Journal of Process Control xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Process Control

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

Deep recurrent Gaussian processes for outlier-robust system identification

César Lincoln C. Mattos [a], Zhenwen Dai [b,1], Andreas Damianou [b,1], Guilherme A. Barreto [a,*], Neil D. Lawrence [b,c]

[a] Federal University of Ceará, Department of Teleinformatics Engineering, Fortaleza, Ceará, Brazil
[b] Amazon.com, Cambridge, UK
[c] University of Sheffield, Department of Computer Science, Sheffield, UK

ARTICLE INFO

Article history:
Received 27 September 2016
Received in revised form 3 May 2017
Accepted 19 June 2017
Available online xxx

Keywords:
Nonlinear system identification
Dynamical modeling
Robust modeling
Gaussian Process
Outliers
Variational Bayes

ABSTRACT

Gaussian Processes (GP) comprise a powerful kernel-based machine learning paradigm which has recently attracted the attention of the nonlinear system identification community, specially due to its inherent Bayesian-style treatment of the uncertainty. However, since standard GP models assume a Gaussian distribution for the observation noise, i.e., a Gaussian likelihood, the learning and predictive capabilities of such models can be severely degraded when outliers are present in the data. In this paper, motivated by our previous work on GP learning with data containing outliers and recent advances in hierarchical (deep GPs) and recurrent GP (RGP) approaches, we introduce an outlier-robust recurrent GP model, the RGP-*t*. Our approach explicitly models the observation layer, which includes a heavy-tailed Student-*t* likelihood, and allows for a hierarchy of multiple transition layers to learn the system dynamics directly from estimation data contaminated by outliers. In addition, we modify the original variational framework of standard RGP in order to perform inference with the new RGP-*t* model. The proposed approach is comprehensively evaluated using six artificial benchmarks, within several outlier contamination levels, and two datasets related to process industry systems (pH neutralization and heat exchanger), whose estimation data undergo large contamination rates. The simulation results obtained by the RGP-*t* model indicates an impressive resilience to outliers and a superior capability to learn nonlinear dynamics directly from highly outlier-contaminated data in comparison to existing GP models.

1. Introduction

Nonlinear system identification, while of fundamental importance for the design and analysis of industrial processes, is a challenging task. Since the quality of the model is usually the bottleneck of the final problem solution [1], complications such as estimation data contaminated with outliers must be carefully considered [2]. Such scenario is very often encountered in everyday practice and is the focus of the present paper. For this purpose, we follow a Bayesian approach to system identification, where the model accounts for the uncertainty in the noisy data and in the learned dynamics [3]. More specifically, we apply Gaussian Processes (GP) models, a principled probabilistic approach to learning in kernel machines [4].

Although GPs have been used for predictions by the geostatistics community for many decades, where it is usually known as *kriging* [5,6], it was only some years later that works such as [7] and [8] indicated that GP models are capable to outperform conventional nonlinear regression frameworks, such as artificial neural networks (ANNs), support vector regression (SVR) and spline methods. GP models are nonparametric data-driven techniques, where instead of a rigid prespecified structure, the model allows for the data to "speak by itself". A typical GP model increases its complexity as more data becomes available and estimation data is used to both optimize the model and make predictions.

Another major distinctive feature of the GP framework over those powerful regression methods is that the outputs of a GP model are probabilistic distributions, i.e., instead of point estimates, each prediction is given by a fully defined distribution. As a consequence, besides computing mean values for predictions, one can easily compute as well prediction intervals associated to those

* Corresponding author.
E-mail addresses: cesarlincoln@terra.com.br (C.L.C. Mattos), zhenwend@amazon.com (Z. Dai), damianou@amazon.com (A. Damianou), gbarreto@ufc.br (G.A. Barreto), lawrennd@amazon.com (N.D. Lawrence).

[1] Work done while Dai and Damianou were at the University of Sheffield.

http://dx.doi.org/10.1016/j.jprocont.2017.06.010
0959-1524/

values, using for instance the standard deviation of the predictive distribution.

Because of the aforementioned appealing features, application of GP models in engineering is gaining momentum. In the context of dynamical system modeling and identification, several authors have applied GP models to, for instance, local approaches for system identification [9,10], system identification with derivative observations [11], nonlinear autoregressive identification [12], learning nonstationary dynamics [13], multi-step ahead prediction with propagation of uncertainty [14,15], transfer function estimation [16] and state space modeling [17,18].

In standard applications of GP models a Gaussian observation noise is usually assumed, and hence Gaussian likelihoods arise naturally within this framework. While this is a reasonable assumption for many applications, it is not the case for modeling scenarios contaminated with non-Gaussian noise, such as impulsive noise.[1] In such scenarios the model's performance considerably deteriorates, compromising its out-of-sample prediction capability.

In the current paper, we pursue GP models that are able to directly learn dynamics from data containing outliers, as opposed to methodologies which remove them from the data before training, by promoting a *data cleaning* step [19]. Such an approach, surveyed in the context of temporal data in [20], can be useful when just a few outliers of high magnitude are present, but it becomes problematic when dealing with outliers [21] which are intermingled with the samples of the system under analysis, or when the outliers have their own “dynamics” apart from that of the system of interest. Huber [22], a pioneer in robust statistics, also argues against the data cleaning approach, emphasizing that dealing with outliers directly avoids erroneous removal of potentially valuable training samples.

Alternatively, robust GP-based modeling approaches for nonlinear regression usually apply heavy-tailed distributions for the model likelihood [23], such as Student-*t* [24–26], Laplace [27], and a mixture of Gaussians [28–30] in order to handle outliers as non-Gaussian noise. However, while the inference process (i.e., learning) in GP models with Gaussian likelihood is tractable,[2] non-Gaussian likelihood models often lead to non-tractable expressions, requiring either stochastic sampling techniques, such as Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) [24,31–33], or deterministic approximation methods, such as variational Bayes (VB) [34–36] and expectation propagation (EP) [37,38].

Although sampling-based algorithms have been extensively used for nonlinear system identification in different scenarios [39–41], in this work we focus on variational approaches, widely used alternatives to approximate posterior probability densities with an optimization-based formulation which tends to be faster and scale up better to large datasets [42,43]. It is worth mentioning that variational methods have been largely used in several non-analytical GP variants such as sparse approximations [44], non-supervised learning [45], state-space models [17] and general hierarchical modeling [46].

Despite the increasing interest in GP-based modeling for outlier-robust nonlinear regression, developing such models specifically for robust dynamical system identification is a relatively new topic. In fact, we were able to find very few contributions other than our own previous work [47,48]. In [32], the authors propose a robust GP model for estimating impulse responses of linear system when the measurements are corrupted by outliers, without addressing nonlinear systems. More recently, in [49] the authors introduce an Expectation-Maximization (EM) algorithm to tackle robust regression tasks, reporting an experiment in system identification, which was evaluated in one-step-ahead prediction scenarios only.

In [47] two robust GP models were applied for dynamical system identification using NARX (nonlinear autoregressive with exogenous inputs) models: a GP model with Student-*t* likelihood and variational inference (GP-tVB) and a GP model with Laplace likelihood and EP inference (GP-LEP). Both models outperformed the standard GP-NARX with Gaussian likelihood, especially GP-tVB, but they were still very sensitive to outliers in some of the studied evaluation scenarios.

A novel robust latent autoregressive model, named GP-RLARX, was proposed by us in [48]. This model incorporated the Student-*t* likelihood and introduced a latent dynamical structure to account for the uncertainty in the regressors, avoiding the explicit feedback of the noisy outputs in the model dynamics. The GP-RLARX followed a variational framework for inference and performed well in several artificial benchmarks with different levels of outlier contamination.

Despite its superior performance in the presence of outliers when compared to GP-tVB, the GP-RLARX model adopts a very simple observation model: the output is equal to the most recent latent variable plus noise. This forces the latent space to be closely related to the output and constrains the latent variables of the model. Furthermore, it relies on a single transition layer. As argued in [50], a hierarchical (or *deep*) structure can be helpful for recurrent modeling. Besides, such multilayer structure has been successfully exploited more recently with the rise of deep GP models [51,46].

The aforementioned restrictions could hinder the GP-RLARX model's capability to learn more complex dynamics from outlier-corrupted data. Bearing this in mind, in the current paper we extend GP-RLARX by including an additional GP to model the observation (or emission) layer, in order to separate the transition and emission nonlinear functions. To further increase the representational capability of the model, we allow the inclusion of more than one transition layer, referred to as *hidden layers*. This approach builds upon our work on Recurrent Gaussian Processes (RGP) model presented in [52], being equivalent to provide it with a Student-*t* likelihood. However, such extension requires a modification in the Recurrent Variational Bayes (REVARB) framework described in [52], since it originally considered a Gaussian likelihood.

In summary, the proposed extension aims to bring together in a synergistic way the best properties of the two previous approaches in order to obtain eventually a more reliable solution for robust system identification: (*i*) *resilience to non-Gaussian noise* provided by the GP-RLARX model due to the Student-*t* likelihood, and (*ii*) the enhanced *representational capability* provided by the hierarchical RGP structure. We name henceforth this new model as the RGP-*t* model and the corresponding modified variational framework as REVARB-*t*, whose details will be presented later on this paper.

We evaluate the proposed RGP-*t* model in system identification tasks with outliers using for that purpose several artificial benchmarks and datasets related to industrial systems. The impressive results obtained by the RGP-*t* model indicate that its powerful multilayer structure is able to learn the system dynamics from data contaminated with non-Gaussian noise and to perform simulation on unseen data better than previous robust GP models. Furthermore, we also illustrate how it can be used not only to deal with outliers (by not being influenced by them), but also to successfully detect them in the estimation data.

The remainder of the paper is organized as follows. Section 2 summarizes the standard GP modeling framework, highlighting how it can be directly applied to NARX models. In Section 3 we get to the proposed RGP-*t* model from the original GP-RLARX and RGP formulations. In Section 4 we detail REVARB-*t*, the modified variational framework that enables robust inference with RGP-*t*. In Section 5 we report and analyze the results obtained with RGP-*t*

[1] Commonly treated as a type of outlier by standard Gaussian models.
[2] A mathematical expression is considered tractable if its solution is analytical and, hence, do not require the use of approximate methods.