Available online at www.sciencedirect.com
ScienceDirect
Automatica 43 (2007) 1134–1144
automatica
www.elsevier.com/locate/automatica

Nonparametric identification of population models via Gaussian processes ☆

M. Neve, G. De Nicolao*, L. Marchesi

Dipartimento di Informatica e Sistemistica, Università di Pavia, via Ferrata 1, 27100 Pavia, Italy

Received 7 June 2005; received in revised form 6 August 2006; accepted 15 December 2006
Available online 16 May 2007

Abstract

Population models are used to describe the dynamics of different subjects belonging to a population and play an important role in drug pharmacokinetics. A nonparametric identification scheme is proposed in which both the average impulse response of the population and the individual ones are modelled as Gaussian stochastic processes. Assuming that the average curve is an integrated Wiener process, it is shown that its estimate is a cubic spline. An empirical Bayes algorithm for estimating both the average and the individual curves is worked out. The model is tested on simulated data sets as well as on xenobiotics pharmacokinetic data.
© 2007 Elsevier Ltd. All rights reserved.

Keywords: Nonparametric identification; Estimation theory; Pharmacokinetic data; Splines; Neural networks; Regularization

1. Introduction

An important problem in biomedicine is that of characterizing the average behaviour as well as the inter-individual variability of a population of subjects. As an example, the analysis of population data is of primary importance in pharmacology, where drug responses measured in multiple subjects are used to obtain average and individual pharmacokinetic and pharmacodynamic models.

When it is possible to collect a sufficient number of observations for each subject, model identification can be performed separately for each individual. However, in many cases there are technical, ethical and cost reasons that limit the number of samples that can be collected in each subject. Some examples are given by toxicokinetic studies as well as pharmacological experiments involving critical patients such as neonatal, pediatric or intensive care unit ones. If the individual models cannot be identified separately, it is necessary to resort to so-called "population methods" that provide the average and individual models from the joint analysis of all the available data (Aarons, 1999; Beal & Sheiner, 1982; Davidian & Giltinan, 1995; Guardabasso, Munson, & Rodbard, 1988; Sheiner, 1994; Sheiner, Rosenberg, & Marathe, 1977; Sheiner & Steimer, 2000; Vozeh et al., 1996; Yuh et al., 1994).

In the drug development process, the use of population approaches has been recommended by the Food and Drug Administration, in order to obtain a reliable assessment of intra- and inter-individual variabilities (Center for Drug Evaluation and Research, 1999). However, the use of such models is not restricted to pharmacology but is being extended to data analysis problems arising in several contexts ranging from medical imaging (Bertoldo, Sparacino, & Cobelli, 2004) and diagnosis of metabolic disorders (Vicini & Cobelli, 2001) to genomics (Ferrazzi, Magni, & Bellazzi, 2003).

Population methods can be divided into three main branches: parametric, semiparametric and nonparametric. In the parametric approach, a structural model is assumed, e.g. a compartmental one, and the model parameters are regarded as random variables extracted from a distribution representative of the given population (Beal & Sheiner, 1998; Jelliffe, Schumitzky, Van Guilder, Wang, & Leary, 2001; Leary, Jelliffe, Schumitzky, & Van Guilder, 2001; Wakefield & Bennett, 1996; Wakefield, Smith, Racine-Poon, & Gelfand, 1994) (note that the term "nonparametric" in the papers by Jelliffe et al. (2001)

☆ This paper was presented at the 16th IFAC World Congress, Prague, Czech Republic, 2005. This paper was recommended for publication in revised form by Associate Editor Kenko Uchida under the direction of Editor Ian Petersen.

* Corresponding author. Tel.: +39 0382 985484; fax: +39 0382 985373.
E-mail addresses: marta.neve@unipv.it (M. Neve), giuseppe.denicolao@unipv.it (G. De Nicolao).

0005-1098/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2006.12.024

and Leary et al. (2001) refers to the estimation of the probability distributions of the parameters of a grey-box model).

In other cases, for instance in the preliminary phases of a study, a structural model is not available and semiparametric or nonparametric techniques must be used. In the semiparametric approach, the response curves are modelled as regression splines (Fattinger & Verotta, 1995a, 1995b; Park, Verotta, Blaschke, & Sheiner, 1997), so that the non-trivial problem of deciding the number and the location of the spline knots arises.

Recently, in order to develop a completely nonparametric approach, the individual curves have been modelled as discrete-time stochastic processes (e.g. random walks), reformulating the problem within the framework of Bayesian estimation (Magni, Bellazzi, De Nicolao, Poggesi, & Rocchetti, 2002). This kind of model has also been used for the analysis of gene expression time series measured using DNA micro-arrays (Ferrazzi et al., 2003). Since the sampling schedules are usually not uniformly spaced in time, it would be more convenient to model the individual curves as continuous-time stochastic processes. In this paper we develop such a continuous-time population model. More precisely, assuming that the average impulse response of the population is an integrated Wiener process, it is shown that its Bayes estimate is a cubic spline. Explicit formulas are worked out also for the estimates of the individual responses. This estimation approach extends the Gaussian processes methodology for the reconstruction of continuous functions given discrete and noisy samples (MacKay, 1998; Smola & Schölkopf, 2003; Williams, 1999) to the case of population models. Remarkably, the overall estimator can be interpreted as a kind of Regularization Network (Poggio & Girosi, 1990) whose weights are the solution of a system of linear equations.

In the last few years there has been a growing interest in smoothing splines within the control community, especially for what concerns their interpretation in an optimal control theoretic context (Egerstedt & Martin, 2001; Sun, Egerstedt, & Martin, 2000). In the present paper, conversely, smoothing splines arise as the solution of an optimal mean-square estimation problem. The method is tested on simulated data sets as well as on pharmacokinetic data related to xenobiotics administration in human subjects.

2. Stochastic population model

Consider the problem of estimating a family of scalar real-valued continuous-time functions $z^j(t)$, $j = 1, \dots, N$, $t \geqslant 0$, on the basis of noisy samples taken at discrete instants. More precisely, assume that the following measurements are available

$$y_k^j = z^j(t_k^j) + v_k^j, \quad k = 1, \dots, n_j, \tag{1}$$

where $t_k^j > 0$ denotes the kth sampling instant ("knot") for the jth curve, and the measurement errors v_k^j are mutually independent and normally distributed with $E[v_k^j] = 0$, $\mathrm{Var}[v_k^j] = (\sigma_k^j)^2$. In an experimental setting, the jth curve $z^j(t)$ will be representative of the jth subject (e.g. an impulse response obtained as a drug concentration profile in plasma after administration of a unit bolus). Note that the number and location of the sampling instants t_k^j may vary from subject to subject. Hereafter, each individual curve will be decomposed as

$$z^j(t) = \bar{z}(t) + \tilde{z}^j(t),$$

where $\bar{z}(t)$ is the "average curve" of the population and $\tilde{z}(t)$ is the "individual shift" with respect to the average behaviour. For ease of notation, the observations will be grouped as follows

$$\mathbf{y} := [y_1^1 \cdots y_{n_1}^1 \;\; y_1^2 \cdots y_{n_2}^2 \cdots y_1^N \cdots y_{n_N}^N]^{\mathrm{T}}.$$

Letting $n = n_1 + n_2 + \cdots + n_N$ be the total number of observations, $\mathbf{y}$ is an n-dimensional column vector. In a similar way, it is possible to define

$$\bar{\mathbf{z}} := [\bar{z}(t_1^1) \dots \bar{z}(t_{n_1}^1) \;\; \dots \bar{z}(t_1^N) \dots \bar{z}(t_{n_N}^N)]^{\mathrm{T}},$$

$$\tilde{\mathbf{z}} := [\tilde{z}^1(t_1^1) \dots \tilde{z}^1(t_{n_1}^1) \;\; \dots \tilde{z}^N(t_1^N) \dots \tilde{z}^N(t_{n_N}^N)]^{\mathrm{T}},$$

$$\mathbf{v} := [v_1^1 \dots v_{n_1}^1 \;\; v_1^2 \dots v_{n_2}^2 \dots v_1^N \dots v_{n_N}^N]^{\mathrm{T}}.$$

Therefore, in vector notation, (1) can be rewritten as

$$\mathbf{y} = \bar{\mathbf{z}} + \tilde{\mathbf{z}} + \mathbf{v},$$

where $\mathbf{v} \sim N(0, \mathbf{\Sigma}_v)$, $\mathbf{\Sigma}_v := \mathrm{diag}\{(\sigma_1^1)^2 \dots (\sigma_{n_N}^N)^2\}$, $\mathbf{\Sigma}_v > 0$.

2.1. Average and individual curves

In the present paper, a stochastic approach is adopted: the unknown functions are modelled as stochastic processes and the aim is to compute their posterior distributions given the observed data (note that the data are processed off-line, so that there is no need for the estimator to satisfy causality constraints).

Assumption 1. *The Gaussian stochastic processes $\bar{z}(t)$ and $\tilde{z}^j(t)$, $j = 1, \dots, N$, are independent of each other and of the noise vector* $\mathbf{v}$.

In the following, $\bar{R}(t, \tau) := \mathrm{Cov}[\bar{z}(t), \bar{z}(\tau)]$ and $\tilde{R}(t, \tau) := \mathrm{Cov}[\tilde{z}^j(t), \tilde{z}^j(\tau)]$, $\forall j$,will denote the auto-covariance functions of the average curve and the individual shifts, respectively. Hereafter, it will be assumed that both $\bar{R}(t, \tau)$ and $\tilde{R}(t, \tau)$ are positive definite operators. Recalling that $\tilde{z}^j(t)$ is a shift with respect to the average response, it is reasonable to assume that $E[\tilde{z}^j(t)] = 0$, $\forall t$, $\forall j$. As for $\bar{z}(t)$, by properly scaling the data, it can be assumed without loss of generality that $E[\bar{z}(t)] = 0$.

Remark 1. The nonparametric approach developed in the present paper is not intended to be an alternative but rather a complement to standard parametric population models currently used in pharmacokinetics. Nonparametric models may be particularly useful when a reliable structural model is not available. This may happen in the early stages of a study, in which case nonparametric modelling may help evaluating the exposure and also checking for misspecification of candidate parametric models. When comparing nonparametric and parametric models, one should be aware that (with the exception of

linear-in-parameter models) the average curve is different from the so-called typical curve, obtained by plugging the average population parameters into the parametric model. Although the typical curve may be preferred for its physiological insight, it depends on the adopted model parametrization. To make an example, referring to poles rather than to time constants would yield different typical curves. Moreover, in absence of a parametric structural model, the average curve, which is uniquely defined in all circumstances, can still be used to characterize the average behaviour of the population. There are also some caveats regarding the average curve. For instance, a multiexponential response may arise as the average of single exponential responses. The average curve may also be misleading when the population is a mixture of subpopulations, e.g. normal and pathological subjects. To avoid these pitfalls one should never trust average features without checking the population distribution.

Since all the involved processes are jointly Gaussian, the posterior distributions are Gaussian as well. The following results provide the point estimates and the confidence intervals for the average curve and the individual ones. In the next proposition and thereafter, Var[**y**] will denote the covariance matrix of the random vector **y**.

Proposition 1.

$$\hat{\bar{z}}(t) := E[\bar{z}(t)|\mathbf{y}] = \sum_{j=1}^{N}\sum_{k=1}^{n_j} c_k^j \bar{R}(t, t_k^j), \tag{2}$$

$$\hat{z}^j(t) := E[z^j(t)|\mathbf{y}] = \hat{\bar{z}}(t) + \sum_{k=1}^{n_j} c_k^j \tilde{R}(t, t_k^j), \tag{3}$$

$$\mathbf{c} = \boldsymbol{\Sigma}_y^{-1}\mathbf{y}, \tag{4}$$

$$\mathbf{c} = [c_1^1 \;\; c_2^1 \dots c_{n_1}^1 \dots c_1^N \dots c_{n_N}^N]^{\mathrm{T}},$$

$$\boldsymbol{\Sigma}_y := \mathrm{Var}[\mathbf{y}] = \mathrm{Var}[\bar{\mathbf{z}}] + \mathrm{Var}[\tilde{\mathbf{z}}] + \boldsymbol{\Sigma}_v,$$

$$\mathrm{Var}[\bar{\mathbf{z}}] = \bar{\mathbf{R}} := \begin{bmatrix} \bar{R}(t_1^1, t_1^1) & \cdots & \bar{R}(t_1^1, t_{n_N}^N) \\ \cdots & \cdots & \cdots \\ \bar{R}(t_{n_N}^N, t_1^1) & \cdots & \bar{R}(t_{n_N}^N, t_{n_N}^N) \end{bmatrix},$$

$$\mathrm{Var}[\tilde{\mathbf{z}}] = \tilde{\mathbf{R}} := \mathrm{blockdiag}\{\tilde{\mathbf{R}}^1, \dots, \tilde{\mathbf{R}}^N\},$$

$$\tilde{\mathbf{R}}^j := \begin{bmatrix} \tilde{R}(t_1^j, t_1^j) & \cdots & \tilde{R}(t_1^j, t_{n_j}^j) \\ \cdots & \cdots & \cdots \\ \tilde{R}(t_{nj}^j, t_1^j) & \cdots & \tilde{R}(t_{n_j}^j, t_{n_j}^j) \end{bmatrix}.$$

Proof. According to a well-known formula for jointly Gaussian random variables, see, e.g. Shiryaev (1996),

$$E[\bar{z}(t)|\mathbf{y}] = E[\bar{z}(t)] + \mathrm{Cov}[\bar{z}(t), \mathbf{y}]\,\mathrm{Var}[\mathbf{y}]^{-1}(\mathbf{y} - E[\mathbf{y}]).$$

Under the given assumptions, $E[\bar{z}(t)] = 0$, $E[\mathbf{y}] = 0$ and

$$\mathrm{Cov}[\bar{z}(t), \mathbf{y}] = \mathrm{Cov}[\bar{z}(t), \bar{\mathbf{z}} + \tilde{\mathbf{z}} + \mathbf{v}] = \mathrm{Cov}[\bar{z}(t), \bar{\mathbf{z}}]$$
$$= [\bar{R}(t, t_1^1) \dots \bar{R}(t, t_{n_N}^N)].$$

Concerning $\tilde{z}(t)$, a completely analogous derivation yields

$$E[\tilde{z}^j(t)|\mathbf{y}] = \mathrm{Cov}[\tilde{z}^j(t), \mathbf{y}]\,\mathrm{Var}[\mathbf{y}]^{-1}\mathbf{y}.$$

Observing that $E[z^j(t)|\mathbf{y}] = E[\bar{z}(t)|\mathbf{y}] + E[\tilde{z}^j(t)|\mathbf{y}]$ and that $E[\tilde{z}^j(t)|y_k^i] = 0$, $\forall i \neq j$, Eq. (3) is obtained. Finally, the expressions for $\boldsymbol{\Sigma}_y$, Var[$\bar{\mathbf{z}}$] and Var[$\tilde{\mathbf{z}}$] follow directly from the assumptions. □

Proposition 2.

$$\mathrm{Var}[\bar{z}(t)|\mathbf{y}] = \bar{R}(t, t) - \bar{\mathbf{r}}\boldsymbol{\Sigma}_y^{-1}\bar{\mathbf{r}}^{\mathrm{T}},$$

$$\bar{\mathbf{r}} := [\bar{R}(t, t_1^1) \dots \bar{R}(t, t_{n_N}^N)],$$

$$\mathrm{Var}[z^j(t)|\mathbf{y}] = \bar{R}(t, t) + \tilde{R}^j(t, t) - (\bar{\mathbf{r}} + \tilde{\mathbf{r}}^j)\boldsymbol{\Sigma}_y^{-1}(\bar{\mathbf{r}} + \tilde{\mathbf{r}}^j)^{\mathrm{T}},$$

$$\tilde{\mathbf{r}}^j := \mathrm{Cov}[\tilde{z}^j(t), \tilde{\mathbf{z}}].$$

Proof. By a well-known formula (Shiryaev, 1996)

$$\mathrm{Var}[\bar{z}(t)|\mathbf{y}] = \mathrm{Var}[\bar{z}(t)] - \mathrm{Cov}[\bar{z}(t), \mathbf{y}]\,\mathrm{Var}[\mathbf{y}]^{-1}\,\mathrm{Cov}[\bar{z}(t), \mathbf{y}]^{\mathrm{T}}.$$

Recalling that $\mathbf{y} = \bar{\mathbf{z}} + \tilde{\mathbf{z}} + \mathbf{v}$ and in view of the independency assumptions, the expression for Var[$\bar{z}(t)|\mathbf{y}$] immediately follows. Analogous considerations hold for Var[$z^j(t)|\mathbf{y}$]. □

2.2. Regularization network interpretation

It is interesting to note from (2) and (3) that the estimates $\hat{\bar{z}}(t)$ and $\hat{z}^j(t)$ are obtained as linear combinations of the functions $\bar{R}(t, t_k^j)$, $\tilde{R}(t, t_k^j)$. This is the typical structure that comes out in the Bayesian estimation of Gaussian processes (Girosi, Jones, & Poggio, 1995; Poggio & Girosi, 1990; Wahba, 1990; Williams & Rasmussen, 1996). Remarkably, the same estimator can also be obtained via Tychonov regularization theory (Girosi et al., 1995; Poggio & Girosi, 1990). This explains why Poggio and Girosi (1990) have introduced the term *regularization network* (RN) to denote such estimators, pointing out their neural network-like structure. Also the estimator of Proposition 1 can be regarded as an RN, although of a special type. Having to do with the identification of a population model, the number of neurons is twice the number n of the data instead of n as in the standard RN, see Fig. 1. A first set of n neurons receive t as input and have $\bar{R}(t, t_k^j)$ as activation function. The estimate $\hat{\bar{z}}(t)$ of the average curve is obtained by linearly combining these outputs through the weights c_k^j. A second set of n neurons, having $\tilde{R}(t, t_k^j)$ as activation functions produce outputs that, combined again through the weights c_k^j, yield the estimates of the individual shifts $\hat{\tilde{z}}^j(t)$. The weight vector **c** is obtained as the solution of a system of n linear equations, see (4). This is an advantage with respect to other kinds of networks, such as multi-layer perceptrons, in which the weights have to be computed using iterative nonlinear optimization (MacKay, 1997).

3. Population splines and hyper-parameters estimation

For the results of the previous section to be of practical use it is necessary to specify the statistics of the stochastic processes

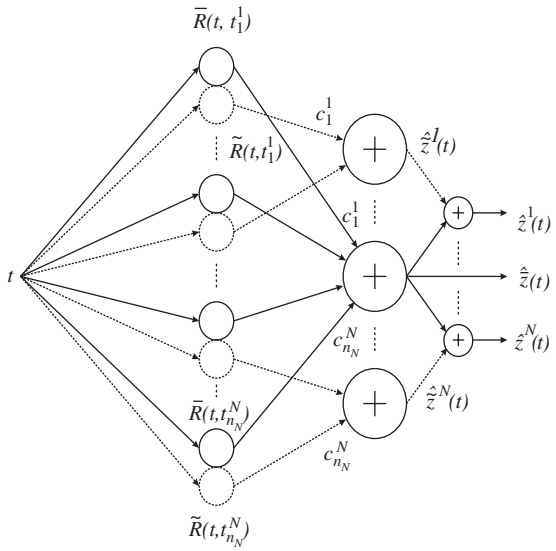

Fig. 1. Regularization network structure of the estimator.

$\bar{z}(t)$, $\tilde{z}^j(t)$. If frequently sampled observations were available, their statistics could be identified by black-box parametric identification methods. On the other hand, population studies are often characterized by the scarcity of samples per subject. Therefore, it is necessary to introduce signal models that reflect the available a-priori knowledge.

3.1. Modelling the average curve

If it is only known that a signal is "smooth", it is a common practice to model it as an integrated Wiener process as done below.

Assumption 2.

$$\dot{\bar{x}}(t) = \bar{\mathbf{A}}\bar{x}(t) + \bar{\mathbf{B}}\bar{w}(t),$$

$$\bar{z}(t) = \bar{\mathbf{C}}\bar{x}(t),$$

$$\bar{\mathbf{A}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \bar{\mathbf{B}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \bar{\mathbf{C}} = [1 \;\; 0],$$

where $\bar{x}(0) \sim N(0, \bar{\mathbf{X}}_0)$, *and* $\bar{w}(t)$ *is a scalar continuous-time white Gaussian noise, independent of* $\bar{x}(0)$ *and the measurement error vector* $\mathbf{v}$, *with* $E[\bar{w}(t)\bar{w}(\tau)] = \bar{\lambda}^2\delta(t-\tau)$.

The model above can describe signals whose initial conditions are deterministically known by setting $\bar{\mathbf{X}}_0 = 0$. The case of completely unknown initial conditions, corresponding to $\bar{\mathbf{X}}_0^{-1} = 0$, will be discussed in Section 4. The parameter $\bar{\lambda}^2$ affects the regularity of the realizations (smaller values correspond to smoother signals). The a-priori knowledge is seldom sufficient to specify $\bar{\lambda}^2$ so that it must be regarded as a "hyperparameter" that will have to be estimated from the data, see Section 3.3. In Assumption 2, an unstable model with two poles in the origin is postulated for the average curve. The advantage of this particular model is that the associated Bayesian estimator can reconstruct linear functions without bias (provided that the initial state has infinite variance). In fact, the next result shows that the Bayes estimate is a cubic spline.

Theorem 1. *Under Assumption* 2, $\hat{\bar{z}}(t)$ *defined in Proposition* 1 *is a cubic spline with knots located in the sampling instants* $\{t_1^1,\ t_2^1, \ldots, t_{n_N}^N\}$.

Proof. It is well known that $\bar{\mathbf{X}}(t) := \mathrm{Var}[\bar{x}(t)]$ is the solution of the differential Lyapunov equation

$$\dot{\bar{\mathbf{X}}}(t) = \bar{\mathbf{A}}\bar{\mathbf{X}}(t) + \bar{\mathbf{X}}(t)\bar{\mathbf{A}}^{\mathrm{T}} + \bar{\lambda}^2\bar{\mathbf{B}}\bar{\mathbf{B}}^{\mathrm{T}},$$

$$\bar{\mathbf{X}}(0) = \bar{\mathbf{X}}_0.$$

Moreover,

$$\bar{R}(t,\tau) = \begin{cases} \bar{\mathbf{C}}\bar{\mathbf{X}}(t)\mathrm{e}^{\bar{\mathbf{A}}^{\mathrm{T}}(\tau-t)}\bar{\mathbf{C}}^{\mathrm{T}}, & t \leqslant \tau, \\ \bar{\mathbf{C}}\mathrm{e}^{\bar{\mathbf{A}}(t-\tau)}\bar{\mathbf{X}}(\tau)\bar{\mathbf{C}}^{\mathrm{T}}, & t > \tau. \end{cases}$$

In view of the definition of $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, $\bar{\mathbf{C}}$, it follows that $\bar{R}(t,\tau)$, seen as a function of t, is a piecewise cubic polynomial. In particular if $\bar{\mathbf{X}}_0 = 0$, the auto-covariance is

$$\bar{R}(t,\tau) = \bar{\lambda}^2 \begin{cases} \dfrac{t^2}{2}\left(\tau - \dfrac{t}{3}\right), & t \leqslant \tau, \\ \dfrac{\tau^2}{2}\left(t - \dfrac{\tau}{3}\right), & t > \tau. \end{cases}$$

Note that $\bar{R}(t,\tau)$ is continuous with all its derivatives everywhere but in $t = \tau$ where it is continuous up to the second derivative. Recalling that $\hat{\bar{z}}(t)$ in (2) is a linear combination of the functions $\bar{R}(t, t_k^j)$ (Proposition 1), the thesis immediately follows. □

In the literature, it is known that the conditional expectation of an integrated Wiener process given discrete observations is a cubic smoothing spline (Wahba, 1990). In some sense, Theorem 1 generalizes such a result to the analysis of a population of signals so that it is natural to define $\hat{\bar{z}}(t)$ a *population smoothing spline*.

3.2. Modelling the individual curves

Concerning the model for the individual shifts $\tilde{z}^j(t)$, the following assumption is in order.

Assumption 3. *For* $j = 1, \ldots, N$,

$$\dot{\tilde{x}}(t) = \tilde{\mathbf{A}}\tilde{x}(t) + \tilde{\mathbf{B}}\tilde{w}^j(t),$$

$$\tilde{z}^j(t) = \tilde{\mathbf{C}}\tilde{x}(t),$$

$$\tilde{\mathbf{A}} = \begin{bmatrix} a_1 & 1 \\ 0 & a_2 \end{bmatrix}, \quad \tilde{\mathbf{B}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \tilde{\mathbf{C}} = [1 \;\; 0],$$

where $a_1 < 0$, $a_2 < 0$, *and* $\tilde{x}(0) \sim N(0, \tilde{\mathbf{X}}_0)$, *and* $\tilde{w}^j(t)$ *is a scalar continuous-time white Gaussian noise* (*independent of* $\mathbf{v}$, $\bar{w}(t)$ *and* $\tilde{w}^i(t)$, $i \neq j$) *with* $E[\tilde{w}(t)\tilde{w}(\tau)] = \tilde{\lambda}^2\delta(t-\tau)$.

The above model encompasses all Gaussian processes obtained by passing a white Gaussian noise through a transfer function with relative degree two, so that the continuity of the process and its first derivative is guaranteed (this expresses the prior knowledge that the sample paths are "smooth"). The statistics of $\tilde{z}(t)$ will depend on the three parameters $a_1, a_2, \tilde{\lambda}^2$. For $\tilde{\lambda}^2$ the same considerations as for $\bar{\lambda}^2$ hold. The two poles a_1 and a_2 provide two more degrees of freedom for shaping the auto-covariance of $\tilde{z}^j(t)$. A possible drawback may be the difficulty in estimating two more hyper-parameters from the data. In this respect, a simpler model describes also the individual shifts as integrated Wiener processes ($a_1 = 0,\ a_2 = 0$). However, observe that the measurements can be rewritten as

$$y_k^j = \bar{z}(t_k) + \bar{v}_k^j,$$

where $\bar{v}_k^j := \tilde{z}^j(t_k) + v_k^j$. In other words, as far as the estimation of $\bar{z}(t)$ is concerned, $\bar{v}_k^j$ acts as measurement noise. If $\tilde{z}^j(t)$ were an integrated Wiener process, its variance would tend to infinity with t, and the confidence intervals for $\bar{z}(t)$ would diverge as t grows. A notable exception occurs when some a-priori knowledge on the asymptotic value of the average curve $\bar{z}(t)$ is available, in which case an integrated Wiener model for the individual shifts can do as well (this is further discussed in Section 5).

In view of Assumption 3, the calculation of $\tilde{R}(t,\tau)$ is completely analogous to that of $\bar{R}(t,\tau)$ described in the proof of Theorem 1. In the particular case $a_1 = a_2 = a$ and $\tilde{\mathbf{X}}_0 = 0$, the auto-covariance is

$$\tilde{R}(t,\tau) = \tilde{\lambda}^2 \begin{cases} \mathrm{e}^{a(\tau-t)}\left(\dfrac{t^2\mathrm{e}^{2at}}{2a} - \dfrac{t\mathrm{e}^{2at}}{2a^2} - \dfrac{\mathrm{e}^{2at}-1}{4a^3}\right) \\ \quad +(\tau-t)\mathrm{e}^{a(\tau-t)}\dfrac{2at\mathrm{e}^{2at} - \mathrm{e}^{2at} + 1}{4a^2}, \quad t \leqslant \tau, \\ \mathrm{e}^{a(t-\tau)}\left(\dfrac{\tau^2\mathrm{e}^{2a\tau}}{2a} - \dfrac{\tau\mathrm{e}^{2a\tau}}{2a^2} - \dfrac{\mathrm{e}^{2a\tau}-1}{4a^3}\right) \\ \quad +(t-\tau)\mathrm{e}^{a(t-\tau)}\dfrac{2a\tau\mathrm{e}^{2a\tau} - \mathrm{e}^{2a\tau} + 1}{4a^2}, \quad t > \tau. \end{cases}$$

3.3. Estimating the hyper-parameters

When one is faced with a Bayesian estimation problem involving unknown hyper-parameters, a simple, yet effective, approach is to resort to the so-called empirical Bayes method (MacKay, 1992). In the first step, a maximum likelihood (ML) estimate of the hyper-parameters is computed. Then, the Bayes estimate is calculated as if the hyper-parameters were deterministically known and equal to their ML estimates. In the problem at hand, this leads to the following estimation algorithm, where $\theta = [\bar{\lambda}^2, \tilde{\lambda}^2, a_1, a_2]$ denotes the hyper-parameters vector.

Algorithm.

1. Let $\theta_{\mathrm{ML}} := \arg\min_\theta\{\ln(\det(\boldsymbol{\Sigma}_y)) + \mathbf{y}^{\mathrm{T}}\boldsymbol{\Sigma}_y^{-1}\mathbf{y}\}$.
2. Let $[\bar{\lambda}^2, \tilde{\lambda}^2, a_1, a_2]^{\mathrm{T}} = \theta_{\mathrm{ML}}$

and compute $\hat{\bar{z}}(t)$ and $\hat{z}^j(t),\ j = 1,\ldots,N$ according to Proposition 1.

If also the individual shifts are modelled as integrated Wiener processes, the only hyper-parameters will be $\bar{\lambda}^2$ and $\tilde{\lambda}^2$. In the present paper, it has been assumed that the measurement error variances $(\sigma_k^j)^2$ are known. If an error model is postulated, e.g. a constant coefficient of variation one, its parameters may well be regarded as hyper-parameters and estimated via likelihood maximization, although more data will be needed to obtain reliable estimates.

4. Completely unknown initial conditions

It is important to be able to estimate average curves whose initial conditions in $t = 0$ are completely unknown. As already mentioned, this would correspond to $\bar{\mathbf{X}}_0^{-1} = 0$. A practical approach is to let $\bar{\mathbf{X}}_0^{-1} = \varepsilon\mathbf{I}$ where ε is a small enough scalar, but this is far from being numerically robust. The rigorous approach calls for the derivation of specific formulas as done in the following. Taking into account average curves whose initial conditions have infinite variance is equivalent to considering a population of the type

$$z^j(t) = \bar{z}^*(t) + \tilde{z}^j(t), \tag{5}$$

$$\bar{z}^*(t) := \boldsymbol{\phi}^{\mathrm{T}}(t)\boldsymbol{\zeta} + \bar{z}(t), \tag{6}$$

where $\bar{z}(t)$ and $\tilde{z}^j(t)$ have finite auto-covariances, $\boldsymbol{\phi}(t) : \mathbb{R}^1 \mapsto \mathbb{R}^{1\times M}$ is a deterministic vector function, and $\boldsymbol{\zeta} \sim N(0, \rho^2\mathbf{I})$, $\rho^2 = \infty$. For instance, with reference to the integrated Wiener process of Assumption 2, letting $\bar{\mathbf{X}}_0 = \rho^2\mathbf{I}$, $\rho^2 = \infty$ would yield $\boldsymbol{\phi}^{\mathrm{T}}(t) = [1\ \ t]$. In other words, handling completely unknown initial conditions amounts to estimating additional parameters with infinite prior variance. In the following, it will be assumed that the measurements are as in (1) and that the $n \times M$ matrix

$$\boldsymbol{\Phi} := [\boldsymbol{\phi}(t_1^1)\ldots\boldsymbol{\phi}(t_{n_1}^1)\ldots\boldsymbol{\phi}(t_1^N)\ldots\boldsymbol{\phi}(t_{n_N}^N)]^{\mathrm{T}}$$

is full column rank.

Proposition 3. *For the model* (5)–(6),

$$\hat{\bar{z}}^*(t) := E[\bar{z}^*(t)|\mathbf{y}] = \sum_{j=1}^{N}\sum_{k=1}^{n_j} c_k^j \bar{R}(t, t_k^j) + \boldsymbol{\phi}^{\mathrm{T}}(t)\mathbf{d},$$

$$\hat{z}^j(t) := E[z^j(t)|\mathbf{y}] = \hat{\bar{z}}^*(t) + \sum_{k=1}^{n_j} c_k^j \tilde{R}(t, t_k^j),$$

$$\mathbf{d} = (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{y},$$

$$\mathbf{c} = \mathbf{M}^{-1}(\mathbf{y} - \boldsymbol{\Phi}\mathbf{d}),$$

$$\mathbf{M} := \bar{\mathbf{R}} + \tilde{\mathbf{R}} + \boldsymbol{\Sigma}_v.$$

Proof. Mutatis mutandis, the proof is completely analogous to that of Theorem 1.5.3 in Wahba (1990) and is therefore omitted.

In the set of sampling instants t_k^j, $j=1,\dots,N$; $k=1,\dots,n_j$, there may be repeated elements as more than one individual curve can be measured at the same time. For the subsequent derivations it is useful to introduce the "minimal set" (i.e. without repetitions) of sampling instants $\{\tau_i\}$, $i=1,\dots,\bar{n}$, where $\tau_{i_1}\neq\tau_{i_2}$, $\forall i_1\neq i_2$, and τ_i is such that there exist j and k such that $\tau_i=t_k^j$. Moreover, let $\mathbf{\Lambda}$ be a matrix whose entries are either 0 or 1 such that

$$[t_1^1\dots t_{n_1}^1\dots t_1^N\dots t_{n_N}^N]^{\mathrm{T}}:=\mathbf{\Lambda}[\tau_1\dots\tau_{\bar{n}}]^{\mathrm{T}}.$$

The next two results provide the posterior variance, and hence the confidence intervals, of the average and individual curves, respectively.

Proposition 4. *For the model* (5)–(6),

$$\mathrm{Var}[\bar{z}^*(t)|\mathbf{y}]=\bar{R}_u(t)+\bar{R}_o(t),$$

$$\bar{R}_u(t)=\bar{R}(t,t)-\bar{\bar{\mathbf{r}}}\bar{\bar{\mathbf{R}}}^{-1}\bar{\bar{\mathbf{r}}}^{\mathrm{T}},$$

$$\bar{R}_o(t)=\bar{\mathbf{L}}\bar{\mathbf{\Sigma}}_\eta\bar{\mathbf{L}}^{\mathrm{T}},$$

$$\bar{\mathbf{\Sigma}}_\eta:=(\bar{\mathbf{F}}^{\mathrm{T}}(\tilde{\mathbf{R}}+\mathbf{\Sigma}_v)^{-1}\bar{\mathbf{F}}+\bar{\mathbf{J}})^{-1},$$

$$\bar{\mathbf{J}}:=\begin{bmatrix}0 & 0\\ 0 & \bar{\bar{\mathbf{R}}}^{-1}\end{bmatrix},$$

$$\bar{\mathbf{F}}:=[\mathbf{\Phi}\ \ \mathbf{\Lambda}],$$

$$\bar{\mathbf{L}}:=[\boldsymbol{\phi}^{\mathrm{T}}(t)\ \ \bar{\bar{\mathbf{r}}}\bar{\bar{\mathbf{R}}}^{-1}],$$

$$\bar{\bar{\mathbf{r}}}:=[\bar{R}(t,\tau_1)\dots\bar{R}(t,\tau_{\bar{n}})],$$

$$\bar{\bar{\mathbf{R}}}:=\begin{bmatrix}\bar{R}(\tau_1,\tau_1) & \cdots & \bar{R}(\tau_1,\tau_{\bar{n}})\\ \cdots & \cdots & \cdots\\ \bar{R}(\tau_{\bar{n}},\tau_1) & \cdots & \bar{R}(\tau_{\bar{n}},\tau_{\bar{n}})\end{bmatrix}.$$

Proof. First of all, observe that the positive definiteness of the operator $\bar{R}(t,\tau)$ implies the invertibility of $\bar{\bar{\mathbf{R}}}$. As for the existence of $\bar{\mathbf{\Sigma}}_\eta$, assume by contradiction that there exists $\mathbf{x}\neq 0$ such that $\mathbf{x}^{\mathrm{T}}(\bar{\mathbf{F}}^{\mathrm{T}}(\tilde{\mathbf{R}}+\mathbf{\Sigma}_v)^{-1}\bar{\mathbf{F}}+\bar{\mathbf{J}})\mathbf{x}=0$. Let $\mathbf{x}$ be partitioned as $\mathbf{x}=[\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{z}^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{\zeta}\in\mathbb{R}^{M\times 1}$. This implies $\bar{\bar{\mathbf{R}}}^{-1}\mathbf{z}=0$ and $\mathbf{\Phi}\boldsymbol{\zeta}+\mathbf{\Lambda}\mathbf{z}=0$, that is $\mathbf{z}=0$ and $\mathbf{\Phi}\boldsymbol{\zeta}=0$, $\boldsymbol{\zeta}\neq 0$, which contradicts the full-rank assumption made on $\mathbf{\Phi}$. In order to apply Lemma 1 (in the Appendix), let $z^*:=\bar{z}^*(t)$ and observe that

$$z^*=\boldsymbol{\phi}^{\mathrm{T}}(t)\boldsymbol{\zeta}+\bar{z}(t),$$

$$\mathbf{y}=\bar{\mathbf{F}}\bar{\boldsymbol{\eta}}+\boldsymbol{\epsilon},$$

$$\boldsymbol{\epsilon}\sim N(0,\mathbf{\Sigma}_\varepsilon),\quad \mathbf{\Sigma}_\varepsilon=\tilde{\mathbf{R}}+\mathbf{\Sigma}_v,$$

$$\bar{\boldsymbol{\eta}}:=[\boldsymbol{\zeta}^{\mathrm{T}}\ \ \bar{\bar{\mathbf{z}}}^{\mathrm{T}}]^{\mathrm{T}},$$

$$\bar{\bar{\mathbf{z}}}:=[\bar{z}(\tau_1)\dots\bar{z}(\tau_{\bar{n}})]^{\mathrm{T}}.$$

Moreover,

$$\mathbf{\Gamma}:=\mathrm{Cov}[z^*,\bar{\boldsymbol{\eta}}]=[\boldsymbol{\phi}^{\mathrm{T}}(t)\,\mathrm{Var}[\boldsymbol{\zeta}]\ \ \bar{\bar{\mathbf{r}}}],$$

$$\mathbf{V}:=\mathrm{Var}[\bar{\boldsymbol{\eta}}]=\begin{bmatrix}\mathrm{Var}[\boldsymbol{\zeta}] & 0\\ 0 & \bar{\bar{\mathbf{R}}}\end{bmatrix},$$

$$\mathbf{\Gamma}\mathbf{V}^{-1}=[\boldsymbol{\phi}^{\mathrm{T}}(t)\ \ \bar{\bar{\mathbf{r}}}\bar{\bar{\mathbf{R}}}^{-1}]=\bar{\mathbf{L}}.$$

Recalling that $\rho^2=\infty$, it is easy to see that

$$\bar{\mathbf{\Sigma}}_\eta=(\bar{\mathbf{F}}^{\mathrm{T}}\mathbf{\Sigma}_\varepsilon^{-1}\bar{\mathbf{F}}+\mathbf{V}^{-1})^{-1}.$$

Then,

$$\begin{aligned}\mathrm{Var}[z^*|\bar{\boldsymbol{\eta}}]&=\mathrm{Var}[z^*]-\mathrm{Cov}[z^*,\bar{\boldsymbol{\eta}}]\,\mathrm{Var}[\bar{\boldsymbol{\eta}}]^{-1}\,\mathrm{Cov}[z^*,\bar{\boldsymbol{\eta}}]^{\mathrm{T}}\\ &=\boldsymbol{\phi}^{\mathrm{T}}(t)\,\mathrm{Var}[\boldsymbol{\zeta}]\boldsymbol{\phi}(t)+\bar{R}(t,t)-\mathbf{\Gamma}\mathbf{V}^{-1}\mathbf{\Gamma}^{\mathrm{T}}\\ &=\bar{R}(t,t)-\bar{\bar{\mathbf{r}}}\bar{\bar{\mathbf{R}}}^{-1}\bar{\bar{\mathbf{r}}}^{\mathrm{T}}.\end{aligned}$$

Finally, the thesis follows straightforwardly from the application of Lemma 1. □

Note that the problem of evaluating the posterior variance is similar to that considered in Wahba (1983). The approach followed herein relies on the decomposition of the noiseless part of the observations as the sum of a contribution $\mathbf{\Phi}\boldsymbol{\zeta}$ due to the (infinite variance) initial conditions and another contribution $\mathit{\Lambda}\bar{\bar{\mathbf{z}}}$ having finite variance. This expedient, together with Lemma 1, helps obtaining a compact expression for the posterior variance.

Proposition 5. *For the model* (5)–(6),

$$\mathrm{Var}[z^j(t)|\mathbf{y}]=R_u^j(t)+R_o^j(t),$$

$$R_u^j(t)=\bar{R}_u(t)+\tilde{R}(t,t)-\tilde{\mathbf{r}}^j\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{r}}^{jT},$$

$$R_o^j(t)=\mathbf{L}^j\mathbf{\Sigma}_\eta\mathbf{L}^{j\,\mathrm{T}},$$

$$\mathbf{\Sigma}_\eta:=(\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}_v^{-1}\mathbf{F}+\mathbf{J})^{-1},$$

$$\mathbf{J}:=\begin{bmatrix}0 & 0\\ 0 & (\bar{\mathbf{R}}+\tilde{\mathbf{R}})^{-1}\end{bmatrix},$$

$$\mathbf{F}:=[\mathbf{\Phi}\ \ \mathbf{I}],$$

$$\mathbf{L}^j:=[\boldsymbol{\phi}^{\mathrm{T}}(t)\ \ (\bar{\mathbf{r}}+\tilde{\mathbf{r}}^j)(\bar{\mathbf{R}}+\tilde{\mathbf{R}})^{-1}].$$

Proof. The invertibility of $\tilde{\mathbf{R}}$ follows from the positive definiteness of the operator $\tilde{R}(t,\tau)$. As for the invertibility of $\mathbf{\Sigma}_\eta$, it can be proved in the same way as the invertibility of $\bar{\mathbf{\Sigma}}_\eta$, demonstrated in the proof of Proposition 4. In order to apply Lemma 1, let $z^*:=z^j(t)$ and observe that

$$z^*=\boldsymbol{\phi}^{\mathrm{T}}(t)\boldsymbol{\zeta}+\bar{z}(t)+\tilde{z}^j(t),$$

$$\mathbf{y}=\mathbf{F}\boldsymbol{\eta}+\mathbf{v},$$

$$\boldsymbol{\eta}=[\boldsymbol{\zeta}^{\mathrm{T}}\ \ (\bar{\mathbf{z}}+\tilde{\mathbf{z}})^{\mathrm{T}}]^{\mathrm{T}}.$$

Moreover,

$$\mathbf{\Gamma} := \mathrm{Cov}[z^*, \boldsymbol{\eta}] = [\phi^{\mathrm{T}}(t)\ \mathrm{Var}[\boldsymbol{\zeta}]\ \ \bar{\mathbf{r}} + \tilde{\mathbf{r}}^j],$$

$$\mathbf{V} := \mathrm{Var}[\boldsymbol{\eta}] = \begin{bmatrix} \mathrm{Var}[\boldsymbol{\zeta}] & 0 \\ 0 & \bar{\mathbf{R}} + \tilde{R} \end{bmatrix},$$

$$\mathbf{\Gamma V}^{-1} = \mathbf{L}^j.$$

The rest of the proof is very similar to the proof of Proposition 4. □

5. Examples

In this section the proposed identification scheme is applied to two case studies, one simulated and the other experimental.

5.1. Simulated example: sparsely sampled data

In this example the proposed nonparametric identification scheme is applied to a problem in which sampling is not uniform between subjects. In particular the number of samples per subject ranges from 1 to 9. In such conditions it is clearly impossible to estimate the average curve by averaging the individual curves estimated by standard identification methods. Conversely, the nonparametric population approach not only reconstructs the average curve but provides also estimates of the individual ones.

The measurements y_k^j in the jth subject were obtained as

$$z^j(t) = \alpha_j \exp(-\beta_j t),$$

$$y_k^j = z^j(t_k) + v_k^j,$$

$$\mathrm{Var}[v_k^j] = (0.1 z^j(t_k^j))^2,$$

where $(\alpha_j,\ \beta_j)$ are the parameters characterizing the jth subject. The population distribution of the individual parameters is as follows

$$\alpha_j = \exp(v_{1j}),$$

$$\beta_j = \exp(v_{2j}),$$

$$v_j = [v_{1j}\ \ v_{2j}]^{\mathrm{T}},$$

$$v \sim N([0\ \ln(0.2)]^{\mathrm{T}}, \mathrm{diag}\{0.01,\ 0.0259\}).$$

The typical curve z^{typ} is obtained in correspondence with the expected values of the parameter vector v: $z^{\mathrm{typ}}(t) = \mathrm{e}^{-0.2t}$. For this model, 500 replicate data sets of $N = 7$ individuals each were generated. The set of possible sampling instants was $\{t_1, \ldots, t_9\} = \{0, 0.5, 1, 1.5, 2, 4, 8, 12, 24\}$. In each data set, subject #1 was fully sampled, #2 was sampled at time points $\{t_1,\ t_3,\ t_5,\ t_6,\ t_7,\ t_8\}$, #3 at $\{t_1,\ t_5,\ t_6,\ t_8,\ t_9\}$, #4 at$\{t_1,\ t_3,\ t_6,\ t_8\}$, #5 at $\{t_2,\ t_4,\ t_6\}$, #6 at $\{t_3,\ t_7\}$ and #7 at $\{t_5\}$ (30 samples in total). For illustrative purposes, in Fig. 2 the noisy measurements and the individual curves of one of the 500 data sets are plotted. In this problem, in order to take into account the prior information that all curves tend to zero,

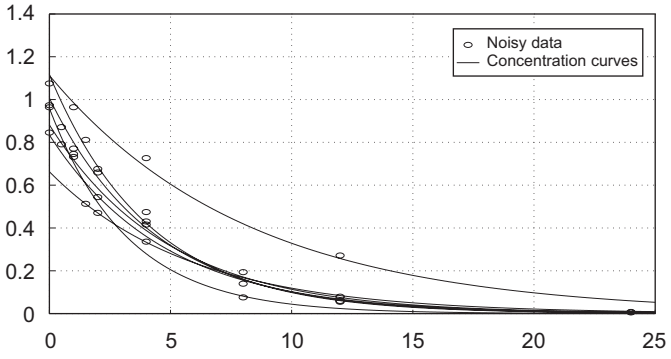

Fig. 2. A simulated data set: noisy measurements and real individual curves.

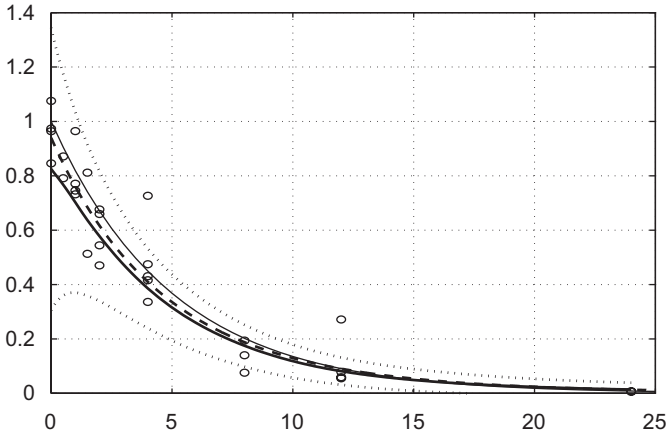

Fig. 3. True (dashed) vs. estimated (solid) average curve with its 95% confidence intervals and available data (open circles). The typical curve (thin solid) is also plotted.

a transformation of the time coordinates was performed. More precisely, $t^{\mathrm{new}} = 1/(1 + t/\gamma)$ so that $t = 0$ and ∞ correspond to $t^{\mathrm{new}} = 1$ and 0, respectively. Then, in the new time coordinates it was assumedthat both the average and individual curves had zero initial conditions (corresponding to zero terminal conditions at $t = \infty$), i.e. $\bar{x}(0) = 0$ and $\tilde{x}(0) = 0$. As the new time range $t^{\mathrm{new}} \in [0, 1]$ is finite, the individual shifts were modelled as integrated Wiener processes (since t^{new} does not go to infinity, the posterior variance of the average curve cannot diverge). Another advantage of the time transformation has to do with its ability to formalize the prior knowledge that the curves become smoother as time increases. In fact, processes whose second derivative is stationary in the new time coordinate correspond to processes whose second derivative has decreasing variance in the original time coordinate. The parameter γ of the time transformation was chosen so as to maximize the minimum distance between each pair of transformed sampling instants, yielding $\gamma = 3.00$. In the estimation algorithm $\mathrm{Var}[v_k^j]$ was approximated by $0.01(\hat{z}(t_k^j))^2$, i.e. by replacing the (unknown) $z^j(t_k^j)$ with the predicted value. The hyper-parameters were estimated via ML. The results of the identification for one of the 500 data sets are given in Figs. 3–5, where the estimated

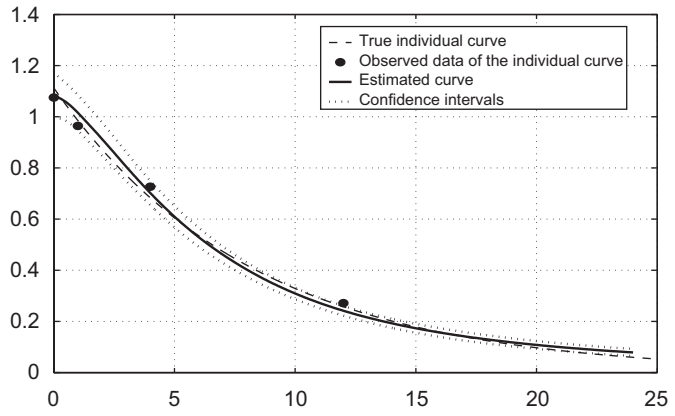

Fig. 4. Individual curve #4: true (dashed) vs. estimated (solid) curve with its 95% confidence intervals and available data.

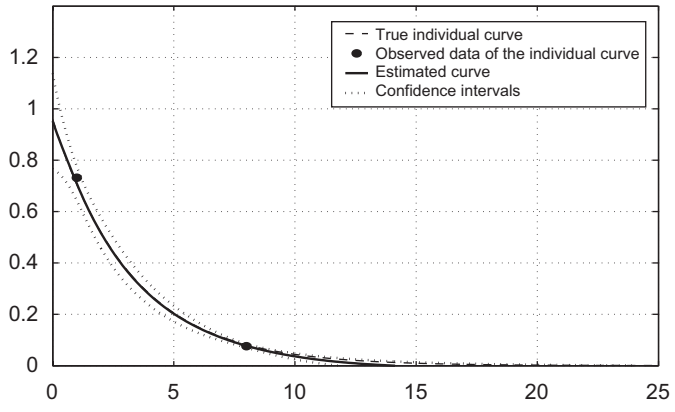

Fig. 5. Individual curve #6: true (dashed) vs. estimated (solid) curve with its 95% confidence intervals and available data.

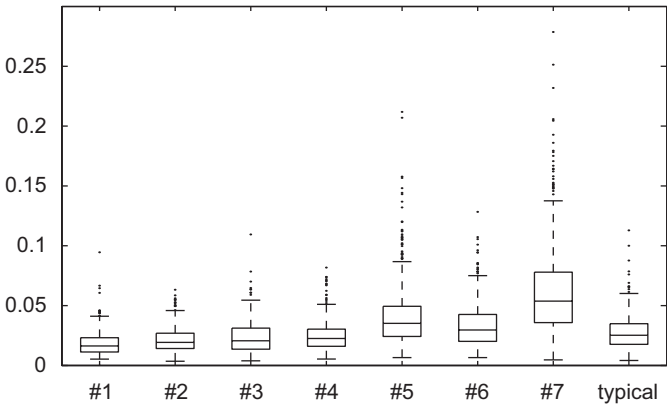

Fig. 6. *RMSE* of each individual curve and of the average curve computed on 500 data sets.

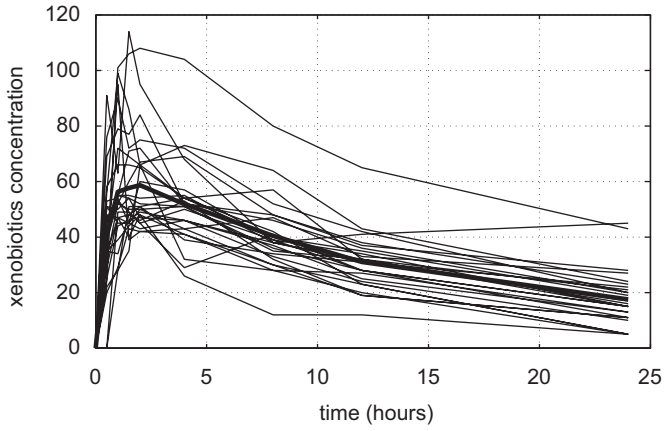

Fig. 7. Xenobiotics concentration data after a bolus in 27 human subjects: average curve (bold) and individual curves.

average curve and two of the seven individual curves are reported together with their confidence intervals. Both the average and individual curves are estimated with reasonable accuracy. Note the difference between the average and typical curve of the population (Fig. 3). The accuracy of the individual curves decreases together with the number of available data. This phenomenon can be appreciated by looking at the boxplots of the RMSEs reported in Fig. 6. The RMSE was computed as

$$RMSE = \left(\frac{1}{t_9} \int_0^{t_9} (\hat{z}^j(\tau) - z^j(\tau))^2 \, \mathrm{d}\tau \right)^{1/2}$$

for the individual curves, and analogously for the average curve $\bar{z}(t)$.

5.2. Analysis of pharmacokinetic data

Finally, the proposed population model was tested on a data set related to xenobiotics administration in 27 human subjects (Rocchetti & Poggesi, 1997). In the experiment, 8 samples were collected in each subject at $\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\} = \{0.5, 1, 1.5, 2, 4, 8, 12, 24\}$ hours after a bolus administration. The data have a 10% coefficient of variation, i.e. $\mathrm{Var}[v_k^j] = (0.1\hat{z}(t_k^j))^2$. To illustrate the population variability, the 27 experimental concentration curves are reported in Fig. 7, together with the average curve which, given the number of subjects, is a reasonable estimate of the average curve. Starting from these experimental data, different sampling schemes can be simulated by choosing proper subsets of the data. In particular, we adopted an example of a sparse sampling protocol: subject #2 is sampled at time points $\{t_6, t_7, t_8\}$, #5 at $\{t_2, t_4, t_8\}$, #7 is fully sampled, #8 at $\{t_3, t_5\}$, #13 at $\{t_1, t_2\}$, #17 at $\{t_7\}$, #19 at $\{t_6\}$, #20 at $\{t_4, t_8\}$, #21 at $\{t_5\}$ and #23 at $\{t_1, t_3\}$ (25 samples in total). Also in this case study the times were transformed by defining a new time axis $t^{\mathrm{new}} = 1/(1 + t/\gamma)$ with $\gamma = 3.00$ (the value of γ coincides with that used for the simulated example because the sampling schedule is the same). In this case, in the new time coordinates all the curves (the average and the individuals) are equal to zero in $t^{\mathrm{new}} = 1$ (corresponds to $t = 0$). This was accommodated by inserting zero-variance null measurements in $t^{\mathrm{new}} = 1$. The hyper-parameters were estimated via ML ($\tilde{\lambda}^2_{\mathrm{ML}} = 102130$, $\tilde{\lambda}^2_{\mathrm{ML}} = 23443$). In Fig. 8, the estimated average curve with its 95% confidence intervals is reported together with the data. The estimated average curve (Fig. 8) appears to be a satisfactory reconstruction, especially if it is taken

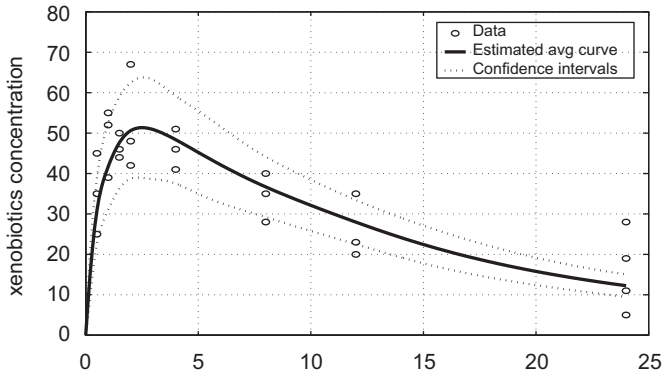

Fig. 8. Estimated average curve (bold) with its 95% confidence intervals. The adopted sparse sampling protocol uses only 25 data (circles) out of the available 216.

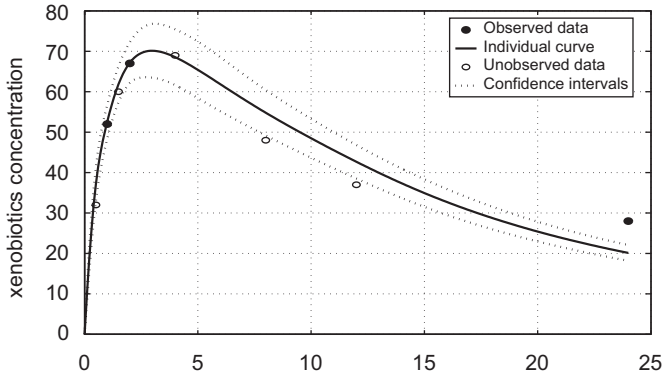

Fig. 9. Estimated individual curve of subject #5 (bold) with its 95% confidence intervals. For this individual curve only three data (full circles) were available. In order to assess the quality of the reconstruction, the other five unobserved data (open circles) are also plotted.

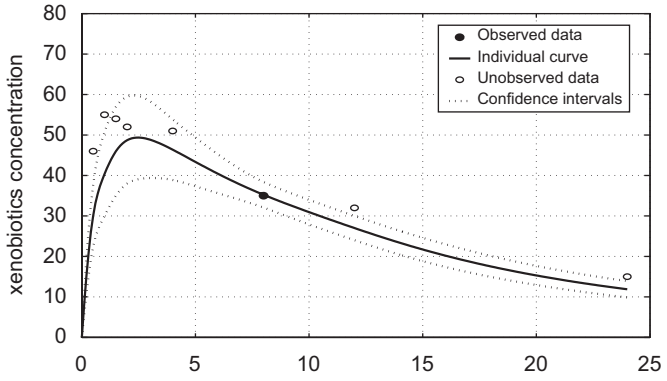

Fig. 10. Estimated individual curve of subject #19 (bold) with its 95% confidence intervals. For this individual curve only one datum (full circle) was available. In order to assess the quality of the reconstruction, the other seven unobserved data (open circles) are also plotted.

into account that it was obtained using only 25 observations. In Figs. 9 and 10 the estimate of the individual curve of subjects #5 and #19, respectively, are shown together with their confidence intervals. For the other individuals, reasonable estimates are obtained (data not shown).

6. Conclusions

A new nonparametric continuous-time model for the population analysis of multiple experiments has been proposed. The average curve as well as the individual ones are modelled as continuous-time Gaussian processes. If the statistics of the processes are known, the posterior expectation given the data (the Bayes estimate) is obtained as the output of an RN, i.e. as the linear combination of auto-covariance functions centred at the sampling knots. The network weights are computed by solving a system of linear equations. Moreover, if the average curve is modelled as an integrated Wiener process, its estimate is a cubic spline. In general, the statistics of the processes are not completely known and depend on some unknown hyper-parameters. Therefore, an empirical Bayes scheme has been proposed: first the hyper-parameters are estimated via ML and subsequently their ML estimates are plugged into the RN. The availability of effective nonparametric methods is of great interest in the population pharmacokinetic field. Especially in the early stages of a study, in absence of reliable parametric models, nonparametric estimation may help both evaluating the exposure, see (Magni et al., 2002) and checking for misspecification of candidate parametric models. Before the present paper, the only approach for identifying continuous-time population models without assuming a parametric model was the semiparametric spline method discussed in Park et al. (1997). Compared to that approach, the proposed method is strongly grounded on a Bayesian paradigm and avoids the nonlinear optimization required to locate the spline knots. On the other hand, some kinds of constraints, such as nonnegativity and negative tail slope, may be more easily handled in the semiparametric approach.

A first direction of future research will focus on the implementation of computationally efficient algorithms. In fact, the proposed scheme requires the solution of a system of linear equations and its computational complexity scales with the cube of the number of observations. By exploiting the state–space model it may be possible to work out algorithms based on Kalman filtering whose complexity scales linearly with the number of data, see, e.g. De Nicolao and Ferrari-Trecate (2001), where the efficient computation of RNs is addressed. The main advantage of a linear complexity algorithm would manifest itself in the hyper-parameter estimation via iterative likelihood maximization. A second topic that is being currently investigated is the development of a truly Bayesian estimation procedure in which the hyper-parameters and the curves are estimated jointly using Markov Chain Monte Carlo algorithms (Neve, De Nicolao, & Marchesi, 2005).

Acknowledgements

This paper was partially supported by the PRIN Project “Metodi e algoritmi innovativi per l’identificazione e il controllo adattativo di sistemi tecnologici”.

Appendix A. Technical lemma

Consider the problem of estimating a scalar random variable z^* given noisy observations $\mathbf{y} = \mathbf{F}\boldsymbol{\eta} + \boldsymbol{\epsilon}$ where the vector $\boldsymbol{\eta}$ is correlated with z^* and ε is an independent noise term. In the next lemma it is shown that the conditional variance $\mathrm{Var}[z^*|\mathbf{y}]$ can be decomposed as the sum of two terms. The first one is the conditional variance when $\boldsymbol{\eta}$ is perfectly known, whereas the second term keeps into account the presence of the measurements noise $\boldsymbol{\epsilon}$. A graphical representation of the lemma in terms of projections in the Hilbert space of jointly normal random variables is provided in Fig. A.1. Lemma 1 in Wahba (1983) is obtained as a particular case letting $\mathbf{F} = \mathbf{I}$.

Lemma 1. *Assume that*

$$\mathbf{y} = \mathbf{F}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad \mathbf{y} \in \mathbb{R}^n,$$

$$\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_\varepsilon), \quad \boldsymbol{\Sigma}_\varepsilon > 0,$$

$$\begin{bmatrix} z^* \\ \boldsymbol{\eta} \end{bmatrix} \sim N(0, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_*^2 & \boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^{\mathrm{T}} & \mathbf{V} \end{bmatrix}, \quad \boldsymbol{\Sigma} > 0,$$

where z^* is a scalar and ε is independent of $[z^*\boldsymbol{\eta}^{\mathrm{T}}]^{\mathrm{T}}$. Then,

$$\mathrm{Var}[z^*|\mathbf{y}] = \mathrm{Var}[z^*|\boldsymbol{\eta}] + \mathrm{Var}[E[z^*|\boldsymbol{\eta}]|\mathbf{y}],$$

$$\mathrm{Var}[z^*|\boldsymbol{\eta}] = \sigma_*^2 - \boldsymbol{\Gamma}\mathbf{V}^{-1}\boldsymbol{\Gamma}^{\mathrm{T}},$$

$$\mathrm{Var}[E[z^*|\boldsymbol{\eta}]|\mathbf{y}] = \boldsymbol{\Gamma}\mathbf{V}^{-1}\,\mathrm{Var}[\boldsymbol{\eta}|\mathbf{y}]\mathbf{V}^{-1}\boldsymbol{\Gamma}^{\mathrm{T}},$$

$$\mathrm{Var}[\boldsymbol{\eta}|\mathbf{y}] = (\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{F} + \mathbf{V}^{-1})^{-1}.$$

Proof. The expression for $\mathrm{Var}[z^*|\boldsymbol{\eta}]$ is a straightforward consequence of well-known properties of jointly Gaussian random variables. As for the computation of $\mathrm{Var}[E[z^*|\boldsymbol{\eta}]|\mathbf{y}]$, observe that $E[z^*|\boldsymbol{\eta}] = \boldsymbol{\Gamma}\mathbf{V}^{-1}\boldsymbol{\eta}$. Therefore,

$$\mathrm{Var}[E[z^*|\boldsymbol{\eta}]|\mathbf{y}] = \boldsymbol{\Gamma}\mathbf{V}^{-1}\,\mathrm{Var}[\boldsymbol{\eta}|\mathbf{y}]\mathbf{V}^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}.$$

On the other hand, in view of standard Bayesian estimation formulas

$$\mathrm{Var}[\boldsymbol{\eta}|\mathbf{y}] = (\mathbf{F}^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{F} + \mathbf{V}^{-1})^{-1}.$$

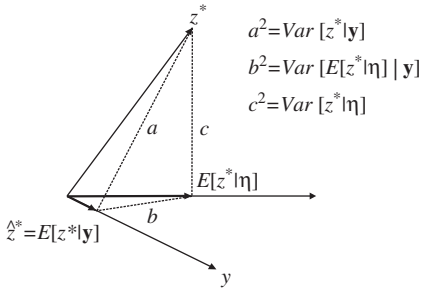

Fig. A.1. Graphical interpretation of Lemma 1 in terms of projections in the Hilbert space of jointly normal random variables.

By applying the matrix inversion lemma, one has that

$$\mathrm{Var}[\boldsymbol{\eta}|\mathbf{y}] = \mathbf{V} - \mathbf{V}\mathbf{F}^{\mathrm{T}}(\mathbf{F}\mathbf{V}\mathbf{F}^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{F}\mathbf{V}.$$

Finally,

$$\begin{aligned}
&\mathrm{Var}[z^*|\boldsymbol{\eta}] + \mathrm{Var}[E[z^*|\boldsymbol{\eta}]|\mathbf{y}] \\
&\quad = \sigma_*^2 - \boldsymbol{\Gamma}\mathbf{V}^{-1}\boldsymbol{\Gamma}^{\mathrm{T}} \\
&\qquad + \boldsymbol{\Gamma}\mathbf{V}^{-1}(\mathbf{V} - \mathbf{V}\mathbf{F}^{\mathrm{T}}(\mathbf{F}\mathbf{V}\mathbf{F}^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{F}\mathbf{V})\mathbf{V}^{-1}\boldsymbol{\Gamma}^{\mathrm{T}} \\
&\quad = \sigma_*^2 - \boldsymbol{\Gamma}\mathbf{F}^{\mathrm{T}}(\mathbf{F}\mathbf{V}\mathbf{F}^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{F}\boldsymbol{\Gamma}^{\mathrm{T}} \\
&\quad = \mathrm{Var}[z^*] - \mathrm{Cov}[z^*, \mathbf{y}]\mathrm{Var}[\mathbf{y}]^{-1}\mathrm{Cov}[z^*, \mathbf{y}]^{\mathrm{T}} \\
&\quad = \mathrm{Var}[z^*|\mathbf{y}]
\end{aligned}$$

so proving the thesis. □

References

Aarons, L. (1999). Software for population pharmacokinetics and pharmacodynamics. *Clinical Pharmacokinetics*, *36*(4), 255–264.

Beal, S. L., & Sheiner, L. B. (1982). Estimating population kinetics. *Critical Review of Biomedical Engineering*, *8*(3), 195–222.

Beal, S. L., & Sheiner, L. B. (1998). *Nonmem users guide*.

Bertoldo, A., Sparacino, G., & Cobelli, C. (2004). "Population" approach improves parameter estimation of kinetic models from dynamic PET data. *IEEE Transactions on Medical Imaging*, *23*(3), 297–306.

Center for Drug Evaluation and Research. (1999). *Guidance for industry*: *Population pharmacokinetics*. United States Department of Health and Human Services, Food and Drug Administration.

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. New York, NY, USA: Chapman & Hall.

De Nicolao, G., & Ferrari-Trecate, G. (2001). Regularization networks: Fast weight calculation via Kalman filtering. *IEEE Transactions on Neural Networks*, *12*(2), 228–235.

Egerstedt, M., & Martin, C. F. (2001). Optimal trajectory planning and smoothing splines. *Automatica*, *37*, 1057–1064.

Fattinger, K. E., & Verotta, D. (1995a). A nonparametric subject-specific population method for deconvolution: I. Description, internal validation and real data examples. *Journal of Pharmacokinetics and Biopharmaceutics*, *23*, 581–610.

Fattinger, K. E., & Verotta, D. (1995b). A nonparametric subject-specific population method for deconvolution: II. External validation. *Journal of Pharmacokinetics and Biopharmaceutics*, *23*, 611–634.

Ferrazzi, F., Magni, P., & Bellazzi, R. (2003). Bayesian clustering of gene expression time series. In *Proceedings of 3rd international workshop on bioinformatics for the management, analysis and interpretation of microarray data* (*NETTAB 2003*) (pp. 53–55).

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, *7*, 219–269.

Guardabasso, V., Munson, P. J., & Rodbard, D. (1988). A versatile method for simultaneous analysis of families of curves. *FASEB Journal*, *2*, 209–215.

Jelliffe, R., Schumitzky, A., Van Guilder, M., Wang, X., & Leary, R. (2001). Population pharmacokinetic and dynamic models: Parametric (P) and nonparametric (NP) approaches. In *14th IEEE symposium on computer-based medical systems* (pp. 407–412). Bethesda, MD, USA.

Leary, R., Jelliffe, R., Schumitzky, A., & Van Guilder, M. (2001). An adaptive grid non-parametric approach to pharmacokinetic and dynamic (PK/PD) population models. In *14th IEEE symposium on computer-based medical systems* (pp. 389–394). Bethesda, MD, USA.

MacKay, D. J. (1997). Gaussian processes: A replacement for supervised neural networks?. In *Lecture notes of neural information processing systems* (*NIPS'97*).

MacKay, D. J. (1998). Introduction to Gaussian processes. In C. M. Bishop (Ed.), *Neural networks and machine learning* (Vol. 168, pp. 133–166). NATO Asi Series, Series F, Computer and Systems Sciences. Dordrecht: Kluwer Academic Press.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation.*, *4*, 415–447.

Magni, P., Bellazzi, R., De Nicolao, G., Poggesi, I., & Rocchetti, M. (2002). Nonparametric AUC estimation in population studies with incomplete sampling: A Bayesian approach. *Journal of Pharmacokinetics and Pharmacodynamics*, *29*(5/6), 445–471.

Neve, M., De Nicolao, G., & Marchesi, L. (June 8–10, 2005). Nonparametric identification of population pharmacokinetic models: An MCMC approach. In *Proceedings of 24th American control conference* (pp. 991–996). Portland, OR, USA.

Park, K., Verotta, D., Blaschke, T. F., & Sheiner, L. B. (1997). A semiparametric method for describing noisy population pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, *25*, 615–642.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of IEEE*, *78*, 1481–1497.

Rocchetti, M., & Poggesi, I. (1997). Comparison of the Bailer and Yeh methods using real data. In L. Aarons, et al. (Ed.), *The population approach*: *Measuring and managing variability in response*, *concentration and dose* (pp. 385–390). Brussels, Belgium: European Cooperation in the Field of Scientific and Technical Research, European Commission.

Sheiner, L. B. (1994). The population approach to pharmacokinetic data analysis: Rationale and standard data analysis methods. *Drug Metabolism Reviews*, *15*, 153–171.

Sheiner, L. B., Rosenberg, B., & Marathe, V. V. (1977). Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. *Journal of Pharmacokinetics and Biopharmaceutics*, *5*(5), 445–479.

Sheiner, L. B., & Steimer, J. L. (2000). Pharmacokinetic/pharmacodynamic modeling in drug development. *Annual Review of Pharmacology and Toxicology*, *40*, 67–95.

Shiryaev, A. N. (1996). *Probability*. New York, NY, USA: Springer.

Smola, A. J., & Schölkopf, B. (2003). Bayesian kernel methods. In S. Mendelson, A. J. Smola, (Eds.), *Machine learning*, *proceedings of the summer school*, *Australian National University* (pp. 65–117). Berlin, Germany: Springer.

Sun, S., Egerstedt, M. B., & Martin, C. F. (2000). Control theoretic smoothing splines. *IEEE Transactions on Automatic Control*, *45*(12), 2271–2279.

Vicini, P., & Cobelli, C. (2001). The iterative two-stage population approach to IVGTT minimal modeling: Improved precision with reduced sampling. *American Journal of Physiology, Endocrinology and Metabolism*, *280*(1), 179–186.

Vozeh, S., Steimer, J. L., Rowland, M., Morselli, P., Mentre, F., Balant, L. P. et al. (1996). The use of population pharmacokinetics in drug development. *Clinical Pharmacokinetics*, *30*(2), 81–93.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross validated smoothing spline. *Journal of Royal Statistical Society, Series B*, *45*(1), 133–150.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia, USA: SIAM.

Wakefield, J., & Bennett, J. (1996). The Bayesian modelling of covariates for population pharmacokinetic models. *JASA*, *91*, 917–927.

Wakefield, J., Smith, A. F. M., Racine-Poon, A., & Gelfand, A. (1994). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. *Applied Statistics*, *41*, 201–221.

Williams, C. K. I. (1999). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. Jordan (Ed.), *Learning and inference in graphical models* (pp. 599–621). Cambridge, MA, USA: MIT Press.

Williams, C. K. I., & Rasmussen, C.E. (1996). Gaussian processes for regression. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo, (Eds.), *Advances in neural information processing systems* (Vol. 8). Cambridge, MA, USA: MIT Press.

Yuh, L., Beal, S., Davidian, M., Harrison, F., Hester, A., Kowalski, K. et al. (1994). Population pharmacokinetic/pharmacodynamic methodology and applications: A bibliography. *Biometrics*, *50*, 566–575.

Marta Neve was born in Codogno, Italy, in 1977. In 2002 she received the Laurea (master degree) cum laude in Informatic Engineering ("Laurea") from the University of Pavia, Italy. In 2006 she obtained a Ph.D. in the same field from the University of Pavia with a thesis entitled "Bayesian Learning Techniques for Nonparametric Identification". In 2004 she visited the Center for Biological and Computational Learning of the Massachusetts Institute of Technology, Boston, MA. During her Ph.D. studies she obtained also the postgraduate degree from the multidisciplinary Advanced School of Integrated Learning (SAFI-IUSS, University of Pavia). She was the recipient of SAFI-IUSS scholarships in 2003, 2004 and 2005. In February 2006 she joined the Motorsport Department of Magneti Marelli, Corbetta, Italy. Since October 2006 she is a research scientist in the Clinical Pharmacokinetics, Modelling & Simulation Department of the GlaxoSmithKline Research Centre in Verona, Italy. Her research interests include Bayesian and neural identification methods applied to physiological systems and internal combustion engines.

Giuseppe De Nicolao was born in Padova, Italy, in 1962. In 1986 he received the Laurea (master degree) cum laude in Electronic Engineering ("Laurea") from the Polytechnic of Milan, Italy. From 1987 to 1988 he was with the Biomathematics and Biostatistics Unit of the Institute of Pharmacological Researches "Mario Negri", Milano. In 1988 he joined the Italian National Research Council (C.N.R.) as a researcher scientist of the Center of System Theory in Milano. From 1992 to 2000 he was associate professor and, since 2000, full professor of Model Identification in the Department of Computer Science and Systems Engineering of the University of Pavia (Italy). In 1991, he held a visiting fellowship at the Department of Systems Engineering of the Australian National University, Canberra, Australia. In 1998 he was a plenary speaker at the workshop on "Nonlinear Model Predictive Control: Assessment and Future Directions for Research", Ascona, Switzerland. He is a senior member of the IEEE, and, from 1999 to 2001, he has been Associate Editor of the IEEE Transactions on Automatic Control. His research interests include model predictive control, optimal and robust filtering and control, Bayesian learning, neural networks, deconvolution techniques, modeling and identification of biomedical systems, statistical process control and fault diagnosis for semiconductor manufacturing. On these subjects he has authored or coauthored 80 journal papers and is coinventor of two patents.

Laura Marchesi was born in Voghera, Italy, in 1978. In 2004 she received the Laurea (master degree) in Informatic Engineering ("Laurea") from the University of Pavia, Italy. From September 2004 to December 2004 she was with the System Identification and Automatic Control Laboratory of the Department of Computer Science and Systems Engineering, University of Pavia. Since February 2005 she is with Accenture Italia.