Computers and Mathematics with Applications 67 (2014) 1171–1185

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

A two-dimensional Poisson equation formulation of non-parametric statistical non-linear modeling

S. Fiori

Dipartimento di Ingegneria dell'Informazione (DII), Facoltà di Ingegneria, Università Politecnica delle Marche, Via Brecce Bianche, Ancona I-60131, Italy

ARTICLE INFO

Article history:
Received 26 July 2013
Received in revised form 29 November 2013
Accepted 3 December 2013

Keywords:
Poisson equation
Statistical modeling
Non-linear system
Non-parametric modeling
Isotonic modeling

ABSTRACT

The present paper deals with a Poisson equation arising in statistical modeling of semi-deterministic non-linear systems with two independent (input) variables and one dependent (output) variable. Statistical modeling is formulated in terms of a differential equation that relates the second-order joint probability density functions of the model's input/output random variables with the sought non-linear model transference. The discussed modeling procedure makes no prior assumptions on the functional structure of the model, except for monotonicity and continuity with respect to both input variables. In particular, the method is non-parametric. Results of numerical tests are presented and discussed in order to get an insight into the behavior of the devised statistical modeling procedure. The results of numerical tests confirm that the proposed statistical modeling approach is able to cope with both synthetic and real-world data sets and, in particular, with underlying systems and data that exhibit strong hidden nuisance variables and measurement disturbances.

1. Introduction

Real-world phenomena may rarely be described accurately by a mathematical model to be evaluated analytically. Statistical modeling provides a useful tool to build up a model of a phenomenon under observation, on the basis of the statistical features of the variables describing such a phenomenon. Statistical modeling finds applications in as diverse fields as social and behavioral sciences [1], biomedical research [2], computer vision and content-based image retrieval [3,4] and econometrics [5]. As intended here, statistical modeling is based on four main assumptions: (1) The physical phenomenon under observation relates two independent variables with a single dependent variable. Namely, it is assumed that three variables of interest in a modeling problem are related by $y = \Phi(x_1, x_2)$, where $y \in \mathcal{Y}$ represents the *dependent variate* and $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 \subset \mathbb{R}^2$ represent the *independent variates*. The function $\Phi : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$ is semi-deterministic as it includes nuisance variables that cannot be directly measured. The joint statistical features of the input/output variates are described by the joint probability density function $p_{y,x_1,x_2}(y, x_1, x_2)$, where, by a slight abuse of notation, the variates have been confused with their realizations. (2) The sought model is monotonically increasing or decreasing with respect to both input variables (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 [6]. (3) Statistical modeling is based on the estimation of second-order joint probability density functions: it is assumed that the number of available observations of the triples (x_1, x_2, y) is large enough to get meaningful statistical estimates. (4) The model is non-linear, non-parametric, namely, there is no assumption on its shape (except for the assumptions of monotonicity and continuity), hence, the devised modeling technique may cope with arbitrary dependencies, albeit restricted to be monotonic.

E-mail address: s.fiori@univpm.it.

0898-1221/$ – see front matter
http://dx.doi.org/10.1016/j.camwa.2013.12.002

Non-parametric modeling is a form of analysis in which the predictor does not take a predetermined form but is constructed according to information derived from the data. Non-parametric modeling requires larger data sets than modeling based on parametric models because the data must supply the model structure as well as the model estimates.

The isotonic modeling problem is a special case of the general modeling problem that arises in various fields, such as production planning, inventory control and psychometry [7]. Isotonic modeling occurs whenever it is known that the dependency between the output variable and the input variables is either monotonically increasing or monotonically decreasing. Typically, isotonic modeling in a single-input/single-output setting is formulated as a constrained quadratic programming problem. The notion of isotonic modeling may be traced back to the seminal contribution [8]. The formulation in terms of least-squares may be generalized to L_1 norm [9] and to L_∞ norm (which gives rise to the notion of strict isotonic modeling) [10]. Note that isotonic modeling is often applied to nonmonotonic data, which justifies the assumption on the underlying system to be semi-deterministic. An example is given in [11]: consider the problem of measuring the viscosity of a fluid at different temperatures. Viscosity is a nonincreasing function of temperature; however, *due to measurement error*, the observed viscosity may not be nonincreasing when ordered by temperature. In this case, the purpose of isotonic modeling is to replace the observed viscosities with a set of values that are nonincreasing when ordered by the temperature.

The present research work aims at extending the previous contribution on one-to-one statistical modeling summarized in the conference paper [12] and explained in detail in the publication [13]. A feature of the present work that is inherited from the above contribution is that the inferred model is non-parametric. An example of the importance of non-parametric isotonic modeling is offered in [10]: researchers are less willing to impose strong assumptions in their modeling. For instance, applied researchers may be willing to make the weak assumption that the expected height of a woman is an increasing function of the height of her father and of her mother, but be unwilling to make parametric assumptions such as linearity.

Statistical isotonic modeling is about determining a relationship between the two independent variates $(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ and the dependent variate $y \in \mathcal{Y}$, described by the model:

$$y = f(x_1, x_2), \tag{1}$$

with $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$ denoting a non-linear deterministic model, by making use of pooled information only, summarized by the joint probability distributions of the variates y, x_1, x_2. The statistical modeling method proposed in the present paper is based on known formulas to quantify the distortion of the statistical distribution of the input values operated by a deterministic monotonic non-linear system.

Paper organization: The present paper is organized as follows. Section 2 explains the fundamental principles behind the proposed bivariate isotonic statistical modeling approach. In particular, it shows that an appropriate choice of a two-input/two-output non-linear system to probe the sought two-input/one-output model allows formulating the statistical isotonic modeling problem in terms of conservation of probability measures in the model's second-order joint probability space. The result is a system of partial differential equations having the sought model as unknown. Section 3 of the present paper moves forward the formulation in terms of a two-dimensional Poisson equation derived on the basis of a variational principle applied to a functional least-squares-error formulation. In such a section, the problem of setting up appropriate boundary conditions is also discussed on and the numerical implementation used to solve the Poisson equation is described briefly. Section 4 summarizes the quantities used for measuring objectively the features of the devised bivariate isotonic statistical modeling technique. Section 4 also illustrates and discusses the results of modeling synthetic data sets as well as real-world data sets, namely, a data set arising from robotic arm dynamics, a data set arising from a food toxicology research and a data set from quantitative palynology research. The results of modeling the synthetic as well as the real-world data sets are encouraging and show that the devised statistical modeling technique can cope with underlying systems and data that include strong hidden nuisance variables and measurement disturbances. Section 5 concludes the paper.

2. Statistical bivariate isotonic modeling: fundamental principles

Monotonic dependencies are common in physical systems. For instance, the rates of biogeochemical processes can be monotonic functions of factors like temperature and humidity. For such systems, it is of prime importance to infer monotonic relationships from a given data set by constructing a model that is consistent with monotonicity, namely, that is isotonic. The majority of the challenging applied isotonic modeling problems is characterized by very large data sets. An example is offered in [9]: in the analysis of large-scale microarray data, which is one of the most important tools in biology, the same procedure is used for studying the fit of tens of thousands of genes to a given partial order. A further example is offered in [14], which concerns the classification of large portions of texts extracted from the World Wide Web.

2.1. System-theoretic grounds of the proposed statistical modeling method

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

$$(y, z) = \varphi(x_1, x_2), \tag{2}$$

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 (2) is supposed to be invertible in the domain of interest and its inverse is denoted by $\varphi^{-1} : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}_1 \times \mathcal{X}_2$.