Comprising Prior Knowledge in Dynamic Gaussian Process Models

Kristjan Ažman and Juš Kocijan

Abstract: *Identification of nonlinear dynamic systems from experimental data can be difficult when, as often happens, more data are available around equilibrium points and only sparse data are available far from those points. The probabilistic Gaussian Process model has already proved to model such systems efficiently. The purpose of this paper is to show how one can relatively easily combine measured data and linear local models in this model. It is shown how uncertainty can be propagated through such models when predicting ahead in time in an iterative manner with Markov Chain Monte Carlo approach. The approach is illustrated with a simple numerical example.*

Key words: *Machine learning, Dynamic system models, Systems identification, Gaussian process models.*

INTRODUCTION

One of the problems frequently met in practice when modelling dynamic systems is the difficulty of constructing a nonlinear model on a reliable and consistent basis from available measured data. The purpose of this paper is to show how linear local models can be incorporated in Gaussian process (GP) models of dynamic systems. The use of Gaussian processes for modelling dynamic systems has recently been studied, e.g. [1,2,3]. A key issue when modelling with this probabilistic model is that, in its simplest form, the computational burden associated with it is cubic in the number of data points used, as it requires the inversion of an $N \times N$ matrix, where N is the number of data points. Although employing approximate inverses can reduce this computational burden, we suggest an alternative approach that summarizes measured data in the vicinity of an equilibrium point with derivative observations, i.e. a local linear model. Therefore, this approach is not only in accord with engineering practice but it can also directly reduce the computational burden. The main contribution of this paper is the propagation of uncertainty ahead in time with Monte Carlo simulation for the GP model with incorporated local models, when such models are used for multiple-step-ahead prediction.

The paper is organized as follows. Gaussian process models are briefly reviewed and the incorporation of derivative observations is then discussed. Afterwards the modelling of dynamic systems with such models is described. An example illustrates the modelling and simulating the dynamic system model. Conclusions are summarized at the end of the paper.

GAUSSIAN PROCESS MODEL

A detailed presentation of Gaussian processes can be found in [7]. A Gaussian process is a random process, fully characterized by its mean and covariance matrix $\mathbf{\Sigma}$. For simplicity, we assume a zero-mean process. Given $\{\mathbf{x}_1,...,\mathbf{x}_n\}$, the corresponding $f(\mathbf{x}_1),...,$ $f(\mathbf{x}_n)$ can be viewed as a collection of random variables which have a joint multivariate Gaussian distribution: $f(\mathbf{x}_1),...,f(\mathbf{x}_n) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where Σ_{pq} gives the covariance between $f(\mathbf{x}_p)$ and $f(\mathbf{x}_q)$ which is a function of the corresponding $\mathbf{x}_p$ and $\mathbf{x}_q$: $\Sigma_{pq}=C(\mathbf{x}_p, \mathbf{x}_q)$. The covariance function $C(.,.)$ can be of any kind, provided that it generates a positive definite covariance matrix $\mathbf{\Sigma}$. The Gaussian Process model fits naturally in the Bayesian modelling framework. It places a prior directly over the space of functions instead of parameterizing $f(\mathbf{x})$. A common choice of covariance function is the squared exponential, i.e. Gaussian function:

$$\mathrm{Cov}\left[f(\mathbf{x}_p), f(\mathbf{x}_q)\right] = C(\mathbf{x}_p, \mathbf{x}_q) = v \ \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d \left(\mathbf{x}_p^d - \mathbf{x}_q^d\right)^2\right] + v_0 \qquad (1)$$

where x_p^d denotes d^{th} component of the D-dimensional input vector $\mathbf{x}_p$, and v, w_1,...,w_D are free parameters. The smoothness assumption holds for covariance function (1), as the points lying closer together in the input space are more correlated as the points lying more far apart. The parameter v controls the vertical scale of variation and the w_d's are inversely proportional to the horizontal length-scale in dimension d ($\lambda_d=1/\sqrt{w_d}$).

Let the input/target relationship be $y=f(\mathbf{x})+\varepsilon$. We assume an additive white noise with variance v_0, $\varepsilon \sim \mathcal{N}(0, v_0)$, and put a GP prior with covariance function (1) and unknown parameters on $f(.)$. Within this probabilistic framework, we can write $y_1,\ldots, y_{N+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{N+1})$, with $K_{pq}=\Sigma_{pq}+v_0\delta_{pq}$, where $\delta_{pq}=1$ if $p=q$ and 0 otherwise. If we split $y_1,\ldots, y_{N+1}$ into two parts, $\mathbf{y}=[y_1,..., y_N]$ and y^*, we can write

$$\mathbf{y}, y^* \sim \mathcal{N}(0, \mathbf{K}_{N+1}), \tag{2}$$

with

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

where $\mathbf{K}$ is an $N \times N$ matrix giving the covariances between y_p and y_q, for $p,q=1 \ldots N$, $\mathbf{k}(\mathbf{x}^*)$ is an $N \times 1$ vector giving the covariances between y^* and y_p ($k_p(\mathbf{x}^*)=C(\mathbf{x}^*, \mathbf{x}_p)$, for $p=1 \ldots N$), and $\kappa(\mathbf{x}^*)=C(\mathbf{x}^*, \mathbf{x}^*)$ is the covariance between the test output and itself.

For our modelling purposes, we can then divide this joint probability into a marginal and a conditional part. Given a set of N training data pairs, $\{\mathbf{x}_p, y_p\}_{p=1}^N$, the marginal term gives us the likelihood of the observed data: $\mathbf{y}|\mathbf{X} \sim \mathcal{N}(0, \mathbf{K})$, where $\mathbf{y}$ is the $N \times 1$ vector of training targets and $\mathbf{X}$ the $N \times D$ matrix of the corresponding training inputs. We can then estimate the unknown parameters of the covariance function, as well as the noise variance v_0, via maximization of the log-likelihood. The conditional part of (2) provides us with the predictive distribution of y^* corresponding to a new given input $\mathbf{x}^*$. We only need to condition the joint distribution on the training data and the new input $\mathbf{x}^*$, $p(y^*|\mathbf{y},\mathbf{X}, \mathbf{x}^*) = p(\mathbf{y}, y^*)/p(\mathbf{y}|\mathbf{X})$. It can be shown that this distribution is Gaussian with mean and variance

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

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

This way, we can use the predictive mean $\mu(\mathbf{x}^*)$ as an estimate for y^* and the predictive variance, or standard deviation $\sigma(\mathbf{x}^*)$, as the uncertainty attached to it.

COMPRISING DERIVATIVE OBSERVATIONS

The Gaussian process modelling framework is readily extended to include situations where derivatives of the function are observed, as well as (or instead of) the values of the function itself. More on this topic can be found in [4,5,6]. Since differentiation is a linear operation, the derivative of a GP remains a GP. Assuming a zero-mean GP for $y=f(\mathbf{x})$, with Gaussian covariance function (1), the mean and covariance functions of the derivative process (in a given dimension) are readily obtained. The output (target) vector $\mathbf{y}$, which before consisted solely of output measurements, now also contains derivative observations. The corresponding inputs are the values of the regressors associated with each function and derivative observation. The covariance matrix is changed accordingly where derivative data is introduced. When using Gaussian covariance function (1) for covariance between two functional observations $y_p=f(\mathbf{x}_p)$ and $y_q=f(\mathbf{x}_q)$, the covariance

between a derivative and functional observation becomes

$$\mathrm{Cov}\left[\frac{\partial y_p}{\partial x_p^d}, y_q\right] = -vw_d(x_p^d - x_q^d)\exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d\left(\mathbf{x}_p^d - \mathbf{x}_q^d\right)^2\right] \quad (7)$$

where $\frac{\partial y_p}{\partial x_p^d}$ denotes the first derivative of y_p in direction of d^{th} component of the D-dimensional input vector $\mathbf{x}_p$. Similary, the covariance between two different derivative observations becomes

$$\mathrm{Cov}\left[\frac{\partial y_p}{\partial x_p^d}, \frac{\partial y_q}{\partial x_q^e}\right] = vw_e(\delta_{d,e} - w_d(x_p^d - x_q^d)(x_p^e - x_q^e)\exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d\left(\mathbf{x}_p^d - \mathbf{x}_q^d\right)^2\right] \quad (6)$$

The GP model acts to integrate and smooth the noisy derivative observations. Derivative observations around an equilibrium point can be interpreted as observations of a local linear model about this equilibrium point. This means that the derivative observations can be synthesized using standard linear regression. Such synthetic derivative observations can then be used to summarize training points in the vicinity of equilibrium points, thereby effectively reducing the number of data points in the model for computational purposes. It is important to note that a local linear input-output model such as a transfer function model only specifies a derivative observation up to a co-ordinate transformation. For simplicity, in this paper, we use lagged input signal samples and lagged output signal samples as our state co-ordinates, although other choices are possible as well.

Given that derivative and function observations are available, the predictive distribution of a function output corresponding to a new $\mathbf{x}$ has mean and variance given by equations (4) and (5), with the matrix $\mathbf{K}$ and the vector $\mathbf{k}^*$ changed adequately. In fact, they can be written so as to reflect the mixed nature of the training data (see [4] for details).

MODELLING OF DYNAMIC SYSTEMS

The above modelling procedure can be readily applied to dynamic systems, within an auto-regressive (AR) representation of the system [1,3]. Consider the following ARX model, where the current output depends on delayed outputs and exogenous control inputs:

$$y(k) = f\left(y(k-1),\ldots,y(k-L),u(k-1),\ldots,u(k-L)\right) + \varepsilon \quad (8)$$

where ε is a white noise and k denotes consecutive number of data sample. Let $\mathrm{x}(k)$ be the state vector at k, composed of the previous outputs y and inputs u, up to a given lag L ($\mathrm{x}(k)=[y(k\text{-}1),y(k\text{-}2),\ldots,y(k\text{-}L),u(k\text{-}1),u(k\text{-}2),\ldots,u(k\text{-}L)]^{\mathrm{T}}$) and $y(k)$ the corresponding output. We can then model this dynamic system using a Gaussian Process.

MULTIPLE-STEP-AHEAD PREDICTIONS WITH MARKOV CHAIN MONTE CARLO APPROACH

Assuming the time-series is known up to, say, k, we wish to predict n steps ahead: That is to say, to find the predictive distribution of $y(k+n)$ corresponding to $\mathbf{x}(k+n) = [y(k+n\text{-}1),\ldots, y(k+n\text{-}L)]^{\mathrm{T}}$. Multiple-step-ahead predictions of a system modelled by (8) can be achieved by iteratively making repeated one-step-ahead predictions, up to the desired horizon $k+n$.

A naive way of doing so is, at each time-step, to feed back the predictive mean (estimate of the output) thus considering $\mathbf{x}(k+n) = [\hat{y}(k+n\text{-}1),\ldots,\hat{y}(k+n\text{-}L)]^{\mathrm{T}}$, where $\hat{y}(k+n\text{-}i)$ is

the point estimate of $y(k+n-i)$.

More realistic approach is, at each time step, to feed back complete output distribution from the GP model. At the time step $k+1$, the output of the GP model is:

$$y_{k+1} \sim \mathcal{N}(\mu(\mathbf{x}_{k+1}), \sigma^2(\mathbf{x}_{k+1})) \tag{9}$$

For the prediction at time step $k+2$ complete output distribution y_{k+1} is taken into account when forming input $\mathbf{x}(k+2) = [\hat{y}(k+1),\ldots,\hat{y}(k-L+1)]^{\mathrm{T}}$. The output can be evaluated using simple Markov Chain Monte Carlo (MCMC) approach, where it is represented as the mixture of Gaussians: $y_{k+2} = \frac{1}{S}\sum_{s=1}^{S} y_{k+2}^{s}$, where y_{k+2}^{s} is the output of the GP model at $\mathbf{x}(k+2) = [\hat{y}^{s}(k+1),\ldots,\hat{y}(k-L+1)]^{\mathrm{T}}$ and $\hat{y}^{s}(k+1)$ are samples from normal distribution (9).

At the next time step the input $\mathbf{x}(k+3)$ contains the mixture of Gaussians y_{k+2}. The next prediction is calculated by numerically integrating over the input distribution $\mathbf{x}(k+3)$ using MCMC methods. The input distribution $\mathbf{x}(k+2)$ is represented with a large number of samples S, and the ouput becomes a mixture of corresponding number of Gaussians, again fed back as the next input. This procedure is repeated until the desired horizont ($k+n$) is reached.

EXAMPLE

The following example is intended to explore the potential of achieving an accurate model of a dynamic system, when using derivative observations at equilibrium points and a small number of function observations at off-equilibrium points.

Consider the nonlinear dynamic system described by

$$y(k) = \frac{y(k-1)}{1+y^2(k-1)} + u^3(k-1) \tag{10}$$

where the sampling time is one step. We select six equilibrium points, uniformly spanning the operating region of interest. At each equilibrium point, we apply a small-scale pseudo-random binary signal with mean 0 and magnitude 0.04 and the corresponding output signal is contaminated with normally distributed measurement noise in the magnitude range [-0.001,0.001]. A linear, first order approximation to the local dynamics at the equilibrium point is identified using the Matlab algorithm IV4. In addition to this equilibrium information, a small sparse set of off-equilibrium input-output data, consisting of only eight points, is selected. A GP model with zero-mean and Gaussian covariance function is then trained using

- six input-output values at equilibrium, spanning the operating region of interest;
- The set of coefficients of the identified first order linear models representing the partial derivatives of the output - 6 times 2 coefficients ;
- The 8 input-output values that were sampled out of equilibrium points.

Figure 1 shows a plot of the output predictive variance. A region with low variance indicates a region where the model is confident about its prediction.

The results of the simulation of the system (that is the n-step-ahead prediction, where n is the length of the validation signal) in the region with pronounced uncertainty is shown in Figure 2.

When simulating GP model with propagation of uncertainty, the MCMC sampling with S=10000 samples was used. It can be seen that propagating the uncertainty causes the standard deviations to become larger in some areas, compared to the naive approach. Also, the means are affected.

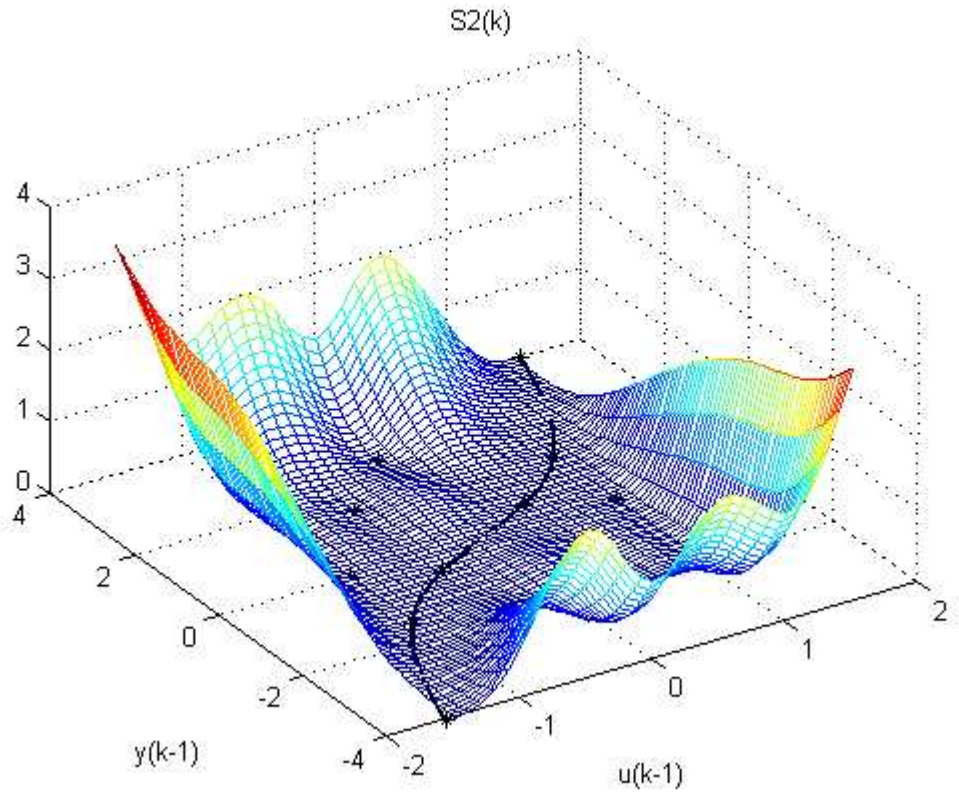

Figure 1: Variance of the GP model together with equilibrium locus represented with solid line and training points represented with dots

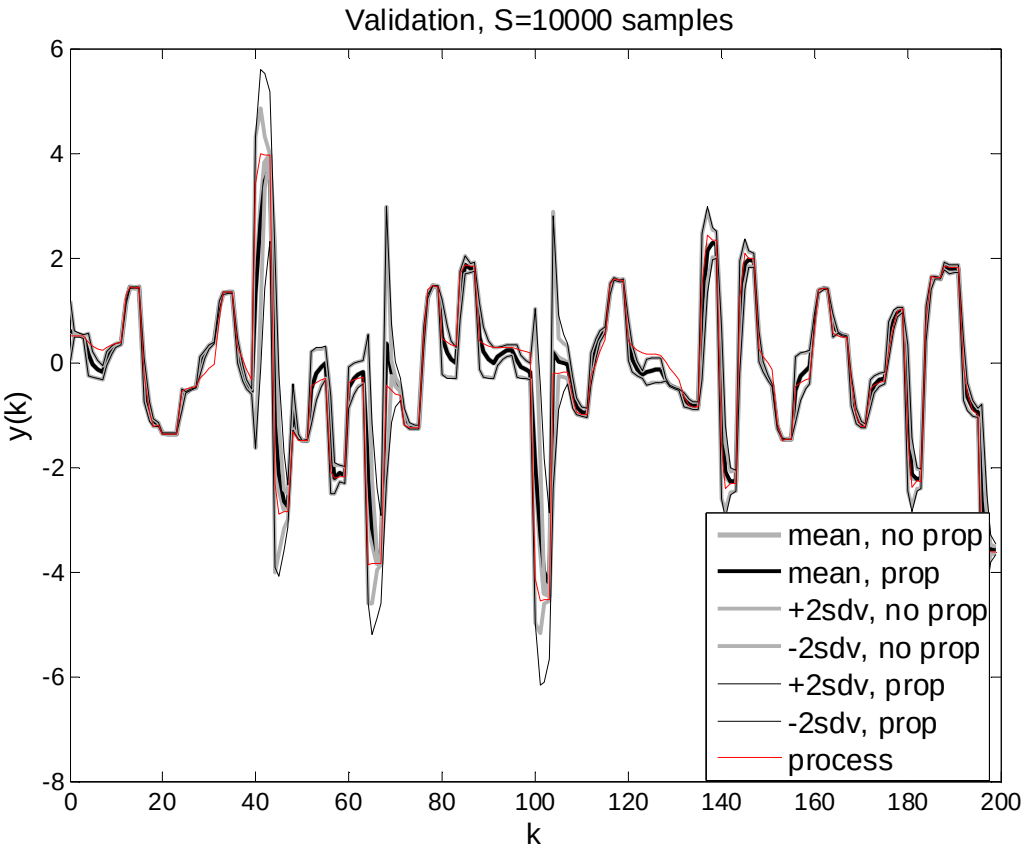

Figure 2: Simulation of GP model with and without propagation of uncertainty

CONCLUSIONS AND FUTURE WORK

This paper describes how linear local models can be incorporated in the Gaussian Process model of dynamic system. Also, we show how one can propagate the uncertainty when making iterative multiple-step-ahead predictions with such a model. Accounting for derivative observations, obtained as coefficients of local linear models in equilibrium points with regular linear regression method, means joining local linear models and GP models. Joining these two sorts of models results in global models containing global and local information, of acceptable dimensions and suited to the kind of data usually available in practice when carrying out experimental modelling (a lot of data in vicinity of equilibrium points and few data far from equilibria).

The main conclusions are as follows:

- The data used to obtain the grey-box model is well suited to the kind of data usually available in practice when carrying out experimental modelling.
- The model obtained is relatively small in comparison with a GP model that does not make use of derivative observations, while the model quality is comparable. This makes it very suitable for applications.

Our simulated example is encouraging and these results offer new possibilities for dynamic system analysis and control, whenever uncertainty information is necessary.

REFERENCES

[1] Girard A., C.E. Rasmussen, R. Murray-Smith. Multi-step ahead prediction for non linear dynamic sytems - A Gaussian Process treatment with propagation of the uncertainty, In: *Advances in Neural Information processing Systems*, (S. Becker and S. Thrun and K. Obermayer, (Eds.)), Vol. 15, pp. 545-552, MIT Press, Cambridge, MA, 2002.

[2] Gregorčič G. and G. Lightbody. Internal model control based on a Gaussian process prior model, In: *Proceedings of ACC'2003*, Denver, 2003, pp. 4981-4986.

[3] Kocijan J., A. Girard, B. Banko and R. Murray-Smith. Dynamic Systems Identification with Gaussian Processes, In: *Proceedings of 4th Mathmod*, Vienna, 2003, pp. 776-784, expanded version accepted for publication in journal *Mathematical and Computer Modelling of Dynamic Systems*.

[4] Kocijan J., A. Girard and D. J. Leith. Incorporating linear local models in Gaussian process models, IJS report DP-8895, Jozef Stefan Institute, Ljubljana, 2003.

[5] Leith, D. J., W. E. Leithead, E. Solak and R. Murray-Smith. Divide & conquer identification: Using Gaussian process priors to combine derivative and non-derivative observations in a consistent manner, In: *Conference on Decision and Control 2002*, Las Vegas, 2002, pp. 624-629.

[6] Solak, E., R. Murray-Smith, W. E. Leithead, D. J. Leith and C. E. Rasmussen. Derivative observations in Gaussian Process models of dynamic systems, In: *Advances in Neural Information processing Systems*, (S. Becker and S. Thrun and K. Obermayer, (Eds.)), Vol. 15, pp. 529-536, MIT Press, Cambridge, MA, 2002.

[7] Williams, C.K.I. Prediction with Gaussian processes: From linear regression to linear prediction and beyond, In: *Learning in Graphical Models* (Jordan, M.I. (Ed.)), pp. 599-621, Kluwer Academic, Dordrecht, 1998.

ABOUT THE AUTHORS

Kristjan Ažman, M.Sc., Jožef Stefan Institute, Ljubljana, Slovenia

Assoc.Prof. Juš Kocijan, PhD, Jožef Stefan Institute, Ljubljana, and Nova Gorica Polytechnic, Nova Gorica, Slovenia,

Phone: +386 5 3315 231, E-mail: jus.kocijan@ijs.si.