An Empirical Evaluation of Robust Gaussian Process Models for System Identification

César Lincoln C. Mattos[1], José Daniel A. Santos[2] and Guilherme A. Barreto[3]

Federal University of Ceará, Department of Teleinformatics Engineering,
Center of Technology, Campus of Pici, Fortaleza, Ceará, Brazil,
[1]cesarlincoln@terra.com.br, [3]gbarreto@ufc.br
Federal Institute of Education, Science and Technology of Ceará,
Department of Industry, Maracanaú, Ceará, Brazil,
[2]jdalencars@gmail.com,

Abstract. System identification comprises a number of linear and nonlinear tools for black-box modeling of dynamical systems, with applications in several areas of engineering, control, biology and economy. However, the usual Gaussian noise assumption is not always satisfied, specially if data is corrupted by impulsive noise or outliers. Bearing this in mind, the present paper aims at evaluating how Gaussian Process (GP) models perform in system identification tasks in the presence of outliers. More specifically, we compare the performances of two existing robust GP-based regression models in experiments involving five benchmarking datasets with controlled outlier inclusion. The results indicate that, although still sensitive in some degree to the presence of outliers, the robust models are indeed able to achieve lower prediction errors in corrupted scenarios when compared to conventional GP-based approach.

Keywords: robust system identification, Gaussian process, approximate Bayesian inference.

1 Introduction

Gaussian processes (GPs) provide a principled, practical, probabilistic approach to learning in kernel machines [1]. Due to is versatility, GP models is receiving considerable attention from the Machine Learning community, leading to successful applications to classification and regression [2], visualization of high dimensional data [3] and system identification [4], to mention just a few.

Of particular interest to the present paper is the application of GP models to nonlinear system identification, which comprises a number of linear and nonlinear tools for black-box modeling of dynamical systems. Contributions to GP-based system identification seem to have started with the work of Murray-Smith et al. [5], who applied it to vehicle dynamics data. Since then, a number of interesting approaches can be found in the literature, such as GP with derivative observations [6], GP for learning non-stationary systems [7], GP-based local models [8], evolving GP models [9], and GP-based state space models [10].

Nevertheless, these previous GP-based system identification approaches have adopted the Gaussian likelihood function as noise model. However, as a light-tailed distribution, this function is not able to suitably handle impulsive noise (a type of outlier). When such outliers are encountered in the data used to tune the model's hyperparameters, these are not correctly estimated. Besides, as a nonparametric approach, the GP model carries the estimation data along for prediction purpose, i.e. the estimation samples containing outliers and the mis-estimated hyperparameters will be used during out-of-sample prediction stage, a feature that may compromise the model generalization on new data.

In this scenario, heavy-tailed distributions are claimed to be more appropriate as noise models when outliers are present. Such distributions are able to account for, or justify, extreme values, as they have higher probability to occur than in light-tailed distributions. This feature prevents the estimation step from being too affected by outliers. However, while inference by GP models with Gaussian likelihood is tractable, non-Gaussian likelihoods models are not, requiring the use of approximation methods, such as Variational Bayes (VB) [**?**] and Expectation Propagation (EP) [11].

Robust GP regression started to draw the machine learning community attention more recently. In Faul and Tipping [12], impulsive noise is modeled as being generated by a second Gaussian distribution with larger variance, resulting in a mixture of Gaussian noise models. Inference is done with the VB method. In Kuss et al. [13], a similar noise model is chosen, but the inference makes use of the EP strategy. In Tipping and Lawrence [14], GP models with Student-t likelihood are also considered in a variational context. The same likelihood is used in Jylänki et al. [15], but it is tackled by a Laplace approximation approach. The same approach is used in Berger and Rauscher [16] to calibrate a diesel engine from data containing outliers. In Kuss' thesis [17], besides reviewing some of the approaches for robust GP regression, a Laplacian noise model is detailed and tackled by an EP-based inference strategy.

From the exposed, the goal of this work is to evaluate some of the aforementioned robust GP models in nonlinear dynamical system identification in the presence of outliers. More specifically, we apply a Student-t noise model likelihood with VB inference, as in [14], and a Laplace noise model with EP inference, following [17]. Our objective is to assess if such algorithms, originally proposed for robust regression, are able to achieve good performance in dynamical system identification scenarios contaminated with outliers and compare them with standard (i.e. non-robust) GP models that have been used in the system identification literature.

The remainder of the paper is organized as follows. In Section 2 we describe the task of nonlinear dynamical system identification with standard GP modeling and the two aforementioned robust variants. In Section 3 we report the results of the performance evaluation of the models for 5 artificial datasets with different levels of contamination with outliers. We conclude the paper in Section 4.

2 GP for Nonlinear Dynamical System Identification

Given a dynamical system modeled by a nonlinear autoregressive with exogenous inputs (NARX) model, its i-th input vector $\boldsymbol{x}_i \in \mathbb{R}^D$ is comprised of L_y past observed outputs $y_i \in \mathbb{R}$ and L_u past exogenous inputs $u_i \in \mathbb{R}$ [4]:

$$y_i = t_i + \epsilon_i, \qquad t_i = f(\boldsymbol{x}_i), \qquad \epsilon_i \sim \mathcal{N}(\epsilon_i|0, \sigma_n^2), \tag{1}$$

$$\boldsymbol{x}_i = [y_{i-1}, \cdots, y_{i-L_y}, u_{i-1}, \cdots, u_{i-L_u}]^T \tag{2}$$

where i is the instant of observation, $t_i \in \mathbb{R}$ is the true (noiseless) output, $f(\cdot)$ is an unknown nonlinear function and ϵ_i is a Gaussian observation noise. 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 $\boldsymbol{y} \in \mathbb{R}^N$.

An estimated model may be used to simulate the output of the identified system. We use a iterative test procedure where past estimated outputs are used as regressors, which is called *free simulation* or *infinite step ahead* prediction.

2.1 Traditional GP Modeling

In the GP framework, the nonlinear function $f(\cdot)$ is given a multivariate Gaussian prior $\boldsymbol{t} = f(\boldsymbol{X}) \sim \mathcal{GP}(\boldsymbol{t}|\boldsymbol{0}, \boldsymbol{K})$, where a zero mean vector was considered and $\boldsymbol{K} \in \mathbb{R}^{N \times N}$, $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, is the covariance matrix, obtained with a *kernel* function $k(\cdot, \cdot)$, which must generate a semidefinite positive matrix $\boldsymbol{K}$. The following function is a common choice and will be used in this paper [18]:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_f^2 \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d^2 (x_{id} - x_{jd})^2\right] + \sigma_l^2 \boldsymbol{x}_i^T \boldsymbol{x}_j + \sigma_c^2. \tag{3}$$

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

Considering a multivariate Gaussian likelihood $p(\boldsymbol{y}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{t}, \sigma_n^2 \boldsymbol{I})$, where $\boldsymbol{I}$ is a $N \times N$ identity matrix, the posterior distribution $p(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{X})$ is tractable. The inference for a new output t_*, given a new input $\boldsymbol{x}_*$, is also tractable

$$p(t_*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}_*) = \mathcal{N}(t_*|\boldsymbol{k}_{*N}(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{y}, k_{**} - \boldsymbol{k}_{*N}(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{k}_{N*}), \tag{4}$$

where $\boldsymbol{k}_{*N} = [k(\boldsymbol{x}_*, \boldsymbol{x}_1), \cdots, k(\boldsymbol{x}_*, \boldsymbol{x}_N)]$, $\boldsymbol{k}_{N*} = \boldsymbol{k}_{*N}^T$ and $k_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$. The predictive distribution of y_* is similar to the one in Eq. (4), but the variance is added by σ_n^2.

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

$$\boldsymbol{\theta}_* = \arg\max\left\{-\frac{1}{2}\ln|\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}| - \frac{1}{2}\boldsymbol{y}^T(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1}\boldsymbol{y} - \frac{N}{2}\log(2\pi)\right\}. \tag{5}$$

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

2.2 Robust GP with Non-Gaussian Likelihood

The previous GP model with Gaussian likelihood is not robust to outliers, due its light tails. An alternative is to consider a likelihood with heavy tails, such as the Laplace and the Student-t likelihoods, respectively given by

$$p_{\text{Lap}}(\boldsymbol{y}|\boldsymbol{t}) = \prod_{i=1}^{N} \frac{1}{2s} \exp\left(-\frac{|y_i - t_i|}{s}\right), \tag{6}$$

and

$$p_{\text{Stu}}(\boldsymbol{y}|\boldsymbol{t}) = \prod_{i=1}^{N} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu}\frac{(y_i - t_i)^2}{\sigma^2}\right)^{-(\nu+1)/2}, \tag{7}$$

where s, ν and σ^2 are likelihood hyperparameters and $\Gamma(\cdot)$ is the gamma function.

However, once a non-Gaussian likelihood is chosen, many of the GP expressions become intractable. In the present paper, we apply approximate Bayesian inference methods to overcome those intractabilities. More specifically, we are interested in the Variational Bayes and the Expectation Propagation algorithms, briefly presented below.

Variational Bayes (VB) In the case of applying VB to the Student-t likelihood, it must be rewritten as follows [17]:

$$p(\boldsymbol{y}|\boldsymbol{t}, \boldsymbol{\sigma}^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{t}, \text{diag}(\boldsymbol{\sigma}^2)), \qquad p(\boldsymbol{\sigma}^2|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{N} \text{Inv}\Gamma(\sigma_i^2|\alpha_i, \beta_i), \tag{8}$$

where $\boldsymbol{t}, \boldsymbol{\sigma}^2 \in \mathbb{R}^N$ are latent variables, $\text{diag}(\cdot)$ builds a diagonal matrix from a vector and σ_i^2 has an inverse gamma prior with parameters α_i and β_i.

The joint posterior of $\boldsymbol{t}$ and $\boldsymbol{\sigma}^2$ is considered to be factorizable as

$$p(\boldsymbol{t}, \boldsymbol{\sigma}^2|\boldsymbol{y}, \boldsymbol{X}) \approx q(\boldsymbol{t})q(\boldsymbol{\sigma}^2) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{m}, \boldsymbol{A}) \left(\prod_{i=1}^{N} \text{Inv}\Gamma(\sigma_i^2|\tilde{\alpha}_i, \tilde{\beta}_i)\right), \tag{9}$$

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

A lower bound $\mathcal{L}(q(\boldsymbol{t})q(\boldsymbol{\sigma}^2))$ to the log-marginal likelihood can be found relating it to the factorized posterior $q(\boldsymbol{t})q(\boldsymbol{\sigma}^2)$ [14]:

$$\ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{L}(q(\boldsymbol{t})q(\boldsymbol{\sigma}^2)) + \text{KL}(q(\boldsymbol{t})q(\boldsymbol{\sigma}^2)||p(\boldsymbol{t}, \boldsymbol{\sigma}^2|\boldsymbol{y}, \boldsymbol{X})), \tag{10}$$

where the last term is the Kullback-Leibler divergence between the approximate distribution and the true posterior. The maximization of the bound $\mathcal{L}(q(\boldsymbol{t})q(\boldsymbol{\sigma}^2))$ also minimizes the KL divergence term, improving the approximation [14].

The optimization of the hyperparameters and the latent variables can be done in an Expectation-Maximization (EM) fashion, as detailed in [17]. Then,

the moments of the prediction $p(t_*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}_*) = \mathcal{N}(t_*|\mu_*, \sigma_*^2)$ 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 $\boldsymbol{\Sigma} = \text{diag}(\tilde{\boldsymbol{\beta}}/\tilde{\boldsymbol{\alpha}})$. Although the calculation of the predictive distribution of y_* is intractable, its mean is equal to the previously calculated μ_*.

Expectation Propagation (EP) EP usually works by approximating the true posterior distribution by a Gaussian which follows a factorized structure [11, 17]:

$$p(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{X}) \approx \frac{\mathcal{N}(\boldsymbol{t}|\boldsymbol{0}, \boldsymbol{K})}{q(\boldsymbol{y}|\boldsymbol{X})} \prod_{i=1}^{N} c(t_i, \mu_i, \sigma_i^2, Z_i) = q(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{m}, \boldsymbol{A}), \tag{12}$$

where $c(t_i, \mu_i, \sigma_i^2, Z_i) = Z_i\mathcal{N}(t_i|\mu_i, \sigma_i^2)$ are called *site functions*. The mean vector $\boldsymbol{m} \in \mathbb{R}^N$ and covariance matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ of the approximate distribution may be computed as $\boldsymbol{m} = \boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{A} = (K^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}$, where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \cdots, \sigma_N^2)$ and $\boldsymbol{\mu} = [\mu_1, \cdots, \mu_N]^T$.

The prediction $p(t_*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}_*) = \mathcal{N}(t_*|\mu_*, \sigma_*^2)$ for a new input $\boldsymbol{x}_*$ is given by

$$\mu_* = \boldsymbol{k}_{*N}\boldsymbol{K}^{-1}\boldsymbol{m}, \quad \text{and} \quad \sigma_*^2 = k_{**} - \boldsymbol{k}_{*N}(\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}\boldsymbol{A}\boldsymbol{K}^{-1})\boldsymbol{k}_{N*}. \tag{13}$$

Although the predictive distribution of y_* is intractable, its mean is also μ_*.

The variables μ_i, σ_i^2 and Z_i are obtained by iterative moment match, which simultaneously minimizes the reverse Kullback-Leibler divergence between the true posterior and the approximate distribution. The convergence is not guaranteed, but it has been reported in the literature that EP works well within GP models [1]. The complete algorithm for a Laplace likelihood is detailed in [17].

3 Experiments

In order to verify the performance of the previously described models in the task of nonlinear system identification in the presence of outliers, we performed computational experiments with five artificial datasets, detailed in Tab. 1. The first four datasets were presented in the seminal work of Narendra et. al. [19]. The fifth dataset was generated following Kocijan et. al. [4].

Besides the Gaussian noise, indicated in the last column of Tab. 1, the estimation data of all datasets was also incrementally corrupted with a number of outliers equal to 5%, 10% and 20% of the estimation samples. Each randomly chosen sample was added by a uniformly distributed value $\text{U}(-M_y, +M_y)$, where M_y is the maximum absolute output. We emphasize that only the output values were corrupted in this step. Such outlier contamination methodology is similar to the one performed in [20]. The orders L_u and L_y chosen for the regressors were set to their largest delays presented in the second column of Tab. 1.

We compare the performances of the following GP models: conventional GP, GP with Student-t likelihood and VB inference (GP-tVB) and GP with Laplace

Table 1. Details of the five artificial datasets used in the computational experiments. The indicated noise in the last column is added only to the output of the estimation data. Note that $\mathrm{U}(A, B)$ is a random number uniformly distributed between A and B.

		Input/Samples		
#	**Output**	**Estimation**	**Test**	**Noise**
1	$y_i = \frac{y_{i-1}y_{i-2}(y_{i-1}+2.5)}{1+y_{i-1}^2+y_{i-2}^2}$	$u_i = \mathrm{U}(-2, 2)$ 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	$\mathcal{N}(0, 0.0025)$

likelihood and EP inference (GP-LEP). The obtained root mean square errors (RMSE) are presented in Tab. 2.

In almost all scenarios with outliers both robust variants presented better performances than conventional GP. Only in one case, *Artificial 3* dataset with 20% of corruption, GP performed better than one of the robust models (GP-tVB). In the scenarios without outliers, i.e., with Gaussian noise only, the GP model achieved the best RMSE for *Artificial 1* and *4* datasets, but it also performed closer to the robust models for the other datasets with 0% of corruption.

A good resilience to outliers was obtained for *Artificial 1* and *2* datasets, with GP-LEP and GP-tVB models being less affected in the cases with outliers. The most impressive performance was the one achieved by the GP-tVB model for all cases of the *Artificial 2* dataset, with little RMSE degradation.

For the *Artificial 3* dataset, only the GP-tVB model with 5% of outliers achieved error values close to the scenario without outliers. In the other cases, both variants, although better than conventional GP model, presented greater RMSE values than their results for 0% of outliers.

Likewise, in the experiments with *Artificial 4* and *5* datasets, we also observed that all models were affected by the corruption of the estimation data, even with lower quantities of outliers. However, it is important to emphasize that both GP-tVB and GP-LEP models achieved better RMSE values than conventional GP, often by a large margin, as observed in the *Artificial 4* dataset for the GP-tVB model. Thus, the robust variants can be considered a valid improvement over the conventional GP model.

Finally, we should mention that during the experiments, the variational approach of the GP-tVB model has been consistently more stable than the EP

Table 2. Summary of simulation RMSE without and with outliers in estimation step.

	Artificial 1				**Artificial 2**			
% of outliers	0%	5%	10%	20%	0%	5%	10%	20%
GP	**0.2134**	0.3499	0.3874	0.4877	0.3312	0.3724	0.5266	0.4410
GP-tVB	0.2455	0.3037	0.2995	**0.2868**	**0.3189**	**0.3247**	**0.3284**	**0.3306**
GP-LEP	0.2453	**0.2724**	**0.2720**	0.3101	0.3450	0.3352	0.3471	0.3963
	Artificial 3				**Artificial 4**			
GP	0.1106	0.4411	0.7022	0.6032	**0.6384**	2.1584	2.2935	2.4640
GP-tVB	0.1097	**0.1040**	**0.3344**	0.8691	0.6402	**0.7462**	2.2220	**2.1951**
GP-LEP	**0.0825**	0.3527	0.4481	**0.5738**	0.9188	1.1297	**2.1742**	2.3762
	Artificial 5							
GP	0.0256	0.0751	0.1479	0.1578				
GP-tVB	**0.0216**	0.0542	**0.0568**	**0.1006**				
GP-LEP	0.0345	**0.0499**	0.0747	0.1222				

algorithm of the GP-LEP model, even with the incorporation of the numerical safeties suggested by Rasmussen and Williams [1] and Kuss [17], which might be a decisive factor when choosing which model to apply for system identification.

4 Conclusion

In this paper we evaluated robust Gaussian process models in the task of nonlinear dynamical system identification in the presence of outliers in the data. The experiments with five artificial datasets considered a GP model with Student-t likelihood and variational inference (GP-tVB) and a model with Laplace likelihood with EP inference (GP-LEP), besides conventional GP with Gaussian likelihood.

Although the robust variants performed better in the scenarios with outliers, we cannot state categorically that they were insensitive to the corrupted data. Both GP-tVB and GP-LEP models obtained considerable lower RMSE for some cases with outliers. Depending on the task in hand, such degradation may or may not be tolerable. This observation, as well as some numerical issues encountered in the EP algorithm, encourages us to further pursue alternative GP-based models which are more appropriate for robust system identification.

Acknowledgments

The authors thank the financial support of FUNCAP, IFCE, NUTEC and CNPq (grant no. 309841/2012-7)

References

1. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. 1 edn. MIT Press (2006)
2. Williams, C.K.I., Barber, D.: Bayesian classification with Gaussian processes. IEEE T Pattern Anal **20**(12) (1998) 1342–1351
3. Lawrence, N.D.: Gaussian process latent variable models for visualisation of high dimensional data. In: Advances in Neural Information Processing Systems. (2004) 329–336
4. Kocijan, J., Girard, A., Banko, B., Murray-Smith, R.: Dynamic systems identification with Gaussian processes. Math Comp Model Dyn **11**(4) (2005) 411–424
5. Murray-Smith, R., Johansen, T.A., Shorten, R.: On transient dynamics, off-equilibrium behaviour and identification in blended multiple model structures. In: European Control Conference (ECC'99), Karlsruhe, BA-14, Springer (1999)
6. Solak, E., Murray-Smith, R., Leithead, W.E., Leith, D.J., Rasmussen, C.E.: Derivative observations in Gaussian process models of dynamic systems. Advances in Neural Information Processing Systems **16** (2003)
7. Rottmann, A., Burgard, W.: Learning non-stationary system dynamics online using Gaussian processes. In: Pattern Recognition. Volume 6373 of Lecture Notes in Computer Science. Springer (2010) 192–201
8. Ažman, K., Kocijan, J.: Dynamical systems identification using Gaussian process models with incorporated local models. Eng Appl Artif Intel **24**(2) (2011) 398–408
9. Petelin, D., Grancharova, A., Kocijan, J.: Evolving Gaussian process models for prediction of ozone concentration in the air. Simul Model Pract Th **33** (2013) 68–80
10. Frigola, R., Chen, Y., Rasmussen, C.E.: Variational Gaussian process state-space models. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., eds.: Advances in Neural Information Processing Systems 27 (NIPS). (2014)
11. Minka, T.P.: Expectation propagation for approximate Bayesian inference. In: Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI'01), Morgan Kaufmann (2001) 362–369
12. Faul, A.C., Tipping, M.E.: A variational approach to robust regression. In: Artificial Neural Networks (ICANN)'2001. (2001) 95–102
13. Kuss, M., Pfingsten, T., Csató, L., Rasmussen, C.E.: Approximate inference for robust Gaussian process regression. Max Planck Inst. Biological Cybern., Tubingen, GermanyTech. Rep **136** (2005)
14. Tipping, M.E., Lawrence, N.D.: Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis. Neurocomputing **69**(1) (2005) 123–141
15. Jylänki, P., Vanhatalo, J., Vehtari, A.: Robust gaussian process regression with a Student-t likelihood. Journal of Machine Learning Research **12** (2011) 3227–3257
16. Berger, B., Rauscher, F.: Robust Gaussian process modelling for engine calibration. In: Proceedings of the 7th Vienna International Conference on Mathematical Modelling (MATHMOD'2012). (2012) 159–164
17. Kuss, M.: Gaussian process models for robust regression, classification, and reinforcement learning. PhD thesis, TU Darmstadt (2006)
18. Rasmussen, C.E.: Evaluation of Gaussian processes and other methods for non-linear regression. PhD thesis, University of Toronto, Toronto, Canadá (1996)
19. Narendra, K.S., Parthasarathy, K.: Identification and control of dynamical systems using neural networks. IEEE T Neural Networ **1**(1) (1990) 4–27

20. Majhi, B., Panda, G.: Robust identification of nonlinear complex systems using low complexity ANN and particle swarm optimization technique. Expert Syst Appl **38**(1) (2011) 321–333