Journal of Process Control 70 (2018) 80–95

Contents lists available at ScienceDirect

Journal of Process Control

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

Set-membership nonlinear regression approach to parameter estimation

Nikola D. Perić[a], Radoslav Paulen[c], Mario E. Villanueva[b], Benoît Chachuat[a,*]

[a] Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, UK
[b] School of Information Science and Technology, ShanghaiTech University, Shanghai, China
[c] Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Slovakia

ARTICLE INFO

Article history:
Received 5 September 2017
Received in revised form 2 April 2018
Accepted 9 April 2018

Keywords:
Parameter estimation
Nonlinear regression
Set-membership estimation
Statistical inference
Semi-infinite programming
Complete-search methods

ABSTRACT

This paper introduces *set-membership nonlinear regression* (SMR), a new approach to nonlinear regression under uncertainty. The problem is to determine the subregion in parameter space enclosing all (global) solutions to a nonlinear regression problem in the presence of bounded uncertainty on the observed variables. Our focus is on nonlinear algebraic models. We investigate the connections of SMR with (i) the classical statistical inference methods, and (ii) the usual set-membership estimation approach where the model predictions are constrained within bounded measurement errors. We also develop a computational framework to describe tight enclosures of the SMR regions using semi-infinite programming and complete-search methods, in the form of likelihood contour and polyhedral enclosures. The case study of a parameter estimation problem in microbial growth is presented to illustrate various theoretical and computational aspects of the SMR approach.

1. Introduction

Mathematical models capable of accurate prediction of physical phenomena have proved to be invaluable tools for engineers and scientists. In the area of process systems engineering, they routinely support the design, control and optimization of production processes, as a means of improving their economical profitability and reducing their environmental footprint. A majority of these models are nonlinear and contain adjustable parameters that need estimating from available experimental data, or else from other, more fundamental, mathematical descriptions. In this context, parameter estimation turns out to be a key step in the verification, and subsequent use, of the mathematical models.

Most commonly, parameter estimation in nonlinear models is cast as a nonlinear regression exercise, where selected parameter values are adjusted so that the model predictions match the available observations as close as possible, for instance in the least-squares or maximum-likelihood sense [1–4]. In order to avoid for the resulting parameter estimates to be biased, one can account for measurement errors in all of the variables, both independent and dependent variable observations, by following the so-called errors-in-variables approach [5,6]. This problem has been widely studied from a computational standpoint over the past decades, including the development of rigorous global optimization approaches for overcoming convergence to local optima [7,8].

Of course, there is more to model identification than just determining values for the unknown parameters. Systematic procedures have been devised to support the development and statistical verification of process models, which include testing structural identifiability, designing experiments for improved parameter precision, and inferring parameter confidence [9–12]. The focus in this paper is on the latter aspect, namely characterizing subregions in parameter space wherein the parameter values can be expected to lie. Other applications of such parameter confidence regions are in design under uncertainty [13,14], robust model predictive control [15–17], robust monitoring [18,19], and robust optimal design of experiments [20–22], to name but a few. For the scope of this paper, the emphasis is on models described by algebraic equations, but these ideas can be extended to dynamic or distributed models described by differential equations too.

Accounting for model mismatch and uncertain observations within the regression problem has spawned several schools of thought. Statistical approaches can be broadly classified as *frequentist* or *Bayesian*. The former seek to determine confidence regions around the regressed parameter values, typically a maximum-likelihood estimate, considered as the 'true' parameter values [1,2,4]. By construction, a $100(1-\alpha)\%$ frequentist confidence region comprises $100(1-\alpha)\%$ of the parameter values that would

* Corresponding author.
E-mail address: b.chachuat@imperial.ac.uk (B. Chachuat).

https://doi.org/10.1016/j.jprocont.2018.04.002

be obtained upon repetition of the parameter estimation using (hypothetical) new observations, considered as random variables. Approximate confidence regions, for instance based on the Wald test or the likelihood-ratio (LR) test, are known to converge to the exact confidence region in the limit of an infinite number of observations under certain conditions. Process modeling environments such as gPROMS and Aspen Custom Modeler have been relying on linear approximation and the Wald test to determine ellipsoidal confidence regions, a computationally efficient procedure for problems having several dozen unknown parameters, but one which may produce inaccurate results with large measurement errors and model mismatch or few measurement points. Confidence regions based on the LR test have been shown to yield superior approximations, but are computationally more involved since the corresponding parameter regions are complex sets in general (e.g., nonconvex, not simply connected) [23,24].

In practice, the term $100(1-\alpha)\%$ confidence region is often misused to refer to the range of parameter values that include $100(1-\alpha)\%$ of their probability distribution [25]. This description corresponds to so-called $100\,(1-\alpha)\%$ credible regions instead, which are defined in the Bayesian inference approach [26]. Bayesian estimation uses the available observations to construct a probability distribution of the parameters, called posterior distribution, based on a likelihood function and a prior probability distribution of the same parameters. In essence, this approach thus considers the unknown parameter values as random variables. Sampling-based techniques such as Markov-Chain Monte-Carlo (MCMC) [27,28] provide a means of constructing (approximate) credible regions, although the computational effort can become prohibitive for problems having upwards of 10 parameters [29]. A most probable estimate can be determined from the posterior distribution, which also corresponds to a maximum-likelihood estimate for a flat prior. Albeit classical frequentist and Bayesian inference regions can be reconciled in special cases, no equivalence can be drawn in general since Bayesian inference incorporates problem specific contextual information from the prior distribution, whereas frequentist inference is solely based on the data; see, e.g., [30, Chapter 5]. The debate on whether to use frequentist or Bayesian statistical inference continues to this day [25,31], but its intricacies are beyond the scope of this paper.

Regardless of whether a mathematical model's structure is correct or not, a frequentist confidence region will normally converge to the maximum- likelihood estimate as the number of observations increases. Likewise, a Bayesian posterior will normally converge to a point mass that corresponds to a most probable estimate, i.e., a point that maximizes the probability of the data given the (possibly wrong) model. An interesting alternative to these statistical approaches is *set-membership* estimation (SME). The traditional SME setting, also called guaranteed parameter estimation (GPE), seeks to determine the set of all possible parameter values for which a model's predictions are consistent with a set of observations subject to bounded errors [32–34]. The fact that this approach does not require a statistical description of the observation errors, solely bounds, is not only less demanding, but also more realistic in many practical applications, including biological systems where the measurements are often scarce and subject to large errors [21]. Beside parameter estimation, the distinctive yes-or-no answer provided by set-membership techniques can also be used for model inconsistency detection [35,36]. One caveat here is that the set of feasible parameter values may be empty in the presence of measurement outliers or due to an inadequate description of the measurement noise, thus calling for remedial strategies [37,38]. Another key challenge in nonlinear set-membership estimation is describing the feasible parameter set accurately, while remaining computationally tractable. This challenge is in fact similar to the one faced by aforementioned statistical inference methods for describing parameter confidence sets, and it may explain why set-membership estimation has not reached a wider diffusion to this day. Existing computational strategies are limited to problem with downwards of a dozen parameters. They range from approximation using sampling-based methods, including stochastic search [39], support vector machines (SVM) [40] and MCMC [41]; to rigorous complete-search methods based on interval analysis and other set arithmetics [42–44]; and to semidefinite relaxation techniques for semi-algebraic problems [45,46].

This paper introduces *set-membership regression* (SMR), a new approach to nonlinear regression. The SMR problem seeks to determine the subregion in parameter space enclosing all (global) solutions to a nonlinear regression problem in the presence of bounded uncertainty on the observed variables. By contrast with the traditional SME setting seeking for parameter values to satisfy certain feasibility constraints, the SMR approach method seeks for parameter values to satisfy an optimality condition. To the best knowledge of the authors, this problem has not been investigated in the general nonlinear setting so far. Milanese [47] studied optimality and convergence properties of least-squares estimates in the presence of unknown bounded disturbance, but their theoretical work is limited to linear problems. This paper sets out to investigate the connections of SMR with both statistical inference and set-membership estimation approaches for nonlinear algebraic models. Another principal contribution is a computational framework to describe tight enclosures of the SMR regions using complete-search methods.

The rest of the paper is organized as follows. Section 2 starts by reviewing classical results from both areas of statistical and set-membership estimation. Section 3 introduces the SMR approach and analyzes its properties, after which numerical solution strategies are developed in Section 4. A simple case study is used throughout Sections 2–4 to illustrate the main concepts and results. Section 5 presents a more challenging estimation problem in microbial growth to demonstrate the SMR approach. Finally, Section 6 concludes the paper and discusses future research opportunities.

2. Background

Our focus throughout this paper is on explicit models in the form

$$\mathbf{y} = \mathbf{g}(\mathbf{p}, \mathbf{u}),$$

where $\mathbf{p} \in \mathbb{R}^{n_p}$ is the vector of unknown parameters; and $(\mathbf{u}, \mathbf{y}) \in \mathbb{R}^{n_u} \times \mathbb{R}^{n_y}$ is the vector of observed variables, denoted collectively by $\mathbf{x}{:=}(\mathbf{u}, \mathbf{y}) \in \mathbb{R}^{n_x}$ for convenience. Notice that $\mathbf{u}$ and $\mathbf{y}$ often correspond to (either controlled or uncontrolled) input and output variables, respectively, in a practical setup. It is also worth pointing out that many of the concepts and methods presented herein can be applied to models described by implicit equation systems, such as $\mathbf{f}(\mathbf{p}, \mathbf{x}) = \mathbf{0}$, and models comprised of differential equations too.

Suppose that n_m observations $\mathbf{x}_k^{\mathrm{m}}{:=}(\mathbf{u}_k^{\mathrm{m}}, \mathbf{y}_k^{\mathrm{m}})$ of the input–output variables are available, and assume that all of these observation errors are independent and described by the probability density functions $p(\cdot \mid \boldsymbol{\psi})$ parameterized by $\boldsymbol{\psi}$. In the error-in-variables approach [6], the reconciled values $\mathbf{u}_1, \ldots, \mathbf{u}_{n_m}$ for the observations are estimated alongside the unknown model parameters $\mathbf{p}$. The joint probability of the prediction-observation mismatch in all data points for the parameter values $\boldsymbol{\theta}{:=}(\mathbf{p}, \mathbf{u}_1, \ldots, \mathbf{u}_{n_m}) \in \mathbb{R}^{n_\theta}$ is described by the following likelihood function:

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}}){:=}\prod_{k=1}^{n_m} p\left(\boldsymbol{\delta}\mathbf{u}_k \middle| \boldsymbol{\psi}_{\mathbf{u}_k}\right) \prod_{k=1}^{n_m} p\left(\boldsymbol{\delta}\mathbf{y}_k \middle| \boldsymbol{\psi}_{\mathbf{y}_k}\right), \tag{1}$$

with $\boldsymbol{\delta}\mathbf{u}_k{:=}\mathbf{u}_k - \mathbf{u}_k^{\mathrm{m}}$ and $\boldsymbol{\delta}\mathbf{y}_k{:=}\mathbf{g}(\mathbf{p}, \mathbf{u}_k) - \mathbf{y}_k^{\mathrm{m}}$. The error-in-equation approach instead, considers the input measurements $\mathbf{u}_k^{\mathrm{m}}$ to be