FAST CLOSED-FORM TRIVARIATE STATISTICAL ISOTONIC MODELING*

SIMONE FIORI†

Abstract. The present Letter introduces a non-iterative, closed-form solution to the problem of statistically modeling the monotonic relationship between a dependent variable and two independent variables through a probability conservation principle.

1. Introduction. A number of data-processing applications deal with modeling the relationship between a dependent variable $U \in \mathbb{R}$ and a pair of independent variables $(X, Y) \in \mathbb{R}^2$. Such kind of modeling is termed *trivariate*. A first noteworthy illustration comes from food toxicology research [1], where the concentration of acrylamide (dependent variable U) produced in the process of cooking French fries is modeled in terms of cooking time (independent variable X) and cooking temperature (independent variable Y). A second interesting example is found in bioenergy analysis [5], where the smoke emission (dependent variable) of a compression ignition engine fed with biomass-derived fuel is modeled in terms of injection timing (independent variable) and biomass blend ratio (independent variable).

A specific modeling problem arises whenever it is known in advance that the underlying relationship between the dependent variable and the two independent variables is of *monotonic* nature. Examples of such situations occur, e.g., in the modeling of electronic devices such as photodiodes in photovoltaic mode (where the independent variables are illumination level and voltage, while the dependent variable is current intensity) and MOSFET transistors (where the independent variables are gate voltage and source voltage, while the independent variable is channel's current intensity). A model that embodies the feature of monotonic dependency is termed *isotonic model* (see, e.g., [4]).

Trivariate modeling consists in inferring a model $U = f(X, Y)$, on the basis of a number of observed triples (U_s, X_s, Y_s), $s = 1, \ldots, S$, where S denotes the total number of observations. The triples of values (U_s, X_s, Y_s) collected in the experimental setting may be affected by measurement errors and the dependency between the three variables may be affected by hidden/unobservable variables that are not taken into account in the modeling process. In such a context, the trivariate modeling process may make use of statistical information, obtained by pooling the available observations (U_s, X_s, Y_s), rather than the observations themselves. The corresponding modeling paradigm is termed *statistical modeling* (see, e.g., [3]).

The present Letter introduces a closed-form instance of trivariate statistical isotonic modeling. The introduced method requires the estimation of the second-order joint probability density functions $p_{X,Y}$ and $p_{U,Y}$ and is based on a probability conservation law of statistics, that constraints the shape of the underlying model f. All the required quantities are stored in tables of numbers and the mathematical operations required to infer the model are simple additions/multiplications and table search.

2. Modeling principle and implementation details. Let $X, Y \in \mathbb{R}^2$ denote two continuous random variables with joint probability density function $p_{X,Y}$. Let

*S. FIORI, "FAST CLOSED-FORM TRIVARIATE STATISTICAL ISOTONIC MODELING", *ELECTRONICS LETTERS*, VOL. 50, NO. 9, PP. 708 – 710, APRIL 2014.

†Dipartimento di Ingegneria dell'Informazione (DII), Facoltà di Ingegneria, Università Politecnica delle Marche, Via Brecce Bianche, Ancona I-60131, Italy, E-mail: s.fiori@univpm.it.

$f : \mathbb{R}^2 \to \mathbb{R}$ denote a regular function. The function f has *no restrictions* except for continuity, continuity of the first-order partial derivatives and monotonicity. In the present Letter, it is assumed that $\frac{\partial f}{\partial X} > 0$. Define the random variables pair $(U, V) \stackrel{\text{def}}{=} (f(X, Y), Y)$. The new random variables are continuous and their joint probability density function is denoted by $p_{U,V}$.

The relationship between the probability density functions $p_{X,Y}$ and $p_{U,V}$ is given by [6]:

$$p_{U,V} = \frac{p_{X,Y}}{\left| \frac{\partial U}{\partial X}\frac{\partial V}{\partial Y} - \frac{\partial U}{\partial Y}\frac{\partial V}{\partial X} \right|}. \tag{2.1}$$

Recalling that $V = Y$ and the assumption on monotonicity, the above relationship becomes:

$$p_{U,Y}\frac{\partial f}{\partial X} = p_{X,Y}. \tag{2.2}$$

In the context of statistical modeling, the quantities $p_{X,Y}$ and $p_{U,Y}$ are estimated by pooling the available observations (U_s, X_s, Y_s). Hence, the above equation has the model f as only unknown. Introduce the following integrals (that denote marginal cumulative distribution functions):

$$P_{X,Y}(x, y) \stackrel{\text{def}}{=} \int_{-\infty}^{x} p_{X,Y}(z, y)dz, \tag{2.3}$$

$$P_{U,Y}(u, y) \stackrel{\text{def}}{=} \int_{-\infty}^{u} p_{U,Y}(z, y)dz. \tag{2.4}$$

Thanks to the above marginal cumulative distribution functions, the solution of the equation (2.2) may be expressed, in intrinsic form, as:

$$P_{U,Y}(f(x, y), y) = P_{X,Y}(x, y). \tag{2.5}$$

Fixing a pair of values (x, y), the left-hand term of the solution (2.5) is an invertible function of the variable $f(x, y)$, while the right-hand term is a known value. Hence, for any fixed pair of values (x, y), the corresponding value of the model $f(x, y)$ may be found by function inversion.

From an implementation viewpoint, the probability density functions $p_{X,Y}$ and $p_{U,Y}$ are estimated through normalized occurrence histograms [7] and are represented by numerical tables (or *look-up tables* – see, e.g., [2] for the univariate case, which can be easily extended to the multivariate case). The integrals $P_{X,Y}(x, y)$ and $P_{U,Y}(u, y)$ are estimated by cumulative sums and are likewise represented by numerical tables. For every given point (x, y) of the look-up table that represents the variables (X, Y), the model value $f(x, y)$ is estimated as the unique value $\hat{u}$ that minimizes the quantity $|P_{U,Y}(u, y) - P_{X,Y}(x, y)|$. Finding such u-value does not require any burdensome calculation, as it may be found through a proximity search on a row of the table that represents the quantity $P_{U,Y}(u, y)$.

The following MATLAB© (version 7.8) implementation of the whole modeling procedure illustrates its simplicity:

```
% Range of variables and width of
% subdivisions for pdf estimation
xmin=min(x); xmax=max(x); wx=std(x)/S^(1/4);
```

```
ymin=min(y); ymax=max(y); wy=std(y)/S^(1/4);
umin=min(u); umax=max(u); wu=std(u)/S^(1/4);
% Variables x, y and u quantized on the
% grid for pdf/model estimation
xg = xmin:wx:xmax; yg = ymin:wy:ymax;
ug = umin:wu:umax;
% Number of subdivisions of the axes for
% pdf and model estimation
Sx = length(xg); Sy = length(yg);
Su = length(ug);
% Estimates pdfs and cumulative functions
pxy = hist3([y' x'],[Sy Sx]);
 pxy = pxy/(sum(sum(pxy))*wy);
 xPxy = cumsum(pxy,2);
puy = hist3([y' u'],[Sy Su]);
 puy = puy/(sum(sum(puy))*wy);
 uPuy = cumsum(puy,2);
% Statistical modeling algorithm
uu = zeros(Sy,Sx);
for iy = 1:Sy,
 for jx = 1:Sx,
  [~,ku] = min(abs(xPxy(iy,jx)-uPuy(iy,:)));
  uu(iy,jx) = ug(ku);
 end
end
```

The above non-iterative procedure inputs the triples (U_s, X_s, Y_s) as three vectors `u`, `x` and `y`, each of length S. The above procedure returns a pair of vectors, namely `xg` and `yg`, and a matrix `uu` that represents the estimated model. In the above version of the statistical isotonic modeling procedure, the number of subdivisions for probability density function estimation is automatically selected and coincides with the number of partitions of the X, Y and U axes for model estimation. The procedure may be modified so as to make the finesse of partition for probability density function estimation and the number of partitions of the X, Y and U axes for model estimation be independent and to choose such partitions manually. In the above code, the instruction `hist3` estimates the joint probability density function of two variable by a histogram method. The instruction `cumsum` estimates the cumulative distribution function with respect to one of the variables.

3. Numerical experiments. Two numerical experiments are discussed, which involve known non-linear models. In both experiments, the number of available samples was set to $S = 10,000$.

The first numerical experiment refers to the model $f(X,Y) = 2X + Y + XY + 1$. The variate X is uniformly generated in $[0,1]$ and the variate Y is uniformly generated in the range $[0,\ 2]$. The variable U is generated by the rule $f(X,Y) + \varepsilon$, with ε zero-mean Gaussian with standard deviation 0.1. The Figure 3.1 compares the estimated model with the actual model. The estimated model looks in excellent agreement with the actual model, as also testified by the low values of the relative point-wise estimation error, defined as

$$\frac{|\hat{u}(x,y) - f(x,y)|}{|f(x,y)|}. \tag{3.1}$$

The second numerical experiment refers to the actual model $f(X,Y) = \log(3X^2 + 4Y + 2)$. In such experiment, the variate X is uniformly generated in $[0,1]$, while the

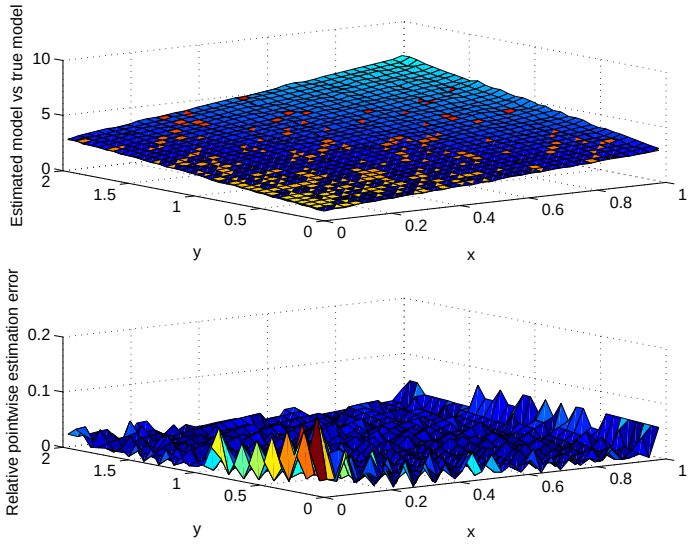

FIG. 3.1. *Numerical experiment performed on the model* $f(X,Y) = 2X+Y+XY+1$. *Top panel: Estimated model versus the actual model. Bottom panel: Relative point-wise estimation error.*

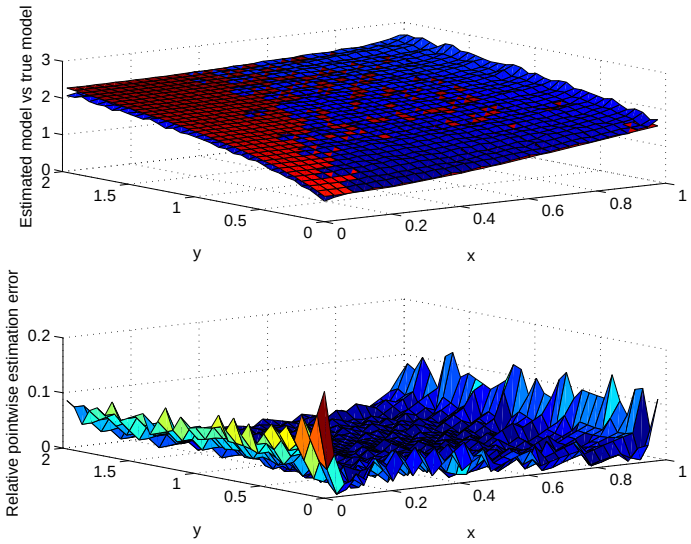

FIG. 3.2. *Numerical experiment performed on the model* $f(X,Y) = \log(3X^2 + 4Y + 2)$. *Top panel: Estimated model versus the actual model. Bottom panel: Relative point-wise estimation error.*

variate Y is uniformly generated in the range $[0,\ 2]$. The variable U is generated by the rule $f(X,Y) + \varepsilon$, with ε zero-mean Gaussian, again with standard deviation 0.1. The Figure 3.2 compares the estimated model and the actual model. The estimated model looks in good agreement with the actual model, at the center of the domain, as also confirmed by the low values of the relative point-wise estimation error. A larger error is observed in the peripheral part of the (X,Y) domain, where the estimation of the probability density functions is less accurate.

4. Conclusion. The present Letter introduced a closed-form method to perform statistical isotonic trivariate modeling. The distinguishing features of the presented statistical modeling technique may be summarized as follows:

- The proposed method does not make direct use of the samples (U_s, X_s, Y_s). The proposed method extracts collective information from the data and represent such information as codified by the second-order joint probability functions $p_{X,Y}$ and $p_{U,Y}$. As a consequence, the model does not try to fit the data samples (which may be unreliable due to measurement errors or other hidden/nuisance variables) but tries to capture the overall structure of the underlying physical system.
- The principle informing the discussed modeling procedure is drawn from a probability conservation law, which differentiates the proposed modeling concept from the classical linear/non-linear fitting schemes. The underlying probability conservation law holds true irrespective of the shape of the involved probability density functions and model, provided continuity and monotonicity hold.
- The proposed procedure does not make any assumption on the shape of the model, except for continuity and monotonicity. As a result, the model is unrestricted and there is no need to choose any functional dependency beforehand, which differentiates the proposed modeling from the parametric/maximum-likelihood estimation methods.
- The involved quantities, namely, the probability density functions and the inferred model, are represented by simple numerical tables. The probability density functions are estimated by constructing occurrence histograms and the associated marginal cumulative distribution functions are estimated by cumulative sums. The model is estimated by proximity search within the tables, which barely require no mathematical operations. The resulting procedure is computationally light and very fast to execute.

The presented numerical experiments confirmed that the inferred statistical model is in good agreement with the actual model at the center of the (X, Y) domain, while a larger relative error is observed in the peripheral part of the domain. Future research endeavors will be oriented toward comparing the performances of the proposed method with other black-box approaches.

REFERENCES

[1] Chen, M.-J., Hsu, H.-T., Lin, C.-L. and Ju, W.-Y., 'A statistical regression model for the estimation of acrylamide concentrations in French fries for excess lifetime cancer risk assessment,' *Food and Chemical Toxicology*, 2012, **50**, pp. 3867-3876

[2] Fiori, S., 'Fast statistical regression in presence of a dominant independent variable', *Neural Computing and Applications*, 2013, **22**, pp. 1367-1378

[3] Fiori, S., 'An isotonic trivariate statistical regression method', *Advances in Data Analysis and Classification*, 2013, **7**, pp. 209-235

[4] Domínguez-Menchero, J.S. and González-Rodríguez, G., 'Analyzing an extension of the isotonic regression problem', *Metrika*, 2007, **66**, pp. 19-30

[5] Maheshwari, N., Balaji C. and Ramesh, A., 'A nonlinear regression based multi-objective optimization of parameters based on experimental data from an IC engine fueled with biodiesel blends', *Biomass and Bioenergy*, 2011, **35**, pp. 2171-2183

[6] Papoulis, A. and Unnikrishna Pillai, S., 'Probability, Random Variables and Stochastic Processes', 2002, McGraw-Hill (Fourth Edition)

[7] Scott, D.W. and Sain, S.R., 'Multi-dimensional density estimation', *Handbook of Statistics, Data Mining and Data Visualization*, 2005, **24**, pp. 229-261