Sparse Physics-based Gaussian Process for Multi-output Regression using Variational Inference

Ankit Chiplunkar[1,3], Emmanuel Rachelson[2], Michele Colombo[1] and Joseph Morlier[3]

[1]*Airbus Operations S.A.S., 316 route de Bayonne 31060, Toulouse Cedex 09, France*

[2]*Université de Toulouse, ISAE, DISC, 10 Avenue Edouard Belin, 31055 Toulouse Cedex 4, France*

[3]*Université de Toulouse, CNRS, ISAE-SUPAERO, Institut Clément Ader (ICA), 31077 Toulouse Cedex 4, France*

Keywords: Gaussian Process, Kernel Methods, Variational Inference, Multi-output Regression, Flight-test data.

Abstract: In this paper a sparse approximation of inference for multi-output Gaussian Process models based on a Variational Inference approach is presented. In Gaussian Processes a multi-output kernel is a covariance function over correlated outputs. Using a general framework for constructing auto- and cross-covariance functions that are consistent with the physical laws, physical relationships among several outputs can be imposed. One major issue with Gaussian Processes is efficient inference, when scaling up-to large datasets. The issue of scaling becomes even more important when dealing with multiple outputs, since the cost of inference increases rapidly with the number of outputs. In this paper we combine the use of variational inference for efficient inference with multi-output kernels enforcing relationships between outputs. Results of the proposed methodology for synthetic data and real world applications are presented. The main contribution of this paper is the application and validation of our methodology on a dataset of real aircraft flight tests, while imposing knowledge of aircraft physics into the model.

1 INTRODUCTION

In this work we consider the problem of modelling multiple output Gaussian Process (GP) regression (Rasmussen and Williams, 2005) correlated through physical laws of the system, while in presence of large number of inputs. In the literature inference on multiple output data is also known as co-kriging (Stein, 1999) or multi-kriging (Boyle and Frean, 2005). Modelling multi-output kernels is particularly difficult because we need to construct auto- and cross-covariance functions between different outputs. We turn to a general framework (Constantinescu and Anitescu, 2013) to calculate these covariance functions while imposing prior information of the physical processes. While a joint model developed using correlated covariance functions gives better predictions, it incurs a huge cost on memory occupied and computational time. The main contribution of this paper is to apply variational inference on these models of large datasets (of the order $O(10^5)$)and reduce the heavy computational costs incurred.

Let us start by defining a P dimensional input space and a D dimensional output space. Such that $\{(x_i^j, y_i^j)\}$ for $j \in [1; n_i]$ are the training datasets for the i^{th} output. Here n_i is the number of measurement points for the i^{th} output, while $x_i^j \in \mathbb{R}^P$ and $y_i^j \in \mathbb{R}$. We next define $x_i = \{x_i^1; x_i^2; \ldots; x_i^{n_i}\}$ and $y_i = \{y_i^1; y_i^2; \ldots; y_i^{n_i}\}$ as the full matrices containing all the training points for the i^{th} output such that $x_i \in \mathbb{R}^{n_i \times P}$ and $y_i \in \mathbb{R}^{n_i}$. Henceforth we define the joint output vector $Y = [y_1; y_2; y_3; \ldots; y_D]$ such that all the output values are stacked one after the other. Similarly, we define the joint input matrix as $X = [x_1, x_2, x_3, \ldots, x_D]$. If $\Sigma n_i = N$ for $i \in [1, D]$. Hence N represents the total number of training points for all the outputs combined. Then $Y \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N \times P}$.

For simplicity take the case of an explicit relationship between two outputs y_1 and y_2. Suppose we measure two outputs with some error, while the true physical process is defined by latent variables f_1 and f_2. Then the relation between the output function, measurement error and true physical process can be written as follows.

$$y_1 = f_1 + \varepsilon_{n1}$$
$$y_2 = f_2 + \varepsilon_{n2} \tag{1}$$

Where, ε_{n1} and ε_{n2} are measurement error sampled from a white noise gaussian $\mathcal{N}(0, \sigma_{n1})$ and

Chiplunkar, A., Rachelson, E., Colombo, M. and Morlier, J.
Sparse Physics-based Gaussian Process for Multi-output Regression using Variational Inference.
DOI: 10.5220/0005700504370445
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 437-445
ISBN: 978-989-758-173-1

$\mathcal{N}(0,\sigma_{n2})$. While the physics based relation can be expressed as,

$$f_1 = g\left(f_2, x_1\right) \tag{2}$$

Here g is an operator defining the relation between f_1 and an independent latent output f_2. A GP prior in such a setting with 2 output variables is:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim GP\left[\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \quad \begin{pmatrix} K_{11}+\sigma_{n1}^2 & K_{12} \\ K_{21} & K_{22}+\sigma_{n2}^2 \end{pmatrix}\right] \tag{3}$$

K_{12} and K_{21} are cross-covariances between the two inputs x_1 and x_2. K_{22} is the covariance function of independent output, σ_{n1}^2 and σ_{n2}^2 are the variance of measurement error, while K_{11} is the auto-covariance of the dependent output variable. m_1 and m_2 are the mean of the prior for outputs 1 and 2. The full covariance matrix is also called the joint kernel, henceforth we will denote the joint-kernel as K_{XX}.

$$K_{XX} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \tag{4}$$

while the joint error matrix will be denoted by Σ;

$$\Sigma = \begin{bmatrix} \sigma_{n1}^2 & 0 \\ 0 & \sigma_{n2}^2 \end{bmatrix} \tag{5}$$

Full joint-kernel of a multi-output GP has huge memory and computational costs. For a multi-output GP as defined earlier the covariance matrix is of size N, needing $O\left(N^3\right)$ calculations for inference and $O\left(N^2\right)$ for storage. (Snelson and Ghahramani, 2006) introduced "Fully independent training conditional" (FITC) and (Quionero-candela et al., 2005) introduced and "Partially Independent Training Conditional" (PITC) approximations on inference of a GP using inducing inputs. Later (Alvarez and Lawrence, 2009) extended the application of FITC and PITC to approximate the inference of multi-output GP's constructed through convolution processes. One problem with FITC and PITC approximation is their tendency to over-fit. In this work we extend the use of a variational approach (Titsias, 2009) to approximate the inference in a joint-kernel for both linear and non-linear relationships. We observe that the current approximation reduces the computational complexity to $O\left(N(MD)^2\right)$ and storage to $O(NMD)$, where M denotes the number of inducing points in the input space.

In Section 2, we start with an introduction to multi-output GP and later derive the multi-output GP regression in presence of correlated covariances. In Section 3 we discuss various methods of approximating inference of a GP and later derive application of variational inference on the problem of multi-output kernels. Finally, in Section 4 we demonstrate the approach on both theoretical and flight-test data.

2 MULTI-OUTPUT GAUSSIAN PROCESS

Choosing covariance kernels for GP regression with multiple outputs can be roughly classified in three categories. In the first case, the outputs are known to be mutually independent, and thus the system can be decoupled and solved separately as two or more unrelated problems. In the second case, we assume that the processes are correlated, but we have no information about the correlation function. In this case, a model can be proposed, or nonparametric inferences can be carried out. In the third situation, we assume that the outputs have a known relationship among them, such as equation 2. The last point forms the scope of this section.

2.1 Related Work

Earlier work developing such joint covariance functions (Bonilla et al., 2008) have focused on building different outputs as a combination of a set of latent functions. GP priors are placed independently over all the latent functions thereby inducing a correlated covariance function. More recently (Alvarez et al., 2009) have shown how convolution processes (Boyle and Frean, 2005) can be used to develop joint-covariance functions for differential equations. In a convolution process framework output functions are generated by convolving several latent functions with a smoothing kernel function. In the current paper we assume one output function to be independent and evaluate the remaining auto- and cross-covariance functions exactly if the physical relation between them is linear (Solak et al., 2003) or use approximate joint-covariance for non-linear physics-based relationships between the outputs (Constantinescu and Anitescu, 2013).

2.2 Multi-output Joint-covariance Kernels

If the two outputs y_1 and y_2 satisfy a physical relationship given by equation 2 with $g\left(.\right) \in \mathcal{C}^2$, and known covariance matrix K_{22}. Then the joint-covariance matrix for a linear operator $g\left(.\right)$ can be derived analytically as (Stein, 1999):

$$K_{XX} = \begin{bmatrix} g(g(K_{22},x_2),x_1) & g(K_{22},x_1) \\ g(K_{22},x_2) & K_{22} \end{bmatrix} \tag{6}$$

Since a non-linear operation on a GP does not result in a GP, for the case of non-linear $g\left(.\right)$ the above joint-covariance matrix as derived in equation 6 is not positive semi-definite. Therefore we will use an approximate joint-covariance as developed by (Constantinescu and Anitescu, 2013) for imposing non-linear relations:

$$K_{XX} = \begin{bmatrix} LK_{22}L^T & LK_{22} \\ K_{22}L^T & K_{22} \end{bmatrix} + O\left(\delta y_2^3\right) \qquad (7)$$

Where $L = \frac{\partial g}{\partial y}\Big|_{y_2=\bar{y}_2}$ is the Jacobian matrix of $g(.)$ evaluated at the mean of independent output y_2. δy_2 is the amplitude of small variations of y_2, introduced by the Taylor series expansion of $g(K_{XX})$ with respect to y_2.

The above kernel takes a parametric form that depends on the mean value process of the independent variable. Equation 7 is basically a Taylor series expansion for approximating related kernels. Since a Taylor series expansion is constructed from derivatives of a function which are linear operations the resulting approximated joint kernel is a gaussian kernel with the non-gaussian part as the error. Higher-order closures can be derived with higher order derivatives of the relation $g(.)$. For simplicity we will restrict ourselves to first order approximation of the auto- and cross-covariance functions leading to an error of the order $O\left(\delta y_2^3\right)$. (Constantinescu and Anitescu, 2013) provide a more detailed derivation of equation 7.

2.3 GP Regression using Joint-covariance

We start with defining a zero-mean prior for our observations and make predictions for $y_1(x_*) = y_{*1}$ and $y_2(x_*) = y_{*2}$. The corresponding prior according to equation 3 and 4 will be:

$$\begin{bmatrix} Y(X)) \\ Y(X_*)) \end{bmatrix} = GP\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_{XX}+\Sigma & K_{XX_*} \\ K_{X_*X} & K_{X_*X_*}+\Sigma \end{bmatrix}\right] \qquad (8)$$

The predictive distribution is then given as a normal distribution with expectation and covariance matrix given by (Rasmussen and Williams, 2005)

$$Y_* \mid X, X_*, Y = K_{X_*X}\left(K_{XX}\right)^{-1} Y \qquad (9)$$

$$Cov\left(Y_* \mid X, X_*, Y\right) = K_{X_*X_*} - K_{X_*X}\left(K_{XX}\right)^{-1} K_{XX_*} \qquad (10)$$

Here, the elements K_{XX}, K_{X_*X} and $K_{X_*X_*}$ are block covariances derived from equations 6 or 7.

The joint-covariance matrix depends on several hyperparameters θ. They define a basic shape of the GP prior. To end up with good predictions it is important to start with a good GP prior. We minimize the negative log-marginal likelihood to find a set of good hyperparameters. This leads to an optimization problem where the objective function is given by equation 11

$$\log(\mathbb{P}(y \mid X, \theta)) = \log[GP(Y|0, K_{XX}+\Sigma)] \qquad (11)$$

With its gradient given by equation 12

$$\frac{\partial}{\partial\theta}\log(\mathbb{P}(y \mid X, \theta)) = \frac{1}{2} Y^T K_{XX}^{-1} \frac{\partial K_{XX}}{\partial\theta} K_{XX}^{-1} Y - \frac{1}{2} tr(K_{XX}^{-1} \frac{\partial K_{XX}}{\partial\theta}) \qquad (12)$$

Here the hyperparameters of the prior are $\theta = \left\{l_2, \sigma_2^2, \sigma_{n1}^2, \sigma_{n2}^2\right\}$. These correspond to the hyperparameters of the independent covariance function K_{22} and errors in the measurements σ_{n1}^2 and σ_{n2}^2. Calculating the negative log-marginal likelihood involves inverting the matrix $K_{XX}+\Sigma$. The size of the $K_{XX}+\Sigma$ matrix depends on total number of input points N, hence inverting the matrix becomes intractable for large number of input points.

In the next section we describe how to solve the problem of inverting huge $K_{XX}+\Sigma$ matrices using sparse GP regression. We also elaborate on how variational approximation overcomes the problem of overfitting by providing a distance measure between two approximate models.

3 SPARSE GP REGRESSION

The above GP approach is intractable for large datasets. For a multi-output GP as defined in section 2.2 the covariance matrix is of size N, where $O\left(N^3\right)$ time is needed for inference and $O\left(N^2\right)$ memory for storage. Thus, we need to consider approximate methods in order to deal with large datasets. Sparse methods use a small set of m function points as support or inducing variables.

Suppose we use m inducing variables to construct our sparse GP. The inducing variables are the latent function values evaluated at inputs x_M. Learning x_M and the hyperparameters θ is the problem we need to solve in order to obtain a sparse GP method. An approximation to the true log marginal likelihood in equation 11 can allow us to infer these quantities.

3.1 Related Work

Before explaining variational inference method, we review FITC method proposed by (Snelson and Ghahramani, 2006). The approximate log-marginal likelihoods have the form

$$F = log[GP(y|0, Q_{nn}+\Sigma)] \qquad (13)$$

where Q_{nn} is an approximation to the true covariance K_{nn}. In FITC approach, the Q_{nn} takes the form,

$$Q_{nn} = diag[K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}] + K_{nm}K_{mm}^{-1}K_{mn} \quad (14)$$

Here, K_{mm} is a $m \times m$ covariance matrix on inducing points x_m, K_{nm} is a $n \times m$ cross-covariance matrix between the training and inducing points.

Hence the position of x_M now defines the approximated marginal likelihood. The maximization of the marginal likelihood in equation 13 with respect to (x_M; θ), is prone to over-fitting especially when the number of variables in x_M is large. Fitting a modified sparse GP model implies that the full GP model is not approximated in a systematic way since there is no distance measure between the two models that is minimized.

3.2 Variational Approximation

During variational approximation, we seek to apply an approximate variational inference procedure where we introduce a variational distribution $q(x)$ to approximate the true posterior distribution $p(x|y)$. We take the variational distribution to have a factorized Gaussian form as given in equation 15

$$q(x) = \mathcal{N}(x|\mu, A) \quad (15)$$

Here, μ and A are parameters of the variational distribution. Using this variational distribution we can express a Jensens lower bound on the $logP(y)$ that takes the form:

$$\begin{aligned} F(q) &= \int q(x) log \frac{p(y|x)p(x)}{q(x)} dx \\ &= \int q(x) log p(y|x) dx - \int q(x) log \frac{q(x)}{p(x)} dx \\ &= \bar{F}_q - KL(q||p) \quad (16) \end{aligned}$$

Where the second term is the negative Kullback Leibler divergence between the variational posterior distribution $q(x)$ and the prior distribution $p(x)$. To determine the variational quantities (x_M, θ), we minimize the KL divergence $KL(q||p)$. This minimization is equivalently expressed as the maximization of the variational lower bound of the true log marginal likelihood

$$F_V = log(\mathcal{N}[y|0, \sigma^2 I + Q_{nn}]) - \frac{1}{2\sigma^2} Tr(\tilde{K}) \quad (17)$$

where $Q_{nn} = K_{nm}K_{mm}^{-1}K_{mn}$ and $\tilde{K} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$. The novelty of the above objective function is that it contains a regularization trace term: $-\frac{1}{2\sigma^2}Tr(\tilde{K})$. Thus, F_V attempts to maximize the log likelihood Q_{nn} and simultaneously minimize the trace $Tr(\tilde{K})$. $Tr(\tilde{K})$ represents the total variance of the conditional prior. When $Tr(\tilde{K})$ = 0, $K_{nn} = K_{mp}K_{mm}^{-1}K_{mn}$, which means that the inducing variables can exactly reproduce the full GP prediction.

3.3 Variational Approximation on Multi-output GP

(Álvarez et al., 2010) derived a variational approximation to inference of multi-output regression using convolution processes. They introduce the concept of variational inducing kernels that allows them to efficiently sparsify multi-output GP models having white noise latent functions. White noise latent functions are needed while modelling partial differential equations. These variational inducing kernels are later used to parametrize the variational posterior $q(X)$.

For our expression of the joint kernel, the auto- and cross-covariance functions are dependent on the independent covariance function K_{22}. Hence the independent output function in our case behaves like a latent function for the case of (Álvarez et al., 2010). Moreover, the latent function or independent output is a physical process and hence will almost never be white noise. Henceforth we will place the inducing points on the input space. We now extend the variational inference method to deal with multiple outputs.

We try to approximate the joint-posterior distribution $p(X|Y)$ by introducing a variational distribution $q(X)$. In the case of varying number of inputs for different outputs, we place the inducing points over the input space and extend the derivation of (Titsias, 2009) to multi-output case.

$$q(X) = \mathcal{N}(X|\mu, A) \quad (18)$$

Here μ and A are parameters of the variational distribution. We follow the derivation provided in section 3.2 and obtain the lower bound of true marginal likelihood.

$$F_V = log(\mathcal{N}[Y|0, \sigma^2 I + Q_{XX}]) - \frac{1}{2\sigma^2} Tr(\tilde{K}) \quad (19)$$

where $Q_{XX} = K_{XX_M}K_{X_MX_M}^{-1}K_{X_MX}$ and $\tilde{K} = K_{XX} - K_{XX_M}K_{X_MX_M}^{-1}K_{X_MX}$. K_{XX} is the joint-covariance matrix derived using equation 6 or 7 using the input vector X defined in section 1. $K_{X_MX_M}$ is the joint covariance function on the inducing points X_M, such that $X_M = [x_{M1}, x_{M2}, ..., x_{M2}]$. We assume that the inducing points x_{Mi} will be same for all the outputs, hence

$x_{M1} = x_{M2} = ... = x_{M2} = x_M$. While K_{XX_M} is the cross-covariance matrix between X and X_M.

Note that this bound consists of two parts. The first part is the log of a GP prior with the only difference that now the covariance matrix has a lower rank of MD. This form allows the inversion of the covariance matrix to take place in $O\left(N(MD)^2\right)$ time. The second part as discussed above can be seen as a penalization term that regularizes the estimation of the parameters.

The bound can be maximized with respect to all parameters of the covariance function; both model hyperparameters and variational parameters. The optimization parameters are the inducing inputs x_M, the hyperparameters θ of the independent covariance matrix K_{22} and the error while measuring the outputs Σ. There is a trade-off between quality of the estimate and amount of time taken for the estimation process. On the one hand the number of inducing points determine the value of optimized negative log-marginal likelihood and hence the quality of the estimate. While, on the other hand there is a computational load of $O\left(N(MD)^2\right)$ for inference. We increase the number of inducing points until the difference between two successive likelihoods is below a predefined quantity.

4 NUMERICAL RESULTS

In this section we provide numerical illustration to the theoretical derivations in the earlier sections. We start with a synthetic problem where we try to learn the model over derivative and quadratic relationships. We compare the error values for the variational approximation with respect to the full GP approach. Finally we look at the results in presence of a real world dataset related to flight loads estimation.

The basic toolbox used for this paper is GPML provided with (Rasmussen and Williams, 2005), we generate covariance functions to handle relationships as described in equations 6 and 7 using the "Symbolic Math Toolbox" in MATLAB 2014b. Since variational approximation was not coded in GPML we have wrapped the variational inference provided in gpStuff (Vanhatalo et al., 2013) in the GPML toolbox. All experiments were performed on an Intel quad-core processor with 4Gb RAM.

4.1 Numerical Results on Theoretical Data

We begin by considering the relationship between two latent output functions as described in equation 2. Such that

$$\begin{aligned} f_2 &\sim GP[0, K_{SE}(0.1,1)] \\ \sigma_{n2} &\sim \mathcal{N}[0, 0.2] \\ \sigma_{n1} &\sim \mathcal{N}[0, 2] \end{aligned} \tag{20}$$

$K_{SE}(0.1,1)$ means squared exponential kernel with length scale 0.1 and variance as 1.

4.1.1 Differential Relationship

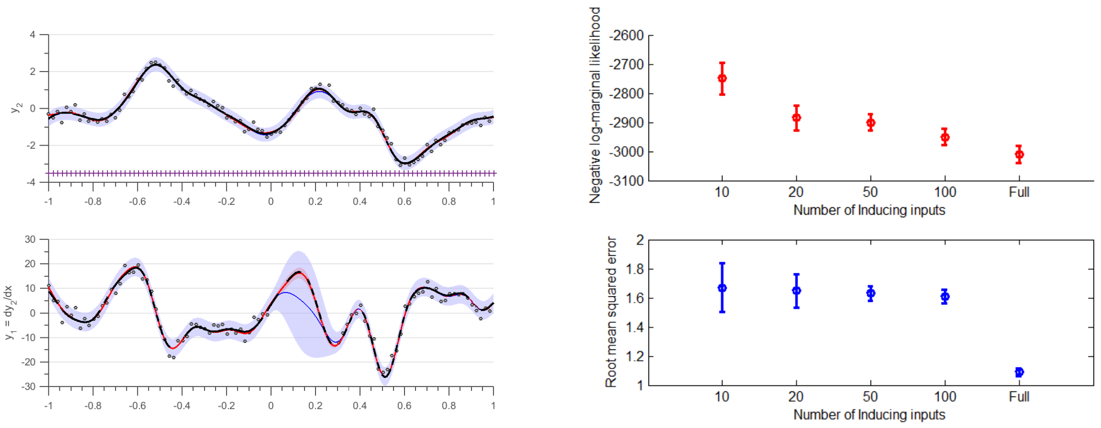
We take the case of a differential relationship $g(.)$, such that $g(f,x) = \frac{\partial f}{\partial x}$. Since the differential relationship $g(.)$ is linear in nature we use the equation 6 to calculate the auto- and cross-covariance functions as shown in table 1.

Table 1: Auto- and cross-covariance functions for a differential relationship.

Initial Covariance	K_{22}	$\sigma^2 exp(\frac{-1}{2}\frac{d^2}{l^2})$
Cross-Covariance	K_{12}	$\sigma^2 \frac{d}{l^2} exp(\frac{-1}{2}\frac{d^2}{l^2})$
Auto-covariance	K_{11}	$\sigma^2 \frac{d^2-l^2}{l^4} exp(\frac{-1}{2}\frac{d^2}{l^2})$

To generate data a single function is drawn from f_2 as described in equation 20 which then is used to calculate y_1 and y_2. 10,000 training points are generated for both the outputs. Values of y_2 for $x \in [0, 0.3]$ are removed from the training points. Next we optimize the lower bound of log-marginal likelihood using variational inference, for independent GP's on y_1 and y_2 as described in 3.2. Later we optimize the same lower bound but with a joint-covariance approach as described in section 3.3 using y_1, y_2 and $g(.)$. As explained in section 3.3 we settled on using 100 inducing points for this experiment because there was negligible increase in lower bound F_V of log-marginal likelihood upon increasing the inducing points.

In figure 1(a) we show the independent (blue shade) and joint fit (red shade) of two GP for the differential relationship. The GP model with joint covariance gives better prediction even in absence of data of y_2 for $x \in [0, 0.3]$ because transfer of information is happening from observations of y_1 present at those locations. Nonetheless we see an increase in variance at masked 'x' values even for joint-covariance kernel. The distribution of inducing points may look even in the diagram but are uneven at places where we have removed the data from y_2.



(a) Independent fit for two GP's in blue variance is represented by light blue region and mean is represented by solid blue line. Variance and mean of the dependent are represented in red region and solid red line. The dashed black line represents the true latent function values; noisy data is denoted by circles. Experiment was run on 10,000 points but only 100 data points are plotted to increase readibility.Here $f_1 = \frac{\partial f_2}{\partial x}$, $y_1 = f_1 + \sigma_{n1}$ and $y_2 = f_2 + \sigma_{n2}$. The $+$ points refer to the location of the inducing points in the inference.

(b) The figure shows progression of different measures upon increasing number of inducing points from 10, 20, 50 to 100 and finally full Multi-output physics based GP. The top figure in red shows the value of mean and variance of negative log-marginal likelihood, while the bottom figure in blue shows the mean and variance of root mean squared error. 10 sets of experiments were run on 75% of the data as training set and 25% of the data as the test set, the training and test sets were chosen randomly.

Figure 1: Experimental results for differential relationship with approximate inference using variational approximation.

For the second experiment we generate 1000 points and compare the Root Mean Squared Error (RMSE) and log-marginal likelihood between full relationship kernel as described in section 2.3 and variational inference relationship kernel as described in section 3.3, using 10, 20, 50 and 100 inducing points. 10 sets of experiments were run on 75% of the data as training set and 25% of the data as the test set, the training and test sets were chosen randomly. We learn the optimal values of hyper-parameters and inducing points for all the 10 sets of experiments of training data. Finally, RMSE values are evaluated with respect to the test set and negative log-marginal likelihood are evaluated for each learned model. The RMSE values are calculated for only the dependent output y_1 and then plotted in the figures below.

In figure 1(b) the mean and variance of negative log-marginal likelihood for 10 sets of experiments are calculated with varying number of inducing inputs at the top in red. As expected upon increasing the number of inducing points we see that the values of mean negative log-marginal likelihood tends to become an asymptote reaching the value of mean negative log-marginal likelihood of a full multi-output GP. In the bottom figure we have the mean and variance of RMSE for 10 sets of experiments. We see that the mean value of RMSE does not vary drastically upon increasing the inducing points but the value of variance gets reduced. One can observe an almost constant RMSE in figure 1(b) with a sudden decrease for a full GP (as M tends to N). Although this behaviour was expected, providing an explanation for the slow decrease with low values of M requires further investigation.

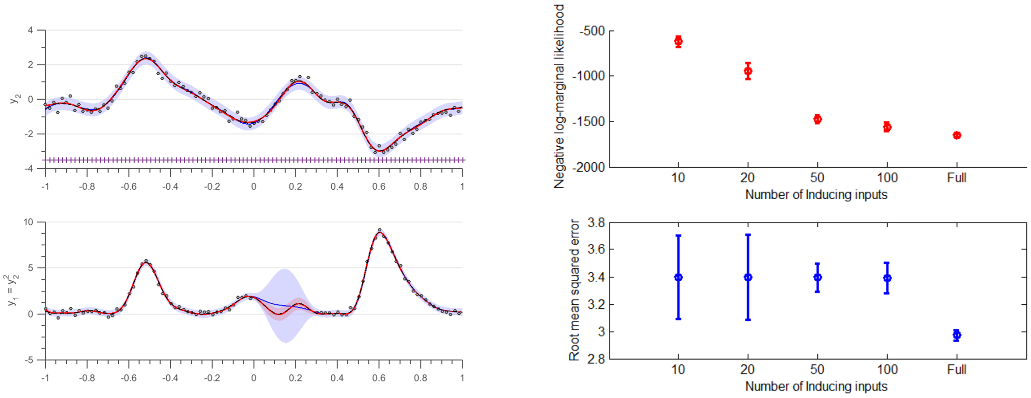
4.1.2 Quadratic Relationship

Now we take the case of a quadratic relationship $g(.)$, such that $g(f,x) = f^2$. Since the quadratic relationship $g(.)$ is non-linear in nature we use the equation 7 to calculate the auto- and cross-covariance functions as shown in table 2. The Jacobin L for this case becomes $2\bar{f}_2(x)$. Here the value of $\bar{f}_2(x)$ is the mean value of the latent-independent output function f_2 calculated at the input point x.

Table 2: Auto- and cross-covariance functions for a quadratic relationship.

Initial Covariance	K_{22}	$\sigma^2 exp(\frac{-1}{2}\frac{d^2}{l^2})$
Cross-Covariance	K_{12}	$2\bar{f}_2(x)K_{22}$
Auto-covariance	K_{11}	$4(\bar{f}_2(x)^2K_{22} + 2K_{22}^2)$

As stated in the earlier section 4.1.1 the output data was generated using the same draw of f_2 as explained in equation 20. 10,000 training points are generated for both the outputs. Values of y_2 for $x \in [0, 0.3]$ are removed from the training points. Note that the above calculated auto- and cross-covariances are third



(a) Independent fit for two GP in blue variance is represented by light blue region and mean is represented by solid blue line. Variance and mean of the dependent are represented in red region and solid red line. The dashed black line represents the true latent function values; noisy data is denoted by circles only 100 data points are plotted to increase readibility. Here $f_1 = f_2^2$, $y_1 = f_1 + \sigma_{n1}$ and $y_2 = f_2 + \sigma_{n2}$. The $+$ points refer to the location of the inducing points in the inference.

(b) The figure shows progression of different measures upon increasing number of inducing points from 10, 20, 50 to 100 and finally full Multi-output physics based GP. The top figure in red shows the value of mean and variance of negative log-marginal likelihood, while the bottom figure in blue shows the mean and variance of root mean squared error. 10 sets of experiments were run on 75% of the data as training set and 25% of the data as the test set, the training and test sets were chosen randomly.

Figure 2: Experimental results for quadratic relationship with approximate inference using variational approximation.

order Taylor series approximations as presented in equation 7.

Calculating the Jacobin $\bar{f}_2(x)$ at the inducing and input points was a requirement of the algorithm. Moreover, during the optimization process the value of x_m keeps on changing, since the $\bar{f}_2(x_m)$ depends on the value of x_m. Many a times we don't have value of latent independent output function at these new points. To solve this problem we first learn an independent model of the output y_2 recover an estimate of $\bar{f}_2(x)$ and use this value to calculate the required auto- and cross-covariances.

In figure 2(a) we show the independent and joint fit of two GP for the quadratic relationship. Even in the presence of the error the joint-covariance GP model gives better prediction because of transfer of information. The GP model with joint covariance gives better prediction even in absence of data of y_2 for $x \in [0, 0.3]$ because transfer of information is happening from the observations of y_1 present at those locations.

Upon performing the second experiment as described in section 4.1.1 we get the figure 2(b). The top part in the figure is the mean and variance of negative log-marginal likelihood for varying number of inducing points. We see that the value mean negative log-marginal likelihood reaches the value of full negative log-marginal likelihood. In the lower figure we have mean and variance of RMSE of the dependent output y_1. As seen in the earlier experiment we observe high amount of variance for lower inducing points due to the approximate nature.

4.2 Numerical Results on Flight Test Data

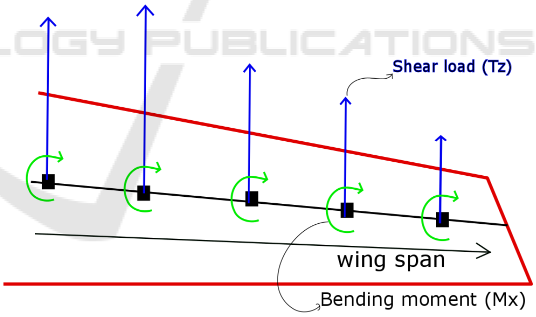
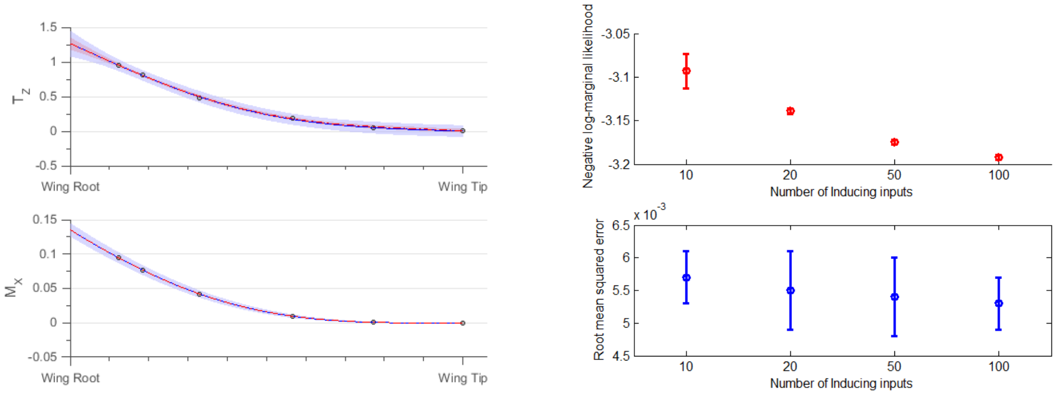


Figure 3: Wing Load Diagram.

In this section we conduct experiments, applying our approach on the flight loads data recovered during flight test campaigns at Airbus. We look at normalized data of a simple longitudinal maneuver. The two outputs in our case are shear load T_z and bending moment M_x as described in figure 3. The input space is 2-dimensional space with wing span η or point of action of forces as he first input variable and angle of attack α as the second input variable. The maneuver is quasi-static which means that airplane is in equilibrium at all time and there are no dynamic effects observed by the aircraft. The relation between T_z and



(a) Independent fit for two GP in blue variance is represented by light blue region and mean is represented by solid blue line. Variance and mean of the dependent are represented in red region and solid red line. The dashed black line represents the true latent function values; noisy data is denoted by circles only 1 α step is plotted.

(b) The figure shows progression of different measures upon increasing number of inducing points from 10, 20, 50 to 100. The top figure in red shows the value of mean and variance of negative log-marginal likelihood, while the bottom figure in blue shows the mean and variance of root mean squared error.

Figure 4: Experimental results for aircraft flight loads data with approximate inference using variational approximation.

M_x can be written as:

$$M_x(\eta,\alpha) = \int_{\eta}^{\eta_{edge}} T_Z(x,\alpha)(x-\eta)dx. \tag{21}$$

Note that the above equation is calculated only on the η axis. Here, η_{edge} denotes the edge of the wing span. The above equation is linear in nature and hence we will use equation 6 to calculate the auto- and cross-covariance functions. The forces are measured at 5 points on the wing span and at 8800 points in the second axis. We follow the procedure described in earlier experiments where we compare plots of relationship-imposed multioutput GP and independent GP. Secondly, we compare the measures of negative-log marginal likelihood and RMSE for varying number of inducing points.

In figure 4(a) we see the plot of independent and dependent T_Z at the top with the dependent plot in red and independent plot in blue. In the bottom part of figure we show the plot for M_X. Only one α is plotted here for better viewing. 100 inducing points in the input space are used to learn and plot the figure. We see that the region in red, has a tighter variance than the one in blue confirming the improvement by our method. The relationship in equation 21 is acting as extra information in tightening the error margins of the loads estimation. This becomes very useful when we need to identify faulty data because the relationship will impose a tighter bound on the variance and push faulty points out of the confidence interval.

In figure 4(b) we see how the negative log-marginal likelihood and RMSE plots improve upon increasing number of inducing points. As expected the likelihood and RMSE improves with more inducing points because of the improvement in the approximate inference. We settle on choosing 100 inducing points for figure 21 because there is not much improvement in RMSE and likelihood upon increasing the inducing points.

5 CONCLUSIONS AND FUTURE WORK

We have presented a sparse approximation for physics-based multiple output GP's, reducing the computational load and memory requirements for inference. We extend the variational inference as derived by (Titsias, 2009) and reduce the computational load for inference from $O\left(N^3\right)$ to $O\left(N(MD)^2\right)$ and load on memory from $O\left(N^2\right)$ to $O(NMD)$.

We have demonstrated our strategy to work on both linear and non-linear relationships, in presence of large amount of data. The strategy of imposing relationships is one way of introducing domain knowledge into the GP regression process. We conclude that due to the support of added information provided by the physical relationship we have a tighter confidence interval for the joint-GP approach, eventually leading to better estimates. Additionally, the results in identification of measurements incoherent with the physics of the system and a more accurate estimation of the latent function, which is crusial in aircraft design and modeling. The approach can be extended to larger number of related outputs giving a richer GP model.

Aircraft flight domain can be divided into various cluster of maneuvers. Each cluster of maneuvers has a specific set of features and mathematically modelled behaviour, which can also be seen as domain knowledge. Recent advancements in approximate inference of GP such as "Distributed Gaussian Process" (Deisenroth and Ng, 2015) distribute the input domain into several clusters. Such kind of approximate inference technique should be explored in the future for our kind of dataset. Future work deals with handling clustered input space with individual features.

REFERENCES

Alvarez, M. and Lawrence, N. D. (2009). Sparse convolved gaussian processes for multi-output regression. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 57–64. Curran Associates, Inc.

Alvarez, M. A., Luengo, D., and Lawrence, N. D. (2009). Latent force models. In Dyk, D. A. V. and Welling, M., editors, *AISTATS*, volume 5 of *JMLR Proceedings*, pages 9–16. JMLR.org.

Álvarez, M. A., Luengo, D., Titsias, M. K., and Lawrence, N. D. (2010). Efficient multioutput gaussian processes through variational inducing kernels. In Teh, Y. W. and Titterington, D. M., editors, *AISTATS*, volume 9 of *JMLR Proceedings*, pages 25–32. JMLR.org.

Bonilla, E., Chai, K. M., and Williams, C. (2008). Multi-task gaussian process prediction. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 153–160. MIT Press, Cambridge, MA.

Boyle, P. and Frean, M. (2005). Dependent gaussian processes. In *In Advances in Neural Information Processing Systems 17*, pages 217–224. MIT Press.

Constantinescu, E. M. and Anitescu, M. (2013). Physics-based covariance models for gaussian processes with multiple outputs. *International Journal for Uncertainty Quantification*, 3.

Deisenroth, M. P. and Ng, J. W. (2015). Distributed gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1481–1490.

Quionero-candela, J., Rasmussen, C. E., and Herbrich, R. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6:2005.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Snelson, E. and Ghahramani, Z. (2006). Sparse gaussian processes using pseudo-inputs. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 1257–1264. MIT press.

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 Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 1057–1064. MIT Press.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.

Titsias, M. K. (2009). Variational learning of inducing variables in sparse gaussian processes. In *In Artificial Intelligence and Statistics 12*, pages 567–574.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). Gpstuff: Bayesian modeling with gaussian processes. *J. Mach. Learn. Res.*, 14(1):1175–1179.