Inference of disjoint linear and nonlinear sub-domains of a nonlinear mapping

D.J. Leith[1,2], W.E. Leithead[1,2], R. Murray-Smith[1,3]

[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.*
[3]*Dept. of Computing Science, University of Glasgow, Glasgow G12 8QQ, U.K.*

Abstract

This paper investigates new ways of inferring nonlinear dependence from measured data. The existence of unique linear and nonlinear sub-spaces which are structural invariants of general nonlinear mappings is established and necessary and sufficient conditions determining these sub-spaces are derived. The importance of these invariants in an identification context is that they provide a tractable framework for minimising the dimensionality of the nonlinear modelling task. Specifically, once the linear/nonlinear sub-spaces are known, by definition the explanatory variables may be transformed to form two disjoint sub-sets spanning, respectively, the linear and nonlinear sub-spaces. The nonlinear modelling task is confined to the latter sub-set, which will typically have a smaller number of elements than the original set of explanatory variables. Constructive algorithms are proposed for inferring the linear and nonlinear sub-spaces from noisy data.

1. Introduction

Methods for inferring nonlinear dependence from measured data are presently almost entirely confined to analysis of the dependence with respect to explanatory variables selected *a priori*. Inference of nonlinear dependence is usually outwith the scope of principal components and analysis of variance techniques. Relevant methods include series expansion approaches whereby the coefficients of the first few terms in some series expansion are estimated, perhaps in a stepwise manner (*e.g.* Korenberg *et al.* 1988, Sjoberg *et al.* 1995). The linearity or non-linearity with respect to each explanatory variable may then be inferred by inspection of the estimated coefficients. Alternatively, when the model has the additive form, $\sum_i \varphi_i(z_i)$(where z_i denotes the i^{th} element of the explanatory variable vector and φ_i is an associated nonlinear, possibly vector, function), back-fitting methods can be used to directly estimate the φ_i, and thereby linearity or nonlinearity with respect to each explanatory variable, z_i, without necessarily postulating a particular series expansion (*e.g.* Hastie & Tibshirani 1990, Young 2000). Similar considerations apply to automatic relevance determination methods in the context of probabilistic neural network and non-parametric Gaussian process prior models (*e.g.* Neal 1996). In the case of blended multiple model representations based on decomposition of the operating space into a number of operating regions, similar considerations again apply when the local models associated with each operating region are sufficiently rich that they can directly embody any linear component (although this excludes the constant local models employed in standard radial basis function networks). In situations such as these, algorithms to search for appropriate operating region decompositions (*e.g.* Johansen & Foss 1995) can indirectly detect linearity with respect to particular explanatory variables.

The effectiveness of such methods in inferring a parsimonious dependence is generally strongly dependent on the choice of co-ordinate axes. For example, when the nonlinearity is dependent on some scalar function of all the chosen explanatory variables, the nonlinear dependence may be inferred to involve every explanatory variable, and thus be far from parsimonious, yet with a different choice of co-ordinate axes the true scalar nature of the dependence would become apparent. In principle, it is, of course, possible to extend the foregoing methods to incorporate estimation of, for example, a coordinate transformation and thereby automatically adjust the choice of explanatory variables as indicated by the data. However, such an approach is generally unattractive. Even a simple linear transformation matrix involves m^2 parameters, where m is the number of explanatory variables, and so estimation can be expected to quickly become unwieldy and intractable introducing, for example, an *additional* 100 parameters into an estimation problem involving 10 explanatory variables. Any attempt, furthermore, to nest current model fitting algorithms, which may already be rather complex and computationally intensive, within an outer axes-estimation iteration which is itself non-trivial are likely to be subject to local minima issues and similar associated difficulties quite apart from computational considerations.

The objective of the present paper is to investigate new ways of inferring the specifically nonlinear dependence (as opposed to linear dependence) from measured data.

Notation. The notation used is essentially standard. For a matrix $M\in\Re^{qxp}$, null(M) denotes the null space of M, *i.e.* null(M) $=\{v\in\Re^p: Mv=0\}$, and comp(M) denotes the orthogonal complement of M, a sub-space of $\Re^{qxp}$. For a twice differentiable vector mapping $F: D\subseteq\Re^p\rightarrow R\subseteq\Re^q$, $H_F(z)$ denotes the Hessian $\mathbf{H}_F(\mathbf{z}) = [\nabla(\nabla F_1(\mathbf{z}))^T \ldots \nabla(\nabla F_q(\mathbf{z}))^T]^T$ with F_i denoting the i^{th} element of the vector mapping F. The derivative of a vector or matrix function, $\Lambda(z)$, in direction v is defined as $(\mathbf{v}^T.\nabla)\mathbf{\Lambda}(\mathbf{z}) = \lim_{h\rightarrow o}\frac{\mathbf{\Lambda}(\mathbf{z}+h\mathbf{v})-\mathbf{\Lambda}(\mathbf{z})}{h}$. The

directional derivative of $\nabla\mathbf{F} = [(\nabla F_1)^T \ldots (\nabla F_q)^T]^T$ in direction $\mathbf{v}$ can be expressed as $H_F(\mathbf{z})\mathbf{v}$.

2. Structural Decomposition

The nonlinear mappings, $\mathbf{F}: \Delta \to P$, with open domain, $\Delta \subseteq \Re^{n+m}$, open range, $P \subseteq \Re^n$, and $\mathbf{F}$ continuously twice differentiable, are considered. While this setting is general, the particular interest here (and reflected in the examples chosen) is in dynamic systems applications where the nonlinear mapping might typically be the right-hand side of a differential/difference equation

$$\partial\mathbf{x}(t) = \mathbf{F}([\mathbf{x}^T(t) \quad \mathbf{r}^T(t)]^T) \qquad (1)$$

where the input is $\mathbf{r} \in \Delta_r \subseteq \Re^m$, the state $\mathbf{x} \in \Delta_x \subseteq \Re^n$ and ∂ denotes an appropriate operator; for example, the derivative operator d/dt (corresponding to continuous-time dynamics), the shift operator q (corresponding to discrete-time dynamics) or perhaps some combination of these.

The nonlinear dependence of the right hand side of (1) can be made explicit by reformulating as

$$\mathbf{F}(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{f}(\mathbf{M}\mathbf{z}) \qquad (2)$$

with $\mathbf{z}=[\mathbf{x}^T \quad \mathbf{r}^T]^T$ and where $\mathbf{A} \in \Re^{n\times(n+m)}$, $\mathbf{M} \in \Re^{q\times(n+m)}$, and $\mathbf{f}(\bullet)$ is a continuously twice differentiable nonlinear function. The decomposition (2), as it stands, is, of course, not unique but uniqueness of the linear term can be imposed without loss of generality, for example, by requiring $\nabla\mathbf{f}^T(\mathbf{M}\mathbf{z}_0) = 0$ for some $\mathbf{z}_0 \in \Delta$. Note, the decomposition (2) can always be trivially achieved by choosing $\mathbf{A}$ to be any matrix and $\mathbf{M}$ the identity. However, what is of interest here is to determine a decomposition, or class of decompositions, that is *minimal*.

Definition (minimality): Let a decomposition for $\mathbf{F}$ on $D \subseteq \Delta$, with D non-empty and open, be defined by

$$\mathbf{F}(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{f}(\mathbf{M}\mathbf{z}) \quad \forall \mathbf{z} \in D \qquad (3)$$

where $\mathbf{M}$ is some matrix. $\mathbf{M}$ is said to be of minimal degree for $\mathbf{F}$ on D when $\mathbf{M}$ is of full rank and an alternative choice of lower rank satisfying the decomposition for some $\mathbf{A}$ and $\mathbf{f}$ does not exist.

This definition of minimality corresponds to the intuitive idea that we would like to choose the rank of $\mathbf{M}$ to be as small as possible. In order to be useful, a testable condition for minimality is required. It is readily verified that the Hessian of a linear or affine mapping is identically zero. Indeed, this is the basis for common regularisation schemes and Bayesian priors. Building on this observation, the following Lemma is obtained.

Lemma (minimal decomposition) Let $\mathbf{M}$ be of full rank and the decomposition (3) exist then $\mathbf{M}$ is of minimal degree for $\mathbf{F}$ on D if and only if $\bigcap_{\mathbf{z}\in D} \text{null}(\mathbf{H}_\mathbf{F}(\mathbf{z})) = \text{null}(\mathbf{M})$.

Proof Note that $\mathbf{H}_\mathbf{F}(\mathbf{z})\mathbf{v} = 0$ whenever $\mathbf{M}\mathbf{v} = 0$. Suppose $\bigcap_{\mathbf{z}\in D} \text{null}(\mathbf{H}_\mathbf{F}(\mathbf{z})) \neq \text{null}(\mathbf{M})$, then there exists a $\mathbf{v}_0$ such that $\mathbf{M}\mathbf{v}_0 \neq 0$ and $\mathbf{H}_\mathbf{F}(\mathbf{z}_0)\mathbf{v}_0 = 0 \ \ \forall \mathbf{z}_0 \in D$. Since the Hessian $\mathbf{H}_\mathbf{F}$ is zero in direction $\mathbf{v}_o$ on an open set, all higher derivatives in this direction also vanish on D. Hence, $\forall \mathbf{z}_0 \in D$, $\exists \varepsilon > 0$: $\mathbf{z}_0 + \lambda\mathbf{v}_0 \in D$ and

$$\mathbf{F}(\mathbf{z}_0+\lambda\mathbf{v}_0) = \mathbf{F}(\mathbf{z}_0) + \lambda(\mathbf{v}_0^T.\nabla)\mathbf{F}(\mathbf{z}_0) \quad \forall \lambda \in [0, \varepsilon]$$

Let $\hat{D} = \{\mathbf{z} \in \Re^{n+m} : \mathbf{z} = \mathbf{z}_0 + \lambda\mathbf{v}_0, \mathbf{z}_0 \in D, \lambda \in \Re\}$. The domains of $\mathbf{F}(\bullet)$ and $\mathbf{f}(\bullet)$ are extended to $\hat{D} \supseteq D$. $\forall \mathbf{z} \in \hat{D}$, define $\mathbf{F}(\mathbf{z}) = \mathbf{F}(\mathbf{z}_0) + \lambda(\mathbf{v}_0^T.\nabla)\mathbf{F}(\mathbf{z}_0)$ and $\mathbf{f}(\mathbf{M}\mathbf{z}) = \mathbf{F}(\mathbf{z}) - \mathbf{A}\mathbf{z}$ where $\mathbf{z} = \mathbf{z}_0 + \lambda\mathbf{v}_0$ with $\mathbf{z}_0 \in D$ and $\lambda \in \Re$. With this extension to the domain of $\mathbf{f}$, $\mathbf{f}(\overline{\mathbf{M}}\mathbf{z})$ is defined $\forall \mathbf{z} \in D$, where

$$\overline{\mathbf{M}} = \mathbf{M}(\mathbf{I} - \mathbf{v}_0(\mathbf{v}_0^T\mathbf{v}_0)^{-1}\mathbf{v}_0^T)$$

Furthermore,

$$\mathbf{G}(\mathbf{z}) = \mathbf{f}(\mathbf{M}\mathbf{z}) - \mathbf{f}(\overline{\mathbf{M}}\mathbf{z})$$
$$= ((\mathbf{v}_0^T.\mathbf{z})(\mathbf{v}_0^T.\nabla)\mathbf{F}(\mathbf{z}) - (\mathbf{A}.\mathbf{v}_0)(\mathbf{v}_0^T.\mathbf{z}))(\mathbf{v}_0^T.\mathbf{v})^{-1}$$

is affine in $\mathbf{z}$, $\forall \mathbf{z} \in D$, since

$$(\mathbf{v}_0^T\mathbf{v}_0)\mathbf{H}_{G_i}(\mathbf{z}) =$$
$$(\mathbf{H}_{F_i}(\mathbf{z})\mathbf{v}_0)\mathbf{v}_0^T + \mathbf{v}_0(\mathbf{H}_{F_i}(\mathbf{z})\mathbf{v}_0)^T + (\mathbf{v}_0^T\mathbf{z})\nabla(\mathbf{H}_{F_i}(\mathbf{z})\mathbf{v}_0)^T = 0$$

Hence, there exists a decomposition (3) for $\overline{\mathbf{M}}$ but $\text{rank}(\overline{\mathbf{M}}) < \text{rank}(\mathbf{M})$. Consequently, when $\bigcap_{\mathbf{z}\in D} \text{null}(\mathbf{H}_\mathbf{F}(\mathbf{z})) \neq \text{null}(\mathbf{M})$, $\mathbf{M}$ must be non-minimal. Now, suppose $\mathbf{M}$ is non-minimal, *i.e.* $\exists\ \overline{\mathbf{M}}$ such that (3) is satisfied for some $\mathbf{A}$ and $\mathbf{f}$ and $\text{rank}(\overline{\mathbf{M}}) < \text{rank}(\mathbf{M})$, then $\dim(\text{null}(\overline{\mathbf{M}})) > \dim(\text{null}(\mathbf{M}))$. Hence, $\exists\ \mathbf{v}_0$: $\mathbf{M}\mathbf{v}_0 \neq 0$, $\overline{\mathbf{M}}\mathbf{v}_0 = 0$ and so $\mathbf{H}_\mathbf{F}(\mathbf{z})\mathbf{v}_0 = 0$, $\forall \mathbf{z}$. Consequently, when $\mathbf{M}$ is non-minimal, $\bigcap_{\mathbf{z}\in D} \text{null}(\mathbf{H}_\mathbf{F}(\mathbf{z})) \neq \text{null}(\mathbf{M})$. It follows immediately that $\mathbf{M}$ is of minimal degree for $\mathbf{F}$ on D if and only if $\bigcap_{\mathbf{z}\in D} \text{null}(\mathbf{H}_\mathbf{F}(\mathbf{z})) = \text{null}(\mathbf{M})$. ■

This lemma enables it to be immediately determined whether a given matrix $\mathbf{M}$ is minimal by inspecting the Hessian $\mathbf{H}_\mathbf{F}$ of the mapping $\mathbf{F}$. Further observe that the minimality test is in terms of the sub-space null($\mathbf{M}$) rather than the matrix $\mathbf{M}$ itself. This is important. Since a non-singular linear transformation applied to $\mathbf{M}$ can be absorbed into the nonlinear function, the mapping $\mathbf{f}: \mathbf{z} \in D \to P$, embodied by a nonlinear function $\mathbf{f}(\mathbf{M}\mathbf{z})$, can be realised by any function $\mathbf{f}_\mathbf{T}(\mathbf{M}_\mathbf{T}\mathbf{z})$ with $\mathbf{M}_\mathbf{T} = \mathbf{T}\mathbf{M}$ and $\mathbf{f}_\mathbf{T} = \mathbf{f} \circ \mathbf{T}^{-1}$. Hence, there does not exist a single, unique $\mathbf{M}$ that is minimal for mapping $\mathbf{F}$. The sub-space null($\mathbf{M}$) is invariant with respect to such transformations. Developing this line of reasoning further, let Ψ_l denote null($\mathbf{M}$) and Ψ_{nl} denote comp($\mathbf{M}$). It follows from (2) that on the domain $\Psi_l \cap D$ the mapping is linear and, conversely, when $\mathbf{M}$ is minimal the mapping $\mathbf{F}$ is nonlinear on the domain $\Psi_{nl} \cap D$; that is, $\Re^{n+m} \cong \Psi_l \oplus \Psi_{nl}$ and $\mathbf{F}(\mathbf{z}) \cong \tilde{\mathbf{F}}(\mathbf{u}, \mathbf{v})$, with $\mathbf{z} \in \Re^{n+m} \cap D$, $\mathbf{u} \in \Psi_l \cap D$ and $\mathbf{v} \in \Psi_{nl} \cap D$, such that $\tilde{\mathbf{F}}(\mathbf{u}_1 + \mathbf{u}_2, \mathbf{v}) = \tilde{\mathbf{F}}(\mathbf{u}_1, \mathbf{v}) + \tilde{\mathbf{F}}(\mathbf{u}_2, \mathbf{v})$ when $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_1 + \mathbf{u}_2 \in \Psi_l \cap D$. The sub-spaces, Ψ_l and Ψ_{nl}, of $\mathbf{M}$ embody the linear and nonlinear dependence of the

mapping **F**. It follows from the minimality lemma that these sub-spaces are identical for all minimal **M** and are structural invariants of the mapping **F**. This is formalised by the following corollary.

Corollary (subspace partitioning) Consider the class of decompositions (3) with **M** minimal for **F** on D. Let $\mathbf{M}_0$ and $\mathbf{M}_1$ denote any two choices of **M**, then comp($\mathbf{M}_0$)=comp($\mathbf{M}_1$); that is, there exists a unique sub-space Ψ_{nl} such that comp(**M**)=Ψ_{nl} for any **M** which is minimal for **F**.

Proof From the minimal decomposition lemma, **M** minimal implies that $\text{null}(\mathbf{M}) = \bigcap_{\mathbf{z}\in D} \text{null}(\mathbf{H}_\mathbf{F}(\mathbf{z}))$; that is, all minimal **M** possess the same null space. Since comp(**M**) is the orthogonal complement of null(**M**), it follows that this is also identical for all minimal **M**.

■

Consider a set, $\{\mathbf{z}_i\}$, i=1,…,K, of values of the explanatory variable, **z**, sufficiently large and disparate that rank($\mathbf{H}_\mathbf{F}^K$)=dim(Ψ_1), where $\mathbf{H}_\mathbf{F}^K=[\mathbf{H}_\mathbf{F}(\mathbf{z}_1),\ldots,\mathbf{H}_\mathbf{F}(\mathbf{z}_K)]^T$. The minimality condition is equivalent to $\mathbf{H}_\mathbf{F}^K\hat{\mathbf{v}}_i = 0,\ \forall\hat{\mathbf{v}}_i \in \text{basis}\{\Psi_1\}$. To determine Ψ_1, assuming r=dim(Ψ_1) is known, it is sufficient to determine a value of $\boldsymbol{\theta}$ such that $\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta}) = 0$ where

$$\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta}) = [(\mathbf{H}_\mathbf{F}(\mathbf{z}_1)\hat{\mathbf{v}}_1(\boldsymbol{\theta}))^T\cdots(\mathbf{H}_\mathbf{F}(\mathbf{z}_1)\hat{\mathbf{v}}_r(\boldsymbol{\theta}))^T,(\mathbf{H}_\mathbf{F}(\mathbf{z}_2)\hat{\mathbf{v}}_1(\boldsymbol{\theta}))^T\cdots(\mathbf{H}_\mathbf{F}(\mathbf{z}_K)\hat{\mathbf{v}}_r(\boldsymbol{\theta}))^T]^T$$

and $\{\hat{\mathbf{v}}_i(\boldsymbol{\theta})$, i=1…r$\}$ is an explicit parameterisation of all the sets of r orthonormal vectors. The non-uniqueness of $\boldsymbol{\theta}$ is immaterial as it is the unique invariant subspace Ψ_1 which is of interest.

3. Nonlinear Structure Identification

The structural decomposition analysis in section 2 is deterministic. In this section, the extension to the probabilistic case with noisy data is considered. Matrices are generically full rank and so under noisy conditions the null space of $\mathbf{H}_\mathbf{F}(\mathbf{z})$ will almost always consist simply of the zero vector. Instead, the requirement must be to determine the largest sub-space within which the range of the estimated Hessian is, in some appropriate sense, close to zero (rather than precisely equal to zero as in the noise-free case).

It is assumed that the joint probability distribution is available for $\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta})$ (for any $\boldsymbol{\theta}$). In the identification context, this probability distribution is inferred from a data set, $\mathbf{X}=[x_1,\ldots,x_N]^T$, and so is conditional on the data set. The objective is to use this probabilistic description to derive relevant information pertaining to the structural decomposition into linear and nonlinear components.

Remarks

(i) An appropriate choice of representation for **F**(**z**) could be by means of a stochastic process model from which is derived a stochastic process model for $\mathbf{H}_\mathbf{F}(\mathbf{z})\mathbf{v}$, for all **v**, and, thereby, the joint probability distribution for $\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta})$. (For example, in the case of Gaussian stochastic process models, the mean and covariance of the Gaussian process model for $\mathbf{H}_\mathbf{F}(\mathbf{z})\mathbf{v}$ are appropriate derivatives of the mean and covariances of the Gaussian process model for **F**(**z**) – see Appendix). It is perhaps worth emphasising that this certainly does not require differentiation of the raw, noisy data. The latter is, of course, highly inadvisable.

(ii) It is important to note that stochastic process descriptions do not necessarily require the imposition of a parametric model structure. Non-parametric descriptions (e.g. Green & Silverman 1994, Neal 1996, Williams 1998) are characterised by drawing inferences directly from the measured data using smoothness information but without assuming an underlying parameterisation. (Various forms of smoothness assumption are typically employed: any specific assumption may of course be more or less appropriate in a particular application context). An example of a non-parametric nonlinear description is a Gaussian process prior model: see the Appendix.

3.1 Summarising Nonlinear Dependence in a Region

It is again assumed again that r=dim(Ψ_1) is known. Let $p_\mathbf{X}(\boldsymbol{\chi})$ be the probability density function for **X** and let $p_\mathbf{V}(\boldsymbol{\omega}|\boldsymbol{\theta},\boldsymbol{\chi})$ be the probability density function for $\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta})$ conditional on $\boldsymbol{\theta}$ and $\boldsymbol{\chi}$. The joint probability density function for ($\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta})$, **X**) is

$$p_{\mathbf{V},\mathbf{X}}(\boldsymbol{\omega},\boldsymbol{\chi}|\boldsymbol{\theta}) = p_\mathbf{V}(\boldsymbol{\omega}|\boldsymbol{\theta},\boldsymbol{\chi})p_\mathbf{X}(\boldsymbol{\chi}) \qquad (4)$$

and $p_{\mathbf{V},\mathbf{X}}(0,\mathbf{X}|\boldsymbol{\theta})$ is the likelihood of $\boldsymbol{\theta}$ with $\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta}) = 0$ and **X**, a specific data set; that is, the likelihood that the $\hat{\mathbf{v}}_i(\boldsymbol{\theta})$, $i = 1,\cdots,r$, are a basis for Ψ_1. A maximum likelihood estimate of the decomposition into linear and nonlinear sub-spaces, Ψ_1 and Ψ_{nl}, is thus provided by any $\boldsymbol{\theta}_M$ for which the likelihood $p_{\mathbf{V},\mathbf{X}}(0,\mathbf{X}|\boldsymbol{\theta})$ is maximal, or equivalently, since $p_\mathbf{X}(\boldsymbol{\chi})$ is independent of $\boldsymbol{\theta}$, $p_\mathbf{V}(0|\boldsymbol{\theta}, \mathbf{X})$ is maximal.

Remark

Suppose, as is the case for the Gaussian process prior models of the appendix, that the joint probability distribution for $\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta})$ and **X** is Gaussian or, more specifically, $N(0,\boldsymbol{\Lambda})$ with

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{\omega\omega} & \boldsymbol{\Lambda}_{\omega\chi} \\ \boldsymbol{\Lambda}_{\chi\omega} & \boldsymbol{\Lambda}_{\chi\chi} \end{bmatrix}$$

then $p_\mathbf{V}(\boldsymbol{\omega}|\boldsymbol{\theta},\mathbf{X}) = N(\bar{\boldsymbol{\omega}},\boldsymbol{\Omega})$ with $\bar{\boldsymbol{\omega}} = \boldsymbol{\Lambda}_{\omega\chi}\boldsymbol{\Lambda}_{\chi\chi}^{-1}\mathbf{X}$ and $\boldsymbol{\Omega} = \boldsymbol{\Lambda}_{\omega\omega} - \boldsymbol{\Lambda}_{\omega\chi}\boldsymbol{\Lambda}_{\chi\chi}^{-1}\boldsymbol{\Lambda}_{\chi\omega}$. Consider any two sets of orthonormal vectors, $\hat{\mathbf{v}}_i(\boldsymbol{\theta}_1)$ and $\hat{\mathbf{v}}_i(\boldsymbol{\theta}_2)$, spanning the same sub-space. There exist t_{ij} such that $\hat{\mathbf{v}}_i(\boldsymbol{\theta}_2) = \sum_{j=1}^{r} t_{ij}\hat{\mathbf{v}}_j(\boldsymbol{\theta}_1)$ and, hence, $\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta}_2) = (I_K \otimes (T\otimes I_r))\mathbf{V}_\mathbf{F}^K(\boldsymbol{\theta}_1)$ where the ij-th element of T is t_{ij}. Since T is clearly non-singular,

$$p_\mathbf{V}(0|\boldsymbol{\theta}_1,\mathbf{X}) = p_\mathbf{V}(0|\boldsymbol{\theta}_2,\mathbf{X})$$

and the non-uniqueness of $\boldsymbol{\theta}_M$ is again immaterial.

As the variance of the measurement noise increases, so do the variances for the posterior probability distributions. It is important to test the statistical significance of any inference made on the basis of the data. A suitable test statistic is discussed below. Let $p_{\mathbf{X}}(\boldsymbol{\chi}\,|\boldsymbol{\omega},\boldsymbol{\theta})$ be the probability density function for the data set conditioned on $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$. The confidence in the estimated decomposition into Ψ_l and Ψ_{nl} can be assessed by the generalised likelihood ratio test; specifically, with $\boldsymbol{\theta}=\boldsymbol{\theta}_M$, the relative tenability for the data set of the hypothesis $\boldsymbol{\omega}=0$ and the hypothesis $\boldsymbol{\omega}\neq 0$ is compared. The test statistic is

$$\eta = p_{\mathbf{X}}(\mathbf{X}|0,\boldsymbol{\theta}_M)\,/\max_{\boldsymbol{\omega}} p_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\omega},\boldsymbol{\theta}_M)$$
$$= \frac{p_{\mathbf{V}}(0|\boldsymbol{\theta}_M,\mathbf{X})}{p_{\mathbf{V}}(0|\boldsymbol{\theta}_M)} \Big/ \max_{\boldsymbol{\omega}} \frac{p_{\mathbf{V}}(\boldsymbol{\omega}|\boldsymbol{\theta}_M,\mathbf{X})}{p_{\mathbf{V}}(\boldsymbol{\omega}|\boldsymbol{\theta}_M)}$$

For data of reasonable quality, the variance of $p_{\mathbf{V}}(\boldsymbol{\omega}|\boldsymbol{\theta}_M, \mathbf{X})$ is much less than the variance $p_{\mathbf{V}}(\boldsymbol{\omega}|\boldsymbol{\theta}_M)$ and, in the vicinity of its maximum, its value is more variable. In these circumstances,

$$\overline{\eta} = p_{\mathbf{V}}(0|\boldsymbol{\theta}_M,\mathbf{X})\,/\max_{\boldsymbol{\omega}} p_{\mathbf{V}}(\boldsymbol{\omega}|\boldsymbol{\theta}_M,\mathbf{X}) \approx \eta$$

is a suitable alternative test statistic.

Remark

For the same situation as in the previous remark, when $\dim(\boldsymbol{\Lambda}_{\chi\chi}) \le \dim(\boldsymbol{\Lambda}_{\omega\omega})$,

$$-2\ln(\eta) = \mathbf{X}^T(\boldsymbol{\Lambda}_{\chi\chi} - \boldsymbol{\Lambda}_{\chi\omega}\boldsymbol{\Lambda}_{\omega\omega}^{-1}\boldsymbol{\Lambda}_{\omega\chi})^{-1}\mathbf{X}$$
$$= \overline{\boldsymbol{\omega}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}_{\omega\chi}(\boldsymbol{\Lambda}_{\chi\omega}\boldsymbol{\Lambda}_{\omega\omega}^{-1}\boldsymbol{\Lambda}_{\omega\chi})^{-1}\boldsymbol{\Lambda}_{\chi\chi}(\boldsymbol{\Lambda}_{\chi\omega}\boldsymbol{\Lambda}_{\omega\omega}^{-1}\boldsymbol{\Lambda}_{\omega\chi})^{-1}\boldsymbol{\Lambda}_{\chi\omega}\boldsymbol{\Lambda}_{\omega\omega}^{-1}\overline{\boldsymbol{\omega}}$$

and, when $\dim(\boldsymbol{\Lambda}_{\chi\chi}) \ge \dim(\boldsymbol{\Lambda}_{\omega\omega})$,

$$-2\ln(\eta) = \mathbf{X}^T(\boldsymbol{\Lambda}_{\chi\chi} - \boldsymbol{\Lambda}_{\chi\omega}\boldsymbol{\Lambda}_{\omega\omega}^{-1}\boldsymbol{\Lambda}_{\omega\chi})^{-1}\boldsymbol{\Lambda}_{\chi\omega}(\boldsymbol{\Lambda}_{\omega\chi}\boldsymbol{\Lambda}_{\chi\chi}^{-1}\boldsymbol{\Lambda}_{\chi\omega})^{-1}\boldsymbol{\Lambda}_{\omega\chi}\boldsymbol{\Lambda}_{\chi\chi}^{-1}\mathbf{X}$$
$$= \overline{\boldsymbol{\omega}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}_{\omega\omega}(\boldsymbol{\Lambda}_{\omega\chi}\boldsymbol{\Lambda}_{\chi\chi}^{-1}\boldsymbol{\Lambda}_{\chi\omega})^{-1}\overline{\boldsymbol{\omega}}$$

Clearly, in both cases, the test statistic, $-2\ln(\eta)$, is the same for all possible $\boldsymbol{\theta}_M$ spanning Ψ_l. Since the means of the probability distributions of the Gaussian process prior model considered here are zero, $\max_{\boldsymbol{\omega}} p_{\mathbf{V}}(\boldsymbol{\omega}|\boldsymbol{\theta}_M) = p_{\mathbf{V}}(0|\boldsymbol{\theta}_M)$ and $-2\ln(\overline{\eta}) = \overline{\boldsymbol{\omega}}^T\boldsymbol{\Omega}^{-1}\overline{\boldsymbol{\omega}}$ is a more conservative test statistic than $-2\ln(\eta)$. For data of reasonable quality, a rejection criterion with significance level α for the hypothesis $\boldsymbol{\omega}=0$ is, thus,

$$-2\ln(\eta) \approx -2\ln(\overline{\eta}) > \chi^2_{r\mathrm{K}}(\alpha)$$

Assuming that the dimension of Ψ_l is known, a set of orthonormal vectors for which the likelihood, $p(0,\mathbf{X}|\boldsymbol{\theta})$, is maximal provides an estimate of the basis for the linear sub-space, Ψ_l. An estimate of the nonlinear sub-space Ψ_{nl} is obtained as the orthogonal complement of Ψ_l. When the dimension of Ψ_l is not known, the dimension of Ψ_l may be sequentially increased and its basis re-estimated. Let Ψ_l^i denote the estimated sub-space of dimension i and let Ψ_l and Ψ_{nl} denote the true linear and nonlinear sub-spaces. It follows that the dimension of Ψ_l is $m+n-q$, where q is the dimension of Ψ_{nl}. It must be that $\Psi_l^i \cap \Psi_{nl} \neq \{0\}$ for $i>m+n-q$. Consequently, as i is increased beyond the true dimension $m+n-q$, the value of η can be expected to abruptly decrease (since $\mathbf{H_F(z)v}\neq 0 \quad \forall \mathbf{v} \in \text{basis}\{\Psi_{nl}\}$) and, consequently, the dimension can be inferred from the data.

In the above sequential estimation approach the re-estimation of Ψ_l at each step can be implemented efficiently by making use of the previous estimate, Ψ_l^{i-1}, when estimating Ψ_l^i. The procedure is the following: search for the direction within Ψ_{nl}^{i-1}, the orthogonal complement of Ψ_l^{i-1}, along which $\mathbf{H_F(z)v}$ is most likely to be zero. Letting $\overline{\mathbf{v}}_i$ denote this direction, Ψ_l^i is then obtained as $\Psi_l^i \oplus \text{span}(\overline{\mathbf{v}}_i)$. This leads to the following algorithm (expressed in terms of matrices, $\mathbf{V}_i$ and $\mathbf{M}_i$, whose columns form orthonormal basis for Ψ_l^i and Ψ_{nl}^i, respectively, such that $\Psi_{nl}^i \cap \Psi_l^i = \{0\}$).

Iterative Estimation Procedure

1. Let i=1, $\Psi_{nl}^i = \Re^{n+m}$.
2. Determine the most likely unit direction $\overline{\mathbf{v}}_i$, lying within the current estimate, Ψ_{nl}^i, of the nonlinear subspace; that is, with $\mathbf{V}_{\mathbf{F}}^{\mathrm{K}}(\boldsymbol{\theta})$ defined using $\hat{\mathbf{v}}(\boldsymbol{\theta})$ an explicit parameterisation of all unit vectors in Ψ_{nl}^i, the unit vector, $\hat{\mathbf{v}}(\boldsymbol{\theta})$, which maximises the likelihood $p_{\mathbf{V},\mathbf{X}}(0,\mathbf{X}|\boldsymbol{\theta})$. Letting the columns of $\mathbf{M}_i$ be an orthonormal basis spanning Ψ_{nl}^i, then $\hat{\mathbf{v}}(\boldsymbol{\theta})$ may be parameterised as $\mathbf{M}_i\boldsymbol{\lambda}$ and the maximisation of $p_{\mathbf{V},\mathbf{X}}(0,\mathbf{X}|\boldsymbol{\theta})$ can be formulated as an optimisation in the elements of the vector, $\boldsymbol{\lambda}$.
3. Let the columns of $\mathbf{V}_i$ be an orthonormal basis spanning null($\mathbf{M}_i$) and $\mathbf{V}_{i+1}=[\mathbf{V}_i \,|\, \overline{\mathbf{v}}_i]$. Let the columns of $\mathbf{M}_{i+1}$ be an orthonormal basis spanning null($\mathbf{V}_{i+1}$), then the columns of $\mathbf{M}_{i+1}$ are also an orthonormal basis of Ψ_{nl}^{i+1}. A diagnostic for the validity of updating the sub-space, Ψ_{nl}^i, to Ψ_{nl}^{i+1} is the test statistic, η, evaluated with the orthonormal set of vectors defining $\mathbf{V}_{\mathbf{F}}^{\mathrm{K}}(\boldsymbol{\theta}_M)$ chosen to be the single vector, $\overline{\mathbf{v}}_i$.
4. If i<n+m then i=i+1, go to 2.

(i) Dimension of the minimal sub-space.

In step (2), η can be expected to abruptly decrease when the rank of $\mathbf{M}_i$ becomes less than the dimension of the minimal nonlinear subspace. Such a transition can be utilised to estimate the dimension, q, of the minimal nonlinear subspace. Transitions can, of course, be obscured by noise but the validity of a choice of dimension, q, can be further assessed/confirmed using the pointwise estimation methods discussed in section 4 below.

(ii) Special case enabling simplified procedure.

Assume that

$$E[(\mathbf{H_F}(\mathbf{z}_i)\mathbf{e}_m - E[\mathbf{H_F}(\mathbf{z}_i)\mathbf{e}_m])(\mathbf{H_F}(\mathbf{z}_j)\mathbf{e}_n - E[\mathbf{H_F}(\mathbf{z}_j)\mathbf{e}_n])^T] \propto \mathrm{I}\delta_{mn}$$

where the j^{th} component of the vector, $\mathbf{e}_i$, is δ_{ij}. It follows that, $\boldsymbol{\Lambda}$, the covariance of $\mathbf{h}_{\mathbf{F}}^i = [(\mathbf{H_F}(\mathbf{z}_1)\hat{\mathbf{v}}_i)^T \cdots (\mathbf{H_F}(\mathbf{z}_K)\hat{\mathbf{v}}_i)^T]^T$, is independent of $\hat{\mathbf{v}}_i$. Since $\boldsymbol{\Lambda}$ is, by definition, positive definite, $\boldsymbol{\Lambda}^{-1}$ can be decomposed as $\mathbf{R}^T\mathbf{R}$, and

$$-\ln p_{\mathbf{V}}(0|\boldsymbol{\theta},\mathbf{X}) \propto E\left(\mathbf{h}_{\mathbf{F}}^{i}\right)^{T}\boldsymbol{\Lambda}^{-1}E\left(\mathbf{h}_{\mathbf{F}}^{i}\right)+\log|\boldsymbol{\Lambda}| = E\left(\mathbf{h}_{\mathbf{F}}^{i}\right)^{T}\mathbf{R}^{T}\mathbf{R}E\left(\mathbf{h}_{\mathbf{F}}^{i}\right)+\log|\boldsymbol{\Lambda}| \quad (5)$$

Hence,

$$-\ln p_{\mathbf{V}}(0|\boldsymbol{\theta},\mathbf{X}) \propto \hat{\mathbf{v}}_i^{T}\mathbf{W}^{\mathbf{T}}\mathbf{W}\hat{\mathbf{v}}_i+\log|\boldsymbol{\Lambda}| \quad (6)$$

with $\mathbf{W}=\mathbf{R}\,E\left[[\mathbf{H}_{\mathbf{F}}(\mathbf{z}_1)\cdots\mathbf{H}_{\mathbf{F}}(\mathbf{z}_K)]^{T}\right]$. (This is available under the assumption that the joint probability distribution for $\mathbf{V}_{\mathbf{F}}^{K}(\boldsymbol{\theta}_M)$ is available for any $\boldsymbol{\theta}$). Since log $|\boldsymbol{\Lambda}|$ is constant it does not affect the minima of $-\ln p_{\mathbf{V}}(0|\boldsymbol{\theta},\mathbf{X})$. The minimum under the constraint that $\hat{\mathbf{v}}_i \in$basis{null($\mathbf{M}$)} can be expressed in closed-form: letting the singular value decomposition of $\mathbf{W}$ be $\mathbf{W}=\mathbf{U}^{T}\boldsymbol{\Sigma}\mathbf{U}$, it follows immediately that (6) is minimised with $\hat{\mathbf{v}}_i \in$basis{null($\mathbf{M}$)} when $\mathbf{M}=\mathbf{U}_q$, where $\mathbf{U}_q$ is the matrix consisting of the first n+m-q rows of $\mathbf{U}$. This can be calculated very efficiently. More generally, this value can be used to initialise the optimisation in the iterative procedure above.

3.2 Examples

(i) Consider the nonlinear dynamic system

$$y(t_{n+1}) = 0.5G(\rho(t_n)) \quad (7)$$

where $G(\rho)=\tanh(\rho)+0.01\rho$ and $\rho=r-y$. The plant output in response to a Gaussian input with mean zero and variance 3 units is measured and 300 data points collected. 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).

The measured data, together with the corresponding predicted fit from a non-parametric Gaussian process prior model of this data, are illustrated in figure 1a (explanatory variables are $(r(t_n),y(t_n))$ and model output is $y(t_{n+1})$). The change in $-2\ln(\overline{\eta})$ as the dimension of the nonlinear sub-space is reduced is shown in Table 1. It can be seen that, as expected, the cost rises abruptly when the dimension falls below unity; that is, the dimension of the minimal nonlinear sub-space. The estimated basis, $\mathbf{M}$, of the minimal nonlinear subspace is [0.697 -0.717]; that is, ρ is estimated to be 0.697r-0.717y. Subject to an arbitrary normalisation factor, it is evident that the identification procedure successfully infers the nonlinear dependence of the plant dynamics.

This example is, of course, simple having been selected to be low order to enable results to be readily visualised. Nevertheless, it should be noted that working directly in terms of the explanatory variables r and y requires the development of a model of the two dimensional mapping relating $(r(t_n),y(t_n))$ to $y(t_{n+1})$; for example, a radial basis function (RBF) model (e.g. see Bishop 1995) with 10 centres per axes has 100 centres in total and 200 parameters. Inference of the scalar nature of the nonlinear dependence during initial data exploration allows the task to be simplified to modelling a one dimensional mapping only: an RBF model with 10 centres per axes now has 10 centres in total and 20 parameters. Hence, even in the case of a simple system the benefits of dimensionality reduction stemming from the identification of the nonlinear structure are potentially considerable.

(ii) Consider the Wiener-Hammerstein nonlinear system illustrated in figure 1b. Reformulating the dynamics in terms of the measured variables (input, r, and output, y) yields

$$y(t_n) = 0.3r_1^3 + 0.165r_2^3$$

where $\boldsymbol{\rho}=\mathbf{M}\left[r(t_n)\quad r(t_{n-1})\quad r(t_{n-2})\quad r(t_{n-3})\right]^{T}$ with

$$\mathbf{M}=\begin{bmatrix}0.9184 & 0.3674 & 0 & 0\\ 0 & 0 & 0.9184 & 0.3674\end{bmatrix}$$

and $\boldsymbol{\rho}_i$, i=1,2 denotes the i^{th} element of vector $\boldsymbol{\rho}$. The plant 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). A non-parametric Gaussian process prior model is used with explanatory variables $[r(t_n)\ r(t_{n-1})\ r(t_{n-2})\ r(t_{n-3})]^{T}$ and model output $y(t_n)$. The change in the test statistic, $-2\ln(\eta)$, as the dimension of the nonlinear sub-space is varied indicates that a minimal nonlinear subspace of dimension two. The associated estimate of the nonlinear dependence is

$$\hat{\mathbf{M}}=\begin{bmatrix}0.9292 & 0.3694 & -0.0008 & 0.0018\\ -0.0015 & 0.0040 & 0.9282 & 0.3719\end{bmatrix}$$

The estimate evidently agrees well with the true nonlinear dependence, particularly in view of the small number of data points on which it is based (150 points from a four dimensional mapping).

Remark Wiener-Hammerstein systems form an important class and the identification of such systems remains a challenging problem in its own right. Consider the transversal Wiener-Hammerstein system

$$x_1 = \left(a_n q^{-n}+\ldots+a_o\right)r$$

$$x_2 = f(x_1)$$

$$y = \left(b_m q^{-m}+\ldots+b_o\right)x_2$$

Reformulating the dynamics in terms of the input, r, and output, y yields

$$y = b_m f(\boldsymbol{\rho}_{m+1})+\ldots+b_o f(\boldsymbol{\rho}_1)$$

where

$$\boldsymbol{\rho}=\begin{bmatrix}a_n & \cdots & a_o & 0 & & \cdots & 0\\ 0 & a_n & \cdots & a_o & 0 & \cdots & 0\\ & & & & \ddots & & \\ 0 & & \cdots & & 0 & a_n \cdots & a_o\end{bmatrix}\begin{bmatrix}q^{-n-m}r\\ q^{-n-m-1}r\\ \vdots\\ r\end{bmatrix} \quad (8)$$

and $\boldsymbol{\rho}_i$, i=1..m+1 denotes the elements of vector $\boldsymbol{\rho}$. (Note, when a coefficients b_i is zero, the corresponding row in (8) is deleted and the dimension of $\boldsymbol{\rho}$ correspondingly reduced, see above example). Using the delayed inputs as explanatory variables, and assuming that the overall order of the system is known (this might be inferred in an iterative manner), it can be seen that the nonlinear dependence has a specific block diagonal structure. By inspection, the coefficients, a_i, of the input filter and the delay taps of the output filter can be directly inferred. As one of the main task with Wiener-

Hammerstein systems is identifying the partitioning into input and output filters, identification of the remaining system elements is now relatively straightforward. Specifically, once the input filter is known, the output filter can be inferred from the transfer function of the linearisation about any equilibrium point and the system nonlinearity then directly estimated.

4. Locally Validating Nonlinear Dependence in a Region

The foregoing methods developed for summarising the nonlinear dependence in a region can be immediately applied to summarise the nonlinear dependence locally to a single point. By studying the local nonlinear dependence at a number of points drawn from a region of interest, D, the validity of the regional estimate of the minimal nonlinear subspace can be assessed in a fairly direct manner. Specifically, for any function (3) we have that

(i) $\dim(\text{null}(\mathbf{H_F}(\mathbf{z}))) \geq \dim \Psi_l$ (ii) $\bigcap_{\mathbf{z}\in D} \text{null}(\mathbf{H_F}(\mathbf{z})) = \Psi_l$

Typically (but not always), the dimension of the null space of the Hessian $\mathbf{H_F}(\mathbf{z})$ is greater than that of ψ_l only for a set of points of measure zero in D. Almost everywhere the $\dim(\text{null}(\mathbf{H_F}(\mathbf{z})))$ is uniformly equal to $\dim(\psi_l)$ with $\text{null}(\mathbf{H_F}(\mathbf{z}))$ necessarily equal to ψ_l. Consequently, good agreement between the local nonlinear dependencies and the regional estimate provides a degree of confidence that the nonlinear dependence is well summarised. Conversely, if, for example, it appears that the domain can be decomposed into sub-regions each exhibiting consistently different local nonlinear dependence, this might indicate limitations in the use of a single summary of the nonlinear dependence over the region.

Remark It is important to note that the regional estimate for the basis of Ψ_l is equivalent to the mean of the pointwise estimates over the region of interest (owing to the correlation that generally exists between the pointwise estimates).

4.2 Examples

(i) Returning to the system, (7), considered in section 3.3 above, figure 2a shows the variation in pointwise test statistic, $-2\ln(\overline{\eta})$, with respect to the dimension of the nonlinear sub-space at 50 operating points selected uniformly from the domain covered by the measured data. (In this case, the rejection criterion, $\chi^2_{rK}(0.99)$, is 9.21 for the dimension of Ψ_{nl} being 0 and 6.63 for the dimension being 1). It can be seen that, in accordance with the previous results, the test statistic rises abruptly when the dimension falls below unity. The corresponding estimates of **M**, a basis for the minimal nonlinear sub-space estimated at each point are shown in figure 2b. Evidently, the pointwise estimates are in good agreement with the overall regional estimate of the nonlinear dependence, indicating that ρ equals r-y, and this helps give some confidence in the regional estimate.

(ii) Consider a system also of the form (7) but with ρ equal to r-sin(ay)/a, and a=1. For values of y close to zero, sin(ay)/a is nearly linear in y and this system accurately approximates the previous system for which ρ=r-y. However, when a wider region is considered, the distinction between the two systems can be expected to become more noticeable as the impact of the difference in dimension of the nonlinear sub-spaces when ρ=r-y and ρ=r-sin(ay)/a (dimension one and dimension two, respectively) becomes significant. A lower level of measurement noise, with standard deviation 0.01 units, is used in this example so as to avoid obscuring the fine detail of the plots, particularly Figure 3a. Applying the techniques developed in section 3, and using the domain considered in Example (i), the estimate of the basis, **M**, of the minimal nonlinear subspace is [0.756 -0.655]. When the input and initial conditions are now constrained such that the data is confined to a region close to the origin, the corresponding estimate of **M** becomes [0.708 -0.703]. The latter agrees well with the results for Example (i), as expected. However, the results for the larger region provide little insight into the nature, or degree, of the difference between the system in Example (i) and that considered here.

With regard to gaining insight into the differences between these systems, consider the pointwise estimates of the local nonlinear sub-space as shown in Figure 3a. This plot uses more data points than the previous plots in order to reveal the detailed structure of the variation in the pointwise estimates across the domain. Measurement noise generally results in *uncorrelated* variations in the pointwise estimates across the domain, while a strong spatial correlation is evident between the estimates in Figure 3a. This structure is visually quite striking, particularly when compared with the corresponding plot for the system in Example (i). In the vicinity of the line y=0, the pointwise estimates of **M** agree well with those for the system of Example (i); this is not unexpected since, as noted previously, sin(ay)/a is nearly linear for small y and so the nonlinear dependence is locally similar near to this line. As the parameter, a, is decreased the pointwise estimates of **M** become more like those observed in Example (i); for example, the pointwise estimates obtained for a=0.1 are shown in Figure 3b. This is in accordance with the fact that sin(ay)/a $\rightarrow$y as a$\rightarrow$0 and thus ρ$\rightarrow$r-y as in Example (i). Detailed diagnostic analysis of pointwise estimates beyond the simple observations noted above is not pursued further here as it is not essential in the present context. That the correct dimension of ρ has been identified, or not, is validated by the uniformity, or otherwise, of the pointwise estimates and this example illustrates that pointwise estimates thereby provide a useful tool for validation.

5. Conclusions

This paper investigates new ways of inferring nonlinear dependence from measured data. The existence of unique linear and nonlinear sub-spaces, that are structural invariants of general nonlinear mappings, is established and necessary and sufficient conditions determining these sub-spaces are derived. The importance of these invariants in an identification context is that they provide a tractable framework for minimising the dimensionality of the nonlinear modelling task. Specifically, once the linear/nonlinear sub-spaces are known, by definition the explanatory variables

may be transformed to form two disjoint sub-sets spanning, respectively, the linear and nonlinear sub-spaces. The nonlinear modelling task is confined to the latter sub-set, which will typically have a smaller number of elements than the original set of explanatory variables. A constructive algorithm is proposed for inferring the linear and nonlinear sub-spaces from noisy data and its application is illustrated in a number of simple examples (as the focus of the present paper is on theoretical issues, large scale applications are not pursued here). Algorithms for inferring pointwise sub-space estimates are proposed and the use of pointwise estimates for validating regional estimates of nonlinear dependence is demonstrated.

Acknowledgement

This work was supported by the Royal Society through a personal resarch fellowship to D.Leith, by Science Foundation Ireland grant 00/PI.1/C067, by the EC through EC TMR grant HPRNCT-1999-00107, and EPSRC grants GR/M76379/01 and GR/R15863/01.

Appendix – Non-parametric Gaussian process priors

Consider a smooth function f(.) dependent on the explanatory variable, $\mathbf{z}\in D\subseteq \Re^p$. To avoid cumbersome notation, f is scalar (the generalisation to vector functions is straightforward). 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(**z**) or, more precisely, to determine a probabilistic description of f(**z**) on the domain, D, containing the data. Note that this is a regression formulation and it is assumed the input **z** is noise free[1].

The probabilistic description of the function, f(**z**), adopted is the stochastic process, $f_{\mathbf{z}}$, with the E[$f_{\mathbf{z}}$], as **z** varies, 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, {$\mathbf{z}_1,\ldots,\mathbf{z}_k$}, from 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 since the measurement noise, n_i, and the 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 particular, in the Gaussian Process prior model considered here, it is assumed that 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(**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 Gaussian assumption 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 (Green & Silverman 1994) subsuming, amongst others, RBF, spline and many neural network models (Williams 1998). 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.

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

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

where $\mathbf{Y}=[y_1,\cdots,y_N]^T$, Λ_{11} is $C(f_{\mathbf{z}},f_{\mathbf{z}})$, the ijth element of the covariance matrix $\boldsymbol{\Lambda}_{22}$ is $C(y_i, y_j)$ and the ith 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

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

with $\hat{f}_{\mathbf{z}} = \boldsymbol{\Lambda}_{21}^{T}\boldsymbol{\Lambda}_{22}^{-1}\mathbf{Y}$, $\Lambda_{\mathbf{z}} = \Lambda_{11} - \boldsymbol{\Lambda}_{21}^{T}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}$. 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

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

In addition, assume that the related stochastic process, $f_{\mathbf{z}}^{\delta\mathbf{e}_i}$, where $f_{\mathbf{z}}^{\delta\mathbf{e}_i} = (f_{(\mathbf{z}+\delta\mathbf{e}_i)} - f_{\mathbf{z}})/\delta$ and $\mathbf{e}_i$ is a unit basis vector, is well-defined in the limit as $\delta\to 0$, i.e. all the necessary probability distributions for a complete description exist. Denote the derivative stochastic process, i.e. the limiting random process, by $f_{\mathbf{z}}^{\mathbf{e}_i}$. The E[$f_{\mathbf{z}}^{\mathbf{e}_i}$] as **z** varies is interpreted as a fit to $\frac{\partial f}{\partial z_i}(\mathbf{z})$ when the partial derivative of f(**z**) in the

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

direction $\mathbf{e}_i$ exists. Provided the covariance $C(f_{\mathbf{z}_i}, f_{\mathbf{z}_j})$ is sufficiently differentiable, it is well known (O'Hagan 1978) that $f_{\mathbf{z}}^{\mathbf{e}_i}$ is itself Gaussian and that

$$E[f_{\mathbf{z}}^{\mathbf{e}_i}] = \frac{\partial}{\partial z_i} h_f(\mathbf{z}) \quad ; \quad h_f(\mathbf{z}) = E[f_{\mathbf{z}}] \qquad (9)$$

where z_i denotes the i^{th} element of $\mathbf{z}$; that is, the expected value of the derivative stochastic process is just the derivative of the expected value of the stochastic process. Furthermore,

$$E[f_{\mathbf{z}_0}^{\mathbf{e}_i} f_{\mathbf{z}_1}^{\mathbf{e}_j}] = \nabla_i^1 \nabla_j^2 C_f(\mathbf{z}_0, \mathbf{z}_1) \; ; C_f(\mathbf{z}_0, \mathbf{z}_1) = E[f_{\mathbf{z}_0} f_{\mathbf{z}_1}] \quad (10)$$

where $\nabla_i^1 Q(\mathbf{z}_o, \mathbf{z}_1)$ denotes the partial derivative of $Q(\mathbf{z}_o, \mathbf{z}_1)$ with respect to the i^{th} element of its first argument, *etc*.

The above procedure can be repeated to construct second derivative stochastic processes. The means and covariances can be determined by recursive application of (9) and (10).

In the examples discussed in sections 3 and 4 of this paper, a straightforward smoothness prior covariance function is used which ensures that measurements associated with nearby values of the explanatory variable should have higher covariance than more widely separated values of the explanatory variable; specifically,

$$C(f_{\mathbf{z}_i}, f_{\mathbf{z}_j}) = \gamma \exp\left[-\sum_k \left((\mathbf{z}_i)_k - (\mathbf{z}_j)_k \right)^2 / 2\alpha_k \right]_k \qquad (11)$$

where $(z_i)_k$ denotes the k^{th} element of vector $\mathbf{z}_i$. The value of α_k characterises the rate of variation of the function in dimension k, thereby, estimating the degree of nonlinearity or the relative smoothness in different directions of the explanatory variable. The corresponding covariance for y_i is

$$C(y_i, y_j) = \gamma \exp\left[-\sum_k \left((\mathbf{z}_i)_k - (\mathbf{z}_j)_k \right)^2 / 2\alpha_k \right] + \beta \delta_{ij} \qquad (12)$$

The parameter β is the variance of the measurement noise, n, on the output. To obtain a model given the data, M, the hyperparameters (β, α_k, γ), whilst constrained to be positive, are adapted to maximise the likelihood $p(M|(\beta, \alpha_k, \gamma))$. The covariance function, (11), is sufficiently smooth for the derivative and second derivative stochastic processes to be well-defined and the relations (9) and (10) to apply (O'Hagan 1978).

References

BISHOP, C., 1995, *Neural Networks for Pattern Recognition*. (Clarendon Press, Oxford)

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

HASTIE, T., TIBSHIRANI, R, 1990, *Generalised Additive Models*. (Chapman & Hall, London).

HUNT, K.J.,JOHANSEN,T.A., 1997, Design and analysis of gain-scheduled control using local controller networks. *Int. J. Contr*, **66**, 619-651.

JOHANSEN,T.A.,FOSS,B.A., 1995, Identification of Nonlinear System Structure and Parameters using Regime Decomposition. *Automatica*, **31**, 321-326.

JOHANSEN, T.A., MURRAY-SMITH, R., 1997, 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).

JUDITSKY,A.,HJALMARSSON,H,BENVENISTE,A.,DEY LON,B.,LJUNG,L.,SJOBERG,J.,ZHANG,Q., 1995, Nonlinear black box models in system identification: mathematical foundations. *Automatica*, **31**, pp1725-1750.

KORENBERG,M., BILLINGS, A., LIU, Y., McILROY,P., 1988, Orthogonal Parameter Estimation Algorithm for Non-linear Stochastic Systems. *Int. J. Contr*, **48**, 193-210.

LEITH, D.J., LEITHEAD, W.E., 1998a, Gain-Scheduled & Nonlinear Systems: Dynamic Analysis by Velocity-Based Linearisation Families. *Int. J. Contr*, **70**, 289-317.

LEITH, D.J., LEITHEAD, W.E., 1998b, Gain-Scheduled Controller Design: An Analytic Framework Directly Incorporating Non-Equilibrium Plant Dynamics. *Int. J. Contr*, **70**, 249-269.

LEITH, D.J., LEITHEAD, W.E, 2000, Survey of Gain-Scheduling Analysis and Design. *Int. J. Contr*, **73**, 1001-1025

MURRAY-SMITH, R., JOHANSEN, T.A., SHORTEN, R., 1999, On Transient Dynamics, Off-Equilibrium Behaviour and Identification in Blended Multiple Model Structures. *Proc. European Control Conference*, Karlsruhe.

NEAL, R., 1996, *Bayesian Learning for Neural Networks*. (Springer, New York).

O'HAGAN, A., 1978, On curve fitting and optimal design for regression*, J. Royal Stat Soc. B*, **40**, 1-42.

SHAMMA, J.S., ATHANS, M., 1990, Analysis of Gain Scheduled Control for Nonlinear Plants. *IEEE Trans Aut Contr.*, **35**, 898-907.

SJOBERGMJ., ZHANG,Q., LJUNG,L., BENVENISTE,A.,DEYLON,B.,GLORRENC,P., HJALMARSEN.H., JUDITSKY, A., 1995, Nonlinear black-box modelling in system identification: a unified overview. *Automatica*, **31**, 1691-1724.

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

YOUNG, P., 2000, Comments on 'A quasi-ARMAX approach to modelling nonlinear systems'. *Int. J. Contr*, in press.

Dimension of Ψ_{nl}	**$-2\ln(\overline{\eta})$**	**$\chi^2_{rK}(0.99)$**
2	0	-
1	99.92	360
0	5963.46	684

Table 1 $-2\ln(\overline{\eta})$ vs dimension of Ψ_{nl} in Example (i) of section 3.2.

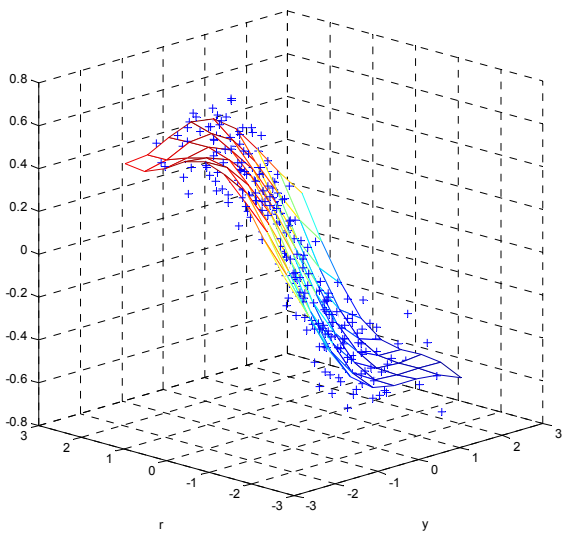

Figure 1a Measured data (+) and associated Gaussian Process model in Example (i) of section 3.3.

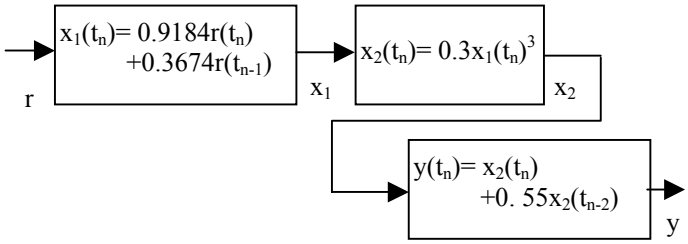

Figure 1b Block diagram representation of system studied in Example (ii) of section 3.3.

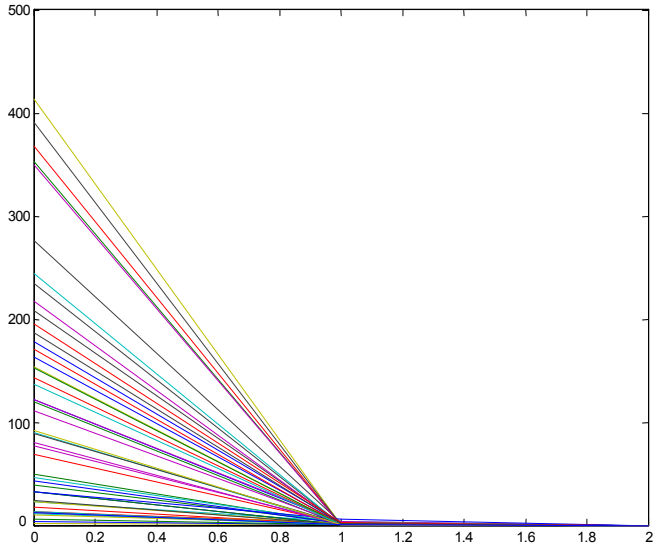

Figure 2a Pointwise $-2\ln(\overline{\eta})$ vs dimension of $\boldsymbol{\Psi}_{\mathbf{nl}}$ in Example (i) of section 4.2.

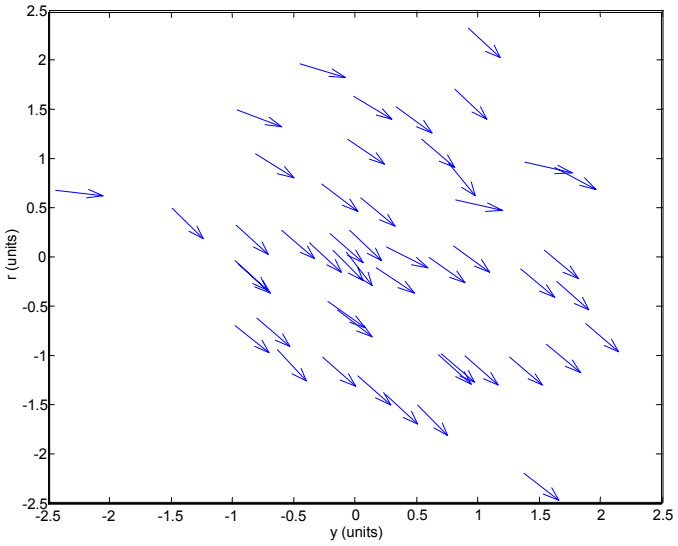

Figure 2b Point estimates of **M** (Example (i) of section 4.2).

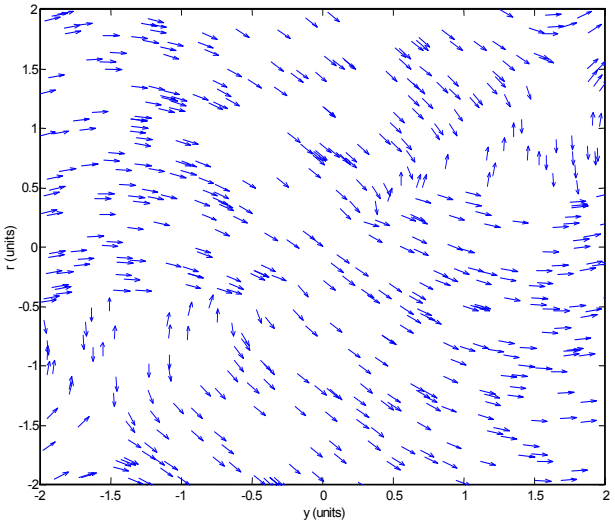

Figure 3a Point estimates of **M** in Example (ii) of section 4.2 with a=1.

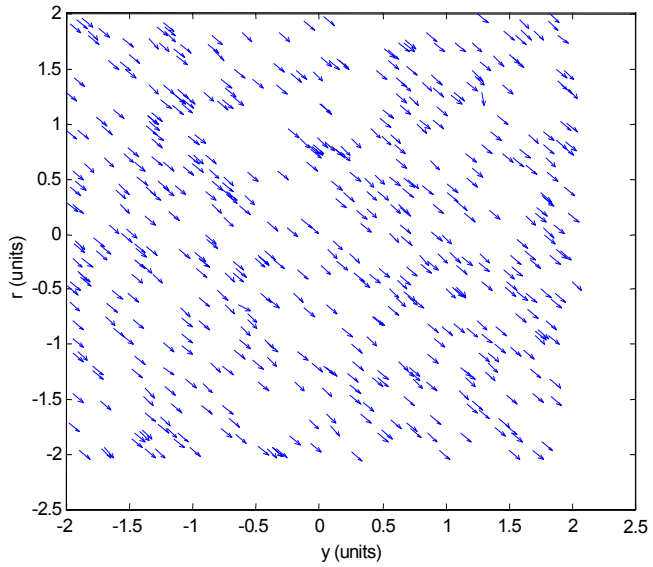

Figure 3b Point estimates of **M** in Example (ii) of section 4.2 with a=0.1.