INCORPORATING PRIOR KNOWLEDGE INTO GAUSSIAN PROCESS MODELS

Kristjan Ažman

Inštitut Jožef Stefan
Jamova 39, 1000 Ljubljana, Slovenia
kristjan.azman@ijs.si

Abstract: Gaussian processes (GP) models form an emerging methodology for modelling nonlinear dynamic systems which tries to overcome certain limitations inherent to traditional methods such as e.g. neural networks, fuzzy models or local model networks.
The GP model seems promising for three reasons — first, smaller number of training parameters, second, the variance of model's output is automatically obtained and third, various prior knowledge, e.g. linear local models, statical characteristic, known hysteresis can be included in the model.
In the paper some of the possibilities of the prior knowledge incorporation will be presented together with illustrative example.

Keywords: Gaussian process model, nonlinear identification, prior knowledge

1 INTRODUCTION

While there are numerous methods for identification of linear dynamical systems from measured data, e.g. (Ljung, 1999), nonlinear systems are more difficult to tackle. Usually nonlinear systems are identified using models such as artificial neural networks (ANN), Takagi-Sugeno fuzzy models (TSFM), local model networks etc. These models usually prove effective, but their wider use is prevented by their drawbacks, namely model's structure determination and training of usually large number of parameters. In this paper the Gaussian processes prior model will be used for dynamic systems modelling instead, as it has the possibilities to overcome these drawbacks.

Gaussian processes model (GP model) is probabilistic, non-parametric black-box model comparable to ANN or TSFM models. The output of the GP model is normally distributed, expressed in terms of the mean and the variance. Mean value represents the most probable value of the predicted output and the variance can be viewed as the measure of confidence in the predicted mean. Obtained variance distinguishes the GP method from ANN or TSFM and can be used as the quality measure of the model. The number of GP model's (hyper)parameters is much smaller than in the comparable ANN or TSFM, which reduces the problem of optimization. Another potentially useful attribute of GP model is the possibility to incorporate prior knowledge into the model. This knowledge can be in various forms, e.g. linear local models (Solak,

et al., 2003; Kocijan and Leith, 2004), static characteristic, prior knowledge about the noise, hysteresis (Ažman and Kocijan, 2005) etc. GP model has been popularized among machine learning community through works of Rasmussen (1996) and Neal (1996), but was only recently used for dynamic system identification, e.g. (Kocijan, *et al.*, 2005).

The purpose of this paper is to briefly present the identification of dynamic systems with GP model and to present the possibilities of prior knowledge incorporation. In the next section the GP model and the identification of dynamic systems with GP model is illustrated. In Section 3 the incorporation of the prior knowledge into the GP model is presented together with illustrative example. Conclusion emphasizes the main points and concludes the paper.

2 INTRODUCTION TO DYNAMIC SYSTEMS IDENTIFICATION WITH THE GP MODEL

2.1 Modelling with the GP model

Here the modelling with the GP model will only be briefly presented, for more detailed introduction see e.g. (Wiliams, 1998).

The idea behind GP modelling is to place the prior directly over the function values instead of parameterizing unknown function $f(\mathbf{x})$. Consider the system

$$y(k) = f(\mathbf{x}(k)) + \epsilon(k) \tag{1}$$

where $\epsilon(k)$ is a white noise with variance v_0 and $\mathbf{x}$ is the vector of system's inputs. To model this system, the GP prior with covariance function

$$C(\mathbf{x}_i, \mathbf{x}_j) \;=\; v \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d (x_i^d - x_j^d)^2 + v_0 \delta_{ij}\right] \tag{2}$$

with unknown hyperparameters $\boldsymbol{\Theta} = [w_1, \ldots, w_D, v]^T$ is put on space of functions $f(.)$, where δ_{ij} is Kronecker operator. This covariance function is common choice when we assume stationarity of the system and smoothness of its output.

Based on a set of N training data pairs $\{\mathbf{x}_i, y_i\}_{i=1}^N$, gathered in $\mathcal{D} = \mathbf{X}|\mathbf{y}$, we wish to find the predictive distribution of y^* corresponding to a new given input $\mathbf{x}^*$. For this collection of (presumably normal) random variables $(y_1, \ldots, y_N, y^*)$ we can write: $(\mathbf{y}, y^*) \sim \mathcal{N}(0, \mathbf{K}_{N+1})$ where $\mathbf{K}_{N+1}$ is the covariance matrix of the process generating outputs $(\mathbf{y}, y^*)$. The elements of the covariance matrix $\mathbf{K}_{N+1}$ are the covariances between values of the function $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, calculated using covariance function $C(.,.)$.

The covariance matrix of the process is:

$$\mathbf{K}_{N+1} = \begin{bmatrix} \begin{bmatrix} \\ \mathbf{K} \\ \\ \end{bmatrix} & \begin{bmatrix} \\ \mathbf{k}(\mathbf{x}^*) \\ \\ \end{bmatrix} \\ \\ \begin{bmatrix} \mathbf{k}(\mathbf{x}^*)^T \end{bmatrix} & \begin{bmatrix} k(\mathbf{x}^*) \end{bmatrix} \end{bmatrix} \tag{3}$$

This joint probability can be divided into a marginal and a conditional part. Hyperparameters of the covariance function are optimized so they maximize the likelihood of the training data $\mathcal{D}$. The conditional part provides us with the (Gaussian) output distribution of the GP model with mean $\mu(\mathbf{x}^*)$ and variance $\sigma^2(\mathbf{x}^*)$:

$$\mu(\mathbf{x}^*) = \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\mathbf{y} \quad (4)$$

$$\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*) - \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}^*) + v_0 \quad (5)$$

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}_1, \mathbf{x}^*), \ldots, C(\mathbf{x}_N, \mathbf{x}^*)]^T$ is the vector of covariances between training inputs and the test input and $k(\mathbf{x}^*) = C(\mathbf{x}^*, \mathbf{x}^*)$ is the autocovariance of the test input.

2.2 Dynamical system identification

Presented GP model was originally used for modelling of static nonlinearities, but it can be extended to model dynamical systems as well (Kocijan, *et al.*, 2005). Our task is to model the dynamical system (1), where $\mathbf{x} = [y(t-1), \ldots, y(t-L), u(t-1), \ldots, u(t-L)]$ is the vector of regressors and be able to make n-step ahead prediction. One way to do n-step ahead prediction is to make iterative one-step ahead predictions up to desired step n whilst feeding back the predicted output.

Two approaches to iterated one-step ahead prediction are possible using the GP model — in first only the mean values of the predicted output are feed back to the input ("naive" approach), in second the complete output distributions are feed back ("exact" approach). More on the GP model simulation can be found e.g. in (Girard, 2004).

3 INCORPORATION OF PRIOR KNOWLEDGE INTO THE GP MODEL

Trained GP model carries information about observed system in two parts:

- input/output data $\mathcal{D} = \mathbf{X}|\mathbf{y}$, describing the input/output behavior of the system and
- the covariance function which expresses the correlation between the data.

Therefor there are two possibilities for prior knowledge incorporation into the GP model. The first possibility is to appropriately change the covariance function, so it expresses our different, stronger prior believes about the system. This can mean different choice of covariance function family or merely the change of the values of its hyperparameters. The second possibility is to either change either add to the input/output data $\mathcal{D}$.

3.1 Changing the covariance function

The role of the covariance function is to correlate the data constituting the GP model. Our *a-priori* knowledge is expressed through the choice of the covariance function family. Function (2) for example is the most widely used covariance function, as it represents common prior believes like stationarity of the process and smoothness of the output and is fairly easy to use.

However if it is known that the unknown system has some other attributes we would like to express with covariance function like periodicity, some kind of non-stationarity or that there are more processes affecting the unknown system, it could be useful to choose the covariance function from different function family, see (Gibbs, 1997) or (Leith, *et al.*, 2005) for the latter case.

There is also a possibility that the noise present at the output of the system is not white Gaussian. If we know the parameters of its dynamical model we can change the "noise part"[1] of the covariance function accordingly as in (Murray-Smith and Girard, 2001).

3.2 Changing input/output data

The second possibility of prior information incorporation is to change the input/output data $\mathcal{D}$ in which the behavior of the unknown system is contained in explicit form, i.e. system's output as the function of corresponding values of input regressors. There are several possibilities:

- A new regressor can be added to already present regressors, increasing the input dimension of the model. Into this regressor additional information about every training data point in data $\mathcal{D}$ is encoded, e.g. the state of hysteresis of the system for particular training point $\mathbf{x}_i|y_i$.
- Second, the new data points $\mathbf{x}_i|y_i$ can be added, reflecting some prior knowledge, e.g. static characteristic, some boundary conditions. Another possibility is to change the nature of the data in the input/output data $\mathcal{D}$ so that it represents derivative instead of functional information. As the derivative of the GP remains a GP (Solak, *et al.*, 2003) this is allowed if we only appropriately change the covariance function for derivative data. An example of data including the derivative information are linear local models.
- The data from several linear local models can be combined with "normal" data representing unknown system, i.e. data samples from system's response. Such GP model could be useful tool for combining local models as it can replace the local models with system's response samples in the regions, where the local models are hard to identify, e.g. in the off-equilibrium regions of the dynamic system (Murray-Smith, *et al.*, 1999). Another advantage of the linear local models incorporation is the reduction of the size of the GP model, which could reduce the training time of the model.

3.3 GP model with incorporated linear local models – an example

In this section the possibility of linear local models incorporation will be presented on simple example. Due to space limitations only the illustration of the concept will be given.

We would like to model following nonlinear dynamic system, given by Narendra and Parthasarathy (1990) to make (for simplicity of illustration) one step-ahead prediction:

$$y(k+1) = \frac{y(k)}{1+y^2(k)} + u^3(k) \tag{6}$$

[1] "Noise part" of the covariance function is the part reflecting the influence of the noise, e.g. $v_0\delta_{ij}$ in the case of covariance function (2).

where the interesting input region is in the vicinity of the centre $[0, 0]$. The training data for the GP model has been composed of six local models on the equilibrium curve and eight samples of system's response to excitation signal in off-equilibrium regions. The comparison between system's (left) and model's (right) response can be seen in Fig. 1. The error of the prediction (left) and the predicted variance of the model (right) can be seen in Fig. 2. From figures it can be concluded, that the prediction is good around equilibrium curve and slightly worse away from it[2]. We can also observe the weaker confidence in prediction in regions represented with less from the right figure in Fig. 2.

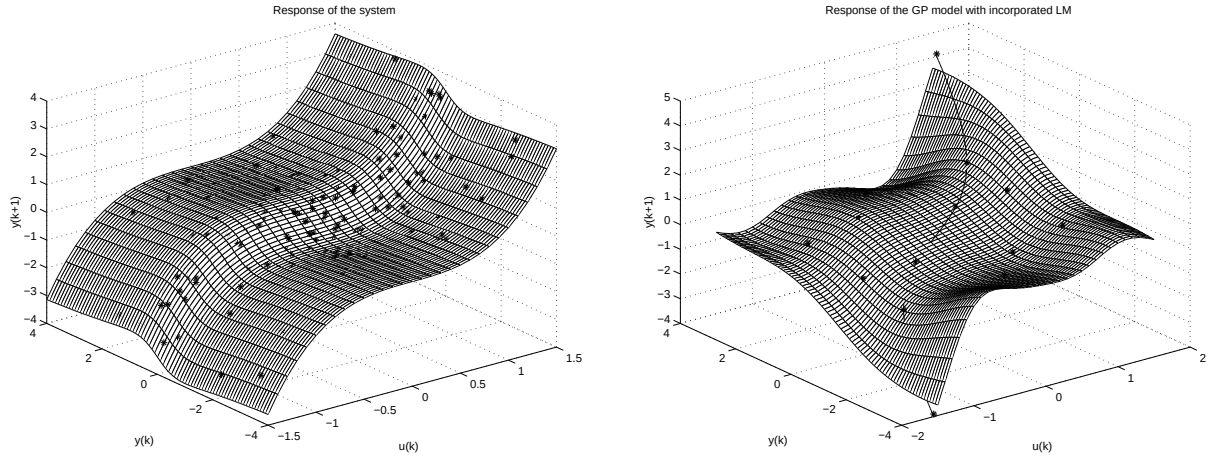

Figure 1: The response of the system (left) and GP model (right)

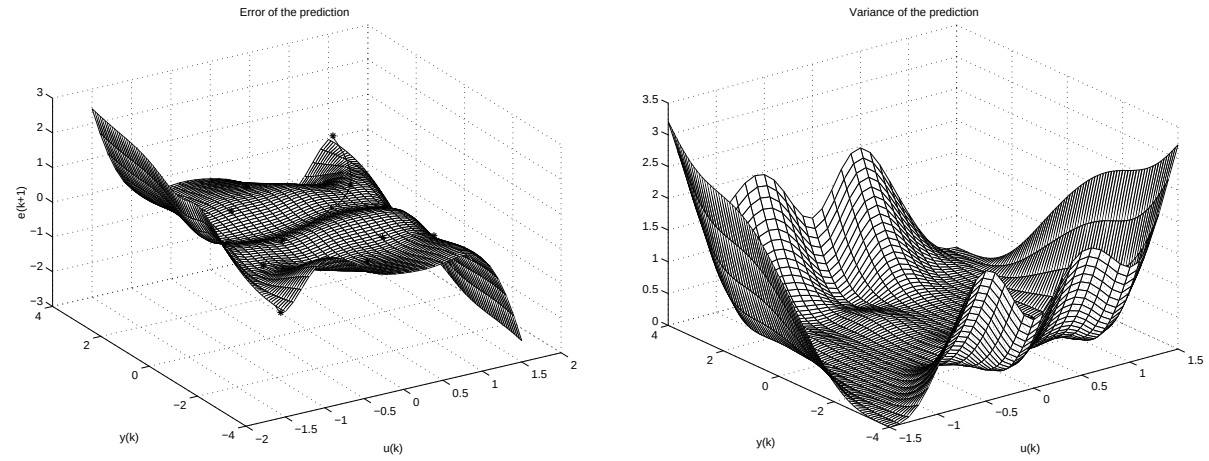

Figure 2: The error (left) and the variance of the prediction (right)

4 Conclusion

In this paper the dynamic system identification with GP model was briefly introduced. Some of the potentially useful advantages of this method for identification over more traditional approaches are: fewer number of parameters, the measure of confidence in prediction, easier

[2]The prediction of the GP model would improve if more training points were added.

use and the possibility to include different kinds of prior knowledge into the model. Some of different possibilities of prior knowledge incorporation were presented in the paper and the incorporation of the linear local models was illustrated on simple example.

Future work should concentrate on finding the other and development of already known possibilities of prior knowledge incorporation and on finding suitable domain application for GP model identification.

REFERENCES

Gibbs, M.N. (1997). *Bayesian Gaussian Processes for Regression and Classification.* PhD thesis, Cambridge University.

Girard, A. (2004). *Approximate Methods for Propagation of Uncertainty with Gaussian Process Models.* PhD thesis, University of Glasgow.

Kocijan J. and Leith, D.J. (2004). Derivative observations used in predictive control. In *Proceedings in Melecon*, Dubrovnik, Croatia, 12–15 May.

Kocijan, J., Girard, A., Banko, B. and Murray-Smith, R. (2005). Dynamic systems identification with gaussian processes. *Mathematical and Computer Modelling of Dynamic Systems*, **11**(4), pp. 411–424.

Leith, D.J., Leithead, W.E., Neo, K.S. (2005). Gaussian regression based on models with two stochastic processes. In *Proceedings in IFAC 16th World Congress 2005*, Prague, Czech Republic.

Ljung, L. (1999). *System Identification – Theory for the User.* Prentice Hall, New Jersey, 2nd edition.

Murray-Smith, R., Johansen, T.A. and Shorten, R. (1999). On transient dynamics, off-equilibrium behaviour and identification in blended multiple model structures. In *Proceedings in European Control Conference*, BA–14, Karslruhe, Germany.

Murray-Smith R. and Girard, A. (2001). Gaussian Process priors with ARMA noise models. In *Proceedings on Irish Signals and Systems Conference*, pp. 147–152, Maynooth, Ireland.

Narendra, K.S. and Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on NN*, **1**(1), pp. 4–27.

Neal, R.M. (1996). *Bayesian learning for neural networks.* Springer-Verlag, New York.

Rasmussen, C.E. (1996). *Evaluation of Gaussian Processes and Other Methods for NonLinear Regresion.* PhD thesis, University of Toronto.

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. In *Advances in Neural Information Processing Systems 15* (S. Becker, S. Thrun, and K. Obermayer. (Ed)), pp. 529–536. MIT Press.

Williams, C.K.I. (1998). *Learning and Inference in Graphical models*, Prediction with Gaussian processes: from linear regression and beyond, pp. 599–621. Kluwer Academic Press.