Direct Identification of Nonlinear Structure Using Gaussian Process Prior Models

W.E. Leithead[1,2], E. Solak[2], D.J. Leith[1]

[1]Hamilton Institute, National University of Ireland, Maynooth, Co. Kildare, Ireland
[2]Dept. of Electronics & Electrical Engineering, University of Strathclyde, Glasgow G1 1QE, U.K.
e-mail: w.leithead@eee.strath.ac.uk

Keywords: Identification, nonlinear structure, Gaussian process prior models, algorithm.

Abstract
When inferring nonlinear dependence from measured data, the nonlinear nature of the relationship may be characterised in terms of all the explanatory variables. However, this is rarely the most parsimonious, or insightful, approach. Instead, it is usually much more useful to be able to exploit the inherent nonlinear structure to characterise the nonlinear dependence in terms of the least possible number of variables. In this paper a new way of inferring nonlinear structure from measured data is investigated. The measured data is interpreted as providing information on a nonlinear map. The space containing the domain of the map is sub-divided into unique linear and nonlinear sub-spaces that are structural invariants. The most parsimonious representation of the map is obtained by the restriction of the map to the nonlinear sub-space. A direct constructive algorithm based on Gaussian process prior models, defined using a novel covariance function, is proposed. The algorithm infers the linear and nonlinear sub-spaces from noisy data and provides a non-parametric model of the parsimonious map. Use of the algorithm is illustrated by application to a Wiener-Hammerstein system.

1. Introduction

Consider a nonlinear dynamic system

$$x_{i+1} = g(x_i, r_i), \quad y_i = h(x_i, r_i) \qquad (1)$$

with state $x \in \Re^n$, input $r \in \Re^m$ and output $y \in \Re$. The input and output are measured but the state is not. When the unmeasured state can be eliminated, the equivalent input-output representation of the system is

$$y_i = H(y_{i-1}, \cdots, y_{i-n}, r_i, r_{i-1}, \cdots r_{i-n}) \qquad (2)$$

and the task of identifying the system from the measured data is equivalent to identifying the nonlinear function H(.). Of course, the nonlinear nature of the dynamics may be characterised in terms of all the explanatory variables (here, all the delayed inputs and outputs). However, this is rarely the most parsimonious, or insightful, approach. Instead, it is usually much more useful to be able to exploit any inherent nonlinear structure to characterise the nonlinear dependence in terms of the least possible number of variables. For example, it is often the case that dynamics involve a significant linear component, in which case knowledge of the nonlinear dependence can considerably reduce the dimensionality of the nonlinear modelling task (*e.g.* [4],[8]). Related to this, the use of appropriate co-ordinate axes (determined by knowledge of the nonlinear dependence) can greatly reduce the number of centres/operating regions required in radial basis function networks, Takagi-Sugeno fuzzy systems, local model networks and other types of blended multiple model representation (*e.g.* [1],[3],[4]).

In this paper, the nonlinear map, f: $D \to R$, with domain $D \subseteq \Re^p$ and range $R \subseteq \Re$ is investigated. Let Ψ_l and Ψ_{nl} be sub-spaces of $\Re^p$ such that the map is linear and nonlinear on $D \cap \Psi_l$ and $D \cap \Psi_{nl}$, respectively. To capture the nonlinear structure of the map, the sub-space Ψ_{nl} is required to be of minimum dimension. The existence of Ψ_l and Ψ_{nl}, sub-spaces such that $\Psi_l \cap \Psi_{nl} = \phi$, $D \subseteq \Psi_l \cup \Psi_{nl} = \Re^p$ and Ψ_{nl} is of minimum dimension, is discussed and the necessary and sufficient conditions determining these sub-spaces are derived in [5]. (In [5], the conditions on the map are that the domain and range are non-empty and open and f is twice continuously differentiable almost everywhere). For a particular nonlinear map, the sub-spaces, Ψ_l and Ψ_{nl}, are structural invariants. The parsimonious representation of the nonlinear component of f is the map F: $D \cap \Psi_{nl} \to R$. While this setting is general, it includes (as reflected in the examples chosen) the situation discussed in the previous paragraph.

The objective considered here is to identify from measured data the linear and nonlinear sub-spaces and the associated parsimonious map. Whereas an indirect method for identifying the nonlinear sub-space is described in [5], a new direct method is presented below.

2. Nonlinear Structure Identification

Denote the nonlinear map by

$$y = f(\mathbf{z}) \qquad (3)$$

Since measured data is generally noisy it is natural to work within a probabilistic framework. Suppose that N measurements, $\{(\mathbf{z}_i, y_i)\}_{i=1}^N$, of the value of the function, f($\mathbf{z}$), with additive Gaussian white measurement noise, i.e. $y_i = f(\mathbf{z}_i) + n_i$, are available and denote them by *S*. 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[1]. The probabilistic description of the nonlinear map adopted is that of a Gaussian process prior model within a Bayesian probability context [7].

2.1 Gaussian process prior models

A brief overview of Gaussian process prior models within a Bayesian probability context is given below. For further details see [7].

The probabilistic description of the function, f(**z**), is the stochastic process, $f_{\mathbf{z}}$, and the $E[f_{\mathbf{z}}]$, as **z** varies, is interpreted to be a fit to f(**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, $\{\mathbf{z}_1,\ldots,\mathbf{z}_k\}$, of D, for all k>1. Of course, the joint probability distributions of lower dimensionality must be the marginal distributions of those of higher dimensionality. 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 as their product since the measurement noise, n_i, and the $f(\mathbf{z}_i)$ (and so the $f_{\mathbf{z}_i}$) are statistically independent. *S* 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, *S*, to determine the posterior probability distributions. In particular, for the Gaussian process prior models considered here, the prior probability distributions for the $f_{\mathbf{z}}$ are chosen to be Gaussian with zero mean (in the absence of any evidence the value of f(**z**) is as likely to be positive as negative). To complete the statistical description, requires only a definition of the covariance function $C(f_{\mathbf{z}_i}, f_{\mathbf{z}_j}) = E[f_{\mathbf{z}_i}, f_{\mathbf{z}_j}]$, for all $\mathbf{z}_i$ and $\mathbf{z}_j$.

The resulting posterior probability distributions are also Gaussian. The choice of Gaussian probability distributions may seem strangely restrictive initially, but recall that this is simply a prior on the relevant stochastic process space and so places few inherent restrictions on the class of nonlinear functions that can be modelled. Indeed, it can be shown that the result is, in fact, a Bayesian form of kernel regression model [2] subsuming, amongst others, RBF, spline and many neural network models [7]. The Gaussian process prior model is non-parametric in the sense that the imposition of a specific parametric structure is avoided. This model is used to carry out inference as follows.

By Bayes theorem, $p(f_{\mathbf{z}} \mid S) = p(f_{\mathbf{z}}, S) / p(S)$, where $p(S)$ acts as a normalising constant. Hence, from the Gaussian nature of the probability distributions

$$p(f_{\mathbf{z}} \mid S) \propto \exp\left[-\frac{1}{2}\begin{bmatrix} f_{\mathbf{z}} & \mathbf{Y}^{\mathrm{T}} \end{bmatrix}\begin{bmatrix} \Lambda_{11} & \boldsymbol{\Lambda}_{21}^{\mathrm{T}} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}^{-1}\begin{bmatrix} f_{\mathbf{z}} \\ \mathbf{Y} \end{bmatrix}\right] \quad (4)$$

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

$$p(f_{\mathbf{z}} \mid S) \propto \exp\left[-\frac{1}{2}(f_{\mathbf{z}} - \hat{f}_{\mathbf{z}})\Lambda_{\mathbf{z}}^{-1}(f_{\mathbf{z}} - \hat{f}_{\mathbf{z}})\right] \quad (5)$$

with $\hat{f}_{\mathbf{z}} = \boldsymbol{\Lambda}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}\mathbf{Y}$, $\Lambda_{\mathbf{z}} = \Lambda_{11} - \boldsymbol{\Lambda}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}$. Therefore, the prediction from this model is that the most likely value of f(**z**) is the mean, $\hat{f}_{\mathbf{z}}$, with variance $\Lambda_{\mathbf{z}}$. Note that $\hat{f}_{\mathbf{z}}$ is simply a **z**-dependent weighted linear combination of the measured data points, **Y**, using weights $\boldsymbol{\Lambda}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}$.

The measurement noise, n_i, has covariance $n\delta_{ij}$ and is statistically independent of $f(\mathbf{z}_i)$. Hence, the covariances for the measured output, y_i, are simply

$$C(y_i,y_j) = (C(f_{\mathbf{z}_i}, f_{\mathbf{z}_j}) + n\delta_{ij}) \quad ; \quad C(y_i, f_{\mathbf{z}}) = C(f_{\mathbf{z}_i}, f_{\mathbf{z}}) \quad (6)$$

2.2 Choice of covariance for $f_{\mathbf{z}}$

A straightforward smoothness prior covariance function for $f_{\mathbf{z}}$, which ensures that outputs associated with nearby inputs should have higher covariance than more widely separated inputs, is [7]

$$C(f_{\mathbf{z}_i}, f_{\mathbf{z}_j}) = a\exp\left(-(\mathbf{z}_i - \mathbf{z}_j)^T D(\mathbf{z}_i - \mathbf{z}_j)/2\right)$$
$$D = \mathrm{diag}[d_1, \cdots, d_p] \quad (7)$$

The value of d_k characterises the rate of variation of the function, f, in the direction of the k^{th}-axis; the smaller the magnitude of d_k, the slower the rate of variation of f with respect to z_k. The matrix, *D*, thus indicates the degree of nonlinearity or relative smoothness in the directions of each explanatory variables. The covariance, (7), is used in [5] to identify the nonlinear maps f(**z**) and in [6] to reconstruct nonlinear systems from locally identified models. The corresponding covariance for y_i is

$$C(y_i, y_j) = a\exp\left(-(\mathbf{z}_i - \mathbf{z}_j)^T D(\mathbf{z}_i - \mathbf{z}_j)/2\right) + n\delta_{ij}\, p \quad (8)$$

To obtain a model given the data, *S*, the hyperparameters (a, d_k, n), whilst constrained to be positive, are adapted to maximise the likelihood, $p(S \mid (a, d_k, n))$.

The covariance (8) has the disadvantage that it provides information about the rates of variation of f projected in the directions of the axes, only. The characteristic signature of the nonlinear sub-space, Ψ_{nl}, is that nonlinear variation occurs only in those directions lying in Ψ_{nl}. Suppose the dimension of Ψ_{nl} is one but the basis vector has a non-zero component along each axes. In this case, there is variation in the direction of each axis and each hyperparameter in (7) needs to be non-zero. The situation is indistinguishable from when the dimension of Ψ_{nl} is equal to the dimension of the domain of

[1]No attempt to being made here to propagate a Gaussian or other distribution through a nonlinear function.

f. Hence, Ψ_{nl} cannot be identified directly using (8). To do so requires a different choice of covariance.

Initially, restrict the function, f, to being constant on Ψ_l. Continuing with the supposition that the dimension of Ψ_{nl} is one, a novel choice of covariance for $f_{\mathbf{z}}$,

$$\alpha \exp\left(-[\gamma^T(\mathbf{z}_i - \mathbf{z}_j)]^2/2\right) = \alpha \exp\left(-(\mathbf{z}_i - \mathbf{z}_j)^T M(\mathbf{z}_i - \mathbf{z}_j)/2\right) \quad (9)$$

$$M = \gamma\gamma^T \quad ; \quad \gamma^T = [\gamma_1, \cdots, \gamma_p]$$

is suggested. Since $(\mathbf{z}_i - \mathbf{z}_j)^T M(\mathbf{z}_i - \mathbf{z}_j) = 0$ whenever $(\mathbf{z}_i - \mathbf{z}_j)$ is orthogonal to γ, the covariance (9) is only capable of indicating the rate of variation of f in the direction of γ. The rate of variation in directions orthogonal to γ, is perforce zero. Consequently, the vector γ directly indicates the direction of variation of the function; that is, γ is a basis vector for Ψ_{nl}. Hence, the covariance, (9), *a priori* incorporates the information that the dimension of Ψ_{nl} is one and, adapting the hyperparameters to maximise the corresponding likelihood, enables Ψ_{nl} to be identified directly.

In general by the same reasoning, when the dimension of Ψ_{nl} is *q*, a covariance that *a priori* incorporates this information is

$$\alpha \exp\left(-(\mathbf{z}_i - \mathbf{z}_j)^T M(\mathbf{z}_i - \mathbf{z}_j)/2\right) \quad ; \quad M = \Gamma\Gamma^T \quad (10)$$

where $\Gamma^T = \{\gamma_{ij}\} \in \Re^{q \times p}$. Without loss of generality, to avoid over-parameterisation $\gamma_{ij} = 0$ for i>j. The corresponding covariance for y_i is

$$C(y_i, y_j) = \alpha \exp\left(-(\mathbf{z}_i - \mathbf{z}_j)^T M(\mathbf{z}_i - \mathbf{z}_j)/2\right) + n\delta_{ij} \quad (11)$$

To identify the nonlinear sub-space, Ψ_{nl}, given the data, the hyperparameters (α, γ_{ij}, n) are adapted, whilst (α, n) are constrained to be positive, to maximise the likelihood, $p(S \mid (\alpha, \gamma_{ij}, n))$. The prediction from the model, so obtained, is that the most likely nonlinear sub-space, Ψ_{nl}, is the sub-space spanned by the rows of Γ. Of course, even though Ψ_l and Ψ_{nl} are sub-spaces, the prediction using (11) is only valid on the region D containing the data and not everywhere.

The probabilistic description of the parsimonious nonlinear map $F(\boldsymbol{\xi})$, $\boldsymbol{\xi} = \Gamma\mathbf{z}$, is the stochastic process, $F_{\boldsymbol{\xi}}$. The probability distributions for $F_{\boldsymbol{\xi}}$ remain Gaussian with covariance functions

$$C(F_{\boldsymbol{\xi}_i}, F_{\boldsymbol{\xi}_j}) = \alpha \exp\left(-(\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)^T(\boldsymbol{\xi}_i - \boldsymbol{\xi}_j)/2\right)$$

$$C(F_{\boldsymbol{\xi}_i}, y_j) = C(F_{\boldsymbol{\xi}_i}, f_{\mathbf{z}}) = \alpha \exp\left(-(\boldsymbol{\xi}_i - \Gamma^T\mathbf{z}_j)^T(\boldsymbol{\xi}_i - \Gamma^T\mathbf{z}_j)/2\right) \quad (12)$$

It follows, similarly to $p(f_{\mathbf{z}} \mid S)$, that

$$p(F_{\boldsymbol{\xi}} \mid S) \propto \exp\left[-\frac{1}{2}(F_{\boldsymbol{\xi}} - \hat{F}_{\boldsymbol{\xi}})\Lambda_{\boldsymbol{\xi}}^{-1}(F_{\boldsymbol{\xi}} - \hat{F}_{\boldsymbol{\xi}})\right] \quad (13)$$

with $\hat{F}_{\boldsymbol{\xi}} = \overline{\boldsymbol{\Lambda}}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}\mathbf{Y}$, $\Lambda_{\boldsymbol{\xi}} = \overline{\Lambda}_{11} - \overline{\boldsymbol{\Lambda}}_{21}^{\mathrm{T}}\boldsymbol{\Lambda}_{22}^{-1}\overline{\boldsymbol{\Lambda}}_{21}$ where $\overline{\Lambda}_{11}$ is $C(F_{\boldsymbol{\xi}}, F_{\boldsymbol{\xi}})$ and the i^{th} element of vector $\overline{\boldsymbol{\Lambda}}_{21}$ is $C(y_i, F_{\boldsymbol{\xi}})$. The prediction from this model is that the most likely value of $F(\boldsymbol{\xi})$ is the mean, $\hat{F}_{\boldsymbol{\xi}}$, with variance $\Lambda_{\boldsymbol{\xi}}$.

To accommodate the function being non-constant on Ψ_l, the term

$$\mathbf{z}_i^T W \mathbf{z}_j \quad ; \quad W = \mathrm{diag}[w_1, \cdots, w_p] \quad (14)$$

which is the covariance corresponding to a general linear function, see [7], can be added to the covariance.

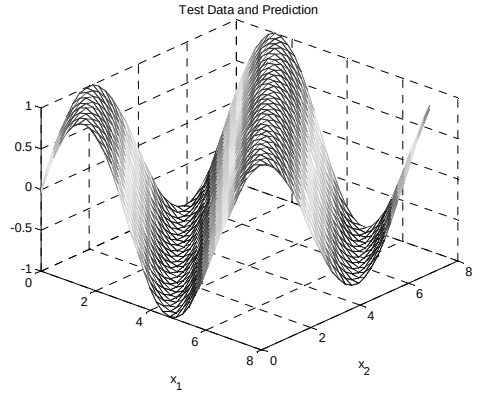

Figure 1 Test data and prediction.

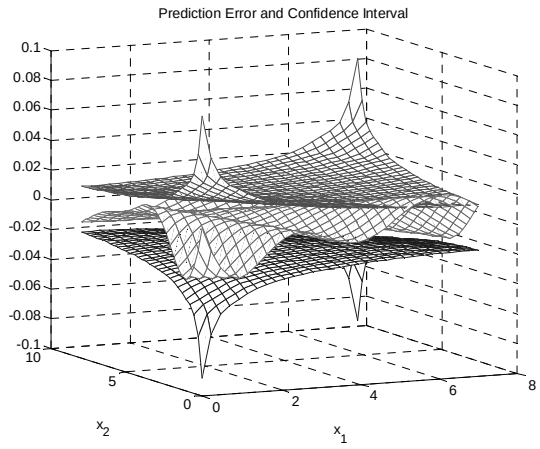

Figure 2 Prediction error and confidence intervals.

Example: $f(\mathbf{z}) = A\sin(a_1 z_1 + a_2 z_2) \quad A = 1, a_1 = 1.0, a_2 = 0.8$

The unit basis vector for Ψ_{nl} is (0.7809,0.6247). Since $f(\mathbf{z})$ is constant on Ψ_l, there is no need to add the linear term, (14), to the covariance. The domain, D, is the rectangle, $0 \le z_1 \le 7, 0 \le z_2 \le 8$ and the training data is the values on a regular 16x16 grid covering D with noise of intensity 0.05. In this example the dimension of Ψ_{nl} is one. On maximising the likelihood, the hyperparameters (typical

for a successful optimisation) obtained with the covariance (11) are

α = 2.1338, γ_1 = -0.4286, γ_2 = -0.3430, n=0.00244

that is, the estimate of noise intensity is 0.0493 and the estimate of the unit basis vector for Ψ_{nl} is (0.7807,0.6249). The model (11) is used to predict the value of f(**z**) on a regular 31x31 grid covering D. The exact and predicted values, respectively f(**z**) and $\hat{\mathrm{f}}_{\mathbf{z}}$, are shown in Figure 1 and the prediction error and confidence interval (± two standard deviations) in Figure 2. It is worthwhile noting, that the confidence interval increases rapidly in the direction of Ψ_{nl} when nearing the boundary of D but not in the direction of Ψ_l. The predicted parsimonious map, F(**ξ**), is depicted in figure 3.

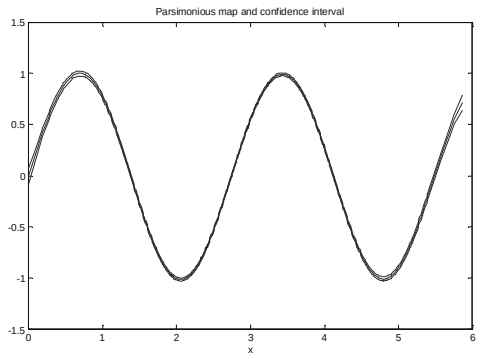

Figure 3 Parsimonious map and confidence intervals.

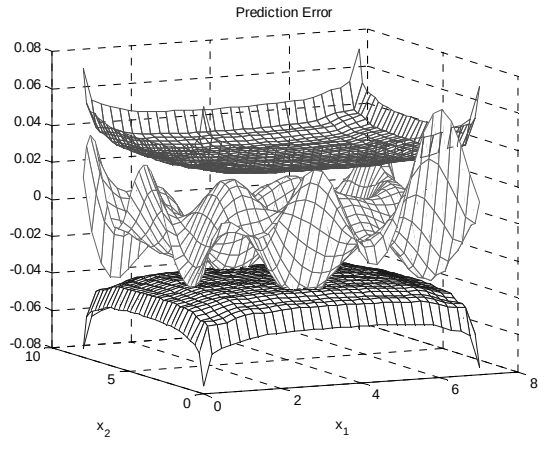

Figure 4 Prediction error and confidence intervals with covariance (8).

A model of the nonlinear map is also constructed for the above example using the covariance (8). The prediction error and the confidence interval for the Gaussian process model of the nonlinear map with this choice of covariance is depicted in Figure 4. With the covariance (11), the measurement data is being used to construct a q-dimensional model; specifically, the parsimonious model from the data projected into Ψ_{nl}. In contrast, with the covariance (8), the measurement data is being used to construct a p-dimensional model dependent on all the explanatory variables. Since $q<p$, the former is generally more accurate than the latter. Comparing Figures 2 and 4, this can be seen to be the case with the errors considerably larger in the latter than the former.

3 Identification algorithm

As in the example in section 2.2, the novel covariance, (11), enables a basis for the nonlinear sub-space, Ψ_{nl}, and the parsimonious map, F, to be identified directly from measured data when the dimension of Ψ_{nl} is known and the hyperparameters optimise successfully. A systematic algorithm utilising the Gaussian process prior models with covariance functions (8) and (11) is described in this section. The two issues of determining the dimension of Ψ_{nl} and ensuring successful optimisation of the hyperparameters are addressed.

The optimisation procedure with the covariance for $\mathrm{f}_{\mathbf{z}}$ chosen to be (10) proves to be rather temperamental in practice. It frequently fails or produces very poor fits when the initial values for the hyperparameters are chosen too distant from those that maximise the likelihood, $\mathrm{p}(S \mid (\alpha, \gamma_{ij}, n))$. This tendency increases rapidly with the number of data points and the degree of the function being fitted. The solution is to determine initial values for the hyperparameters sufficiently close to the optimal ones by the following procedure that simultaneously indicates whether the dimension of Ψ_{nl} is chosen correctly. First, the covariance for $\mathrm{y_i}$, (8), is used and the hyperparameters (a, d_{k}, n), whilst constrained to be positive, are adapted to maximise the likelihood. Second, the covariance is modified to

$$\begin{aligned}&\beta a \exp\left(-(\mathbf{z}_{\mathrm{i}}-\mathbf{z}_{\mathrm{j}})^T D(\mathbf{z}_{\mathrm{i}}-\mathbf{z}_{\mathrm{j}})/2\right)\\&\quad+\alpha\exp\left(-(\mathbf{z}_{\mathrm{i}}-\mathbf{z}_{\mathrm{j}})^T M(\mathbf{z}_{\mathrm{i}}-\mathbf{z}_{\mathrm{j}})/2\right)+n\delta_{\mathrm{ij}}\end{aligned} \qquad (15)$$

with the hyperparameters (a, d_{k}) constant and set to the values found in the first step. The hyperparameters (β, α, γ_{ij}, n) are then adapted, with (β, α, n) constrained to be positive, to maximise the likelihood for a particular choice of the dimension of Ψ_{nl}. The presence of the first term in (15) successfully moderates the optimisation procedure. The magnitude of β is an indicator of the correctness of the choice of dimension of Ψ_{nl}. When β is close to zero, a model of the form (11) with the chosen dimension of Ψ_{nl} is a good representation of the data. When β is not close to zero, no model of the form (11) can be found that is a good representation of the data; that is the chosen dimension of Ψ_{nl} is too small. The values for the hyperparameters (α, γ_{ij},

n) are suitable initial values for the optimisation procedure with the covariance chosen to be (11).

The full algorithm is as follows:

1. Choose the covariance function to be (8) with the initial values for the hyperparameters chosen randomly. The hyperparameters (a, d_k, n), whilst constrained to be positive, are adapted to maximise the likelihood, $p(S \mid (a, d_k, n))$.
2. Let $u_1 = \sqrt{d_{k_1}}\, e_{k_1}$, where d_{k_1} is the largest of the hyperparameters, d_k, and e_{k_1} is the unit vector in the direction of the k_1-st axis; $u_2 = \sqrt{d_{k_2}}\, e_{k_2}$, where d_{k_2} is the next largest of the hyperparameters, d_k, and e_{k_2} is the unit vector in the direction of the k_2-nd axis; etc.
3. Choose q, the dimension of Ψ_{nl} to be one.
4. Assign the rows of $G = \{g_{ij}\} \in \Re^{q \times p}$ to be $u_1^T, \cdots, u_q^T$, and re-order them so that G is upper-triangular.
5. Change the covariance to (15). The hyperparameters (a, d_k) are constant and set to the values found in 1. The initial values for the other hyperparameters are as follows: β is set to (1-ε), α is set to (εa), γ_{ij} are set to g_{ij} and n is set to the value found in 1. The value chosen for ε is arbitrary, but positive and small, say 0.1. The hyperparameters (β, α, γ_{ij}, n) are then adapted, with (β, α, n) constrained to be positive, to maximise the likelihood.
6. Repeat the procedure from 4 with q incremented by one until β<<1.
7. Change the covariance to (11) with the initial values for the hyperparameters (α, γ_{ij}, n) set to values found in 5. When the value of β is very small this last optimisation may be omitted. The hyperparameters (α, γ_{ij}, n) are then adapted, with (α, n) constrained to be positive, to maximise the likelihood.

This algorithm is very reliable and well-behaved in practice. The most likely sub-space, Ψ_{nl}, is of dimension, q, and is spanned by the rows of Γ.

4. Application to a Wiener-Hammerstein System

Consider the identification of a transversal Wiener-Hammerstein system

$$\begin{aligned} x_i &= a_n r_{i-n} + \cdots + a_0 r_i \\ z_i &= f(x_i) \\ y_i &= b_m z_{i-m} + \cdots + b_0 z_i \end{aligned} \tag{16}$$

Assume that N>n+m input-output measurements $\{(r_i, \hat{y}_i)\}_{i=1}^N$ are available, where $\hat{y}_i = y_i + n_i$ with n_i zero mean Gaussian noise. Let $\mathbf{R}_i$ denote the delayed input vector

$$\mathbf{R}_i = \begin{bmatrix} r_{i-(n-m)} & r_{i-(n-m-1)} & \cdots & r_i \end{bmatrix}^T \tag{17}$$

Reformulating the dynamics in terms of $\mathbf{R}_i$ and y_i yields

$$y_i = F(\mathbf{R}_i) \tag{18}$$

where

$$F(\mathbf{R}_i) = b_m f(\mathbf{H}_{m+1}\mathbf{R}_i) + \cdots + b_0 f(\mathbf{H}_1 \mathbf{R}_i) \tag{19}$$

$$\mathbf{H} = \begin{bmatrix} a_n & \cdots & a_0 & 0 & & \cdots & 0 \\ 0 & a_n & \cdots & a_0 & 0 & \cdots & 0 \\ & & & & \ddots & & \\ 0 & & \cdots & & 0 & a_n \cdots & a_0 \end{bmatrix} \tag{20}$$

and $\mathbf{H}_k$, k=1..m+1 denotes the k^{th} row of $\mathbf{H}$. At least one of the coefficients b_k, k=0..m is non-zero. Clearly, the rows of $\mathbf{H}$ constitute a basis for the nonlinear sub-space, Ψ_{nl}, associated with the map, F($\mathbf{R}_i$). The algorithm discussed in Section 3, can be applied to determine Ψ_{nl} and to identify the map, F($\mathbf{R}_i$).

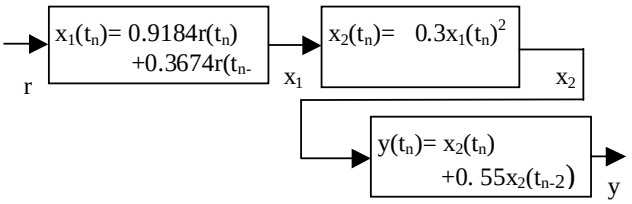

Figure 5 Block diagram representation of system.

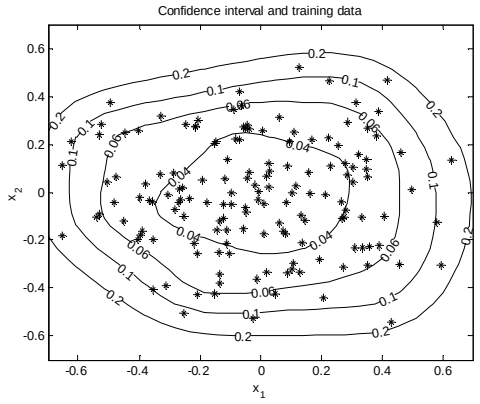

Figure 6 Contours of constant confidence intervals and data points.

Example:

Consider the Wiener-Hammerstein nonlinear system illustrated in Figure 5. Reformulating the dynamics in terms of the measured variables (input r and output y) yields

$$y_i = 0.3(\mathbf{H}_1\mathbf{R})^2 + 0.165(\mathbf{H}_3\mathbf{R})^2 \tag{21}$$

where $\mathbf{R} = [r(t_i) \quad r(t_{i-1}) \quad r(t_{i-2}) \quad r(t_{i-3})]^T$ with

$$\mathbf{H} = \begin{bmatrix} 0.9184 & 0.3674 & 0 & 0 \\ 0 & 0.9184 & 0.3674 & 0 \\ 0 & 0 & 0.9184 & 0.3674 \end{bmatrix} \tag{22}$$

and $\mathbf{H}_i$ is the i-th row of $\mathbf{H}$. The sub-space Ψ_{nl} has dimension two and its unit basis functions are (0.9284,0.3714,0,0) and (0,0,0.9284,0.3714). The output in response to a Gaussian input is measured: data is collected for 15 seconds with a sampling interval of 0.1 seconds (150 data points). Gaussian white noise of standard deviation 0.1 units is added to the output measurement (the underlying signal has a peak magnitude of 0.5, so this represents a substantial level of noise). Again for clarity, f($\mathbf{z}$) is constant on Ψ_l and there is no need to add the linear term (14) to the covariance.

Set the dimension of Ψ_{nl} to *one*. On maximising the likelihood, the hyperparameters for the covariance, (15), are

γ_1 = -0.3084, γ_2 = -0.1478, γ_3 = -0.2283, γ_4 = -0.0920

α = 15.2976, n=0.00958, β = 0.4415

that is, the estimate of noise intensity is 0.0979 and the estimate of the unit basis vector for Ψ_{nl} is (0.7321,0.3507,0.5418,0.2183). However, β is large, 0.4415, indicating that the dimension of Ψ_{nl} is greater than one.

Set the dimension of Ψ_{nl} to *two*. On maximising the likelihood, the hyperparameters are

γ_{11} = 0.3244, γ_{12} = 0.1286, γ_{13} = -0.0081, γ_{14} = -0.0061

γ_{21} = 0, γ_{22} = -0.0102, γ_{23} = 0.2628, γ_{24} = 0.1072

α = 55.4304, n=0.00917, β = 2e-012

that is, the estimate of noise intensity is 0.0958 and the estimate of the unit basis vectors for Ψ_{nl} are (0.9292,0.3685,-0.0231,-0.0173) and (0,-0.0358,0.9254, 0.3775). Since the value of β is very small, 2e-012, it can be concluded that the dimension of Ψ_{nl} is two and step 7 of the algorithm can be omitted. The sub-space, Ψ_{nl}, is spanned by the above two vectors. Clearly the prediction for the two basis vectors is good, particularly in view of the low signal to noise ratio and small number of data points on which it is based (150 points from a four dimensional map). In figure 6, the contours of constant confidence intervals for the parsimonious nonlinear map are shown together with the measurement data points projected onto Ψ_{nl}. The confidence intervals are small where the data points are dense but increase as the data points become scarce.

Identification of the Wiener-Hammerstein system is not pursued to completion as that is not the objective of this paper. However, having identified Ψ_{nl} and the hyperparameters defining the parsimonious nonlinear map, F($\boldsymbol{\xi}$), the rest is straightforward.

5. Conclusions

In this paper a new way of inferring nonlinear structure from measured data is investigated. The measured data is interpreted as providing information on a nonlinear map. The space containing the map is sub-divided into unique linear and nonlinear sub-spaces that are structural invariants. The most parsimonious representation of the map is obtained by the restriction of the map to the nonlinear sub-space; that is, the dimensionality of the nonlinear component of the map is minimised. A new direct constructive algorithm based on Gaussian process prior models, defined using a novel covariance function, is proposed. The algorithm infers the linear and nonlinear sub-spaces from noisy data and provides a non-parametric model of the parsimonious map. Use of the algorithm is illustrated by application to a Wiener-Hammerstein system.

Acknowledgements

This work was supported by Science Foundation Ireland grant 00/PI.1/C067, by EC TMR grant HPRNCT-1999-00107 and by EPSRC grant GR/M76379.

References

[1] C. BISHOP, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, (1995).

[2] P. J. GREEN, B. W. SILVERMAN, Nonparametric Regression and Generalised Linear Models, Chapman & Hall, London, (1994).

[3] T. A. JOHANSEN, B. A. FOSS, "Identification of Nonlinear System Structure and Parameters using Regime Decomposition", *Automatica*, **31**, pp. 321-326, (1995).

[4] T. A. JOHANSEN, R. MURRAY-SMITH, "The Operating Regime Approach to Nonlinear Modelling and Control", In *Multiple Model Approaches to Modelling and Control* (Murray-Smith,R, Johansen,T.A.), Taylor& Francis, London, (1997).

[5} D. J. LEITH, W. E. LEITHEAD, R. MURRAY-SMITH, "Nonlinear structure identification & dimensionality reduction", submitted to *Automatica*, (2003).

[6] D. J. LEITH, W. E. LEITHEAD, S. SOLAK, R. MURRAY-SMITH, "Divide & conquer identification: using gaussian process prior models to combine derivative & non-derivative observation in a consistent manner", Proc. CDC2002, Los Vegas, (2002).

[7] C. K. I. WILLIAMS, "Prediction with Gaussian Processes: From linear regression to linear prediction and beyond", In *Learning and Inference in Graphical Models* (M. I. Jordan, Ed.), Kluwer, (1998).

[8] P. YOUNG, "Comments on 'A quasi-ARMAX approach to modelling nonlinear systems", *Int. J. Contr*, **74**, pp. 1767-1771, (2001).