Latent Autoregressive Gaussian Process Models for Robust System Identification

César Lincoln C. Mattos * **Andreas Damianou** **
Guilherme A. Barreto * **Neil D. Lawrence** **

* *Federal University of Ceará, Dept. of Teleinformatics Engineering, Center of Technology, Campus of Pici, Fortaleza, Ceará, Brazil (e-mail: cesarlincoln@terra.com.br; gbarreto@ufc.br).*
** *Dept. of Computer Science & SITraN, The University of Sheffield, Sheffield, UK (e-mail: andreas.damianou@sheffield.ac.uk; N.Lawrence@dcs.sheffield.ac.uk)*

Abstract: We introduce GP-RLARX, a novel Gaussian Process (GP) model for robust system identification. Our approach draws inspiration from nonlinear autoregressive modeling with exogenous inputs (NARX) and it encapsulates a novel and powerful structure referred to as *latent* autoregression. This structure accounts for the feedback of uncertain values during training and provides a natural framework for free simulation prediction. By using a Student-t likelihood GP-RLARX can be used in scenarios where the estimation data contain non-Gaussian noise in the form of outliers. Further, a variational approximation scheme is developed to jointly optimize all the hyperparameters of the model from available estimation data. We perform experiments with five widely used artificial benchmarking datasets with different levels of outlier contamination and compare GP-RLARX with the standard GP-NARX model and its robust variant, GP-tVB. GP-RLARX is found to outperform the competing models by a relatively wide margin, indicating that our latent autoregressive structure is more suitable for robust system identification.

Keywords: Modelling and system identification, dynamic modelling, Gaussian process, outliers, autoregressive models.

1. INTRODUCTION

System identification is classically defined as the task of creating mathematical models of dynamical systems based on their inputs and observed outputs (Ljung, 1998). This general definition can be further complicated if we consider the analysis of nonlinear systems and *noisy* data, possibly containing outliers. In this paper we are interested in the later problem, which is very often encountered in practice.

In order to account for the uncertainty in the noisy data and in the dynamics learned by the model, we follow a Bayesian approach to system identification (Peterka, 1981). In this context, Gaussian Process (GP) models provide a principled, practical, probabilistic approach to learning in kernel machines (Rasmussen and Williams, 2006) and are the main subject of our work.

Since the early research on modeling dynamics with GPs, e.g. by Murray-Smith et al. (1999) and Solak et al. (2003), several contributions to GP-based system identification have been published, such as autoregressive models (Kocijan et al., 2005), non-stationary systems (Rottmann and Burgard, 2010), local modeling (Ažman and Kocijan, 2011), evolving models (Petelin et al., 2013) and state space models (Frigola et al., 2014).

Most work on GP-based system identification has been limited to the case of Gaussian noise, which implies a Gaussian likelihood. However, when one expects to have non-Gaussian observations in the form of outliers, such as impulsive noise, the estimation of the model's hyperparameters can be severely compromised. Furthermore, because of the nonparametric nature of the GP model, the estimation data is carried along the prediction phase, i.e. the estimation samples containing outliers and the misestimated hyperparameters will be used during the prediction stage, something which can deteriorate the model capability to generalize for unseen test data.

In (Mattos et al., 2015) we reviewed some recent work on GP regression in the presence of outliers. Such models replace the Gaussian likelihood by heavy-tailed distributions, such as Student-t and Laplace. While inference by GP models with Gaussian likelihood is tractable, non-Gaussian likelihood models are not, requiring the use of approximation methods, such as variational Bayes (VB) (Jordan et al., 1999) and expectation propagation (EP) (Minka, 2001). We then evaluated two robust models in the task of robust system identification: a GP model with Student-t likelihood and variational inference (GP-tVB) and a GP model with Laplace likelihood and EP inference (GP-LEP). The experimental results indicated that although the robust models performed better than the standard GP, especially GP-tVB, they were still sensitive to the outliers in some scenarios.

As in (Mattos et al., 2015), here we are interested in nonlinear autoregressive models with exogenous inputs (NARX) and in performance evaluation by free simulation

on test data. However, the autoregressive structure and the free simulation procedure result in the feedback of noisy values, and even outliers, into the model. Standard GP-based NARX models do not consider the additional uncertainty associated with this kind of noise. The standard variational approximation used in GP-tVB, detailed by Tipping and Lawrence (2005) and Kuss (2006), covers only the intractabilities from a non-Gaussian likelihood, but we are still left with intractabilities from uncertain inputs.

The problem of GP modeling with uncertain inputs has been studied before. Girard et al. (2003) consider prediction under uncertain inputs and apply it to multi-step ahead prediction. In (McHutchon and Rasmussen, 2011) a Taylor expansion is proposed to tackle the problem of training with noisy inputs. Frigola and Rasmussen (2013) propose a model with an integrated pre-processing step and optimize the filter's and the model's hyperparameters jointly. More recently, the thesis of Damianou (2015) has considered the broad problem of uncertainty propagation by variational techniques, inspired by the framework of Bayesian GP latent variable models (GP-LVM) (Titsias and Lawrence, 2010).

Following ideas presented in those recent works, we introduce a new robust latent NARX model named GP-RLARX. Our model includes a Student-t likelihood and a modified variational approach which enables it to learn dynamics from data containing outliers and account for uncertainty in both training and test steps. Moreover, the latent autogressive structure of GP-RLARX is able to tackle the problem of noisy feedback. We compare the proposed approach to standard GP-NARX with Gaussian likelihood and the mentioned GP-tVB model in five artificial datasets with different levels of corruption by outliers.

The remainder of the paper is organized as follows. In Section 2 we briefly describe the standard NARX model. In Section 3 we explain GPs for regression tasks. In Section 4 we summarize robust GP modeling and the GP-tVB model. In Section 5 we introduce our new approach, GP-RLARX. In Section 6 we report the results of the performance evaluation on computational experiments. We conclude the paper in Section 7.

2. SYSTEM IDENTIFICATION WITH AUTOREGRESSIVE MODELS

A standard NARX model considers an input vector $\boldsymbol{x}_i \in \mathbb{R}^D$ comprised of L_y past observed outputs $y_i \in \mathbb{R}$ and L_u past exogenous inputs $u_i \in \mathbb{R}$ (Kocijan et al., 2005):

$$\boldsymbol{x}_i = [y_{i-1}, \cdots, y_{i-L_y}, u_{i-1}, \cdots, u_{i-L_u}]^\top, \tag{1}$$

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i^{(y)}, \qquad \epsilon_i^{(y)} \sim \mathcal{N}(\epsilon_i^{(y)}|0, \sigma_y^2), \tag{2}$$

where i is the instant of observation, $f(\cdot)$ is an unknown nonlinear function and $\epsilon_i^{(y)}$ is a zero mean Gaussian observation noise with variance σ_y^2. After N instants, we have the dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N = (\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ is the so-called *regressor matrix* and $\boldsymbol{y} \in \mathbb{R}^N$.

Although one could use the estimated model of an identified system for one-step-ahead prediction, i.e. use previous inputs and observed outputs until the present instant to predict the output in the next instant, this validation methodology can be misleading, because even "poor models can look good" (Billings, 2013).

We follow the more adequate procedure where only given inputs and past estimated outputs are used for prediction, which is called *free simulation*, *infinite step ahead prediction* or *model predicted output* (Billings, 2013).

For the rest of the paper we focus on single exogenous input and one-dimensional output systems, but the multiple inputs case is straightforward and the multivariate output can be implemented using separate models for each output.

3. GP FOR REGRESSION

In the GP framework, a MISO (multiple input single output) nonlinear function $f(\cdot)$, as in Eq. 2, is assigned a multivariate Gaussian prior:

$$\boldsymbol{f} = f(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{f}|\boldsymbol{0}, \boldsymbol{K}), \tag{3}$$

where a zero mean vector was considered, $\boldsymbol{f} \in \mathbb{R}^N$ and $\boldsymbol{K} \in \mathbb{R}^{N \times N}$ is the covariance matrix. Each element of $\boldsymbol{K}$ is obtained from $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where $k(\cdot, \cdot)$ is a *kernel* function, which must generate a semidefinite positive matrix. In our experiments we use the so-called squared exponential kernel:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d^2 (x_{id} - x_{jd})^2\right], \tag{4}$$

where the vector $\boldsymbol{\theta} = [\sigma_f^2, w_1^2, \ldots, w_D^2]^\top$ is comprised of the hyperparameters which characterize the covariance of the model.

If we consider a Gaussian likelihood $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_y^2 \boldsymbol{I})$, where $\boldsymbol{I}$ is a $N \times N$ identity matrix, the posterior distribution $p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X})$ is tractable and the inference for a new output f_*, given a new input $\boldsymbol{x}_*$, is calculated by:

$$\begin{aligned} p(f_*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}_*) &= \mathcal{N}(f_*|\mu_*, \sigma_*^2), \\ \mu_* &= \boldsymbol{k}_{*N}(\boldsymbol{K} + \sigma_y^2 \boldsymbol{I})^{-1} \boldsymbol{y}, \\ \sigma_*^2 &= k_{**} - \boldsymbol{k}_{*N}(\boldsymbol{K} + \sigma_y^2 \boldsymbol{I})^{-1} \boldsymbol{k}_{N*}, \end{aligned} \tag{5}$$

where $\boldsymbol{k}_{*N} = [k(\boldsymbol{x}_*, \boldsymbol{x}_1), \cdots, k(\boldsymbol{x}_*, \boldsymbol{x}_N)]$, $\boldsymbol{k}_{N*} = \boldsymbol{k}_{*N}^\top$ and $k_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$. The predictive distribution of y_* is similar to the one in Eq. (5), but σ_y^2 is added to the variance.

The vector of hyperparameters $\boldsymbol{\theta}$ can be extended to include the noise variance σ_y^2 and be determined with the maximization of the marginal log-likelihood $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ of the observed data, the so-called *evidence* of the model:

$$\begin{aligned} \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = &-\frac{1}{2}\log|\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}| \\ &- \frac{1}{2}\boldsymbol{y}^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \boldsymbol{y} - \frac{N}{2}\log(2\pi). \end{aligned} \tag{6}$$

The optimization is guided by the gradients of the marginal log-likelihood with respect to each component of the vector $\boldsymbol{\theta}$. The optimization of the hyperparameters can be seen as the model selection step of obtaining a plausible GP model from the estimation data.

4. ROBUST GP WITH STUDENT-T LIKELIHOOD

The Gaussian likelihood cannot deal with outliers, due its light tails. An alternative is to consider a likelihood with heavy tails, such as the Student-t or Laplace distributions,

since they are able to account for extreme values. In this paper we focus on the Student-t distribution. However, by changing the Gaussian likelihood, the standard GP equations become intractable.

In (Kuss, 2006), a variational framework is used to tackle the intractability. First, the Student-t likelihood is rewritten as follows [1]:

$$p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{\tau}^{-1}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f},\mathrm{diag}(\boldsymbol{\tau}^{-1})), \tag{7}$$

$$p(\boldsymbol{\tau}|\alpha,\beta) = \prod_{i=1}^{N}\Gamma(\tau_i|\alpha,\beta), \tag{8}$$

where $\boldsymbol{\tau} \in \mathbb{R}^N$ are the precisions (inverse variances), $\mathrm{diag}(\cdot)$ builds a diagonal matrix from a vector and τ_i has a gamma prior with hyperparameters α and β.

The joint posterior of $\boldsymbol{f}$ and $\boldsymbol{\tau}$ is considered to be factorized as

$$p(\boldsymbol{f},\boldsymbol{\tau}|\boldsymbol{y},\boldsymbol{X}) \approx q(\boldsymbol{f})q(\boldsymbol{\tau}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{m},\boldsymbol{A})\prod_{i=1}^{N}\Gamma(\tau_i|\hat{\alpha}_i,\hat{\beta}_i), \tag{9}$$

where $\boldsymbol{m} \in \mathbb{R}^N$, $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and $\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}} \in \mathbb{R}^N$ are unknown variational parameters.

A lower bound $\mathcal{L}(q(\boldsymbol{f})q(\boldsymbol{\tau}))$ to the log-marginal likelihood can be found relating it to the factorized approximate posterior $q(\boldsymbol{f})q(\boldsymbol{\tau})$ (Tipping and Lawrence, 2005):

$$\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = \mathcal{L}(q(\boldsymbol{f})q(\boldsymbol{\tau}))+\mathrm{KL}(q(\boldsymbol{f})q(\boldsymbol{\tau})||p(\boldsymbol{f},\boldsymbol{\tau}|\boldsymbol{y},\boldsymbol{X})), \tag{10}$$

where the last term is the Kullback-Leibler (KL) divergence between the approximate distribution and the true posterior. The maximization of the bound $\mathcal{L}(q(\boldsymbol{t})q(\boldsymbol{\tau}))$ also minimizes the KL term, improving the approximation (Tipping and Lawrence, 2005). We name this robust GP model with Student-t likelihood the GP-tVB model.

The optimal values of $\boldsymbol{m}$ and $\boldsymbol{A}$ can be written in terms of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, which themselves are optimized with the help of the gradients of the bound, as detailed by Kuss (2006). Then, the moments of the prediction $p(f_*|\boldsymbol{y},\boldsymbol{X},\boldsymbol{x}_*) = \mathcal{N}(f_*|\mu_*,\sigma_*^2)$ for a new input $\boldsymbol{x}_*$ are given by

$$\mu_* = \boldsymbol{k}_{*N}(\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{y}, \quad \text{and} \quad \sigma_*^2 = k_{**}-\boldsymbol{k}_{*N}(\boldsymbol{K}+\boldsymbol{\Sigma})\boldsymbol{k}_{N*}, \tag{11}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\hat{\boldsymbol{\beta}}/\hat{\boldsymbol{\alpha}})$.

5. ROBUST LATENT AUTOREGRESSIVE GP

In the present work we propose an alternative to the NARX approach where the autoregressive vector contains, besides the control inputs, only latent variables:

$$x_i = f(x_{i-1},\cdots,x_{i-L_x}u_{i-1},\cdots,u_{i-L_u}) + \epsilon_i^{(x)}, \tag{12}$$

$$y_i = x_i + \epsilon_i^{(y)}, \tag{13}$$

$$\epsilon_i^{(x)} \sim \mathcal{N}(\epsilon_i^{(x)}|0,\sigma_x^2), \tag{14}$$

$$\epsilon_i^{(y)} \sim \mathcal{N}(\epsilon_i^{(y)}|0,\tau_i^{-1}), \quad \tau_i \sim \Gamma(\tau_i|\alpha,\beta), \tag{15}$$

where L_x is the number of considered past latent variables and the observation noise $\epsilon_i^{(y)}$ follows a Student-t distribution, which is once again written as a mixture of Gaussians with gamma distributed precisions.

[1] In Kuss (2006) an inverse gamma prior is chosen for the variance of the Guassian distribution, which is equivalent to our analysis.

We emphasize that Eqs. 12 and 13 are distinct from Eqs. 1 and 2, which are used in the GP-tVB. In our model the autoregression is made with the latent (unobserved) variable x_i, instead of the observed outputs y_i. This feature avoids the feedback of possibly corrupted observations into the dynamics.

Differently from the inputs of standard NARX models, the latent variables x_i have a probability distribution, which allows the propagation of uncertainty during free simulation, as will be explained later.

Moreover, the separate transition and observation functions allow the use of distinct noise models, a Gaussian for the transition and a Student-t for the observation. Henceforth, we call this proposed robust latent autoregressive with exogenous inputs approach the GP-RLARX model.

The features proposed for GP-RLARX make it more powerful, but also introduce additional intractabilities not covered by the variational framework of GP-tVB. These additional intractabilities come from the difficulty in propagating uncertainty of latent inputs through the nonlinear GPs. To overcome this issue, we build on the variational approach of the Bayesian GP latent variable model (Titsias and Lawrence, 2010) and extend it to account for the autoregressive structure and the Student-t likelihood of our model, as explained in the next section.

5.1 Variational Lower Bound

We can rewrite Eqs. 12-15 in terms of distributions:

$$p(\boldsymbol{f}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{0},\boldsymbol{K}_f), \tag{16}$$

$$p(x_i) = \mathcal{N}(x_i|\mu_i^{(0)},\lambda_i^{(0)}), \quad \forall i \in \{1,\cdots,l_x\}, \tag{17}$$

$$p(x_i|f_i) = \mathcal{N}(x_i|f_i,\sigma_x^2), \quad \forall i \in \{l_x+1,\cdots,N\}, \tag{18}$$

$$p(y_i|x_i,\tau_i) = \mathcal{N}(y_i|x_i,\tau_i^{-1}), \quad \forall i \in \{l_x+1,\cdots,N\}, \tag{19}$$

$$p(\tau_i) = \Gamma(\tau_i|\alpha,\beta), \quad \forall i \in \{l_x+1,\cdots,N\}, \tag{20}$$

where $\boldsymbol{K}_f$ is the covariance matrix of the GP and we have put Gaussian priors to the initial latent variables $x_i|_{i=1}^{l_x}$. The joint distribution of all the variables is given by

$$p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{f},\boldsymbol{\tau}) = \left(\prod_{i=l_x+1}^{N} p(y_i|x_i,\tau_i)p(\tau_i) \right. \\ \left. p(x_i|f_i)p(f_i|\bar{\boldsymbol{x}}_i)\right)\prod_{i=1}^{l_x}p(x_i), \tag{21}$$

where we introduce the notation $\bar{\boldsymbol{x}}_i = [x_{i-1},\ldots,x_{i-l_x},u_{i-1},\ldots,u_{i-l_u}]^\top$. It can be seen from the joint distribution that we can not integrate out the latent variables $\boldsymbol{x}$, since they appear in the terms $|\boldsymbol{K}_f|$ and $\boldsymbol{K}_f^{-1}$ in $p(\boldsymbol{f}|\boldsymbol{x})$.

We tackle the intractabilities of our model with the variational sparse framework introduced by Titsias (2009). First, we include M inducing points $\boldsymbol{z} \in \mathbb{R}^M$ evaluated in M pseudo-inputs $\boldsymbol{x}_i^{(z)}|_1^M \in \mathbb{R}^D$ so that $\boldsymbol{z}$ are extra samples of the GP that models $f(\cdot)$, i.e. $\boldsymbol{f}$ and $\boldsymbol{z}$ are jointly Gaussian. Thus, we replace in Eq. 21 the term $p(f_i|\bar{\boldsymbol{x}}_i)$ by $p(f_i|\boldsymbol{z},\bar{\boldsymbol{x}}_i)p(\boldsymbol{z})$. Note that if we integrate out $\boldsymbol{z}$ we recover exactly the original expression.

By applying Jensen's inequality within the variational approach, we can obtain a lower bound to the log-likelihood

$p(\boldsymbol{y})$ (Bishop, 2006):

$$p(\boldsymbol{y}) \geq \int Q \log \left[\frac{p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{f}, \boldsymbol{z}, \boldsymbol{\tau})}{Q} \right] d\boldsymbol{x} d\boldsymbol{f} d\boldsymbol{z} d\boldsymbol{\tau}, \quad (22)$$

where Q is the variational distribution. We choose $Q = q(\boldsymbol{x})q(\boldsymbol{z})q(\boldsymbol{\tau}) \prod_{i=l_y+1}^{N} p(f_i|\boldsymbol{z}, \bar{\boldsymbol{x}}_i)$, such that

$$q(\boldsymbol{x}) = \prod_{i=1}^{N} q(x_i) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu_i^{(x)}, \lambda_i^{(x)}), \quad (23)$$

$$q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}^{(z)}, \boldsymbol{\Sigma}^{(z)}), \quad (24)$$

$$q(\boldsymbol{\tau}) = \prod_{i=l_x+1}^{N} q(\tau_i) = \prod_{i=l_x+1}^{N} \Gamma(\tau_i|\hat{a}_i, \hat{b}_i), \quad (25)$$

$$p(f_i|\boldsymbol{z}, \bar{\boldsymbol{x}}_i) = \mathcal{N}(f_i|[\boldsymbol{a}_f]_i, [\boldsymbol{\Sigma}_f]_{ii}), \quad (26)$$

$$\boldsymbol{a}_f = \boldsymbol{K}_{fz}\boldsymbol{K}_z^{-1}\boldsymbol{z},$$

$$\boldsymbol{\Sigma}_f = \boldsymbol{K}_f - \boldsymbol{K}_{fz}\boldsymbol{K}_z^{-1}\boldsymbol{K}_{fz}^{\top},$$

where $\mu_i^{(x)}$, $\lambda_i^{(x)}$, $\boldsymbol{\mu}^{(z)}$, $\boldsymbol{\Sigma}^{(z)}$, $\hat{a}_i$ and $\hat{b}_i$ are variational parameters which will be optimized with the maximization of the lower bound, $\boldsymbol{K}_z \in \mathbb{R}^{M \times M}$ is the sparse kernel matrix calculated from $\boldsymbol{x}^{(z)}$ and $\boldsymbol{K}_{fz} = k(\bar{\boldsymbol{x}}, \boldsymbol{x}^{(z)}) \in \mathbb{R}^{N \times M}$ is the cross-covariance matrix calculated from $\boldsymbol{x}$ and $\boldsymbol{x}^{(z)}$.

We replace the factorized variational distribution Q back to Eq. 22 and by working the expressions with a strategy similar to (Titsias and Lawrence, 2010) we are able to optimally eliminate the variational parameters $\boldsymbol{\mu}^{(z)}$ and $\boldsymbol{\Sigma}^{(z)}$ and obtain the final form of the lower bound:

$$\begin{aligned}
p(\boldsymbol{y}) \geq & -\frac{N - l_x}{2} \log 2\pi\sigma_x^2 + \frac{1}{2} \sum_{i=l_x+1}^{N} (\psi(\hat{a}_i) - \log \hat{b}_i) \\
& - \frac{1}{2} \sum_{i=l_x+1}^{N} \left(\frac{\hat{a}_i}{\hat{b}_i} \left(y_i^2 - 2y_i\mu_i^{(x)} + \lambda_i^{(x)} + (\mu_i^{(x)})^2 \right) \right) \\
& - \frac{1}{2\sigma_x^2} \left(\sum_{i=1}^{N} \left(\lambda_i^{(x)} + (\mu_i^{(x)})^2 \right) + \Psi_0 - \mathrm{Tr}(\boldsymbol{K}_z^{-1}\boldsymbol{\Psi}_2) \right) \\
& + \frac{1}{2}|\boldsymbol{K}_z| - \frac{1}{2} \left| \boldsymbol{K}_z + \frac{1}{\sigma_x^2}\boldsymbol{\Psi}_2 \right| \\
& + \frac{1}{2(\sigma_x^2)^2} \boldsymbol{\mu}_x^{\top} \boldsymbol{\Psi}_1 \left(\boldsymbol{K}_z + \frac{1}{\sigma_x^2}\boldsymbol{\Psi}_2 \right)^{-1} \boldsymbol{\Psi}_1^{\top} \boldsymbol{\mu}_x \\
& - \sum_{i=l_x+1}^{N} \int q(x_i) \log q(x_i) dx_i \\
& + \sum_{i=1}^{l_x} \int q(x_i) \log p(x_i) dx_i - \mathrm{KL}(q(\boldsymbol{\tau}) \,\|\, p(\boldsymbol{\tau})), \quad (27)
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function, the last term is the KL divergence between two gamma distributions, $\Psi_0 = \mathrm{Tr}(\langle \boldsymbol{K}_f \rangle_{q(\boldsymbol{x})})$, $\boldsymbol{\Psi}_1 = \langle \boldsymbol{K}_{fz} \rangle_{q(\boldsymbol{x})}$ and $\boldsymbol{\Psi}_2 = \langle \boldsymbol{K}_{fz}^{\top} \boldsymbol{K}_{fz} \rangle_{q(\boldsymbol{x})}$, where $\langle \cdot \rangle_{q(\boldsymbol{x})}$ denotes expectation with respect to the distribution $q(\boldsymbol{x})$. When a squared exponential kernel is chosen, all the expectations can be solved in closed form, as detailed by Titsias and Lawrence (2010). The lower bound includes some entropy integrals in the last two lines of Eq. 27, which involve Gaussians and are, hence, also tractable.

It is worth mentioning that the second line in the Eq. 27 shows the only term that includes the observations. The value of each y_i is weighted by the fraction $\frac{\hat{a}_i}{\hat{b}_i}$, which comes from the expectation of a gamma distribution. For observations containing outliers, the value of this fraction is much lower than regular observations, reducing their influence in the bound. Thus, the inspection of the optimized variational parameters *per se* can be used as a method to detect outliers in the estimation data.

The kernel hyperparameters and the variational parameters, including the pseudo-inputs $\boldsymbol{x}^{(z)}$, are jointly optimized by maximization of the lower bound, using the analytical gradients of Eq. 27.

5.2 Predictions with Approximate Uncertainty Propagation

Although the computation of the exact predictive distribution is intractable, since the input is uncertain, we can calculate its moments. Given a new regressor vector $\bar{\boldsymbol{x}}_*$ obtained from past latent variables and a sequence of external control inputs, the mean and variance of the prediction f_* are calculated by following the results presented by Girard et al. (2003):

$$\mathbb{E}\{p(f_*|\bar{\boldsymbol{x}}_*)\} = \boldsymbol{B}^{\top}(\boldsymbol{\Psi}_1^*)^{\top}, \quad (28)$$

$$\begin{aligned}
\mathbb{V}\{p(f_*|\bar{\boldsymbol{x}}_*)\} = & \boldsymbol{B}^{\top}(\boldsymbol{\Psi}_2^* - (\boldsymbol{\Psi}_1^*)^{\top}\boldsymbol{\Psi}_1^*)\boldsymbol{B} + \Psi_0^* \\
& - \mathrm{Tr}((\boldsymbol{K}_{zz}^{-1} - (\boldsymbol{K}_{zz} + \sigma_x^{-2}\boldsymbol{\Psi}_2)^{-1})\boldsymbol{\Psi}_2^*), \quad (29)
\end{aligned}$$

where $\boldsymbol{B} = \sigma_x^{-2}(\boldsymbol{K}_{zz} + \sigma_x^{-2}\boldsymbol{\Psi}_2)^{-1}\boldsymbol{\Psi}_1^{\top}\boldsymbol{y}$, $\Psi_0^* = \mathrm{Tr}(\langle \boldsymbol{K}_* \rangle_{q(\boldsymbol{x}_*)})$, $\boldsymbol{\Psi}_1^* = \langle \boldsymbol{K}_{*z} \rangle_{q(\boldsymbol{x}_*)}$ and $\boldsymbol{\Psi}_2^* = \langle \boldsymbol{K}_{*z}^{\top}\boldsymbol{K}_{*z} \rangle_{q(\boldsymbol{x}_*)}$.

As opposed to GP-tVB and other NARX models, our framework allows for a natural way of approximate propagation of the uncertainty during free simulation, since it uses the distribution $q(\boldsymbol{x}_*)$ of the input, instead of just its mean value or single point estimates.

6. EXPERIMENTS

We perform computational experiments on five artificial datasets which constitute popular benchmarks. The first four were presented in the seminal work of Narendra and Parthasarathy (1990). The fifth was generated following Kocijan et al. (2005). Tab. 1 summarizes the datasets.

Besides the Gaussian noise indicated in Tab. 1 we further corrupt the estimation data with different amounts of outliers equal to 0%, 5%, 10%, 15%, 20%, 25% and 30% of the estimation samples. The outliers were sampled from $\sigma(\boldsymbol{y}) \times \mathcal{T}(0, 2)$, where $\sigma(\boldsymbol{y})$ is the standard deviation of the original estimation data and $\mathcal{T}(0, 2)$ is a Student-t distribution with zero mean and 2 degrees of freedom. We emphasize that, although we use the same datasets of Mattos et al. (2015), the results can not be directly compared due the different outlier generation scheme.

We evaluate the performances of the following GP models: standard GP-NARX with Gaussian likelihood, GP with Student-t likelihood and VB inference, GP-tVB, and the proposed robust latent autoregressive model, GP-RLARX.

The orders L_u and L_y (or L_x for GP-RLARX) chosen for the regressors were set to their largest delays presented in the second column of Tab. 1. For GP-RLARX we use a number of pseudo-inputs $\boldsymbol{x}^{(z)}$ equal to 15% of the estimation data size and initialize them by applying the K-Means clustering algorithm in the regressor matrix.

Table 1. Details of the five artificial datasets used in the computational experiments. Note that $\mathrm{U}(A, B)$ is a random number uniformly distributed between A and B.

#	Output	Input/Samples		Noise
		Estimation	**Test**	
1	$y_i = \frac{y_{i-1}y_{i-2}(y_{i-1}+2.5)}{1+y_{i-1}^2+y_{i-2}^2} + u_{i-1}$	$u_i = \mathrm{U}(-2, 2)$ 300 samples	$u_i = \sin(2\pi i/25)$ 100 samples	$\mathcal{N}(0, 0.29)$
2	$y_i = \frac{y_{i-1}}{1+y_{i-1}^2} + u_{i-1}^3$	$u_i = \mathrm{U}(-2, 2)$ 300 samples	$u_i = \sin(2\pi i/25) + \sin(2\pi i/10)$ 100 samples	$\mathcal{N}(0, 0.65)$
3	$y_i = 0.8y_{i-1} + (u_{i-1} - 0.8)u_{i-1}(u_{i-1} + 0.5)$	$u_i = \mathrm{U}(-1, 1)$ 300 samples	$u_i = \sin(2\pi i/25)$ 100 samples	$\mathcal{N}(0, 0.07)$
4	$y_i = 0.3y_{i-1} + 0.6y_{i-2} + 0.3\sin(3\pi u_{i-1}) + 0.1\sin(5\pi u_{i-1})$	$u_i = U(-1, 1)$ 500 samples	$u_i = \sin(2\pi i/250)$ 500 samples	$\mathcal{N}(0, 0.18)$
5	$y_i = y_{i-1} - 0.5\tanh(y_{i-1} + u_{i-1}^3)$	$u_i = \mathcal{N}(u_i\|0, 1)$ $-1 \leq u_i \leq 1$ 150 samples	$u_i = \mathcal{N}(u_i\|0, 1)$ $-1 \leq u_i \leq 1$ 150 samples	$\mathcal{N}(0, 0.0025)$

The obtained root mean square errors (RMSE) are presented in Fig. 1. As expected, both GP-tVB and GP-RLARX were better than GP-NARX in the scenarios containing outliers. However, for most cases, with the exception of some scenarios of *Artificial 1* and *2* datasets, GP-tVB obtained considerably worse RMSE after the addition of outliers, when compared to the outlier-free experiments.

On the other hand, GP-RLARX was able to avoid large error increments even with the addition of several outliers and this behavior is observed for all datasets. In fact, GP-RLARX is the only method that achieved an almost constant RMSE in three of the datasets, and its gains compared to baselines are significant in almost all cases.

The impressive results obtained by GP-RLARX indicate that in many scenarios, only the inclusion of a Student-t likelihood, as in GP-tVB, is not enough. The proposed latent autoregressive structure of our model showed to have a great impact on its capability to learn dynamics from data containing outliers and perform free simulation.

7. CONCLUSION

We presented GP-RLARX, a new robust GP model for system identification which includes a Student-t likelihood and a latent autoregressive structure. We also introduced a modified variational approximation framework to tackle the intractabilities of our model and to enable it to account for uncertainty during both training and free-simulation. In several computational experiments GP-RLARX obtained better overall performance over standard GP-NARX and its robust variant GP-tVB.

In future work we intend to explore variations of the base structure of GP-RLARX, for example by including a nonlinear observation function modeled by a second GP prior between the latent variable x_i and the output y_i to further increase its flexibility.

ACKNOWLEDGEMENTS

The authors thank the financial support of NUTEC, FUNCAP, CAPES, CNPq (grant no. 309841/2012-7) and the EU (research project FP7-ICT).

REFERENCES

Ažman, K. and Kocijan, J. (2011). Dynamical systems identification using Gaussian process models with incorporated local models. *Eng Appl Artif Intel*, 24(2), 398–408.

Billings, S.A. (2013). *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer.

Damianou, A. (2015). *Deep Gaussian processes and variational propagation of uncertainty*. Ph.D. thesis, University of Sheffield.

Frigola, R., Chen, Y., and Rasmussen, C. (2014). Variational Gaussian process state-space models. In *NIPS*, 3680–3688.

Frigola, R. and Rasmussen, C.E. (2013). Integrated preprocessing for Bayesian nonlinear system identification with gaussian processes. In *IEEE CDC*, 5371–5376.

Girard, A., Rasmussen, C., Quiñonero-Candela, J., and Murray-Smith, R. (2003). Multiple-step ahead prediction for non linear dynamic systems: A Gaussian process treatment with propagation of the uncertainty. In *NIPS*, 529–536. MIT Press, Cambridge, MA, USA.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2), 183–233.

Kocijan, J., Girard, A., Banko, B., and Murray-Smith, R. (2005). Dynamic systems identification with Gaussian processes. *Math Comp Model Dyn*, 11(4), 411–424.

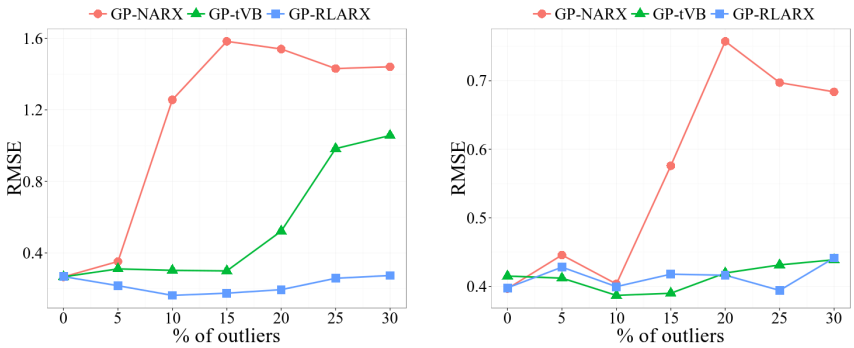
Kuss, M. (2006). *Gaussian process models for robust regression, classification, and reinforcement learning*. Ph.D. thesis, TU Darmstadt.

Ljung, L. (1998). *System identification*. Springer.

Mattos, C.L.C., Santos, J.D.A., and Barreto, G.A. (2015). An empirical evaluation of robust Gaussian process models for system identification. In *16th International Conference on Intelligent Data Engineering and Automated Learning (to appear)*.

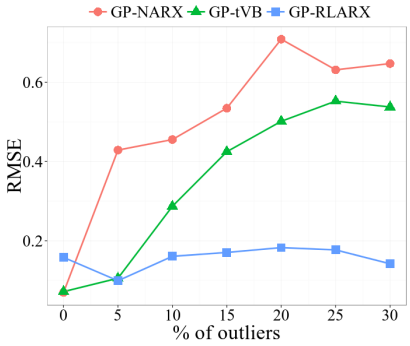
McHutchon, A. and Rasmussen, C.E. (2011). Gaussian process training with input noise. In *NIPS*, 1341–1349.

Minka, T.P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th UAI)*, 362–369. Morgan Kaufmann.

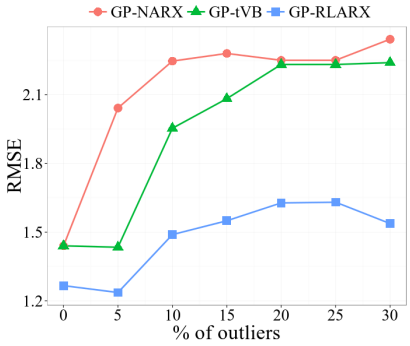


(a) Artificial 1.

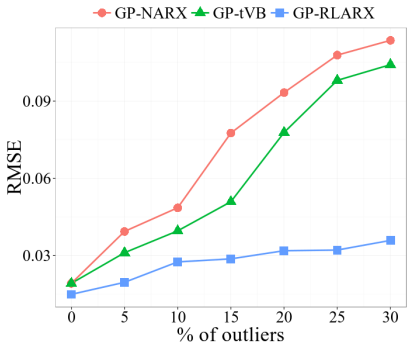
(b) Artificial 2.



(c) Artificial 3.



(d) Artificial 4.



(e) Artificial 5.

Fig. 1. RMSE values for free simulation on test data with different levels of contamination by outliers.

Murray-Smith, R., Johansen, T.A., and Shorten, R. (1999). On transient dynamics, off-equilibrium behaviour and identification in blended multiple model structures. In *European Control Conference (ECC'99), Karlsruhe, BA-14*. Springer.

Narendra, K.S. and Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE T Neural Networ*, 1(1), 4–27.

Petelin, D., Grancharova, A., and Kocijan, J. (2013). Evolving Gaussian process models for prediction of ozone concentration in the air. *Simul Model Pract Th*, 33, 68–80.

Peterka, V. (1981). Bayesian approach to system identification. *Trends and Prog in Syst ident*, 1, 239–304.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press, 1 edition.

Rottmann, A. and Burgard, W. (2010). Learning non-stationary system dynamics online using Gaussian processes. In *Pattern Recognition*, volume 6373 of *Lecture Notes in Computer Science*, 192–201. Springer.

Solak, E., Murray-Smith, R., Leithead, W.E., Leith, D.J., and Rasmussen, C.E. (2003). Derivative observations in Gaussian process models of dynamic systems. *NIPS*, 16.

Tipping, M.E. and Lawrence, N.D. (2005). Variational inference for student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(1), 123–141.

Titsias, M.K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*, 567–574.

Titsias, M.K. and Lawrence, N.D. (2010). Bayesian Gaussian process latent variable model. In *AISTATS*, 844–851.