NONLINEAR STRUCTURE IDENTIFICATION: A NON-PARAMETRIC/VELOCITY-BASED APPROACH

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

[1]Dept. of Electronics & Electrical Engineering, University of Strathclyde,
50 George St., Glasgow G1 1QE, U.K., Email. doug@eee.strath.ac.uk

[2]Dept. of Computing Science, University of Glasgow,
Glasgow G12 8QQ, Email. rod@dcs.gla.ac.uk

Abstract

This paper investigates new ways of inferring nonlinear dependence parsimoniously from measured data. The existence of unique linear and nonlinear sub-spaces which are structural invariants of general nonlinear maps is established for the first time. Necessary and sufficient conditions determining these sub-spaces are derived. In addition to being of considerable interest in their own right, 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

This paper is concerned with the structure identification task for nonlinear dynamic systems. Model structure selection/identification is widely recognised as a key aspect of any system identification campaign, impacting directly on the bias/variance trade-off which lies at the heart of empirical modelling theory and practice (*e.g* Ljung 1987) and largely determining the degree of interpretability and transparency achieved. One particularly important aspect of model structure in many contexts is the nonlinear dependence of the dynamics. Of course, the nonlinearity of any dynamics may be characterised in terms of all the explanatory variables (for example, all the inputs and states). However, this is rarely the most parsimonious, or insightful, approach. Instead, it is usually much more useful to be able 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*. Chen *et al.* 1991, Johansen & Foss 1993, Johansen & Murray-Smith 1997, Young 2000). 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.* Bishop 1995, Johansen & Foss 1995, Johansen & Murray-Smith 1997). Parsimonious knowledge of the nonlinear dependence of dynamics also plays a key role in analysis and design: it is, for example, known that the use of an appropriate scheduling variable is often of great importance in gain-scheduling controllers (*e.g.* Shamma & Athans 1990, Hunt & Johansen 1997, Leith & Leithead 2000), and similarly with regard to the parameter dependence in LPV/quasi-LPV modelling and design approaches (*e.g.* Leith & Leithead 2000).

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 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, $\sum_i \boldsymbol{\varphi}_i(u_i)$ (where u_i denotes the i^{th} element of the input vector and $\boldsymbol{\varphi}_i$ is an associated nonlinear, possibly vector, function), back-fitting methods can be used to directly estimate the $\boldsymbol{\varphi}_i$, and thereby linearity or nonlinearity with respect to each explanatory variable, u_i, without necessarily postulating a particular series expansion (*e.g.* Hastie & Tibshirani 1990, Young 2000). Similarly with 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, a similar situation also holds 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 such situations, algorithms to search for appropriate operating region decompositions (*e.g.* Johansen & Foss 1995) can indirectly detect linearity with respect to particular input elements.

As noted previously, existing methods are focussed on situations where the nonlinear dependence is determined with respect to variables which are aligned with some particular choice of co-ordinate axes that has been selected in advance. Obviously, the effectiveness of such methods in inferring a parsimonious dependence may be 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 axes aligned methods to incorporate estimation of, for example, an input transformation in order to adjust the axes 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 nonlinear dependence parsimoniously from measured data. A key enabling technology for this work are the recent developments in systems theory relating to non-parametric nonlinear representations. In an identification context only discrete measured data points are available and it is therefore necessary to determine a suitable representation for the underlying nonlinear function which is tractable yet does not *a priori* assume the nonlinear dependence. Standard parameter estimation techniques necessarily require the postulation of a parametric model with a specific structure. In particular, parametric models inevitably entail structural assumptions regarding the nonlinear dependence (as noted previously, when these assumptions are inappropriate, for example a poor choice of explanatory variables in a radial basis function network, a great many parameters may be needed in order to obtain an accurate model). Parametric models do not, therefore, seem well suited to the present structure identification context. In contrast, non-parametric approaches are characterised by drawing inferences directly from the measured data using smoothness information but without assuming an underlying parameterisation (*e.g.* Green & Silverman 1994). Non-parametric approaches are thus well suited to initial data analysis and exploration due to their ability to model data well while making few structural assumptions. Non-parametric approaches are also attractive from the viewpoint that they permit direct inference of model structure, reducing the need for iterative postulation of model structure followed by hypothesis testing. An example of a non-parametric model for nonlinear dynamics is a Gaussian Process prior, as reviewed in Williams (1998) and initially proposed in O'Hagan (1978). This is a Bayesian form of kernel regression model (Green & Silverman 1994). We concentrate on Gaussian Process priors in this paper in order to fix ideas and because of their high performance and analytic properties, but other non-parametric representations could also be used (*e.g.* support vector machines, locally weighted regression; see also Juditsky *et al*. 1995).

2. Structural Decomposition

This paper studies nonlinear maps, $\mathbf{F}: D \rightarrow R$, with open domain $D \subseteq \Re^{n+m}$, range $R \subseteq \Re^n$ and $\mathbf{F}$ continuously twice differentiable. While this setting is general, the particular interest here (and reflected in the examples chosen) is in dynamic systems applications where the nonlinear map might typically be the right-hand side of a differential/difference equation

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

where the input is $\mathbf{r} \in D_r \subseteq \Re^m$, the state $\mathbf{x} \in D_x \subseteq \Re^n$ and $\mathbf{D}$ 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 can be made explicit by reformulating as

$$\mathbf{F}(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{f}(\boldsymbol{\rho}(\mathbf{z})) \tag{2}$$

with $\mathbf{z}=[\mathbf{x}^T(t) \quad \mathbf{r}^T(t)]^T$ and where $\mathbf{A}$ is an appropriately dimensioned constant matrix, $\mathbf{f}(\bullet)$ is a nonlinear function, and $\boldsymbol{\rho}(\mathbf{z})\in D_\rho \subseteq \Re^q$, $q \leq m+n$. As it stands, the decomposition (2) is, of course, not unique. Non-uniqueness is, for example, associated with the freedom available in the choice of $\boldsymbol{\rho}$ and the assignment of the linear component of the dynamics between the $\mathbf{A}$ matrix and the mapping $\mathbf{f}$. The requirement is therefore to determine a canonical decomposition, or class of decompositions, for which the dimension of $\boldsymbol{\rho}$ is, in some sense, minimal.

Consider the class of decompositions, (2), for which $\boldsymbol{\rho}$ is a linear function of $\mathbf{z}$; that is, $\boldsymbol{\rho}(\mathbf{z})=\mathbf{M}\mathbf{z}$ with $\mathbf{M}$ a constant matrix. Trivially, such a reformulation can always be achieved by letting $\boldsymbol{\rho} = \mathbf{z}$, in which case $q=m+n$. However, the nonlinearity of the system is frequently dependent on only a subset of the states and inputs, in which case the dimension, q, of $\boldsymbol{\rho}$ may be less than $m+n$.

Proposition (minimal decomposition) Consider a twice differentiable nonlinear function **F** and a decomposition

$$\mathbf{F}(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{f}(\mathbf{M}\mathbf{z}) \tag{3}$$

where $\mathbf{M}$ has full row rank. $\mathbf{M}$ is said to be of minimal degree when there exists no choice of $\mathbf{M}$ with lower full row rank such that a nonlinear function $\mathbf{f}$ can be found satisfying the equality in (3). Let $\mathbf{H_F}(\mathbf{z_1})$ denote the Hessian

$$\mathbf{H_F}(\mathbf{z_1}) = \begin{bmatrix} \nabla\left(\nabla \mathbf{F}_1(\mathbf{z_1})\right)^T \\ \vdots \\ \nabla\left(\nabla \mathbf{F}_n(\mathbf{z_1})\right)^T \end{bmatrix} \tag{4}$$

with $\mathbf{F}_i$ denoting the i^{th} element of the vector function $\mathbf{F}$. Then $\mathbf{M}$ is of minimal degree if an only if there exists no vector $\mathbf{v}\in\{\mathbf{v}\in\Re^{n+m}: \mathbf{Mv}\neq 0\}$ for which $\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0\ \forall \mathbf{z}\in \mathrm{D}$.

Proof First of all, observe that $\mathbf{H_F}(\mathbf{z}) = \mathbf{H_f}(\mathbf{z})$ since $\mathbf{H_{Az}}(\mathbf{z})$ is identically zero. It is then straightforward to verify that the conditions of the proposition are satisfied when $\mathbf{M}$ is minimal. Conversely, suppose $\mathbf{M}$ is non-minimal, then by definition $\mathbf{f}$ can be decomposed as $\bar{\mathbf{f}}(\overline{\mathbf{M}}\mathbf{z}) + \overline{\mathbf{A}}\mathbf{z}$ where $\overline{\mathbf{M}}$ has full column rank lower than that of $\mathbf{M}$. Evidently, $\bar{\mathbf{f}}(\overline{\mathbf{M}}\mathbf{z})$ is constant for $\mathbf{z}=\mathbf{z}_o+\mathbf{v}$: $\mathbf{v}\in \overline{\mathrm{V}}_o = \{\mathbf{v}\in\Re^{n+m}: \mathbf{z}_o+\mathbf{v}\in \mathrm{D}, \overline{\mathbf{M}}\mathbf{v}=0\}$ while $\mathbf{f}(\mathbf{Mz})$ is constant for $\mathbf{z}=\mathbf{z}_o+\mathbf{v}$: $\mathbf{v}\in \mathrm{V}_o=\{\mathbf{v}\in\Re^{n+m}: \mathbf{z}_o+\mathbf{v}\in \mathrm{D}, \mathbf{Mv}=0\}$ but may vary for $\mathbf{z}=\mathbf{z}_o+\mathbf{v}$: $\mathbf{v}\in\overline{\mathrm{V}}_o$. This variation is necessarily linear and so there exists $\mathbf{v}\in \overline{\mathrm{V}}_o \supset \mathrm{V}_o$ for which $\mathbf{H_f}(\mathbf{z})\mathbf{v} = 0\ \forall \mathbf{z}\in \mathrm{D}$, thus violating the assumed conditions. ■

Corollary (subspace partitioning) Consider the class of decompositions

$$\mathbf{F}(\mathbf{z}) = \mathbf{Az} + \mathbf{f}(\mathbf{Mz}) \tag{5}$$

with $\mathbf{M}$ minimal. Let $\overline{\Psi}_{nl}(\mathbf{M})$ denote the sub-space spanned by $\mathbf{M}$. This sub-space is identical for every minimal $\mathbf{M}$; that is, $\exists$ a unique sub-space Ψ_{nl} defined by $\overline{\Psi}_{nl}(\mathbf{M})=\Psi_{nl}$ $\forall\mathbf{M}$: $\exists$ no vector $\mathbf{v}\in \mathrm{V}_o=\{\mathbf{v}\in\Re^{n+m}: \mathbf{Mv}\neq 0\}$ for which $\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0\ \forall \mathbf{z}\in \mathrm{D}$.

Proof This corollary is a direct consequence of the definition of minimality and the foregoing proposition. Proceeding by contradiction, consider two minimal decompositions, $\mathbf{M_o}$ and $\mathbf{M_1}$, spanning different sub-spaces, $\overline{\Psi}_o$ and $\overline{\Psi}_1$, of $\Re^{n+m}$. Assume, for the moment, that the domain D is $\Re^{n+m}$. Assume also that $\overline{\Psi}_o / \{\overline{\Psi}_o \cap \overline{\Psi}_1\}$ and $\overline{\Psi}_1 / \{\overline{\Psi}_o \cap \overline{\Psi}_1\}$ are not empty. Bearing in mind that both decompositions embody the same mapping $\mathbf{F}$:$\mathbf{z}\to$R, with regard to case (i) it follows that in the region $\overline{\Psi}_o / \{\overline{\Psi}_o \cap \overline{\Psi}_1\}$ the realisation of the mapping $\mathbf{F}$ provided by the decomposition in terms of $\mathbf{M_1}$ is linear. Similarly, the mapping must also be linear in the region $\overline{\Psi}_1 / \{\overline{\Psi}_o \cap \overline{\Psi}_1\}$. Since $\overline{\Psi}_o \cap \overline{\Psi}_1$ is itself a linear sub-space, it follows that there exists a decomposition in terms of $\mathbf{M}$ spanning $\overline{\Psi}_o \cap \overline{\Psi}_1$ which is also a realisation of the mapping $\mathbf{F}$ but with $\mathbf{M}$ lower full row rank than both $\mathbf{M_o}$ and $\mathbf{M_1}$; that is, $\mathbf{M_o}$ and $\mathbf{M_1}$ are not minimal. By a similar argument, when $\overline{\Psi}_o / \{\overline{\Psi}_o \cap \overline{\Psi}_1\}$ (respectively $\overline{\Psi}_1 / \{\overline{\Psi}_o \cap \overline{\Psi}_1\}$) is empty, $\mathbf{M_o}$ (respectively $\mathbf{M_1}$) is minimal but $\mathbf{M_1}$ (respectively $\mathbf{M_o}$) is not. Hence, when $\mathbf{M_o}$ and $\mathbf{M_1}$ are minimal then $\overline{\Psi}_o$ and $\overline{\Psi}_1$ must be the same sub-space. A similar argument applies when the domain D is open. ■

Let Ψ_l denote the complement of the sub-space Ψ_{nl}. The sub-spaces Ψ_l and Ψ_{nl} are *structural invariants* which capture the linear and nonlinear dependencies of the function; specifically, the function can always be decomposed into linear and nonlinear components with $\Psi_{nl}\cap$D and $\Psi_l\cap$D the domains of, respectively, the linear and nonlinear components (with, of course, $(\Psi_{nl}\cap \mathrm{D}) \cup (\Psi_l\cap \mathrm{D})=\mathrm{D}$).

2.1 Remarks

(i) Equivalent minimality condtions

The minimality condition in the above proposition is that that there exists no vector $\mathbf{v}\in \mathrm{V}_o=\{\mathbf{v}\in\Re^{n+m}: \mathbf{Mv}\neq 0\}$ for which $\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0\ \forall \mathrm{z}\in \mathrm{D}$. First of all, since $\mathbf{H_F}(\mathbf{z})\mathbf{v}$ is linear in $\mathbf{v}$, it is sufficient to confine consideration to the class of normalised vectors $\mathbf{v}\in\{\mathbf{v}\in\Re^{n+m}: \mathbf{Mv}\neq 0, \mathbf{v}^T\mathbf{v}=1\}$. It is this linearity, moreover, which focuses attention on the sub-spaces Ψ_l and Ψ_{nl} as the quantities generally of interest rather than the regions $\Psi_{nl}\cap$D and $\Psi_l\cap$D. Secondly, let N denote the intersection of the null spaces of the Hessian maps, $\mathbf{H_F}(\mathbf{z})$ as $\mathbf{z}$ ranges over D. Hence, the minimality condition $\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0\ \forall \mathbf{z}\in \mathrm{D}, \mathbf{v}\in\Re^{n+m}:\mathbf{Mv}=0$ is equivalent to the requirement that N is the null space of $\mathbf{M}$.

(ii) Non-uniqueness of minimal realisations

While the linear and nonlinear sub-spaces Ψ_l and Ψ_{nl} are unique, there exist many possible *realisations* of minimal decompositions. This arises from the freedom which exists (a) in the decomposition of the nonlinearity into functions $\mathbf{f}(\bullet)$ and $\mathbf{Mz}$ and (b) in the assignment of the linear component of the dynamics between the $\mathbf{A}$ matrix and the mapping $\mathbf{f}$. With regard to (a), since a non-singular linear transformation applied to $\mathbf{M}$ can be absorbed into the nonlinear function, the mapping $\mathbf{f}:\mathbf{z}\in D\rightarrow R$ embodied by a nonlinear function $\mathbf{f}(\mathbf{Mz})$ can be realised by any function $\mathbf{f_T}(\mathbf{M_T z})$ with $\mathbf{M_T}=\mathbf{TM}$ and such that $\mathbf{f_T}=\mathbf{f}\circ\mathbf{T}^{-1}$. It is straightforward to verify that $\mathbf{H_F}$ is invariant with respect to such transformations. One consequence of this non-uniqueness is a shift in emphasis, evident also in the previous analysis, away from a particular choice of basis $\mathbf{M}$ for this sub-space and towards more geometric viewpoint. With regard to (b), consider two minimal decompositions

$$\mathbf{F}(\mathbf{z}) = \mathbf{A}_o\mathbf{z}+\mathbf{f}_o(\mathbf{Mz}) \tag{6}$$

and

$$\mathbf{F}(\mathbf{z}) = \mathbf{A}_1\mathbf{z}+\mathbf{f}_1(\mathbf{Mz}) \tag{7}$$

Provided

$$\left(\mathbf{A}_o - \mathbf{A}_1\right) = \mathbf{XM} \tag{8}$$

then the domains of $\mathbf{f_0}$, $\mathbf{f_1}$ are the intersection with D of the subspace spanned by $\mathbf{M}$ as required yet the mappings $\mathbf{f_o}:\mathbf{z}\in D\rightarrow R$, $\mathbf{f_1}:\mathbf{z}\in D\rightarrow R$ may differ by a linear term. However, this is an essentially trivial non-uniqueness which may be removed by, for example, calibrating $\mathbf{f}(\boldsymbol{\rho}_o)$ to be zero for some suitable value of $\boldsymbol{\rho}$, say $\boldsymbol{\rho}_o$.

(iii) Relationship to Regularisation.

A certain complementarity can be observed between the methods discussed here and the regularisation techniques which are widely employed in nonlinear curve fitting and regression analysis. Linearity is usually defined in terms of satisfying superposition, with affine maps defined as translates of the linear maps. Nevertheless, the class of affine scalar maps can also be defined in terms of the solutions to the ordinary differential equation, $\frac{d^2 f(z)}{dz^2} = 0\ \forall z \in \Re$. The corresponding vector generalisation is $\mathbf{H_F}(\mathbf{z})=0\ \forall\mathbf{z}\in \Re^n\times\Re^m$, where $\mathbf{H_F}$ is the Hessian defined by (4). In regularisation theory, a penalty term involving $\mathbf{H_F}(\mathbf{z})$ is commonly included in the objective function measuring the "goodness of fit" achieved so as to penalise any nonlinearity of the fitted function while leaving the cost of linear/affine terms unchanged. Conversely, the present objective is not fitting but the analysis of a function to infer its decomposition into linear and nonlinear components. A further important aspect here is that we only seek *sub-spaces* on which linearity/affinity holds, corresponding to determining directions, $\mathbf{v}$, in which $\mathbf{H_F}(\mathbf{z})\mathbf{v}$ is zero rather than seeking to uniformly minimise all elements of $\mathbf{H_F}(\mathbf{z})$.

(iv) Interpretation in terms of Velocity-based Linearisations.

The tangent map corresponding to an equilibrium operating point is associated with the classical series expansion linearisation and provides rich information concerning the dynamic characteristics in the vicinity of the specific equilibrium point considered. Indeed, because it enables the considerable wealth of methods developed for the analysis of linear systems to be brought to bear on the nonlinear analysis task, it is standard engineering practice to investigate the dynamic behaviour of a nonlinear system, at least initially, by studying the dynamic characteristics of representative equilibrium linearisations. The limitations of classical equilibrium linearisations are, however, well known. In particular, they provide little information regarding the dynamics during transitions between operating points or during operation far from equilibrium. These limitations are directly addressed by the recently developed velocity-based (VB) linearisation framework (Leith & Leithead 1998a,b, 1999) which utilises tangent map information for *every* operating point, not just equilibrium points. When $\mathbf{D}$ denotes the continuous-time d/dt operator, an alternative representation of the nonlinear system (1), obtained by differentiating , is

$$\dot{\mathbf{x}}=\mathbf{w} \tag{9}$$

$$\dot{\mathbf{w}} = \nabla_{\mathbf{x}}\mathbf{F}(\mathbf{x},\mathbf{r})\,\mathbf{w} + \nabla_{\mathbf{r}}\mathbf{F}(\mathbf{x},\mathbf{r})\,\dot{\mathbf{r}} \tag{10}$$

(note the minor change in notation here to accord with that of Leith & Leithead 1998a,b, 1999). The relationship between (9)-(10) and (1) is evidently direct and, furthermore, extends rather more deeply than might initially be expected. Consider the linear system, obtained by "freezing" (9)-(10) at an operating point $(\mathbf{x_1},\mathbf{r_1})$,

$$\dot{\hat{\mathbf{x}}} = \hat{\mathbf{w}} \tag{11}$$

$$\dot{\hat{\mathbf{w}}} = \nabla_{\mathbf{x}}\mathbf{F}(\mathbf{x_1},\mathbf{r_1})\,\hat{\mathbf{w}} + \nabla_{\mathbf{r}}\mathbf{F}(\mathbf{x_1},\mathbf{r_1})\,\dot{\mathbf{r}} \tag{12}$$

The system (11)-(12) is referred to as the velocity-based (VB) linearisation of (1) associated with the operating point $(\mathbf{x_1},\mathbf{r_1})$. Evidently, the coefficients, $\nabla_{\mathbf{x}}\mathbf{F}(\mathbf{x_1},\mathbf{r_1})$ and $\nabla_{\mathbf{r}}\mathbf{F}(\mathbf{x_1},\mathbf{r_1})$, of the linearisations are simply the appropriate elements of the tangent map associated with the operating point ($\mathbf{x_1},\mathbf{r_1}$) (which, it is emphasised, need not be an equilibrium point). The linear system (11)-(12) has a direct interpretation in relation to the nonlinear system (1); namely, the solution to (11)-(12) is an accurate approximation to the solution of (1) locally to the operating point ($\mathbf{x}_1,\mathbf{r}_1$) (Leith & Leithead 1998a). Furthermore, while the solution to an individual VB linearisation is only a locally accurate approximation, there exists a VB linearisation, (11)-(12), for every operating point ($\mathbf{x},\mathbf{r}$) and thus a VB linearisation family, with members defined by (11)-(12), can be associated with the nonlinear system, (1). The solutions to the members of the family of VB linearisations may be pieced together to approximate the solution to the nonlinear system (1) to an arbitrary degree of accuracy (Leith & Leithead 1998a).

With regard to the present structure identification context, it seems natural to expect that linearity of the dynamics with respect to some function or combination of $\mathbf{x}$ and $\mathbf{r}$ might manifest itself in terms of a simplification of the structure of the associated VB linearisation family and indeed this turns out to be the case. Reformulating (1) as a minimal decomposition, $\mathbf{Dx} = \mathbf{A}\begin{bmatrix}\mathbf{x}\\ \mathbf{r}\end{bmatrix} + \mathbf{f}(\mathbf{M}\begin{bmatrix}\mathbf{x}\\ \mathbf{r}\end{bmatrix})$, the VB linearisation, (11)-(12), becomes

$$\dot{\hat{\mathbf{x}}} = \hat{\mathbf{w}} \tag{13}$$

$$\dot{\hat{\mathbf{w}}} = \left(\mathbf{A} + \nabla\mathbf{f}(\boldsymbol{\rho}_1)\mathbf{M}\right)\begin{bmatrix}\hat{\mathbf{w}}\\ \dot{\mathbf{r}}\end{bmatrix} \tag{14}$$

Evidently, the quantity, $\boldsymbol{\rho}$, embodying the nonlinear dependence of the dynamics, also serves to parameterise the VB linearisation family and indeed it is this observation which originally motivated much of the development in the present paper.

The structure of the VB linearisation family implies that t he VB linearisations are identical at all points, ($\mathbf{x}_1,\mathbf{r}_1$), such that $\boldsymbol{\rho}(\mathbf{x}_1,\mathbf{r}_1)$ has the same value; that is, at all points lying on a surface of constant $\boldsymbol{\rho}$. The derivative of a vector or matrix function, $\boldsymbol{\Lambda}(\mathbf{z})$, in direction $\mathbf{v}$ is defined as $\lim_{h\to o}\frac{\boldsymbol{\Lambda}(\mathbf{z}+h\mathbf{v})-\boldsymbol{\Lambda}(\mathbf{z})}{h}$. It follows from the foregoing discussion that, at any point ($\mathbf{x}_1,\mathbf{r}_1$) the directional derivative of the matrix $\mathbf{A}+\nabla\mathbf{f}(\boldsymbol{\rho})\mathbf{M}$ *vanishes* in directions aligned with the surface of constant $\boldsymbol{\rho}$ passing through that point. Equivalently, since the directional derivative can be expressed (stacking the elements) as $\mathbf{H_F}(\mathbf{z}_1)\mathbf{v}$, where $\mathbf{H_F}(\mathbf{z}_1)$ is the Hessian map (4) and $\mathbf{z}_1$ is ($\mathbf{x}_1,\mathbf{r}_1$), the null sub-space of $\mathbf{H_F}$ ($\mathbf{z}_1$) is precisely the surface of constant $\boldsymbol{\rho}$ passing through the point $\mathbf{z}_1$ (*i.e.* ($\mathbf{x}_1,\mathbf{r}_1$)). The complement of this sub-space thus embodies the nonlinear dependence of the dynamics.

3. Nonlinear Structure Identification

The foregoing analysis is deterministic in nature and, since measured data can, in general, be expected to be noisy it is necessary to extend the analysis to include probabilistic considerations.

3.1 Probabilistic Structural Decomposition

Consider the output y of a stochastic process, the pdf of which is conditional on explanatory variable $\mathbf{z}\in\Re^n$. To avoid cumbersome notation, assume that y is scalar (the generalisation of the results which follow to processes with multiple outputs is straightforward). Let $\mu(\mathbf{z})$ denote the mean of the process (*i.e.* $\mu(\mathbf{z})$=E(y($\mathbf{z}$))) and assume that $\mu(\mathbf{z})$ is differentiable. It is a standard results that the mean of the associated derivative process is

$$E(\frac{\partial y}{\partial \mathbf{z}_i}(\mathbf{z})) = \frac{\partial}{\partial \mathbf{z}_i}E(y(\mathbf{z})) \tag{15}$$

where $\mathbf{z}_i$ denotes the ith element of $\mathbf{z}$; that is, the expected value of the derivative process is just the derivative of the mean of y (assuming this exists). Let Q($\mathbf{z}_o,\mathbf{z}_1$) denote the covariance of y (*i.e.* Q($\mathbf{z}_o,\mathbf{z}_1$)=E(y($\mathbf{z}_o$)y($\mathbf{z}_1$))) and assume that Q is continuously twice differentiable. Then the variance of the derivative process is

$$E(\frac{\partial y}{\partial \mathbf{z}_i}(\mathbf{z}_o)\frac{\partial y}{\partial \mathbf{z}_j}(\mathbf{z}_1)) = \nabla_i^1\nabla_j^2 Q(\mathbf{z}_o,\mathbf{z}_1) \tag{16}$$

where ∇_i^1Q denotes the partial derivative of Q with respect to the ith element of its first argument, *etc*. The mean and variance of the Hessian maps associated with y can be similarly derived by application of (15) and (16) to (4)..

At this point it is perhaps worth emphasising that, as in the case of parametric models, differentiation of smooth non-parametric models is entirely admissible and certainly does not require differentiation of the raw, noisy data. The latter is, of course, highly inadvisable. The model is a *smooth* fit through the measured data and the associated derivative model is well-defined. Consequently, once a model has been fitted to measured data, the probability distributions of the tangent and Hessian maps associated with the nonlinear dynamics are also immediately available. The results of section 2 can now be directly generalised to the probabilistic context. In particular, while the probability distribution of the minimal nonlinear subspace may be summarised/visualised in many ways, a useful metric is provided by the following definition.

Definition (MAP, or maximum *a posteriori*, minimal nonlinear subspace). Ψ_{nl}^{MAP} is defined as the subspace spanned by the matrix **M** for which the posterior probability is greatest that there exists no vector $\mathbf{v} \in \{\mathbf{v} \in \Re^{n+m}: \mathbf{Mv} \neq 0, \mathbf{v}^T\mathbf{v}=1\}$ for which $\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0 \ \forall \mathbf{z} \in D$.

3.3 Estimation of Nonlinear Dependence in a Region

The task considered in this section is the inference of Ψ_{nl}^{MAP}. Assuming that the dimension, q, of the minimal nonlinear subspace is known, by straightforward application of Bayes' Rule, the posterior probability distribution is

$$p(\Psi_{nl}|\Pi) = \frac{p(\Pi|\Psi_{nl})p(\Psi_{nl})}{p(\Pi)} \tag{17}$$

where Π is the information at $\mathbf{z_1}$ provided by the non-parametric model embodying the measured data. The denominator, $p(\Pi)$, is invariant with respect to Ψ_{nl}. The prior distribution, $p(\Psi_{nl})$, embodies prior knowledge of the likely nonlinear dependency (assigned large variance when little prior knowledge is available). The likelihood $p(\Pi|\Psi_{nl})$, embodying the information contained within the measured data, can be expressed as

$$p(\Pi|\Psi_{nl}) = p\left(\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0 : \mathbf{v} \in \left\{\mathbf{v} : \mathbf{Mv} = 0, \mathbf{v}^T\mathbf{v} = \mathbf{1}\right\}, \mathbf{z} \in D\right) \tag{18}$$

The MAP estimate, Ψ_{nl}^{MAP} is the subspace which minimises the risk

$$J(\Psi_{nl}) = -\log p(\Pi|\Psi_{nl}) - \log p(\Psi_{nl}) \tag{19}$$

(taking the negative log of (17) and neglecting the term involving $p(\Pi)$ since this does not alter the location of the minimum).

Since $\mathbf{v}$ and $\mathbf{z}$ range over a continuum of points, evaluation of (18) may be relatively difficult. However, a useful approximation is

$$\widehat{p(\Pi|\Psi_{nl})} = p\left(\mathbf{H_F}(\mathbf{z}_j)\mathbf{v}_i = 0, i = 1..N_v, j = 1..N_x\right) \tag{20}$$

where $\mathbf{v}_i$, i=1,2,..N $_v$ are N $_v$ representative unit vectors from the null space of **M**, $\mathbf{z}_j$, j=1,2,..N $_x$ are N $_x$ representative operating points. This approximation is readily calculated directly from a Gaussian Process model and under mild continuity conditions, converges to (18) as the number of points used is increased. This can be seen as follows. Let C denote the compact set $\left\{\mathbf{v} : \mathbf{Mv} = 0, \mathbf{v}^T\mathbf{v} = \mathbf{1}\right\}$) and $V_i^{N_v}$, i=1..N$_v$ denote a collection of open sets with $\cup_i V_i^{N_v}$ =C (it follows from the compactness of C that N may be assumed finite) and let $\mathbf{v}_i$ be a point in $V_i^{N_v}$. Similarly, let D be a compact region of $\Re^n \times \Re^m$ and $X_j^{N_x}$, j=1..N$_x$ denote a collection of open sets with $\cup_j X_j^{N_x}$ =D (it follows from the compactness of D that N $_x$ may be assumed finite) and let $\mathbf{z}_j$ be a point in $X_i^{N_x}$ It follows that

$$\begin{aligned} p\left(\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0 : \mathbf{v} \in C_v, \mathbf{z} \in C_x\right) &= p\left(\mathbf{H_F}(\mathbf{z}_j)\mathbf{v}_i = 0, i = 1..N_v, j = 1..N_x \ \& \ \mathbf{H_F}(\mathbf{z})\mathbf{v} = 0, \mathbf{v} \in \cup_i\left(V_i^{N_v} / \{\mathbf{v}_i\}\right), \mathbf{z} \in \cup_j\left(X_j^{N_x} / \{\mathbf{z}_j\}\right)\right) \\ &= p\left(\mathbf{H_F}(\mathbf{z}_j)\mathbf{v}_i = 0, i = 1..N_v, j = 1..N_x\right) p\left(\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0, \mathbf{v} \in \cup_i\left(V_i^{N_v} / \{\mathbf{v}_i\}\right), \mathbf{z} \in \cup_j\left(X_j^{N_x} / \{\mathbf{z}_j\}\right) \middle| \mathbf{H_F}(\mathbf{z}_j)\mathbf{v}_i = 0, i = 1..N_v, j = 1..N_x\right) \end{aligned} \tag{21}$$

Hence, provided the limit exists and

$$\lim_{N_x \to \infty} \lim_{N_v \to \infty} p\left(\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0, \mathbf{v} \in \cup_i\left(V_i^{N_v} / \{\mathbf{v}_i\}\right), \mathbf{z} \in \cup_j\left(X_j^{N_x} / \{\mathbf{z}_j\}\right) \middle| \mathbf{H_F}(\mathbf{z}_j)\mathbf{v}_i = 0, i = 1..N_v, j = 1..N_x\right) = 1 \tag{22}$$

then $p\left(\mathbf{H_F}(\mathbf{z}_j)\mathbf{v}_i = 0, i = 1..N_o, j = 1..N_1\right) \to p\left(\mathbf{H_F}(\mathbf{z})\mathbf{v} = 0, \mathbf{v} \in \mathbf{C}, \mathbf{z} \in D\right)$ as N $_x$, N $_v$ $\to\infty$. The condition (22) imposes a continuity requirement that the values of $\mathbf{H_F}$ are similar in similar directions and at nearby operating points. It should be noted that this is a mild requirement which is, in general, automatically satisfied by the regularisation conditions employed in non-parametric models such as the Gaussian Process approach. The operating region D is required to be compact and so cannot encompass the whole of $\Re^n \times \Re^m$. However, in practice, it seems unrealistic to expect to estimate the global nonlinear dependence over the whole space $\Re^n \times \Re^m$ on the basis of a finite number of measured data points and approximation in terms of a compact operating region is certainly not restrictive.

Even with the simplification provided by the approximation (20), however, the computational complexity of determining the mean and/or MAP estimates using (19) clearly rises quite rapidly as the dimension, q, of the scheduling variable increases. An alternative iterative approach which, by exploiting orthonormal basis' of Ψ_{nl}^{MAP}, scales better is obtained by determining candidate basis vectors for Ψ_{nl}^{MAP} in turn, leading to the following iterative estimation procedure.

Iterative Estimation Procedure

1. Let i=1, $\Psi_{nl}^{i}=\Re^{n+m}$.
2. Determine the most likely unit direction $\bar{\mathbf{v}}_i$, lying within the current estimate, Ψ_{nl}^{i}, of the nonlinear subspace; that is, the direction which minimises

$$J_i(\mathbf{v}_i) = -\log p\left(\mathbf{H}_\mathbf{F}(\mathbf{z})\mathbf{v}_i = 0 \middle| \mathbf{v}_i \in \Psi_{nl}^{i}, \mathbf{v}_i^{\mathbf{T}}\mathbf{v}_i = \mathbf{1}, \mathbf{z} \in D\right) - \log p(\mathbf{v}_i) \tag{23}$$

   Letting the rows of $\mathbf{M}_i$ be an orthonormal basis spanning Ψ_{nl}^{i}, then $\mathbf{v}_i$ may be parameterised as $\boldsymbol{\lambda}\mathbf{M}_i$ and the minimisation of $J_i(\mathbf{v}_i)$ can formulated as an *unconstrained* minimisation in $\boldsymbol{\lambda}$. This optimisation requires the estimation of the elements of vector $\boldsymbol{\lambda}$; that is, only i parameter values are estimated at each iteration.
3. Let $\Psi_{nl}^{i+1} = \Psi_{nl}^{i}/v_i$, where v_i is the subspace spanned by $\bar{\mathbf{v}}_i$. Specifically, let $\mathbf{V}_i$ be a orthonormal basis spanning the null space of $\mathbf{M}_i$ and $\mathbf{V}_{i+1}=[\mathbf{V}_i \ \ \bar{\mathbf{v}}_i]$. Letting $\mathbf{M}_{i+1}$ be an orthonormal basis spanning the null space of $\mathbf{V}_{i+1}$, then $\mathbf{M}_{i+1}$ is an orthonormal basis of Ψ_{nl}^{i+1}. The risk function, (19) (or its approximation (20)), may be evaluated with $\mathbf{M}$ equal to $\mathbf{M}_{i+1}$ as a diagnostic to confirm the validity of the updated sub-space, Ψ_{nl}^{i+1}.
4. If i<n+m then i=i+1, go to 2

Remarks

(i) Dimension of the minimal sub-space.

In step (2), the incremental risk, $J_i(\bar{\mathbf{v}}_i)$, can be expected to abruptly increase when the row 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 and the validity of a choice of dimension, q, can be further assessed/confirmed using the pointwise estimation methods discussed in section 4 below.

(ii) Simplified procedure applicable when the Hessian is approximately Gaussian with diagonal covariance.

As noted in section 2, the minimal nonlinear subspace is just the complement of N, where N is the intersection of the null spaces of the Hessian maps, $\mathbf{H}_\mathbf{F}(\mathbf{z})$ as z ranges over D. In other words, the minimal nonlinear subspace is the complement of the null space of $\mathbf{H}_\mathbf{F}(\mathbf{z})$. Of course, the situation is not so straightforward in the case of noisy data. Matrices are generically full rank and so under noisy conditions the null space of the Hessian map associated with each operating point z will almost always consist simply of the zero vector. Instead, the requirement must be to determine the largest sub-space within with 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). More formally, Ψ_{nl}^{MAP} is the complement of the null space of the most probable Hessian map having null space of dimension n+m-q.

When the probability distribution of the Hessian, $\mathbf{H}_\mathbf{F}(\mathbf{z})$, is normal (or can be approximated by a normal distribution),

$$\begin{aligned} J(\Psi_{nl}(\mathbf{z})) &= -\log p\left(\mathbf{H}_\mathbf{F}(\mathbf{z})\mathbf{v} = 0 : \mathbf{v} \in \left\{\mathbf{v}: \mathbf{M}\mathbf{v} = 0, \mathbf{v}^\mathbf{T}\mathbf{v} = \mathbf{1}\right\}, \mathbf{z} \in D\right) - \log p(\Psi_{nl}) \\ &\cong -\log p\left(\mathbf{H}_\mathbf{F}(\mathbf{z}_j)\mathbf{v} = 0 : \mathbf{v} \in \left\{\mathbf{v}: \mathbf{M}\mathbf{v} = 0, \mathbf{v}^\mathbf{T}\mathbf{v} = \mathbf{1}\right\}, j = 1, \ldots N_x\right) - \log p(\Psi_{nl}) \\ &\propto E\left(\mathbf{H}_\mathbf{F}^{N_x}\mathbf{v}\right)^T \boldsymbol{\Lambda}^{-1}(\mathbf{v}) E\left(\mathbf{H}_\mathbf{F}^{N_x}\mathbf{v}\right) + \log|\boldsymbol{\Lambda}| - \log p(\Psi_{nl}) \end{aligned} \tag{24}$$

where $\mathbf{H}_\mathbf{F}^{N_x}$ is the stacked matrix $\left[\mathbf{H}_\mathbf{F}(\mathbf{z}_1)^T \ \cdots \ \mathbf{H}_\mathbf{F}(\mathbf{z}_{N_x})^T\right]^T$ and $\boldsymbol{\Lambda}(\mathbf{v})$ is the covariance of $\mathbf{H}_\mathbf{F}^{N_x}\mathbf{v}$. In general, $\boldsymbol{\Lambda}(\mathbf{v})$ depends on $\mathbf{v}$. However, when the covariance of the Hessian map is diagonal, straightforward algebraic manipulation confirms that $\boldsymbol{\Lambda}$ is invariant with respect to $\mathbf{v}$ (and also diagonal). Since $\boldsymbol{\Lambda}$ is, by definition, positive definite, it can be decomposed as $\mathbf{R}^T\mathbf{R}$ yielding

$$J(\Psi_{nl}) \propto \mathbf{v}^T\mathbf{W}^\mathbf{T}\mathbf{W}\mathbf{v} + \log|\boldsymbol{\Lambda}| - \log p(\Psi_{nl}(\mathbf{x}_1, \mathbf{r}_1)) \tag{25}$$

with $\mathbf{W} = \mathbf{R}\, E\left(\mathbf{H}_\mathbf{F}^{N_x}\right)$. Since log | $\boldsymbol{\Lambda}$| is constant it does not affect the minima of J(Ψ_{nl}). When logp(Ψ_{nl}) is sufficiently small that it can be neglected (*i.e*. there is insignificant prior knowledge, corresponding to the maximum likelihood situation) then

$$J(\Psi_{nl}) \propto \mathbf{v}^T\mathbf{W}^\mathbf{T}\mathbf{W}\mathbf{v} \tag{26}$$

The minimum of the risk function, (26), under the linear constraint that $\mathbf{v} \in \left\{\mathbf{v} \in \Re^q : \mathbf{M}\mathbf{v} = 0, \mathbf{v}^\mathbf{T}\mathbf{v} = \mathbf{1}\right\}$ 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 (26) is minimised with $\mathbf{v} \in \left\{\mathbf{v} \in \Re^q : \mathbf{M}\mathbf{v} = 0, \mathbf{v}^\mathbf{T}\mathbf{v} = \mathbf{1}\right\}$ when $\mathbf{M}=\mathbf{U}_q$, where $\mathbf{U}_q$ is the matrix consisting of the first n+m-q rows of $\mathbf{U}$. When the foregoing conditions are satisfied, the subspace spanned by $\mathbf{U}_q$ is precisely Ψ_{nl}^{MAP} and can be calculated very efficiently. More generally, this value can be used to initialise the optimisation in the iterative procedure above, although experience suggests that, for many purposes, $\mathbf{U}_q$ is in fact a sufficiently accurate estimate in its own right with the refinement provided by further optimisation often of relatively minor significance.

3.4 Examples

As noted previously, non-parametric approaches are well suited to initial data analysis and exploration and are a key enabling technology of the approach proposed here. This is due to their ability to model the data well with few structural assumptions, particularly with regard to the nonlinear dependence. All of the examples studied in the present paper make use of non-parametric Gaussian process prior models, a Bayesian form of kernel regression model, but other non-parametric representations could also be used (*e.g.* support vector machines, locally weighted regression). For simplicity, the explanatory variables are assumed to be noise-free; that is, uncertainty is confined to the output of the nonlinear map.

(i) Consider the nonlinear dynamic system

$$y(t_{n+1}) = 0.5G(\rho(t_n)) \tag{27}$$

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: data is collected for 20 seconds with a sampling interval of 0.1 seconds (200 data points). The measured data, together with the corresponding predicted mean 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 the risk function as the dimension of the nonlinear sub-space is reduced is shown in figure 1b. It can be seen that, as expected, the risk 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 [0.706 -0.709]; that is, ρ is estimated to be 0.706r-0.709y. Subject to an arbitrary normalisation factor, it is evident that the identification procedure successfully infers the nonlinear dependence of the plant dynamics. This is, of course, a simple example selected to have 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 map relating $(r(t_n),y(t_n))$ to $y(t_{n+1})$; for example, an RBF model 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 map 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) *Wiener-Hammerstein System*

Consider the Wiener-Hammerstein nonlinear system illustrated in figure 2a. Reformulating the dynamics in terms of the measured variables (the input, r, and output, y) yields

$$y(t_n) = 0.3\boldsymbol{\rho}_1^2 + 0.165\boldsymbol{\rho}_2^2 \tag{28}$$

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

$$\mathbf{M} = \begin{bmatrix} 0.9184 & 0.3674 & 0 & 0 \\ 0 & 0 & 0.9184 & 0.3674 \end{bmatrix} \tag{29}$$

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 risk function 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

$$\hat{\mathbf{M}} = \begin{bmatrix} 0.9292 & 0.3694 & -0.0008 & 0.0018 \\ -0.0015 & 0.0040 & 0.9282 & 0.3719 \end{bmatrix} \tag{30}$$

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 map).

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

$$\begin{aligned} x_1 &= \left(a_n q^{-n} + \ldots + a_o\right) r \\ x_2 &= f(x_1) \\ y &= \left(b_m q^{-m} + \ldots + b_o\right) x_2 \end{aligned} \tag{31}$$

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

$$y = b_m f(\boldsymbol{\rho}_{m+1}) + \ldots + b_o f(\boldsymbol{\rho}_1) \tag{32}$$

where

$$\boldsymbol{\rho} = \begin{bmatrix} a_n & \cdots & a_o & 0 & & \cdots & 0 \\ 0 & a_n & \cdots & a_o & 0 & \cdots & 0 \\ & & & \ddots & & & \\ 0 & & \cdots & 0 & a_n & \cdots & a_o \end{bmatrix} \begin{bmatrix} q^{-n-m} r \\ q^{-n-m-1} r \\ \vdots \\ r \end{bmatrix} \tag{33}$$

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 (33) 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. Validating Nonlinear Dependence in a Region

The foregoing methods developed for the estimation of the nonlinear dependence in a region can also be immediately applied to estimate the nonlinear dependence locally to a single operating point. By studying the local nonlinear dependence at a number of points drawn from the operating region of interest, the validity of the MAP estimate, Ψ_{nl}^{MAP}, of the minimal nonlinear subspace in the region can be assessed in a fairly direct manner. Specifically, for any function (3) we have that

(i) $\dim \mathrm{null}(\mathbf{H_F}(\mathbf{z})) \geq \dim \Psi_l$

(ii) $\bigcap_{\mathbf{z}\in D} \mathrm{null}(\mathbf{H_F}(\mathbf{z})) = \Psi_l$

where null($\mathbf{H}$) denotes the null sub-space of matrix $\mathbf{H}$ and, as before, ψ_l denotes the complement of the nonlinear subspace ψ_{nl}. The dimension of the null sub-space of the Hessian $\mathbf{H_F}(\mathbf{z})$ is greater than that of ψ_l only at points where $\mathbf{H_f}$ (*i.e.* $\nabla(\nabla\mathbf{f})^T$) vanishes. Typically (but not always), these points form a set of measure zero and almost everywhere dim null($\mathbf{H_F}(\mathbf{z})$) is uniformly equal to dim ψ_l with null($\mathbf{H_F}(\mathbf{z})$) necessarily equal to ψ_l. Consequently, good agreement between the local nonlinear dependencies and Ψ_{nl}^{MAP} provides a degree of confidence that the nonlinear dependence is well summarised by Ψ_{nl}^{MAP}. Conversely, if, for example, it appears that the operation region can be decomposed into sub-regions each exhibiting consistently different local nonlinear dependence, this might indicate limitations in the use of Ψ_{nl}^{MAP} as a summary of the nonlinear dependence over the region.

4.1 Estimation of Local Nonlinear Dependence

Let $\Psi_{nl}^{MAP}(\mathbf{z}_1)$ denote the subspace spanned by the matrix $\mathbf{M}$ for which the posterior probability is greatest that there exists no vector $\mathbf{v}\in\{\mathbf{v}\in\Re^{n+m}: \mathbf{M}\mathbf{v}\neq 0, \mathbf{v}^T\mathbf{v}=1\}$ for which $\mathbf{H_F}(\mathbf{z}_1)\mathbf{v} = 0$. This is just a pointwise version of Ψ_{nl}^{MAP}, the MAP minimal nonlinear subspace in a region. Assuming that the dimension, q, of the minimal nonlinear subspace is known, then following a similar approach to that used in section 3 it follows that the MAP estimate, $\Psi_{nl}^{MAP}(\mathbf{z}_1)$ is the subspace which minimises the risk

$$J(\Psi_{nl}(\mathbf{z}_1)) = -\log p(\Pi|\Psi_{nl}(\mathbf{z}_1)) - \log p(\Psi_{nl}(\mathbf{z}_1)) \tag{34}$$

where Π is the information at $\mathbf{z_1}$ provided by the non-parametric model embodying the measured data, the prior distribution, $p(\Psi_{nl}(\mathbf{z}_1))$, embodies prior knowledge and $p(\Pi|\Psi_{nl}(\mathbf{z}_1))$ is the likelihood. This can be expressed as

$$p(\Pi|\Psi_{nl}(\mathbf{z}_1)) = p\left(\mathbf{H_F}(\mathbf{z}_1)\mathbf{v} = 0 : \mathbf{v}\in\left\{\mathbf{v}:\mathbf{M}\mathbf{v}=0, \mathbf{v}^T\mathbf{v}=\mathbf{1}\right\}\right) \tag{35}$$

While $\mathbf{v}$ generally ranges over a continuum of points in the subspace defined by $\mathbf{M}$, a useful approximation is

$$\widehat{p(\Pi|\Psi_{nl}(\mathbf{z}_1))} = p\left(\mathbf{H_F}(\mathbf{z}_1)\mathbf{v}_i = 0, i=1..N\right) \tag{36}$$

where $\mathbf{v}_i$, i=1,2,..N_i are N representative unit vectors from the null space of $\mathbf{M}$. This approximation is readily calculated directly from a Gaussian Process model and, following analysis precisely analogous to that in section 3 it can be shown that, under mild continuity conditions, converges to (35) as the number of points used is increased. Similarly to section 3, an efficient iterative estimation procedure can be derived for estimating $\Psi_{nl}^{MAP}(\mathbf{z_1})$.

Iterative Estimation Procedure

1. Let i=1, $\Psi_{nl}^{i}=\Re^{n+m}$.
2. Determine the most likely unit direction $\bar{\mathbf{v}}_i$, lying within the current estimate, Ψ_{nl}^{i}, of the nonlinear subspace; that is, the direction which minimises

$$J_i(\mathbf{v}_i) = -\log p\left(\mathbf{H_F}(\mathbf{z}_1)\mathbf{v}_i = 0 \middle| \mathbf{v}_i \in \Psi_{nl}^i, \mathbf{v}_i^T\mathbf{v}_i = \mathbf{1}\right) - \log p(\mathbf{v}_i) \quad (37)$$

3. Let $\Psi_{nl}^{i+1} = \Psi_{nl}^{i}$-$v_i$, where v_i is the subspace spanned by $\bar{\mathbf{v}}_i$.
4. If i<n+m then i=i+1, go to 2

Remarks

(i) *Dimension of the minimal sub-space*.

In step (2), the incremental risk, $J_i(\bar{\mathbf{v}}_i)$, can be expected to abruptly increase when the row 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.

(ii) *Simplified procedure applicable when the Hessian is approximately Gaussian with diagonal covariance*.

Specialising the minimality proposition in section 2 to the pointwise case, t he minimal nonlinear subspace is just the complement of the null space of $\mathbf{H_F}(\mathbf{z}_1)$. That is, in probabilistic terms, $\Psi_{nl}^{MAP}(\mathbf{z_1})$ is the complement of the null space of the most probable Hessian map having null space of dimension n+m-q. Similarly to the analysis in section 3.2, when the Hessian map when the probability distribution of the Hessian map, $\mathbf{H_F}(\mathbf{z}_1)$, is normal (or can be approximated by a normal distribution) with diagonal covariance, $\Psi_{nl}^{MAP}(\mathbf{z_1})$ is the subspace spanned by $\mathbf{U}_q$ where $\mathbf{U}_q$ is the matrix consisting of the first n+m-q rows of $\mathbf{U}$ with $\mathbf{W}=\mathbf{U}^T\mathbf{\Sigma U}$ (the singular value decomposition of $\mathbf{W}$) and $\mathbf{W} = \mathbf{R}\,E(\mathbf{H_F}(\mathbf{z}_1))$, $\mathbf{R}^T\mathbf{R}=\mathbf{\Lambda}$, the covariance of $\mathbf{H_F}(\mathbf{z}_1)\mathbf{v}$. More generally, this SVD approach can be used to quickly obtain an estimate of $\Psi_{nl}^{MAP}(\mathbf{z_1})$ which can be used to initialise the optimisation in the iterative procedure above, although experience suggests that, for many purposes, this estimate is in fact a sufficiently accurate estimate in its own right with the refinement provided by further optimisation often of relatively minor significance.

(iii) It is important to note that Ψ_{nl}^{MAP} is not equivalent to the mean of the pointwise sub-spaces $\Psi_{nl}^{MAP}(\mathbf{z_1})$ over the operating region, C.

4.2 Example

(i) Returning to the system, (27), considered in Example (i) above, figure 3a shows the variation in pointwise risk function vs dimension of nonlinear sub-space at 30 operating points selected uniformly from the operating space covered by the measured data. It can be seen that, in accordance with the previous results, the risk uniformly rises abruptly when the dimension falls below unity. The corresponding estimates of $\mathbf{M}$, a basis for the minimal nonlinear sub-space estimated at each point are shown in figure 3b. 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 (27) but with ρ satisfying

$$\rho = r - \sin(ay)/a \quad (38)$$

with 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 operating 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. Applying the techniques developed in section 3, and using the operating region considered in Example (i), the variation in the cost function J with the dimension of the nonlinear sub-space is shown in Figure 4a. The increase in risk when the dimension is reduced from 2 to 1 is somewhat greater than in Example (i), as might be expected in view of the additional nonlinear dependence in (38). The corresponding estimate of the basis, $\mathbf{M}$, 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 an operating region close to the origin, the corresponding estimate of $\mathbf{M}$ becomes [0.708 -0.703]. The latter agrees well with the results for Example (i), as expected. However, the results for the larger operating 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 4b. 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 operating space. Measurement noise generally results in *uncorrelated* variations in the pointwise estimates across the operating space, while a strong spatial correlation is evident between the estimates in Figure 4b. This structure is visually quite striking, particularly when compared with the corresponding plot, Figure 3b, for the system in Example (i). In the vicinity of the line y=0, the pointwise estimates of $\mathbf{M}$ 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 in (38) 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 4c. This is in accordance with the fact that $\sin(ay)/a \rightarrow y$ as $a\rightarrow 0$ and thus $\rho\rightarrow r-y$ 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

The is paper investigates new ways of inferring nonlinear dependence parsimoniously from measured data. The existence of unique linear and nonlinear sub-spaces which are structural invariants of general nonlinear maps is established for the first time. Necessary and sufficient conditions determining these sub-spaces are derived. In addition to being of considerable interest in their own right, 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

D.J. Leith gratefully acknowledges the generous support provided by the Royal Society for the work presented.

References

CHEN, S., COWAN,C.F., GRANT,P., 1991, Orthogonal Least Squares Learning Algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*.

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).

HU,J., KUMAMARU,K., HIRASAWA,K., 2000, A quasi-ARMAX approach to modelling nonlinear systems. *International Journal of Control*, in press.

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

JOHANSEN,T.A.,FOSS,B.A., 1993, Constructing NARMAX models using ARMAX models. *Int. J. Control*, **58**, 1125-1153.

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* (Eds. Murray-Smith, R, Johansen, T.A.) (Taylor & Francis).

JUDITSKY,A.,HJALMARSSON,H,BENVENISTE,A.,DEYLON,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. *International Journal of Control*, **48**, 193-210.

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

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

LEITH, D.J., LEITHEAD, W.E., 1999, Input-Output Linearisation by Velocity-Based Gain-Scheduling. *International Journal of Control*, **72**, 229-246.

LEITH, D.J., LEITHEAD, W.E, 2000, Survey of Gain-Scheduling Analysis and Design. *International Journal of Control* , **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).

NEAL, R, 1997 Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *Technical Report 9702*. Department of Statistics, University of Toronto.
O'HAGAN, A., 1978, On curve fitting and optimal design for regression , *Journal of the Royal Statistical Society B* , **40**, 1-42.
SHAMMA, J.S., ATHANS, M., 1990, Analysis of Gain Scheduled Control for Nonlinear Plants. *IEEE Transactions on Automatic Control*, **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'. *International Journal of Control*, in press.

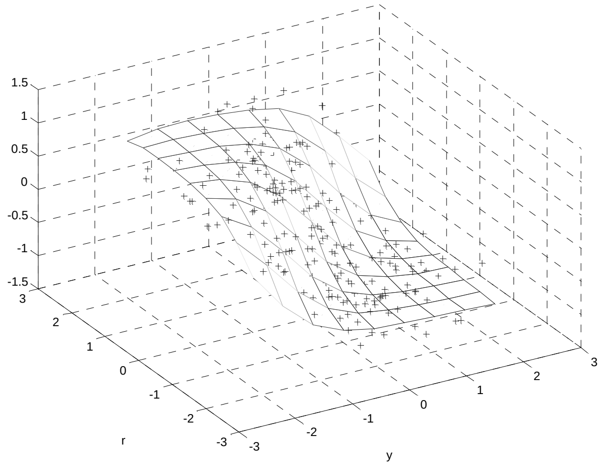

Figure 1a Measured data (+) and associated Gaussian Process model in Example (i) of section 3.4.

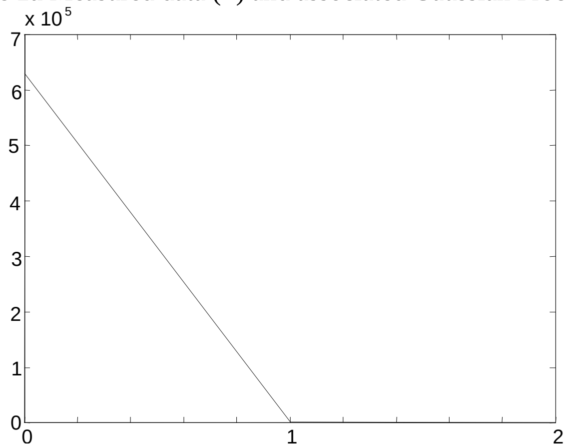

Figure 1b Risk function, J, vs dimension of ρ in Example (i) of section 3.4.

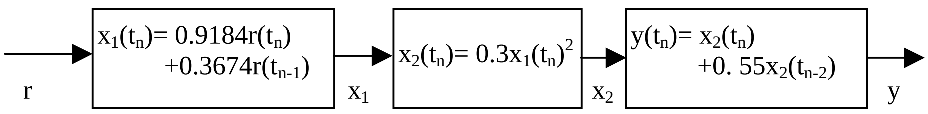

Figure 2 Block diagram representation of system studied in Example (ii) of section 3.4.

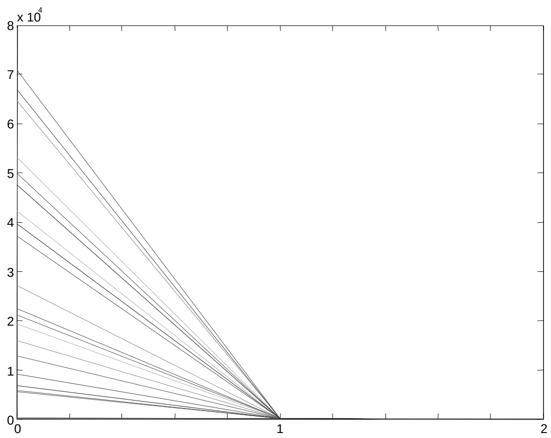

Figure 3a Pointwise risk function, J, vs dimension of ρ in Example (i) of section 3.4.

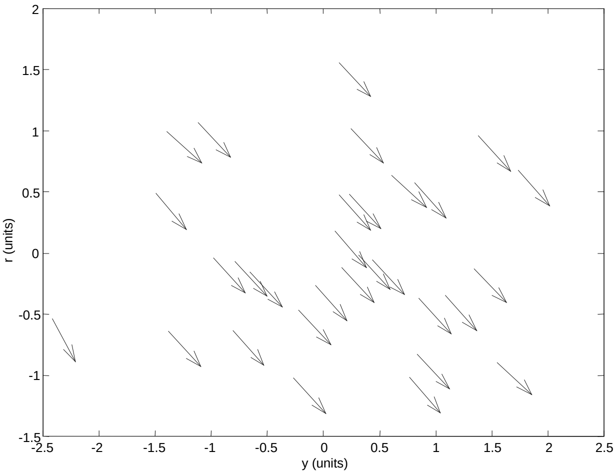

Figure 3b Point estimates of **M**, the mapping relating the scheduling variable to the state and input (Example (i) of section 3.4).

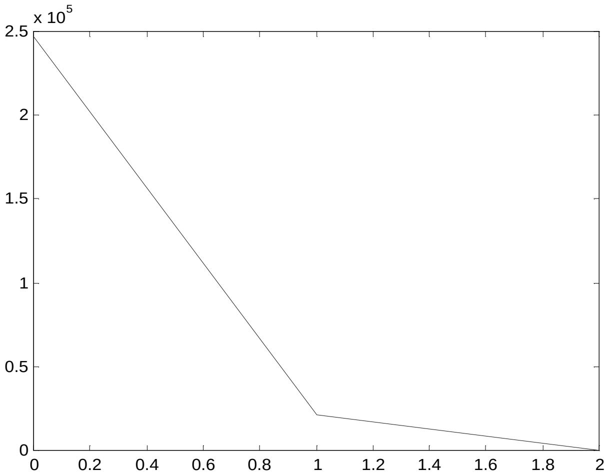

Figure 4a Risk function, J, vs dimension of ρ in Example (ii) of section 4.2 with *a*=1.

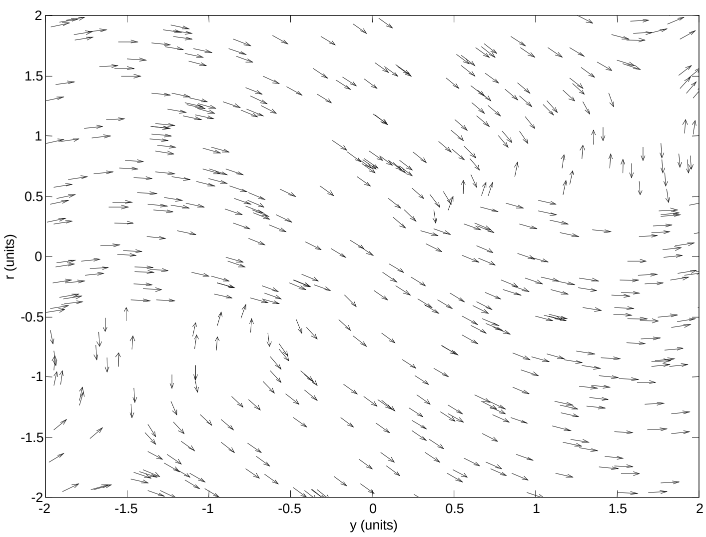

Figure 4b Point estimates of **M**, the mapping relating the scheduling variable to the state and input in Example (ii) of section 4.2 with *a*=1.

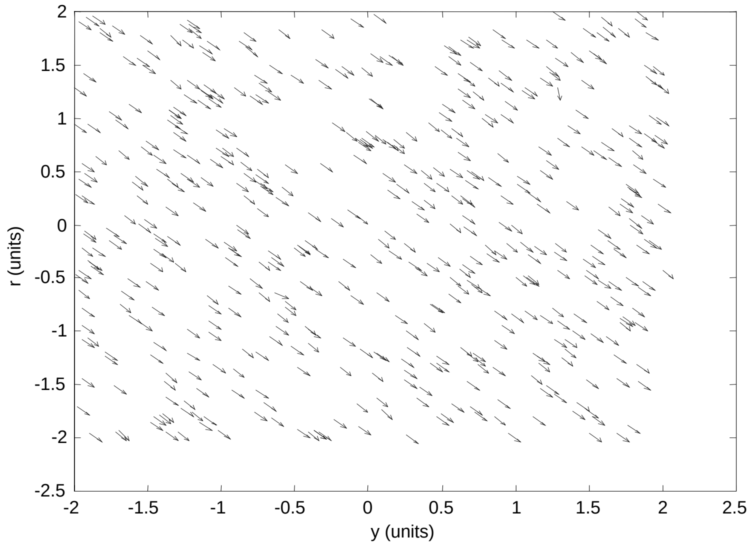

Figure 4c Point estimates of **M**, the mapping relating the scheduling variable to the state and input in Example (ii) of section 4.2 with *a*=0.1.