Automatica 45 (2009) 173–179

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Fast algorithms for nonparametric population modeling of large data sets☆

Gianluigi Pillonetto [a,*], Giuseppe De Nicolao [b], Marco Chierici [c], Claudio Cobelli [a]

[a] Dipartimento di Ingegneria dell'Informazione, University of Padova, Italy
[b] Dipartimento di Informatica e Sistemistica, University of Pavia, Italy
[c] Predictive Models for Biomedicine and Environment Unit, Fondazione Bruno Kessler, Trento, Italy

ARTICLE INFO

Article history:
Received 11 April 2007
Received in revised form
28 April 2008
Accepted 5 June 2008
Available online 5 December 2008

Keywords:
Nonparametric identification
Bayesian estimation
Glucose metabolism
Gaussian processes
Estimation theory

ABSTRACT

Population models are widely applied in biomedical data analysis since they characterize both the average and individual responses of a population of subjects. In the absence of a reliable mechanistic model, one can resort to the Bayesian nonparametric approach that models the individual curves as Gaussian processes. This paper develops an efficient computational scheme for estimating the average and individual curves from large data sets collected in standardized experiments, i.e. with a fixed sampling schedule. It is shown that the overall scheme exhibits a "client–server" architecture. The server is in charge of handling and processing the collective data base of past experiments. The clients ask the server for the information needed to reconstruct the individual curve in a single new experiment. This architecture allows the clients to take advantage of the overall data set without violating possible privacy and confidentiality constraints and with negligible computational effort.

1. Introduction

One of the most interesting identification problems arising in biomedical data analysis is the characterization of a population of subjects. Classical examples are found in pharmacokinetics (PK) and pharmacodynamics (PD), where multiple subjects are sampled in order to obtain both the average and individual response to the administered drug. If a sufficiently large number of samples are collected in each individual, it is possible to identify a distinct model for each subject. The typical response of the population could then be obtained from the distribution of the individual models. However, the specific nature of biomedical experiments often poses technological, cost or ethical constraints that permits to collect only few data in each single subject. When the separate identification of individual models is not viable, an effective solution is provided by the so-called population modeling approaches (Beal & Sheiner, 1982; Davidian & Giltinan, 1995; Sheiner, 1994). Such methods process all the data simultaneously in order to achieve both the typical and individual models. Although originated in the PK/PD field, population modeling is becoming more and more popular also in other scenarios as metabolic systems, medical imaging and even genomics (Bertoldo, Sparacino, & Cobelli, 2004; Ferrazzi, Magni, & Bellazzi, 2003; Vicini & Cobelli, 2001).

The standard population model is a continuous-time dynamical system containing a finite number of unknown parameters, typically a compartmental model (Jacquez, 1985). This leads to a nonlinear-in-parameter identification problem that can be tackled resorting to various iterative algorithms. Among them, one may mention the celebrated NONMEM software (Beal & Sheiner, 1992), which relies on maximum likelihood estimation, but also Bayesian algorithms that compute the posterior distribution of parameters exploiting the Markov chain Monte Carlo (MCMC) machinery (Gilks, Richardson, & Spiegelhalter, 1996; Lunn, Best, Thomas, Wakefield, & Spiegelhalter, 2002).

At the early stages of a study or when the mechanistic model of a physiological phenomenon is not available, it may be difficult to formulate a reliable parametric model. Hence the need for flexible nonparametric population approaches that reduce the structural assumptions to a minimum (Ibragimov & Khasminskii, 1981). Along this direction, an example is provided by the so-called semiparametric methods that model the response curves as regression splines (Fattinger & Verotta, 1995; Park, Verotta, Blaschke, & Sheiner, 1997). A potential difficulty underlying the use of these techniques is the optimization of the number and location of the knots of regression splines, which could suffer from the presence of local minima. More recently, in order to develop a fully nonparametric approach, within a Bayesian paradigm it

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor George Yin under the direction of Editor Ian R. Petersen.

* Corresponding author. Tel.: +39 0498277607; fax: +39 049 827 7699.
E-mail addresses: giapi@dei.unipd.it (G. Pillonetto), giuseppe.denicolao@unipv.it (G. De Nicolao), chierici@fbk.eu (M. Chierici), cobelli@dei.unipd.it (C. Cobelli).

0005-1098/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2008.06.003

has been proposed to model the individual curves as realizations of discrete- or continuous-time stochastic processes, e.g. random walks or integrated Wiener processes (Magni, Bellazzi, DeNicolao, Poggesi, & Rocchetti, 2002; Neve, Nicolao, & Marchesi, 2007). In these works, each individual curve is seen as the sum of an average curve (common to all subjects) and an individual shift (varying from subject to subject). In particular, both the average curve and the individual shifts are assumed to be Gaussian processes whose statistics are specified by few hyperparameters. For instance, if the curve is an integrated Wiener process, the hyperparameter is the corresponding intensity. Hyperparameter tuning can be carried out via likelihood maximization. For a given choice of the hyperparameters, the posterior expectations of the processes given the data provide point estimates of the average and individual curves. In particular, when the prior is formulated in terms of integrated Wiener processes, the estimated curves are cubic splines (Neve et al., 2007). This Bayesian nonparametric approach has strong connections with kernel methods, Gaussian processes estimation, regularization networks (Evgeniou, Micchelli, & Pontil, 2005; Rasmussen & Williams, 2006). Recently, a Bayesian MCMC approach able to return the full posterior of hyperparameters and unknown functions has been also worked out, see Neve, Nicolao, and Marchesi (2008).

In this paper, attention is focused on the nonparametric population analysis of standardized experiments which involve a large number of subjects. Herein, the term standardized is used to denote an experiment that is repeated in multiple subjects adopting a standard sampling schedule. A notable example, treated later in the paper, is the intravenous glucose tolerance test (IVGTT), where glucose plasma concentration is measured after intravenous administration of a glucose bolus. This test is widely employed in the diagnosis of metabolic disorders, see e.g. Bergman, Bowden, and Cobelli (1981).

In the Bayesian nonparametric approach the computation of the posterior expectations calls for the solution of an algebraic linear system of order n_T, where n_T is the total number of observations. This is a potential drawback because the complexity scales with the cube of n_T. The burden may seem even worse in the case of standardized experiments involving a large number of subjects. As a matter of fact, in the present paper it is shown that the fixed sampling schedule can be exploited to design an algorithm whose complexity scales with the cube of the number of samples collected in each individual. This holds for evaluation of both the posterior expectations and confidence intervals.

The new algorithms pave the way to the implementation of a *client–server* architecture for managing the identification of population models for standardized experiments. The server computes and stores the sufficient statistics abstracted from a large historical data set. The client, whose aim is analyzing a single new experiment (not necessarily standardized), interrogates the server to get the information needed to compute the posterior expectation of the individual curve given all the historical data. The client can also send its data to the server in order to update the centralized sufficient statistics. As an example, the server could be managed by a reference research center, whereas the clients could be laboratories collecting and processing clinical data. According to this architecture, the local laboratories benefit from the information contained in the collective database in a computationally efficient way and without accessing individual data subject to privacy and confidentiality constraints. To the authors' knowledge, the client–server architecture is a novel contribution of this paper. In fact, most population modeling approaches cannot be decentralized because of their intrinsic nonlinear-in-parameter structure.

The paper is organized as follows. In Section 2 the problem is given its mathematical formulation. In Section 3 the computational algorithms are derived. In Section 4 the proposed methodology is tested on a large data set of IVGTT experiments. Some conclusions end the paper.

2. Statement of the problem

In what follows, $E[.]$ is used to denote the expectation operator and vectors are column vectors, unless otherwise specified. Further, given two random vectors q and w, let $\mathrm{cov}[q, w] = E[(q - E[q])(w - E[w])^{\mathrm{T}}]$ and $\mathrm{Var}[q] = E[(q - E[q])(q - E[q])^{\mathrm{T}}]$.

We consider the problem of estimating realizations of continuous-time stochastic processes $x^j(t)$, $j = 1, 2, \ldots, m+1$, from a finite number of noisy samples. The curves $x^j(t)$ represent the responses of $m+1$ subjects randomly drawn from a population. It is assumed that number and location of the sampling instants do not vary from subject to subject except for what concerns the last one. To be more specific, for $j = 1, \ldots, m$ the curves are sampled at instants $\{t_k\}, k = 1, 2, \ldots, n$, while the $(m+1)$-th curve is sampled at instants $\{t_k^*\}, k = 1, 2, \ldots, n^*$. The measurement model is

$$y_k^j = x^j(t_k) + v_k^j, \quad k = 1, \ldots, n,\ j = 1, \ldots, m$$
$$y_k^{m+1} = x^{m+1}(t_k^*) + v_k^{m+1}, \quad k = 1, \ldots, n^*$$

where $\{v^j\} = [v_1^j \ldots v_n^j]^{\mathrm{T}}$, $j = 1, \ldots, m$, and $\{v^{m+1}\} = [v_1^{m+1} \ldots v_{n^*}^{m+1}]^{\mathrm{T}}$ are Gaussian and independent random vectors such that for every k and j

$$E[v_k^j] = 0, \qquad \mathrm{Var}[v^j] = \Sigma_v^j.$$

We assume that the individual curves can be decomposed as

$$x^j(t) = \bar{x}(t) + \tilde{x}^j(t), \quad j = 1, \ldots, m+1$$

where $\bar{x}(t)$ and $\tilde{x}^j(t)$ are zero-mean normal stochastic processes that represent the average curve and the individual shift from the average, respectively. We also assume that processes $\{v_k^j\}_{j=1}^{m+1}$, $\bar{x}(t)$ and $\{\tilde{x}^j(t)\}_{j=1}^{m+1}$ are all mutually independent. For the sake of simplicity, it is assumed that $\{\tilde{x}^j(t)\}_{j=1}^{m+1}$ are identically distributed. Define now

$$y^j = [y_1^j \quad y_2^j \quad \ldots \quad y_n^j]^{\mathrm{T}}, \quad j = 1, 2, \ldots, m$$
$$y^{m+1} = [y_1^{m+1} \quad y_2^{m+1} \quad \ldots \quad y_{n^*}^{m+1}]^{\mathrm{T}}$$
$$y = \left[(y^1)^{\mathrm{T}} \ldots (y^m)^{\mathrm{T}}\right]^{\mathrm{T}} \qquad y^+ = \left[y^{\mathrm{T}} \quad (y^{m+1})^{\mathrm{T}}\right]^{\mathrm{T}}.$$

The paper is concerned with the solution of the following two estimation problems.

- Given y, for any t compute efficiently the continuous-time minimum variance estimate of the average curve $\bar{x}(t)$, i.e. $E[\bar{x}(t)|y]$, as well as the variance of the reconstruction error, i.e. $\mathrm{Var}[\bar{x}(t)|y]$.
- Assuming that a new data set y^{m+1} is available, for any t compute efficiently $E[x^{m+1}(t)|y^+]$ and $\mathrm{Var}[x^{m+1}(t)|y^+]$.

3. Computational algorithms

3.1. Computing $E[\bar{x}(t)|y]$ and $\mathrm{Var}[\bar{x}(t)|y]$

The aim is to derive efficient algorithms to compute the estimates $E[\bar{x}(\tau)|y]$, $E[x^{m+1}(\tau)|y^+]$, where τ is a generic temporal instant, together with their confidence intervals.

We start by introducing the following notation

$$\bar{x} = [\bar{x}(t_1) \ldots \bar{x}(t_n)]^{\mathrm{T}} \qquad \bar{x}_\tau = [\bar{x}(\tau) \quad \bar{x}(t_1) \ldots \bar{x}(t_n)]^{\mathrm{T}} \tag{1}$$
$$\tilde{x}^j = [\tilde{x}^j(t_1) \quad \ldots \quad \tilde{x}^j(t_n)]^{\mathrm{T}}, \quad j = 1, 2, \ldots, m$$

and

$$\bar{R} = \mathrm{Var}[\bar{x}] \qquad \bar{R}_\tau = \mathrm{cov}[\bar{x}_\tau, \bar{x}]$$
$$\bar{R}_{\tau\tau} = \mathrm{Var}[\bar{x}_\tau] \qquad \bar{r}_\tau = \mathrm{cov}[\bar{x}(\tau), \bar{x}] \tag{2}$$
$$\hat{R}_{\tau\tau} = \mathrm{Var}[\bar{x}_\tau|y] \qquad \tilde{R} = \mathrm{Var}[\tilde{x}^j], \quad j = 1, 2, \ldots, m.$$