iterative statistical regression algorithm converges in a few iterations. A numerical test about modeling the dependency of acrylamide concentration from frying time and temperature in the cooking of French fries further illustrated the behavior of the modeling procedure as well as its current limitations.

The present paper illustrated the initial step of a research work on statistical regression. In order to extend the content of the present work, the following issues are presently under consideration: 1) Extension of the devised trivariate regression method to a multivariate case with more than two predictors; 2) Selection of few predictors from a set of many on the basis of an analysis

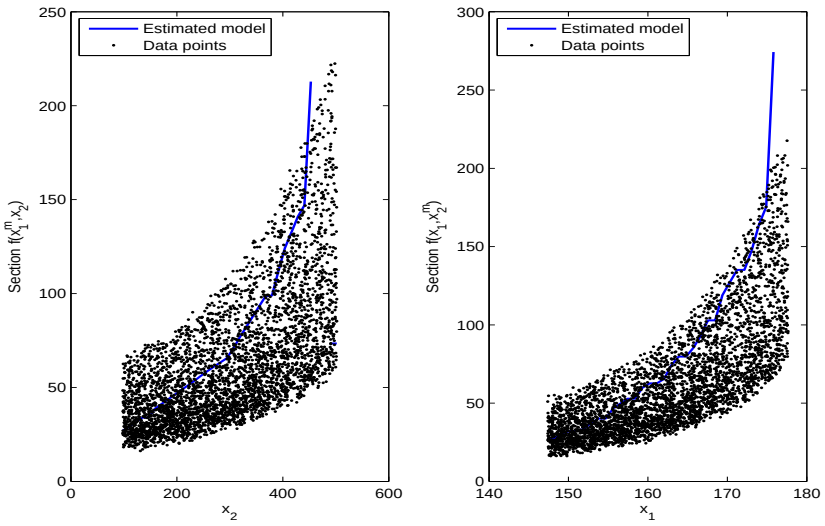

Fig. 12 Numerical behavior of the statistical regression algorithm on the regression problem about estimation of acrylamide formation during the process of cooking French fries. The variable x_1 denotes the cooking temperature (°C), the variable x_2 denotes the cooking time (sec), the variable y denotes the concentration of acrylamide (μg/kg).

of relevance tailored to the present regression model; 3) Estimation of joint probability density functions by a kernel method in presence of a limited number of available data; 4) Extension of the present procedure to non-monotonic models.

Acknowledgments

The author wishes to gratefully thank the anonymous referees and the associate editor who coordinated the review of the present paper for their thorough and stimulating comments and suggestions that helped improving and enriching the presentation of the technical content conveyed by the present manuscript.

References

1. S. Ahuja, A. Lakshminarayana and S.K. Shukla, "Statistical regression based power models," in *Low Power Design with High-Level Power Estimation and Power-Aware Synthesis*, pp. 59 – 70, Springer New York, 2012
2. P.K. Andersen and R.D. Gill, "Cox's regression model for counting processes: A large sample study," *The Annals of Statistics*, Vol. 10, No. 4, pp. 1100 – 1120, December 1982
3. R.E. Barlow, D.J. Bartholomew, J.M. Bremner and H.D. Brunk, *Statistical Inference Under Order Restrictions*, New York: Wiley, 1972
4. M. Birke and H. Dette, "Testing strict monotonicity in nonparametric regression," *Mathematical Methods of Statistics*, Vol. 16, pp. 110 – 123, 2007

5. M.-J. Chen, H.-T. Hsu, C.-L. Lin and W.-Y. Ju, "A statistical regression model for the estimation of acrylamide concentrations in French fries for excess lifetime cancer risk assessment," *Food and Chemical Toxicology*, Vol. 50, No. 10, pp. 3867 – 3876, October 2012
6. A. Colubi, J.S. Domínguez-Menchero and G. González-Rodríguez, "Testing constancy for isotonic regressions," *Scandinavian Journal of Statistics*, Vol. 33, No. 3, pp. 463 – 475, September 2006
7. A. Colubi, J.S. Domínguez-Menchero and G. González-Rodríguez, "A test for constancy of isotonic regressions using the L_2-Norm," *Statistica Sinica*, Vol. 17, pp. 713 – 724, 2007
8. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, Vol. 47, pp. 547 – 533, 2009
9. J.S. Domínguez-Menchero and G. González-Rodríguez, "Analyzing an extension of the isotonic regression problem," *Metrika*, Vol. 66, No. 1, pp. 19 – 30, July 2007
10. J.S. Domínguez-Menchero, G. González-Rodríguez and M.J. López-Palomo, "An L_2 point of view in testing monotone regression," *Journal of Nonparametric Statistics*, Vol. 17, No. 2, pp. 135 – 153, 2005
11. C. Durot, "A Kolmogorov-type test for monotonicity of regression," *Statistics & Probability Letters*, Vol. 63, pp. 425 – 433, 2003
12. S. Fiori, "Statistical nonparametric bivariate isotonic regression by look-up-table-based neural networks," Proceedings of the *2011 International Conference on Neural Information Processing* (ICONIP 2011, Shanghai (China), November 14-17, 2011), B.-L. Lu, L. Zhang, and J. Kwok (Eds.), Part III, LNCS 7064, pp. 365 – 372, Springer, Heidelberg, 2011
13. S. Fiori, "Fast statistical regression in presence of a dominant independent variable," *Neural Computing and Applications* (Springer). (Special issue of the 2011 International Conference on Neural Information Processing - ICONIP'2011). Accepted for publication (available online: June 2012)
14. D.R. Forrest, R.D. Hetland and S.F. DiMarco, "Multivariable statistical regression models of the areal extent of hypoxia over the Texas-Louisiana continental shelf," *Environmental Research Letters*, Vol. 6, No. 4, 045002 (10 pp), October-December 2011
15. M.A. Kulkarni, S. Patil, G.V. Rama and P.N. Sen, "Wind speed prediction using statistical regression and neural network," *Journal of Earth and System Sciences*, Vol. 117, No. 4, pp. 457 – 463, August 2008
16. X. Li and H.-Z. Liu, "Statistical regression for efficient high-dimensional modeling of analog and mixed-signal performance variations," Proceedings of the 45$^{\text{th}}$ *ACM/IEEE Design Automation Conference* (DAC 2008, Anaheim Convention Center, California, USA, June 9-13, 2008), pp. 38 – 43, June 2008
17. J. Liu and H. Li, "Application research of a statistical regression algorithm in the IVR system," Proceedings of the *2010 International Conference on Educational and Network Technology* (ICENT, Qinhuangdao (China), June 25-27, 2010), pp. 358 – 360, 2010
18. S. Liu, R.X. Gao, Q. He, J. Staudenmayer and P. Freedson, "Development of statistical regression models for ventilation estimation," Proceedings of the 31$^{\text{st}}$ *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBC, Minneapolis (Minnesota, USA,) September 2-6, 2009), pp. 1266 – 1269, 2009
19. N. Maheshwari, C. Balaji and A. Ramesh, "A nonlinear regression based multi-objective optimization of parameters based on experimental data from an IC engine fueled with biodiesel blends," *Biomass and Bioenergy*, Vol. 35, pp. 2171 – 2183, 2011
20. K.-M. Marstokk, H. Møllendal and S. Samdal, "Microwave spectrum, conformational equilibrium, ^{14}N quadrupole coupling constants, dipole moment, vibrational frequencies and quantum chemical calculations for acrylamide," *Journal of Molecular Structure*, Vol. 524, No.s 13, pp. 69 – 85, June 2000
21. S. Qian and W.F. Eddy, "An algorithm for isotonic regression on ordered rectangular grids," *Journal of Computational and Graphical Statistics*, Vol. 5, No. 3, pp. 225 – 235, September 1996
22. T. Robertson, F.T. Wright and R.L. Dykstra, *Order Restricted Statistical Inference*, New York: Wiley, 1988

23. R. Roelant, D. Constales, R. Van Keer and G.B. Marin, "Second-order statistical regression and conditioning of replicate transient kinetic data," *Chemical Engineering Science*, Vol. 63, No. 7, pp. 1850 – 1865, April 2008
24. D.W. Scott and S.R. Sain, "Multi-dimensional density estimation", In *Handbook of Statistics, Data Mining and Data Visualization*, Vol. 24, pp. 229 – 261, 2005
25. T. Thierfelder, "Empirical/statistical modeling of water quality in dimictic glacial/boreal lakes," *Journal of Hydrology*, Vol. 220, pp. 186 – 208, 1999
26. M.V. Velikova, "Monotone models for prediction in data mining", *Ph.D. Dissertation*, Dutch Graduate School for Information and Knowledge Systems and Graduate School of the Faculty of Economics and Business Administration of Tilburg University, 2006
27. K. White Vugrin, L. Painton Swiler, R.M. Roberts, N.J. Stucky-Mack and S.P. Sullivan, "Confidence region estimation techniques for nonlinear regression: Three case studies," Sandia report SAND2005-6893 (Unlimited Release), October 2005
28. R. Woodhouse, *Statistical Regression Line-Fitting in the Oil and Gas Industry*, PennWell Books, 2003
29. A. Žilinskas and J. Žilinskas, "Interval arithmetic based optimization in nonlinear regression," *INFORMATICA*, Vol. 21, No. 1, pp. 149 – 158, 2010
30. D.H. Zou, "Statistical regression applied to borehole strain measurements data analysis," *Geotechnical and Geological Engineering*, Vol. 13, No. 1, pp. 17 – 27, 1995