GAUSSIAN REGRESSION BASED ON MODELS WITH TWO STOCHASTIC PROCESSES

W. E. Leithead[1], **Kian Seng Neo, D. J. Leith**

Hamilton Institute, National University of Ireland,
Maynooth, Co. Kildare, Ireland

Abstract: When data contains components with different characteristics and it is required to identify both, standard Gaussian regression, based on a model with a single stochastic process, is inadequate. In this paper, a novel adaptation of Gaussian regression, based on models with two stochastic processes, is presented. In both the prior and posterior joint probability distributions, the Gaussian processes for the two components are independent. The effectiveness of the revised Gaussian regression method is demonstrated by application to wind turbine time series data.

Keywords: Identification, Gaussian processes, independent priors, independent posteriors

1. INTRODUCTION[2]

Following some initial publications in the late 1990s (e.g., MacKay (1998), Williams (1999)), interest has grown quickly into the application of Gaussian processes prior models to data analysis; e.g. Gibbs and MacKay (2000), Sambu, et al. (2000), Yoshioka and Ishii (2001), Leithead *et al*. (2003). When the data contains components with different characteristics and it is required to identify both, the standard model, consisting of a single Gaussian process, is inadequate. In this paper, a novel adaptation of the Gaussian regression methodology, based on models with two stochastic processes, is proposed (Section 4) and its effectiveness is demonstrated by its application (Section 5) to wind turbine time series data, specifically, site measurements of rotor speed for a commercial 1MW machine.

2. GAUSSIAN PROCESS PRIOR MODELS

A brief explanation of the standard Gaussian regression methodology is given below. Consider a smooth scalar nonlinear function f(.) dependent on the explanatory variable, $\mathbf{z} \in D \subseteq \Re^p$. Suppose N measurements, $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, of the value of the function with additive Gaussian white measurement noise, i.e. $y_i=f(\mathbf{z}_i)+n_i$, are available and denote them by M. It is of interest here to use this data to learn the mapping $f(\mathbf{z})$ or, more precisely, to determine a probabilistic description of $f(\mathbf{z})$ on the domain, D, containing the data. Note that this is a regression formulation and it is assumed the input, $\mathbf{z}$, is noise free. The probabilistic description of the function, $f(\mathbf{z})$, adopted is the stochastic process, $f_{\mathbf{z}}$, with the $E[f_{\mathbf{z}}]$, as $\mathbf{z}$ varies, interpreted to be a fit to $f(\mathbf{z})$. By necessity, to define the stochastic process, $f_{\mathbf{z}}$, the probability distributions of $f_{\mathbf{z}}$ for every choice of value of $\mathbf{z} \in D$ are required together with the joint probability distributions of $f_{\mathbf{z}_i}$ for every choice of finite sample, $\{\mathbf{z}_1,\ldots,\mathbf{z}_k\}$, from D, for all k>1. Given the joint probability distribution for $f_{\mathbf{z}_i}$, i=1..N, and the joint probability distribution for n_i, i=1..N, the joint probability distribution for y_i, i=1..N, is readily obtained since the measurement noise, n_i, and the

[1] Tel.: +44-141-5482408, Fax: +44-141-5484203; e.mail: w.leithead@eee.strath.ac.uk.
[2] This work was supported by Science Foundation Ireland grant, 00/PI.1/C067, and by the EPSRC, GR/M76379/01.

$f(\mathbf{z}_i)$ (and so the $f_{\mathbf{z}_i}$) are statistically independent. M is a single event belonging to the joint probability distribution for y_i, i=1..N.

In the Bayesian probability context, the prior belief is placed directly on the probability distributions describing $f_{\mathbf{z}}$ which are then conditioned on the information, M, to determine the posterior probability distributions. In the Gaussian process prior model, the prior probability distributions for the $f_{\mathbf{z}}$ are all Gaussian with zero mean (in the absence of any evidence the value of $f(\mathbf{z})$ is as likely to be positive as negative). To complete the statistical description, requires only a definition of the covariance function $C_f(\mathbf{z}_i,\mathbf{z}_j)=E[f_{\mathbf{z}_i}, f_{\mathbf{z}_j}]$, for all $\mathbf{z}_i$ and $\mathbf{z}_j$. The resulting posterior probability distributions are also Gaussian. This model is used to carry out inference as follows.

Clearly $p(f_{\mathbf{z}} \mid M) = p(f_{\mathbf{z}}, M)/p(M)$ where p(M) acts as a normalising constant. Hence, with the Gaussian prior assumption,

$$p(f_{\mathbf{z}} \mid M) \propto \exp\left[-\tfrac{1}{2}\begin{bmatrix} f_{\mathbf{z}} & \mathbf{Y}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\Lambda_{11} & \boldsymbol{\Lambda}_{21}^{\mathrm{T}} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22}\end{bmatrix}^{-1}\begin{bmatrix} f_{\mathbf{z}} \\ \mathbf{Y}\end{bmatrix}\right] \quad (1)$$

where $\mathbf{Y}=[y_1,\cdots y_N]^T$, Λ_{11} is $E[f_{\mathbf{z}}, f_{\mathbf{z}}]$, the ijth element of the covariance matrix $\boldsymbol{\Lambda}_{22}$ is $E[y_i, y_j]$ and the i^{th} element of vector $\boldsymbol{\Lambda}_{21}$ is $E[y_i, f_{\mathbf{z}}]$. Both Λ_{11} and $\boldsymbol{\Lambda}_{21}$ depend on $\mathbf{z}$. Applying the partitioned matrix inversion lemma, it follows that

$$p(f_{\mathbf{z}} \mid M) \propto \exp\left[-\tfrac{1}{2}(f_{\mathbf{z}}-\hat{f}_{\mathbf{z}})\Lambda_{\mathbf{z}}^{-1}(f_{\mathbf{z}}-\hat{f}_{\mathbf{z}})\right] \quad (2)$$

with $\hat{f}_{\mathbf{z}} = \boldsymbol{\Lambda}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}\mathbf{Y}$ and $\Lambda_{\mathbf{z}} = \Lambda_{11} - \boldsymbol{\Lambda}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}$. Therefore, the prediction from this model is that the most likely value of $\mathbf{f}(\mathbf{z})$ is the mean, $\hat{f}_{\mathbf{z}}$, with variance $\Lambda_{\mathbf{z}}$. Note that $\hat{f}_{\mathbf{z}}$ is simply a $\mathbf{z}$-dependent weighted linear combination of the measured data points, $\mathbf{Y}$, using weights $\boldsymbol{\Lambda}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}$. The measurement noise, n_i, i=1,..N, is statistically independent of $f(\mathbf{z}_i)$, i=1,..N, and has covariance matrix $\mathbf{B}$. Hence, the covariances for the measurements, y_i, are simply

$$E[y_i,y_j] = E[f_{\mathbf{z}_i}, f_{\mathbf{z}_j}] + B_{ij} \;;\; E[y_i, f_{\mathbf{z}}] = E[f_{\mathbf{z}_i}, f_{\mathbf{z}}] \quad (3)$$

The prior covariance function is generally dependent on a few hyperparameters, θ. To obtain a model given the data, M, the hyperparameters are adapted to maximise the likelihood, $p(M|\theta)$, or equivalently minimise the negative log likelihood, $L(\theta)$, where

$$L(\theta) = \frac{1}{2}\log\det C(\theta) + \frac{1}{2}\mathbf{Y}^T C(\theta)^{-1}\mathbf{Y} \quad (4)$$

with $C(\theta) = \boldsymbol{\Lambda}_{22}$.

3. MODELS WITH COMPOUND COVARIANCE FUNCTIONS

The procedure outlined in Section 2 is very effective when used to identify a single function. However, suppose that the measurements are not of a single function but of the sum of two functions with different characteristics; that is, the measured values are $y_i=f(\mathbf{z}_i)+g(\mathbf{z}_i)+n_i$. A possible probabilistic description of $h(\mathbf{z})=f(\mathbf{z})+g(\mathbf{z})$ is by means of the sum of two independent Gaussian processes, $f_{\mathbf{z}}$ and $g_{\mathbf{z}}$. Let the covariance functions for $f_{\mathbf{z}}$ and $g_{\mathbf{z}}$ be $C_f(\mathbf{z}_i,\mathbf{z}_j)$ and $C_g(\mathbf{z}_i,\mathbf{z}_j)$, respectively, then $h_{\mathbf{z}}=(f_{\mathbf{z}}+g_{\mathbf{z}})$ is itself a stochastic process with covariance function, $C_h=(C_f+C_g)$, since $f_{\mathbf{z}}$ and $g_{\mathbf{z}}$ are independent.

Following Section 2, the prior joint probability distribution for $\mathbf{H}=[h_{\mathbf{z}_1},\cdots h_{\mathbf{z}_N}]^T$ and $\mathbf{Y}$ is Gaussian with mean zero and covariance matrix,

$$E\left[\begin{bmatrix}\mathbf{H} \\ \mathbf{Y}\end{bmatrix}\begin{bmatrix}\mathbf{H}^T & \mathbf{Y}^T\end{bmatrix}\right] = \begin{bmatrix}\boldsymbol{\Lambda}_{\mathbf{HH}} & \boldsymbol{\Lambda}_{\mathbf{HH}} \\ \boldsymbol{\Lambda}_{\mathbf{HH}} & \mathbf{Q}_{\mathbf{H}}\end{bmatrix} \quad (5)$$

with $\boldsymbol{\Lambda}_{\mathbf{HH}} = E[\mathbf{H}\mathbf{H}^T]$ and $\mathbf{Q}_{\mathbf{H}} = \mathbf{B} + \boldsymbol{\Lambda}_{\mathbf{HH}}$. Conditioning on the data set, the posterior probability distribution for $\mathbf{H}$ remains Gaussian with mean and covariance matrix, respectively,

$$\boldsymbol{\Lambda}_{\mathbf{HH}}\mathbf{Q}_{\mathbf{H}}^{-1}\mathbf{Y} \text{ and } \overline{\boldsymbol{\Lambda}}_{\mathbf{HH}} = \boldsymbol{\Lambda}_{\mathbf{HH}} - \boldsymbol{\Lambda}_{\mathbf{HH}}\mathbf{Q}_{\mathbf{H}}^{-1}\boldsymbol{\Lambda}_{\mathbf{HH}} \quad (6)$$

The prediction for $\mathbf{H}$ is the mean ($\boldsymbol{\Lambda}_{\mathbf{HH}}\mathbf{Q}_{\mathbf{H}}^{-1}\mathbf{Y}$) with confidence interval ± 2 standard deviations ($\pm 2\sqrt{\mathbf{D}}$, where the diagonal matrix $\mathbf{D} = \mathrm{diag}(\overline{\boldsymbol{\Lambda}}_{\mathbf{HH}})$).

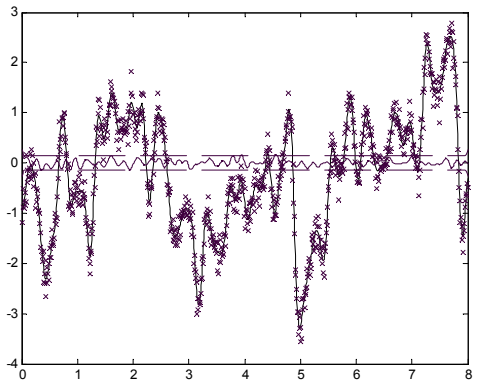

Fig. 1. Two length-scale data (x x), prediction (—), error and confidence interval (= =)

Example 1: A commonly used prior covariance function for a Gaussian process with scalar explanatory variable is

$$a\exp\left[-\tfrac{1}{2}d(z_i - z_j)^2\right] \quad (7)$$

It ensures that measurements associated with nearby values of the explanatory variable should have higher covariance than more widely separated values of the explanatory variable; a is related to the amplitude of the Gaussian process and d inversely related to its length-scale. Let the covariance function for $f_{\mathbf{z}}$ be (7) with a=1.8 and d=2.5, and the covariance function for $g_{\mathbf{z}}$ be (7) with a=0.95 and d=120; that is, $f_{\mathbf{z}}$ has a long length-scale and $g_{\mathbf{z}}$ a short length-scale. Also, let the measurement noise be Gaussian white noise with variance b=0.04, i.e. $B_{ij}=b\delta_{ij}$., where δ_{ij} is the Kronecker delta. Gaussian regression is applied to a

set of 800 measurements, $y_i=f(\mathbf{z}_i)+g(\mathbf{z}_i)+n_i$, at constant interval, 0.01 with the $f(\mathbf{z}_i)$ and $g(\mathbf{z}_i)$ the sample values for the above stochastic processes $f_{\mathbf{z}}$ and $g_{\mathbf{z}}$, respectively. The data values are shown in figure 1 together with the prediction, error and confidence intervals obtained using (6).

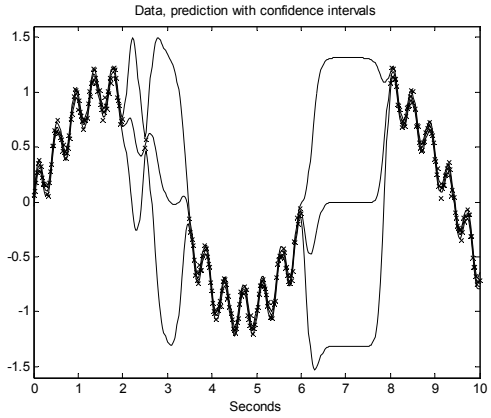

Fig. 2. Variable density data, prediction and confidence interval.

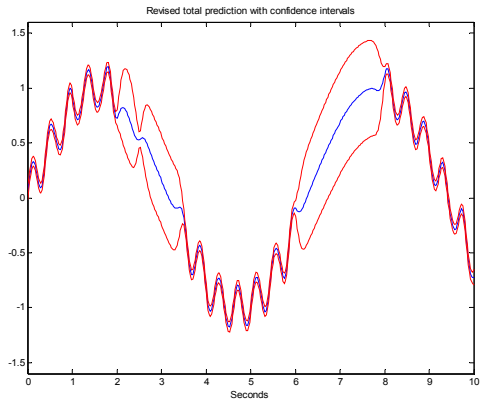

Fig. 3. Prediction and confidence interval with long and short length scale components.

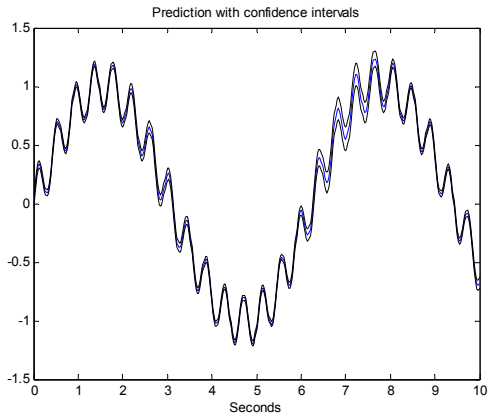

Fig. 4. Prediction and confidence interval with long length scale and periodic components.

Remark 1: In Example 1, the probabilistic description for $h(\mathbf{z})$ is by means of a single Gaussian process, $h_{\mathbf{z}}$, with the compound covariance function, $C_h=(C_f+C_g)$. An alternative simpler probabilistic description would be by means of a Gaussian process, $\widetilde{h}_{\mathbf{z}}$, with the covariance function, $\widetilde{C}_h$, of the form (7). A suitable value of the length scale hyperparameter, d, is the same as that for the short length-scale in Example 1, i.e. $\widetilde{h}_{\mathbf{z}}$ has the same short length-scale as $g_{\mathbf{z}}$ in Example 1, but a suitable value of the amplitude hyperparameter, a, is larger, i.e. the value maximising the likelihood of the data. A suitable value of the length scale hyperparameter, d,. This simpler probabilistic description is almost equally as effective as the probabilistic description with covariance function C_h, since the prediction and confidence interval at any point depend primarily on nearby data values rather than remoter values.

The benefits for prediction of using a compound covariance function such as C_h, become apparent when the density of the data varies. Consider the data in Figure 2. It clearly contains a long length-scale component and a short length-scale component. Both are sinusoids. However, there are large gaps in the data between 2 and 3.5 (except for two values at 2.5) and between 6 and 8. First, consider the situation when, the covariance function is chosen to be (7) with the hyperparameters adapted such that the value of the length scale hyperparameter, d, corresponds to the short length-scale. The prediction and confidence interval obtained are shown in figure 2. Since it now depends only on nearby values, the prediction over the data gaps is poor. Indeed, no prediction is made over the second gap between 6 and 8. Over the data gaps, the confidence interval, reflecting the uncertainty in the prediction, is much enlarged. Now, consider the situation when the covariance function is chosen to be similar to that of Example 1; that is, it is the sum of two functions, for the long length-scale and short length-scale, respectively. The prediction and confidence interval are shown in figure 3. Over the data gaps, the prediction is improved due to the inclusion of the long length-scale component in the covariance function. The confidence interval, reflecting that the uncertainty is now mainly in the short length scale component, is considerably reduced. Nevertheless, the periodic nature of the short length-scale component in the data can be exploited to further improve the prediction over the gaps. A suitable prior covariance function for a periodic Gaussian process with scalar explanatory variable is

$$a\exp\left[-\tfrac{1}{2}d\sin^2(\pi\lambda(z_i-z_j))\right] \qquad (8)$$

Finally, consider the situation when the covariance function is chosen to be the sum of (7) and (8), the former being for the long length-scale component and the latter for the periodic short-term component in the data. The prediction and confidence interval are shown in figure 4. Over the data gaps, the prediction is much improved and the confidence interval much narrower.

4. MODELS WITH TWO GAUSSIAN PROCESSES

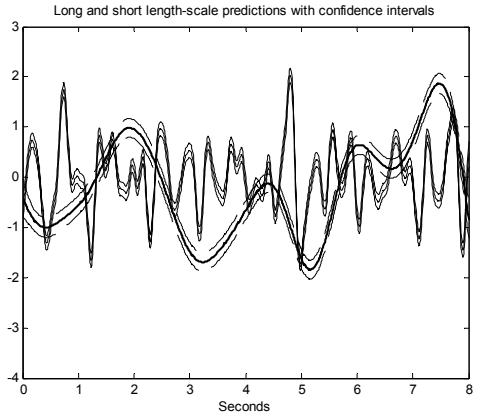

Fig. 5. Ad hoc prediction and confidence intervals.

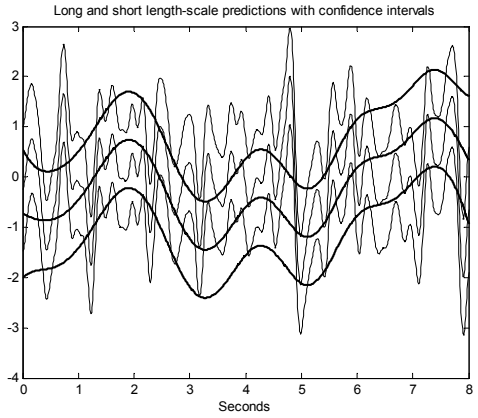

Fig. 6. Prediction and confidence intervals from joint probability distribution.

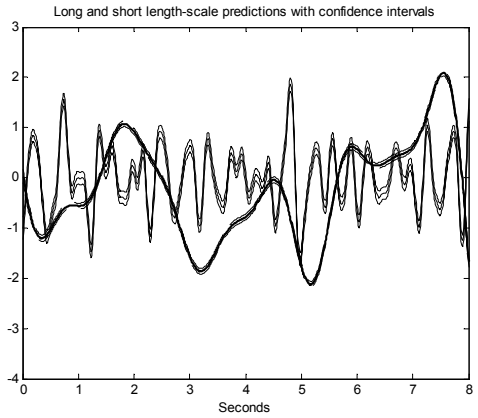

Fig. 7. Prediction and confidence intervals from modified probability distribution of theorem 1.

Rather than the probabilistic description for $h(\mathbf{z})=f(\mathbf{z})+g(\mathbf{z})$ by means of a single stochastic process, $h_{\mathbf{z}}$, see Section 3, the requirement here is to determine a separate probabilistic description for both $f(\mathbf{z})$ and $g(\mathbf{z})$ by means of the two independent Gaussian processes, $f_{\mathbf{z}}$ and $g_{\mathbf{z}}$.

4.1 Ad hoc identification of f(z) and g(z)

A simple procedure to identify the two contributions to the data values might be as follows. Firstly, interpret the measurements to be of the form $f(\mathbf{z}_i)+n_i^*$; that is, to be due solely to the $f_{\mathbf{z}}$ plus white noise with covariance matrix $\mathbf{B}^*=b^*I$. The variance of the noise, b^*, is adapted to be sufficiently large to account for $g(\mathbf{z}_i)+n_i$. Conditioning on the $\mathbf{Y}$, the mean and covariance matrix for $\mathbf{F}=[f_{\mathbf{z}_1},\cdots f_{\mathbf{z}_N}]^T$ are $\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y}$ and $\overline{\mathbf{\Lambda}}_{\mathbf{FF}}=\mathbf{\Lambda_{FF}}-\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{\Lambda_{FF}}$, where $\mathbf{\Lambda_{FF}}=\mathrm{E}[\mathbf{FF}^T]$ and $\hat{\mathbf{Q}}_{\mathbf{F}}=\mathbf{B}^*+\mathbf{\Lambda_{FF}}$. Secondly, interpret the residues with respect to the mean of $\mathbf{F}$, $\overline{\mathbf{R}}=\mathbf{Y}-\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y}=\mathbf{B}^*\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y}$, to be of the form $g(\mathbf{z}_i)+n_i$; that is, to be due to $g_{\mathbf{z}}$. Conditioning on $\overline{\mathbf{R}}$, the mean and covariance matrix for $\mathbf{G}=[g_{\mathbf{z}_1},\cdots g_{\mathbf{z}_N}]^T$ are $\mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\overline{\mathbf{R}}$ and $\overline{\mathbf{\Lambda}}_{\mathbf{GG}}=\mathbf{\Lambda_{GG}}-\mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{\Lambda_{GG}}$, where $\mathbf{\Lambda_{GG}}=\mathrm{E}[\mathbf{GG}^T]$ and $\mathbf{Q_G}=\mathbf{B}+\mathbf{\Lambda_{GG}}$. The predictions and confidence intervals for $\mathbf{F}$ and $\mathbf{G}$ are calculated as for $\mathbf{H}$ in Section 3.

Example 1 (cont.): The above procedure is applied to Example 1 with $f_{\mathbf{z}}$ the long length-scale component and $g_{\mathbf{z}}$ the short length-scale component. The long length-scale (solid line) and short length-scale (dashed line) predictions together with their confidence intervals are shown in figure 5. The value for b^* is 0.48249. The confidence interval for the short length-scale component is narrow and similar in magnitude to the confidence interval in figure 5, as would be expected (see Remark 1). The confidence interval in the long length-scale component is much broader since it has to account through b^* for both the short length-scale component and the measurement noise.

Unfortunately, the probabilistic description for $f(\mathbf{z})$ and $g(\mathbf{z})$, obtained by the above simple procedure, is not coherent. Other than the somewhat dubious procedure of accounting for $g(\mathbf{z}_i)+n_i$ by white noise, the concern is over combining the separate probabilistic descriptions for $\mathbf{F}$ and $\mathbf{G}$ to obtain a similar description for $\mathbf{H}=(\mathbf{F}+\mathbf{G})$ to that of Section 3.

In the context of Example 1 when $f_{\mathbf{z}}$ has a long length-scale and $g_{\mathbf{z}}$ a short length-scale, suppose the prediction, $\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y}$, for $\mathbf{F}$ is interpreted as a de-trending of the data. The description for $\mathbf{G}$ is, then, probabilistic, a Gaussian distribution with mean $\mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\overline{\mathbf{R}}$ and covariance matrix $\overline{\mathbf{\Lambda}}_{\mathbf{GG}}$, whilst the description for $\mathbf{F}$ is deterministic, $\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y}$. The covariance matrix for $\mathbf{H}$ is $\overline{\mathbf{\Lambda}}_{\mathbf{GG}}$ (the short length-scale covariance matrix) and the confidence intervals are narrow as required; see the Remark 1. However, when $f(\mathbf{z})$ need be explicitly identified, this description being deterministic is inadequate.

An alternative is to consider the residues, $\mathbf{R}$, for all possible values of the contribution to the data due to

f_z rather than only the mean. Since the probability distribution for **F** conditioned on **Y** is Gaussian with mean $\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y}$ and covariance matrix $\overline{\mathbf{\Lambda}}_{\mathbf{FF}}$, the probability distribution for **R** conditioned on **Y**, p(**R**|**Y**), is Gaussian with mean $\mathbf{B}^*\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y}$ and covariance matrix $\overline{\mathbf{\Lambda}}_{\mathbf{FF}}$. For each **R** belonging to p(**R**|**Y**), the probability distribution for **G** conditioned on **R**, p(**G**|**R**), remains Gaussian with mean $\mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{R}$ and covariance matrix $\overline{\mathbf{\Lambda}}_{\mathbf{GG}}$. Suppose the hyperparameters are unchanged for each **R** (again somewhat dubious), then the joint probability distribution for $[\mathbf{R}^T,\mathbf{G}^T]^T$ conditioned on **Y**, p(**R**,**G**|**Y**)=p(**G**|**R**)p(**R**|**Y**), is Gaussian with mean ,

$$\begin{bmatrix} \mathbf{B}^*\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y} \\ \mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{B}^*\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y} \end{bmatrix}$$

and covariance matrix,

$$\begin{bmatrix} \overline{\mathbf{\Lambda}}_{\mathbf{FF}} & \overline{\mathbf{\Lambda}}_{\mathbf{FF}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{\Lambda_{GG}} \\ \mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\overline{\mathbf{\Lambda}}_{\mathbf{FF}} & \overline{\mathbf{\Lambda}}_{\mathbf{GG}}+\mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\overline{\mathbf{\Lambda}}_{\mathbf{FF}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{\Lambda_{GG}} \end{bmatrix}$$

It follows immediately that the joint probability distribution for $[\mathbf{F}^T,\mathbf{G}^T]^T$ conditioned on **Y**, is Gaussian with mean vector,

$$\begin{bmatrix} \mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y} \\ \mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{B}^*\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{Y} \end{bmatrix}$$

and covariance matrix,

$$\begin{bmatrix} \overline{\mathbf{\Lambda}}_{\mathbf{FF}} & -\overline{\mathbf{\Lambda}}_{\mathbf{FF}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{\Lambda_{GG}} \\ -\mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\overline{\mathbf{\Lambda}}_{\mathbf{FF}} & \overline{\mathbf{\Lambda}}_{\mathbf{GG}}+\mathbf{\Lambda_{GG}}\mathbf{Q}_{\mathbf{G}}^{-1}\overline{\mathbf{\Lambda}}_{\mathbf{FF}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{\Lambda_{GG}} \end{bmatrix}$$

Consequently, the covariance matrix for **H** is $\overline{\mathbf{\Lambda}}_{\mathbf{GG}}+\mathbf{B}\mathbf{Q}_{\mathbf{G}}^{-1}\overline{\mathbf{\Lambda}}_{\mathbf{FF}}\mathbf{Q}_{\mathbf{G}}^{-1}\mathbf{B}$. In the context of Example 1, this covariance matrix is very similar to $\overline{\mathbf{\Lambda}}_{\mathbf{GG}}$ as required. However, the description for f(**z**) remains inadequate since its confidence interval is unnecessarily broad through accounting for both the short length-scale component and the measurement noise, see figure 5.

4.2 Systematic identification of f_z and g_z

A systematic and statistically correct procedure to identify the two contributions is required. Since f_z and g_z are independent, i.e. $E[\mathbf{F}\mathbf{G}^T]=0$, it follows that

$$\mathbf{\Lambda} = \mathrm{E}\left[\begin{bmatrix}\mathbf{F}\\ \mathbf{G}\\ \mathbf{Y}\end{bmatrix}\begin{bmatrix}\mathbf{F}^T & \mathbf{G}^T & \mathbf{Y}^T\end{bmatrix}\right] = \begin{bmatrix}\mathbf{\Lambda_{FF}} & 0 & \mathbf{\Lambda_{FF}}\\ 0 & \mathbf{\Lambda_{GG}} & \mathbf{\Lambda_{GG}}\\ \mathbf{\Lambda_{FF}} & \mathbf{\Lambda_{GG}} & \mathbf{Q}\end{bmatrix} \quad (9)$$

with $\mathbf{Q}=\mathbf{B}+\mathbf{\Lambda_{FF}}+\mathbf{\Lambda_{GG}}$. The prior joint probability distribution for **F**, **G** and **Y** is Gaussian with mean zero and covariance matrix **Λ**. The posterior joint probability distribution for **F** and **G** conditioned on the data **Y** remains Gaussian with mean, $\overline{\mathbf{M}}$, and covariance matrix, $\overline{\mathbf{\Lambda}}$, where

$$\overline{\mathbf{M}} = \begin{bmatrix}\mathbf{\Lambda_{FF}}\mathbf{Q}^{-1}\mathbf{Y}\\ \mathbf{\Lambda_{GG}}\mathbf{Q}^{-1}\mathbf{Y}\end{bmatrix} \quad (10)$$

$$\overline{\mathbf{\Lambda}} = \begin{bmatrix}\mathbf{\Lambda_{FF}}-\mathbf{\Lambda_{FF}}\mathbf{Q}^{-1}\mathbf{\Lambda_{FF}} & -\mathbf{\Lambda_{FF}}\mathbf{Q}^{-1}\mathbf{\Lambda_{GG}}\\ -\mathbf{\Lambda_{GG}}\mathbf{Q}^{-1}\mathbf{\Lambda_{FF}} & \mathbf{\Lambda_{GG}}-\mathbf{\Lambda_{GG}}\mathbf{Q}^{-1}\mathbf{\Lambda_{GG}}\end{bmatrix}$$

The covariance matrix for **H=F+G** is $(\mathbf{\Lambda_{FF}}+\mathbf{\Lambda_{GG}})+(\mathbf{\Lambda_{FF}}+\mathbf{\Lambda_{GG}})\mathbf{Q}^{-1}(\mathbf{\Lambda_{FF}}+\mathbf{\Lambda_{GG}})$ which is identical, as required, to $\overline{\mathbf{\Lambda}}_{\mathbf{HH}}$ in Section 3.

Example 1 (cont.): The above procedure is applied to Example 1. The long length-scale and short length-scale predictions together with their confidence intervals are shown in figure 6. The large breadth of the confidence intervals reflects uncertainty over attributing part of the data values to either f_z or g_z.

The requirement is to obtain the posterior probability distribution for **F** and **G** conditioned on the data set, M, subject to the condition that they remain independent. Of course, the posterior probability distribution remains Gaussian. The mean and covariance matrix is provided by theorem 1.

Theorem 1: Given that the prior joint probability distribution for **F**, **G** and **Y** is Gaussian with mean zero and covariance matrix **Λ**, the posterior joint probability distribution for $[\mathbf{F}^T, \mathbf{G}^T]^T$ conditioned on the M, subject to the condition that they remain independent, is Gaussian with

$$\text{mean} = \begin{bmatrix}\mathbf{\Lambda_{FF}}\mathbf{Q}_{\mathbf{F}}^{-1}\mathbf{Y}\\ \mathbf{B}\mathbf{Q}_{\mathbf{F}}^{-1}\mathbf{\Lambda_{GG}}\mathbf{Q}^{-1}\mathbf{Y}\end{bmatrix}$$
$$\text{cov} = \begin{bmatrix}\mathbf{\Lambda_{FF}}\mathbf{Q}_{\mathbf{F}}^{-1}\mathbf{B} & 0\\ 0 & \mathbf{B}\mathbf{Q}_{\mathbf{F}}^{-1}\mathbf{\Lambda_{GG}}\mathbf{Q}^{-1}\mathbf{B}\end{bmatrix} \quad (11)$$

where $\mathbf{Q_F}=\mathbf{B}+\mathbf{\Lambda_{FF}}$.

Proof: The prediction, from (10), for the contribution of g_z to the data, $\mathbf{\Lambda_{GG}}\mathbf{Q}^{-1}\mathbf{Y}$, may alternatively be in part explainable by f_z. An estimate of that part of the data values is assumed to be of the form $(\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1})\mathbf{\Lambda_{GG}}\mathbf{Q}^{-1}\mathbf{Y}$. It should be transferred from the prediction for **G** to the prediction for **F**. The appropriate modification to (10) is achieved by

$$\overline{\mathbf{M}} \to \overline{\mathbf{M}}_0 = T\overline{\mathbf{M}} \quad ; \quad \overline{\mathbf{\Lambda}} \to \overline{\mathbf{\Lambda}}_0 = T\overline{\mathbf{\Lambda}}T^T \quad (12)$$

where $T=\begin{bmatrix}I & \mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\\ 0 & \mathbf{B}^*\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\end{bmatrix}$. Should the above assumption be correct then there exists a $\mathbf{B}^*$ such that the off-diagonal elements of $\overline{\mathbf{\Lambda}}_0$ are zero. The off-diagonal elements are

$$(\overline{\mathbf{\Lambda}}_0)_{12}=(\overline{\mathbf{\Lambda}}_0)_{21}=\mathbf{\Lambda_{FF}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}[\mathbf{B}-\mathbf{B}^*]\mathbf{Q}^{-1}\mathbf{\Lambda_{GG}}\hat{\mathbf{Q}}_{\mathbf{F}}^{-1}\mathbf{B}^*$$

Hence, the correct choice for $\mathbf{B}^*$ is **B**. It follows that the required posterior joint probability distribution is Gaussian with mean and covariance matrix given by (11). (Note, the likelihood of the data is not affected by this adjustment of the posterior probability distribution).

The covariance matrix for **H=F+G** is again $(\mathbf{\Lambda_{FF}} + \mathbf{\Lambda_{GG}}) + (\mathbf{\Lambda_{FF}} + \mathbf{\Lambda_{GG}})\mathbf{Q}^{-1}(\mathbf{\Lambda_{FF}} + \mathbf{\Lambda_{GG}})$ as required.

Example 1 (cont.): Applying Theorem 1 to Example 1, the modified long length-scale and short length-scale predictions together with their confidence intervals are shown in figure 7. The confidence interval for the long length scale is much narrower than in figure 5.

5. WIND TURBINE DATA

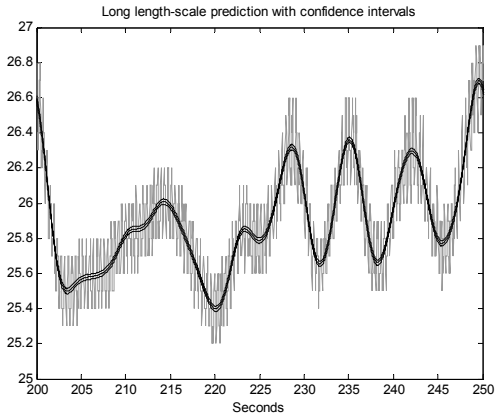

Fig. 8. Wind turbine data, long length scale prediction and confidence intervals.

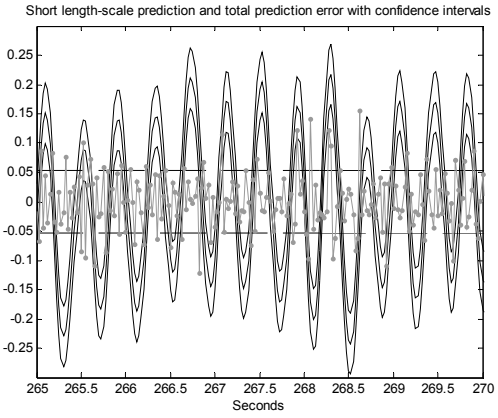

Fig. 9. Wind turbine short length scale prediction with confidence intervals, total error and confidence interval.

The identification procedure of Section 4.2, based on models with two Gausian processes, is applied to wind turbine time series data, specifically, site measurement of rotor speed for a commercial 1MW machine. The data consist of a run of 600 second sampled at 40Hz. A typical section, from 200*s* to 250*s*, is shown in Figure 8 (grey line). It has a long length-scale component due to variations in the aerodynamic torque, caused by changes in the wind speed and the pitch angle of the rotor blades, and a short length-scale component due to the structural and electro-mechanical dynamics of the machine. These two components can be clearly seen in figure 8 as can the poor quality of the data. The purpose is to identify both components, an initial yet important part of identifying the aerodynamics and drive-train dynamics of variable speed wind turbines (Leithead *et al*. 2003).

A typical section, from 200*s* to 250*s*, of the long length-scale component prediction with confidence intervals is shown in figure 8 (black lines) and, a typical section from 265*s* to 270*s*, of the short length- scale component prediction and confidence interval in figure 9 (solid lines). In addition, the confidence interval for the total prediction of the combined long and short length-scale components is shown in figure 9 (dashed lines) together with the residue between the data values and the total prediction (dotted line). The long and short length-scale components are successfully extracted from the measurement data.

6. CONCLUSION

To extract two components of different characteristics from data, a novel adaptation of the Gaussian regression methodology, based on models with two stochastic processes, is developed. In the prior and posterior joint probability distributions, the two components are independent. The effectiveness of the revised Gaussian regression method is demonstrated by application to wind turbine time series data. A long and a short length-scale component are successfully identified.

REFERENCES

MacKay, D. J. C.(1998). Introduction to Gaussian processes, *Neural Networks and Machine Learning, F: Computer and Systems Sciences*), **168**, 133-165.

Williams, C. K. I. (1999) Prediction with Gaussian processes: from linear regression to linear prediction and beyond, *Learning in Graphical Models*, 599-621,.

Gibbs, M. N. and D. J. C. Mackay (2000) Variational Gaussian process classifiers, *IEEE Trans.on Neural Networks*, **11**, pp. 1458-1464.

Sambu, S., M. Wallat, T. Graepel, T., and K. Obermayer, (2000). Gaussian process regression: active data selection and test point rejection, *Proc. IEEE International Joint Conference on Neural Networks*, **3**, 241-246.

Yoshioka T. and S. Ishii, (2001), Fast Gaussian process regression using representative data, *Proc. International Joint Conference on Neural Networks*, **1l** , 132-137.

Leithead, W. E., S. Solak, and D. Leith (2003), Direct identification of nonlinear structure using Gaussian Process prior models, *Proc. of European Control Conf.*, Cambridge.

Leithead, W. E., F. Hardan, and D. J. Leith (2003), Identification of aerodynamics and drive-train dynamics for a variable speed wind turbine, *Proc. European Wind Energy Conf.*, Madrid.