LETTER Communicated by Maxwell Stinchcombe

Modeling Multivariate Time Series on Manifolds with Skew Radial Basis Functions

Arta A. Jamshidi
jamshidi@math.colostate.edu
Michael J. Kirby
kirby@math.colostate.edu
Department of Mathematics, Colorado State University, Fort Collins, CO 80523, U.S.A.

We present an approach for constructing nonlinear empirical mappings from high-dimensional domains to multivariate ranges. We employ radial basis functions and skew radial basis functions for constructing a model using data that are potentially scattered or sparse. The algorithm progresses iteratively, adding a new function at each step to refine the model. The placement of the functions is driven by a statistical hypothesis test that accounts for correlation in the multivariate range variables. The test is applied on training and validation data and reveals nonstatistical or geometric structure when it fails. At each step, the added function is fit to data contained in a spatiotemporally defined local region to determine the parameters—in particular, the scale of the local model. The scale of the function is determined by the zero crossings of the autocorrelation function of the residuals. The model parameters and the number of basis functions are determined automatically from the given data, and there is no need to initialize any ad hoc parameters save for the selection of the skew radial basis functions. Compactly supported skew radial basis functions are employed to improve model accuracy, order, and convergence properties. The extension of the algorithm to higher-dimensional ranges produces reduced-order models by exploiting the existence of correlation in the range variable data. Structure is tested not just in a single time series but between all pairs of time series. We illustrate the new methodologies using several illustrative problems, including modeling data on manifolds and the prediction of chaotic time series.

1 Introduction

In data-driven modeling problems, generally the goal is to construct a representation of the data that will be both descriptive and predictive. The

Arta Jamshidi is now at the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K.

requirement that the model be descriptive dictates that it accurately reproduces the relationships in the observed data, while the requirement that the model be predictive means that unobserved inputs, or domain values, be mapped to meaningful outputs, or range values. In general terms, it is the predictive quality, or ability to generalize from observations, of the model that leads to added value and the potential for knowledge discovery.

When data may be viewed as being sampled from a manifold, we may appeal to some powerful theorems from differential geometry to guide our modeling approach. For example, Whitney's (1936) (easy) embedding theorem pertains to representing data as the graph of a nonlinear function and provides guidance concerning the dimension of the domain when it is not known a priori. More precisely, Whitney's theorem states that an m_d-dimensional manifold residing in an n-dimensional ambient space may be embedded in a $2m_d + 1$ dimensional subspace.[1] The proof of the theorem is constructive and shows that there is a large set of projections such that there exists a nonlinear inverse that affords perfect reconstruction of the data.[2] Whitney's theorem pertains not to data but to manifolds. When data may be viewed as a sampled manifold, the question arises concerning how to construct the nonlinear function that Whitney's theorem proves exists.

In contrast, Takens' theorem (1980) indicates that data sampled from an m_d-dimensional smooth manifold may be reconstructed (up to a diffeomorphic copy) by using a time-delay embedding in a space of dimension $2m_d + 1$. What is surprising about this theorem is that it states that only a single component, a scalar function, of the time-dependent trajectory must be observed to permit this reconstruction. This permits a now-well-known framework for the time-series prediction problem (Packard, Crutchfield, Farmer, & Shaw, 1980). These geometric theorems suggest that data on manifolds may possess a structure that is richer than simple nonlinearity and may be exploited in the modeling process. This is the primary motivation for the algorithm proposed in this letter.

The objective of this letter is to present a new approach for building nonlinear models for multivariate data that is particularly useful for modeling time-evolving trajectories on manifolds. The main contribution is a practical algorithm to model geometric structure not just in the individual components of the range of model but also to identify and model structure that exists between all pairs of components. The proposed algorithm is an extension to our previous work (Jamshidi, 2004; Jamshidi & Kirby, 2007; Anderle & Kirby, 2001), which deals with modeling where the range dimension is one. Similar to the univariate case, the proposed algorithm progresses iteratively, adding a new function at each step to refine the model.

[1] This theorem is particularly helpful when $n >> 2m_d + 1$.

[2] In mathematical terms, there is an open and dense set of projections that allow one to reduce the dimension to $2m_d + 1$.

The placement of the functions is driven by a statistical hypothesis test, but now in higher dimensions that reveals geometric structure when it fails. At each step, the added function is fit to data contained in a spatiotemporally defined local region to determine the model parameters. A great advantage of the proposed algorithm is that it does not require ad hoc parameters save for the selection of the skew radial basis functions. Thus, the number of basis functions required for an accurate fit is determined automatically by the algorithm, making it a powerful tool for modeling multivariate time series, especially when the additional manifold structure may be exploited.

This letter is organized as follows. In section 2, the geometric theorems are discussed in the context of data analysis. In section 3.1, basic facts about radial basis functions (RBFs) are reviewed, and a recent variation, skew RBFs (sRBFs), is described. In section 4, the theoretical aspects of the multivariate RBF algorithm are presented. In section 5, the basic multivariate RBF algorithm is presented in detail. Numerical experiments are described in section 6, illustrating applications of Whitney's and Takens' theorems. Section 7 summarizes the main contributions of this letter.

2 Geometric Framework

Here we outline the geometric framework of the algorithms. We shall see that at the heart of each problem is a need for construction of a nonlinear mapping of multivariate data.

2.1 Representing Data as a Graph of a Function. Previously, we proposed a framework for representing data as a graph of a function (Broomhead & Kirby, 2000, 2001). We summarize this approach here. Given a data set $\mathcal{A}$ sampled from a smooth manifold $\mathcal{M}$, the foundation for this representation is provided by the decomposition of a data point $x \in \mathcal{A} \subset \mathcal{M}$,

$$x = p + q, \tag{2.1}$$

where $\mathbb{P}$ is a projection matrix, $p = \mathbb{P}x$, $q = \mathbb{Q}x$, and $\mathbb{Q}$ is complementary to $\mathbb{P}$, that is, $\mathbb{Q} = I - \mathbb{P}$. In this setting, we view any element x as being the sum of the portion of x in the range of $\mathbb{P}$, that is, $p \in \mathcal{R}(\mathbb{P})$, and the portion in the null space of $\mathbb{P}$, that is, $q \in \mathcal{N}(\mathbb{P})$.

Following Broomhead and Kirby (2000, 2001), let $\hat{U} = [u_1|u_2|\dots|u_n]$ be an orthonormal basis for $\mathbb{R}^n$. Then a rank d projector may be defined by

$$\mathbb{P} = \hat{U}_1 \hat{U}_1^T,$$

where $\hat{U}_1 = [u_1|\dots|u_d]$ is an $n \times d$ matrix. Similarly, the complementary projector may be defined as

$$\mathbb{Q} = \hat{U}_2 \hat{U}_2^T,$$

where $\hat{U}_2 = [u_{d+1}|\dots|u_n]$. We note that the quantity $\mathbb{P}x \in \mathcal{R}(\mathbb{P})$ is an n-tuple in the ambient basis: $\mathbb{P}x = (u_1^T x)u_1 + \cdots + (u_d^T x)u_d$. It is the expansion coefficients that provide the d-dimensional representation, and these are given as

$$\hat{p} = \hat{U}_1^T x.$$

Similarly, the reduced representation for q consists of $n - d$ components:

$$\hat{q} = \hat{U}_2^T x.$$

Taking these together, we rewrite equation 2.1 in terms of the reduced variables $\hat{p}, \hat{q}$ as

$$x = \hat{U}_1 \hat{p} + \hat{U}_2 \hat{q}.$$

The nonlinear representation of the data comes from writing the residuals of the projection in terms of a function of the projected data values:

$$\hat{q} = f(\hat{p}). \tag{2.2}$$

It is Whitney's (1936) embedding theorem that ensures the existence of this mapping. Note that f is now a mapping from the d-dimensional projected data to the $n - d$–dimensional set of residuals of the projection. The reconstruction of the original data then may be written in terms of two components—one linear and one nonlinear:

$$x = \hat{U}_1 \hat{p} + \hat{U}_2 f(\hat{p}). \tag{2.3}$$

Another way to view this representation is as the graph of a function

$$x = (\hat{p}, f(\hat{p})), \tag{2.4}$$

where the coordinates of the representation are given with respect to the columns of the basis matrix $\hat{U}$. Thus, each point x is given as a mapping from the d-dimensional linear space to the $n - d$-dimensional residual. In general, one does not know the domain of the graph, and in these cases, one must identify a good projection to determine this.

If the rank of $\mathbb{P}$ is d where $d > 2m_d$, then Whitney's theorem ensures the existence of a global map from the range of the projector to its null space, that is, there exists an f such that equation 2.2 will hold. We have referred to the representation in equation 2.3 as the Whitney reduction network (Broomhead & Kirby, 2001). Note that the special case of principal component analysis is obtained if the linear term $\hat{U}_1 \hat{p}$ contains enough

dimensions to fully reconstruct, or encapsulate, the data. However, if the data have the structure of a manifold, then it is more efficient to represent by taking the rank of $\hat{U}_1$ to be as small as possible and to have the nonlinear term $\hat{U}_2 f(\hat{p})$ encode the manifold structure. Determining parsimonious representations for f in this expression is one of our main motivations for developing efficient algorithms for multivariate RBFs.

2.2 Time-Delay Embedding. Another opportunity for modeling using multivariate RBFs arises in the context of time-series prediction. It is of particular interest to attempt to construct mappings from past to future values given a time-evolving system,

$$x_{k+1} = f(x_k, x_{k-1}, \ldots, x_{k-n+1}).$$

Further, one may be interested not in predicting simply the next value but many future values,

$$(x_{k+T_s}, x_{k+2T_s}, x_{k+(m-1)T_s}) = f(x_k, x_{k-\tau}, \ldots, x_{k-(n-1)\tau}),$$

where T_s corresponds to a separation interval between predicted samples, τ corresponds to a separation interval between domain samples, and m indicates the number of ambient dimension of the range of the map f. Typically the values of T_s and τ are problem dependent and must be determined for each data set. In general, it is not possible to say, a priori, whether such a function f exists for a given data set. One may answer this question empirically by attempting to construct f empirically from the data.

It has been shown that if a system explores a smooth manifold $\mathcal{M}$ of dimension m_d, then it is sufficient to observe a scalar value from this system, say $x(t)$, to be able to reconstruct geometry of the manifold up to a diffeomorphism; this result is known as Takens' theorem (Takens, 1980). In particular, the time-delay coordinates of the scalar time series,

$$z(t) = (x(t), x(t-\tau), \ldots, x(t-(n-1)\tau)),$$

create a time series $z(t) \in \mathbb{R}^n$ (Packard et al., 1980). This time series traces out a manifold structure that is effectively a reconstruction of the original manifold as long as, according to Takens' theorem, n is taken large enough and guaranteed for $n > 2m_d$. In this context, n is referred to as the global embedding dimension, and τ is the time delay (see, e.g., Abarbanel, Gilpin, & Rotenberg, 2005, for additional implementation details).

In some cases, we are interested in modeling multiple time series; we may seek to predict the future values of foreign currencies from past values.

Here again we may also employ a time-delay embedding of the data,

$$z(t) = (x_1(t), x_1(t-\tau_1), \ldots, x_1(t-n_1\tau_1), \ldots, x_{n_r}(t), \\ x_{n_r}(t-\tau_{n_r}), \ldots, x_{n_r}(t-n_{n_r}\tau_{n_r})),$$

where n_r is the number of time series of interest and τ_i, n_i are the time delay and the embedding dimension of the ith time series. Again, a priori it is often not possible to know whether the data of interest possess a manifold structure or if a prediction function f exists.

3 Relationship to Other Work

Here we describe the background of RBFs, as well as provide an overview of recent work in the general area of manifold learning

3.1 Data Fitting with Radial Basis Functions. The mappings described in the previous section were of the form

$$f : U \in \mathbb{R}^n \to V \in \mathbb{R}^m, \tag{3.1}$$

where we assume that in general, both m and n may be greater than 1. In the data-fitting problem, we assume that we have samples $x^{(k)} \in U$ and $y^{(k)} \in V$ indexed by k and related via f, a nonlinear function, as

$$y^{(k)} = f(x^{(k)}). \tag{3.2}$$

A standard RBF seeks to represent $f(x)$ as

$$f(x) = \sum_{i=1}^{N_c} w_i \phi(\|x - c_i\|_{W_i}), \tag{3.3}$$

where x is an input pattern, ϕ is an RBF centered at location c_i, and w_i denotes the weight for ith RBF. The term W denotes the parameters in the weighted inner product,

$$\|x\|_W = \sqrt{x^T W x}.$$

RBFs have been widely used for data approximation. Multiquadrics, that is, functions of the form $\phi(r) = \sqrt{1+r^2}$, were introduced by Hardy (1997) for modeling scattered data on topographical surfaces. Thin plate splines, $\phi(r) = r^2 \ln r$, were introduced by Harder and Desmarais (1972) for surface interpolation. Gaussian RBFs, $\phi(r) = \exp(-r^2/\sigma^2)$, were proposed by Broomhead and Lowe (1988) for data fitting and classification. There is a

significant literature treating theoretical aspects of RBFs, including universal approximation theorems (see, e.g., Powell, 1992; Park & Sandberg, 1991, 1993; Schaback & Wendland, 2000). A large number of algorithms have been proposed in the literature for computing the model parameters (Platt, 1991; Karayiannis & Mi, 1997; Holmes & Mallick, 1998; Yingwei, Sundararajan, & Saratchandran, 1998; Nabney, McLachlan, & Lowe, 1996). The research monographs (Lee & Haykin, 2001; Buhmann, 2003; Wendland, 2005) contain references treating additional theory, algorithms, and applications.

Alternatively, we also consider the sRBFs introduced in Jamshidi and Kirby (2008, 2010), which are defined as

$$f(x) = \sum_{i=1}^{n} w_i z(x; \nu_i) \phi(\|x - c_i\|_{W_i}). \tag{3.4}$$

In equation 3.4, the function $z(x; \nu_i)$ is a skew component that makes the representation nonradial, and ν_i contains the parameters for the symmetric breaking term. This has the advantage of being able to represent asymmetric data, such as data near boundaries, much more efficiently. An example of an sRBF is the skew gaussian RBF,

$$f(x) = \sum_{i=1}^{N_c} w_i \exp(-\|x - c_i\|_{W_i}^2) \int_{-\infty}^{-\lambda_i^T (x - c_i)} \exp(-y^2) dy, \tag{3.5}$$

where ϕ is the usual radial gaussian term,

$$\phi(\|x - c_i\|_{W_i}) = \exp(-\|x - c_i\|_{W_i}^2),$$

and the integral is responsible for skewing the output where

$$z(x; \nu_i) = \int_{-\infty}^{\lambda_i (x - c_i)} \exp(-y^2) dy.$$

Note that in this representation, $\nu_i = (\lambda, c_i)$.

In this letter, we employ a truncated cosine function in the same fashion as a Hanning filter to produce an RBF with compact support (Jamshidi & Kirby, 2006),

$$f(x) = \sum_{i=1}^{N_c} w_i \big(\cos(\|x - c_i\|_{W_i} \pi) + 1 \big) H(1 - \|x - c_i\|_{W_i}), \tag{3.6}$$

where H is the heaviside function. To create an sRBF, we employ the Arctan function as the skewing term z. These functions, taken together, result in

the Arctan-Hanning sRBF,

$$f(x) = \sum_{i=1}^{N_c} w_i \left(\frac{1}{\pi} \arctan\left(\lambda_i^T (x - c_i)\right) + \frac{1}{2} \right) \times \left(\cos(\|x - c_i\|_{W_i} \pi) + 1 \right) H(1 - \|x - c_i\|_{W_i}). \tag{3.7}$$

The issue now is to determine the parameters associated with the data-fitting problem. Of particular importance are the locations of the fitting functions $\{c_i\}$ and the width of the basis $\{W_i\}$, as well as the number of fitting functions N_c. A method for initializing these using the correlation and cross-correlation structure is described in section 4. Once these parameters are initialized, together with the remaining parameters they are determined using a variety of nonlinear optimization routines.

3.2 Manifold Learning. This letter particularly concerned with modeling multivariate time series on manifolds—where locally the structure behaves like Euclidean space. This setting is attractive given the existence of Whitney's theorem and Takens' theorem described in section 2, which indicate that data on manifolds are particularly amenable to nonlinear modeling by constructing an empirical representation of a function f that captures the structure of the data from examples. This review presented here is not intended to be comprehensive but rather a sampling of the literature on this topic.

The modeling procedure presented here represents a special case of a more general area now referred to as manifold learning. We note that this field comprises a collection of approaches to model data assuming a special structure, in the same way lines of best fit are used to model data where we assume some linear trend exists. It is not essential that the data actually reside on a manifold to make these techniques useful in the same way that data need not fit exactly on a line when we use linear least squares. Of course, the more like a manifold the data are, the better we can expect the results to be.

There are a variety of ways to approach the manifold learning problem. A local approach based on data charts models the data in patches that, when taken together, form a complete manifold model. Each chart consists of a connected neighborhood of the manifold, equipped with a projection mapping from the manifold to a linear space (whose dimension is the topological dimension of the manifold) and an inverse function from the linear space to the manifold.

The composition of these functions behaves as the identity mapping. (For additional details on this approach, see Hundley, 1998; Hundley, Kirby, & Miranda, 1999.) A probabilistic approach for generating an atlas of charts is provided in Brand (2003). This algorithm employs a mapping that preserves

local geometric relations in the manifold and is pseudo-invertible. The algorithm decomposes the sample data into locally linear low-dimensional patches that are merged into a single low-dimensional coordinate system by stochastic optimization.

A manifold learning approach referred to as local linear embedding (Roweis & Saul, 2000) represents each point in the ambient space as a weighted sum of its nearest neighbors. This weighting matrix is then used to generate a configuration of points with the same spatial relationships in a space of reduced dimension. We note that this approach results in an eigenvector problem where the matrix has the same size as the number of points in the sample. Although strictly speaking, this is not an isometric embedding of the data, it may be viewed loosely as topology preserving.

Multidimensional scaling (MDS) approaches this problem from a different perspective (Mardia, Kent, & Bibby, 1980). Given a set of distances between points, is it possible to locate these points in Euclidean space such that they have the same interpoint distances. MDS provides a solution in terms of the eigenvectors of the interpoint distance matrix. An extension to this approach for data on manifolds, referred to as ISOMAP, is based on computing geodesic distances numerically from sampled data and using these distances in the MDS eigenvector problem (Tenenbaum, de Silva, & Langford, 2000). The result is an embedding of the data in Euclidean space with distances between the data points, as measured on the manifolds, preserved. Other techniques have been proposed for constructing isometric embeddings, notably Belkin and Niyogi (2003), Donoho and Grimes (2003), and Coifman and Lafon (2006).

In Roweis, Saul, and Hinton (2002), the local linear models are represented by a mixture of factor analyzers, and the global coordination of these models is achieved by adding a regularizing term to the standard maximum likelihood objective function favoring models whose internal coordinate systems are aligned in a consistent way. As a result, the internal coordinates change smoothly as one traverses a connected path on the manifold even when the path crosses the domains of many different local models. To get an efficient algorithm that allows separate local models to learn consistent global representations, an automatic alignment procedure that maps the disparate internal representations learned by several local dimensionality-reduction experts into a single coherent global coordinate system for the original data space is studied in Teh and Roweis (2002).

Kernel principal component analysis (KPCA; Verbeek, Roweis, & Vlassis, 2004) treats the processing of data on manifolds by first flattening the data out using an appropriately selected kernel function, for example, a Veronese embedding Schölkopf, Smola, and Müller (1998). This approach has the advantage of turning nonlinear problems into tractable linear problems via the kernel trick. KPCA is also related to the well-known classification technique of support vector machines (Vapnik, 1995).

Projection pursuit was proposed as a linear approach for reducing the dimension of data to reveal clusters in data and their separation (Friedman & Tukey, 1974). It is often used in the context of an interactive data visualization system such as the projection pursuit guided tour (Cook, Buja, Cabrera, & Hurley, 1995).

A global approach to manifold learning that involves reducing data such that it is constrained to lie on a prescribed manifold in the reduced space was presented in Kirby and Miranda (1996). For example, one can construct the best topological circle through a data set using this approach. This idea was extended to spheres in Hundley, Kirby, and Miranda (1995).

4 Statistical Background for the Multivariate sRBF Algorithm

We denote the set of residuals for a model of order N_c as

$$R^{N_c} = \{e_k\}_{k=1}^{L}, \tag{4.1}$$

where $e_k = y_k - f(x_k)$ is the m-variate residual of the kth data point. L is the cardinality of the training set. In addition, μ is the mean vector $E(e_k)$, and $\Gamma(h) = E(e_{k+h}e_k') - \mu\mu'$ is the covariance matrix at lag h. An unbiased estimate for μ is given by $\bar{e} = \frac{1}{L}\sum_{k=1}^{L} e_k$. An estimate of the covariance matrix $\Gamma(h) = E[(e_{k+h} - \mu)(e_k - \mu)'] = [\gamma_{ij}(h)]_{i,j=1}^{m}$ is given by

$$\widehat{\Gamma}(h) = \begin{cases} \dfrac{1}{L}\displaystyle\sum_{k=1}^{L-h} \alpha(h, e_k), & \text{if } 0 \le h \le L-1 \\ \widehat{\Gamma}'(-h), & \text{if } -L+1 \le h \le 0. \end{cases} \tag{4.2}$$

Similar to the univariate case, we decompose the autocovariance function (ACVF) into its components as $\alpha(h, e_k) = (e_{k+h} - \bar{e})(e_k - \bar{e})'$. Furthermore, $\alpha(h, e_k^i, e_k^j) = (e_{k+h}^i - \bar{e}^i)(e_k^j - \bar{e}^j)$ is the (i, j)-component of $\alpha(h, e_k)$. In other words,

$$\widehat{\gamma}_{ij}(h) = Cov(e_{k+h}^i, e_k^j) = \frac{1}{L}\sum_{k=1}^{L-h} \alpha(h, e_k^i, e_k^j).$$

For a fixed lag h, the quantity $\alpha(h, e_k)$ is the contribution of the kth residual to the autocorrelation function. And the quantity $\alpha(h, e_k^i, e_k^j)$ is the contribution of the i and jth time series at the kth residual of the autocovariance function. Later we focus on this quantity α and illustrate that it reveals critical information concerning where new basis functions should be placed.

The estimate of the correlation matrix function $R(.)$ is then given by

$$\widehat{R}(h) = [\widehat{\rho}_{ij}(h)]_{i,j=1}^{m} = [\widehat{\gamma}_{ij}(h)(\widehat{\gamma}_{ii}(0)\widehat{\gamma}_{jj}(0))^{\frac{-1}{2}}]_{i,j=1}^{m}, \tag{4.3}$$

where $\widehat{\gamma}_{ij}(h)$ is the $(i,\ j)$-component of $\widehat{\Gamma}(h)$. If $i = j, \widehat{\rho}_{ij}$ reduces to the sample autocorrelation function of the ith series. (For the asymptotic behavior and the convergence properties of the sample mean and covariance functions see Brockwell & Davis, 1991).

We seek to terminate the addition of new basis functions when the residuals appear to have no further structure. As a test for structure, we consider whether the residuals are independent and identically distributed (i.i.d.). We need to extend our definition of white noise to the multivariate case (see Jamshidi & Kirby, 2007, for details concerning the univariate test). The m-variate series $\{e_t\}, t \in \mathbb{Z}$ is said to be white noise with mean 0 and covariance matrix Σ, written as $\{e_t\} \sim WN(0, \Sigma)$, if and only if e_t is stationary with mean vector 0 and covariance matrix function:

$$\Gamma(h) = \begin{cases} \Sigma, & \text{if } h = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

We use the notation $\{e_t\} \sim IID(0, \Sigma)$ to indicate that the random vectors $\{e_t\}$ are i.i.d. with mean 0 and variance Σ.

In general, the derivation of the asymptotic distribution of the sample cross-correlation function is quite complicated even for multivariate moving averages (Brockwell & Davis, 1991). The methods employed for the univariate case are not immediately adaptable to the multivariate case. An important special case arises when the two component time series have independent moving averages. The asymptotic distribution of $\widehat{\rho}_{12}(h)$ for such a process is given in the following theorem:

Theorem 1 (Brockwell & Davis, 1991)**.** *Suppose that*

$$X_{t1} = \sum_{j=-\infty}^{\infty} \alpha_j Z_{t-j,1}, \{Z_{t1}\} \sim IID(0, \sigma_1^2), \tag{4.5}$$

$$X_{t2} = \sum_{j=-\infty}^{\infty} \beta_j Z_{t-j,2}, \{Z_{t2}\} \sim IID(0, \sigma_2^2), \tag{4.6}$$

where the two sequences $\{Z_{t1}\}$ *and* $\{Z_{t2}\}$ *are independent,* $\Sigma_j|\alpha_j| < \infty$ *and* $\Sigma_j|\beta_j| < \infty$*. If* $h \geq 0$*, then*

$$\widehat{\rho}_{12}(h) \text{ is } AN\left(0, L^{-1} \sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)\right). \tag{4.7}$$

If $h, k \geq 0$ and $h \neq k$, then the vector $(\widehat{\rho}_{12}(h), \widehat{\rho}_{12}(h))'$ is asymptotically normal (AN) with mean 0, variances as above and covariance,

$$L^{-1} \sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j+k-h). \tag{4.8}$$

As reported in Brockwell and Davis (1991), without knowing the correlation function of each of the processes, it is impossible to decide if the two processes are uncorrelated with one another. The problem is resolved by prewhitening the two series before computing the cross-correlation $\widehat{\rho}_{12}(h)$, that is, transfer the two series to white noise by application of suitable filters. In other words, any test for independence of the two component series cannot be based solely on estimated values of the cross-correlation without taking into account the nature of the two component series. Note that since in practice the true model is nearly always unknown and since the data X_{tj}, $t \leq 0$, are not available, it is convenient to replace the sequences $\{Z_{tj}\}$ by the residuals, which, if we assume that the fitted models are in fact the true models, are white noise sequences. To test the hypothesis H_0 that $\{X_{t1}\}$ and $\{X_{t2}\}$ are independent series, we observe that under H_0, the corresponding two prewhitened series $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are also independent. Under H_0, theorem 1 implies that the sample autocorrelations $\widehat{\rho}_{12}(h)$ and $\widehat{\rho}_{12}(k)$, $h \neq k$, of $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are asymptotically independent normal with mean 0 and variances L^{-1}. An appropriate test for independence can therefore be obtained by comparing the values of $|\widehat{\rho}_{12}(h)|$ with $1.96L^{\frac{-1}{2}}$. If we prewhiten only one of the two original series, say, $\{X_{t1}\}$, then under H_0, theorem 1 implies that the sample cross-correlations $\widehat{\rho}_{12}(h)$ and $\widehat{\rho}_{12}(k)$, $h \neq k$, of $\{Z_{t1}\}$ and $\{X_{t2}\}$ are asymptotically independent normal with mean 0 and variances L^{-1} and covariance $L^{-1}\rho_{22}(k-h)$. Hence, for any fixed h, $\widehat{\rho}_{12}(h)$ also falls (under H_0) between the bounds $\pm 1.96L^{\frac{-1}{2}}$ with a probability of approximately 0.95.

Therefore, if one computes the sample cross-correlations up to lag h and finds that more than $0.05h$ of the samples fall outside the bound, or that one value falls far outside the bounds, the i.i.d. hypothesis is rejected. This test can equivalently be written in terms of χ^2 distribution. Given

$$Q = L\widehat{\rho}_{12}^{T}\widehat{\rho}_{12} = L\sum_{j=1}^{L-1}\widehat{\rho}_{12}^{2}(j),$$

Brockwell and Davis (1991) showed that Q has a χ^2 distribution with $L-1$ degrees of freedom. The adequacy of the model is therefore rejected at level α if

$$Q > \chi^2_{1-\alpha}(L-1).$$

5 Multivariate Algorithm Implementation

The main difference with the univariate algorithm is the statistical hypothesis test. Again, the question of whether a new basis function should be added is answered by the i.i.d. test. We shall see that this test also indicates where the new basis function should be initialized. First, we compute the autocorrelation functions of all the m time series. If all of these pass the white noise (WN) or i.i.d. test, then the cross-correlations among the time series are considered. If there is structure in the autocorrelations or cross-correlations of the time series, then the i.i.d. will be rejected.

As in the univariate case, the next requirement is to determine where the new basis function should be located to optimally reduce the structure in the model residuals. In our extension, we look for the point in the domain that makes the largest contribution to the auto- or cross-correlation that has caused the test to fail.

Given this information, we use the fact that the residuals are associated with the data in the domain bijectively, that is, there is a mapping, say ψ, from a data point to its higher-dimensional residual of the form $e_k = \psi(x_k)$. Thus, by identifying the residual associated with the largest contribution to auto- or cross-correlation, we may identify the location in the domain where the basis function should be added. To actually find this point, first we determine the exact lag for which the correlation function, $\widehat{\gamma}_{ij}(h)$ reaches its maximum value h^*:

$$h^* = \arg\max \widehat{\gamma}_{ij}(h), h > 0. \tag{5.1}$$

Then we find the point in the spatial domain that has the maximum contribution to the associated ACF for lag $h = h^*$ by solving

$$i^* = \arg \max_{k=1,\ldots,L-h^*} \alpha(h^*, e_k^i, e_k^j). \tag{5.2}$$

Thus, the center for the new basis function is given by

$$x_{i^*} = \psi^{-1}(e_{i^*}),$$

where ψ^{-1} is the inverse of the function ψ. For simplicity, we refer to this center location as x^*.

Now that the center of the new basis function has been found, it is necessary to determine what data should be used to determine the scale and weight of the new RBF. Proceeding in a manner analogous to the univariate case (Jamshidi & Kirby, 2007), consider the function $\beta_k^{i,j} = \alpha(h^*, e_k^i, e_k^j)$. The index k is inherited from the data labels and, in the case of a time series, corresponds to a time ordering. For simplicity, we assume that $\beta_k^{i,j}$ decreases monotonically for both increasing and decreasing values of k until it crosses

zero at the indices $l^* < i^*$ and $r^* > i^*$. Here we use l, r to indicate left and right, respectively. We now compute the distances

$$d_l = d(x_{i^*}, x_{l^*})$$

and

$$d_r = d(x_{i^*}, x_{r^*}),$$

as these indicate the size of the data ball around the center x^*. The subset of the data employed to update the added basis function is then

$$\mathcal{X}_{local} = \{x \in \mathcal{X} : \|x - x^*\| \leq d_c\},$$

where $\mathcal{X}$ is the entire training set. The distance d_c can be selected in a variety of ways, and here we select

$$d_c = \max\{d_l, d_r\}.$$

Note that $\mathcal{X}_{local}$ now may contain data whose indices have values that are substantially different from i^*, l^*, and r^*.

The new RBF added to the expansion is initialized and optimized similar to the univariate case. The center c_0 is initialized at the point of most structure according to our test: $c_0 = x^*$. The vector of widths σ is very effectively initialized using the diagonal elements of the covariance matrix of the local data,

$$\sigma_0 = \sqrt{\text{diag}(\text{cov}(\mathcal{X}_{local}))}.$$

Note here that $W = \text{diag}(\sigma_0)$. The initial value for the multivariate weight, α_0, is calculated via least squares using the initial values for center location and widths. We have initialized the skew parameters at zero. Then the parameters associated with the new basis function are optimized by solving the nonlinear optimization procedure using BFGS. The optimization cost function is

$$E(v) = \min_{v} \sum_{x \in \mathcal{X}_{local}} \|h(x; v) - y\|_2^2,$$

where v contains the new sRBFs parameters: $v = [c_0, \sigma_0, \alpha_0, \lambda_0]^T$. Note that all the multivariate range values associated with $\mathcal{X}_{local}$ contribute to the optimization procedure.

Similar to the univariate case, we could use one of the statistical tests, root mean square error (RMSE),

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{m} e_{ij}^2},$$

or normalized prediction error (NPE),

$$NPE_1 = \frac{\sum_{i=1}^{T}\sum_{j=1}^{m} |e_{ij}|}{\sum_{i=1}^{T}\sum_{j=1}^{m} |y_{ij} - \overline{y}_j|},$$

$$NPE_2 = \frac{\sum_{i=1}^{T}\sum_{j=1}^{m} e_{ij}^2}{\sum_{i=1}^{T}\sum_{j=1}^{m} (y_{ij} - \overline{y}_j)^2},$$

or another measure of structure as stopping criteria. The pseudocode of this algorithm is provided in algorithm 1.

6 Numerical Results

In this section, we provide numerical evidence that the proposed k-variate algorithm results in a model that significantly outperforms k univariate models. The enhanced parsimony of the model is result of accounting for correlations between the multivariate ranges.

6.1 Multivariate *Pringle* Data Set. To begin, in this section, we consider the application of the multivariate RBF algorithm to the Pringle data set (Broomhead & Kirby, 2000, 2001). As described below, we will build a two-variate data set with correlation one between the first and the second time series, as a first illustration of the methodology. This problem illustrates the challenges of modeling data as the graph of a function with range dimension two.

This Pringle data set, as proposed in Broomhead and Kirby (2005), can be generated as the solution to the following systems of ordinary differential equations,

$$\frac{dx}{dt} = y$$

$$\frac{dy}{dt} = -x - (x^2 + y^2 - 1)y$$

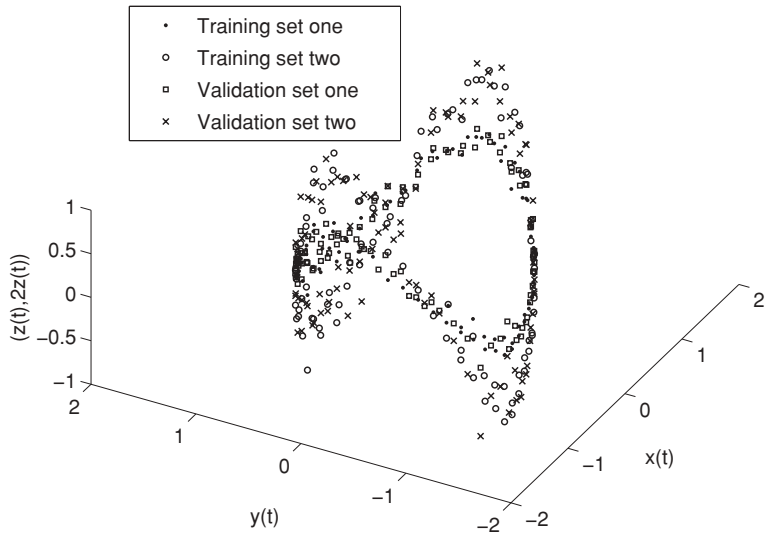
$$\frac{dz}{dt} = -\lambda z + 2(\lambda xy + \omega(x^2 - y^2)),$$

Algorithm 1: A Multivariate sRBF Algorithm Using a Pairwise Hypothesis Test on Time Series.

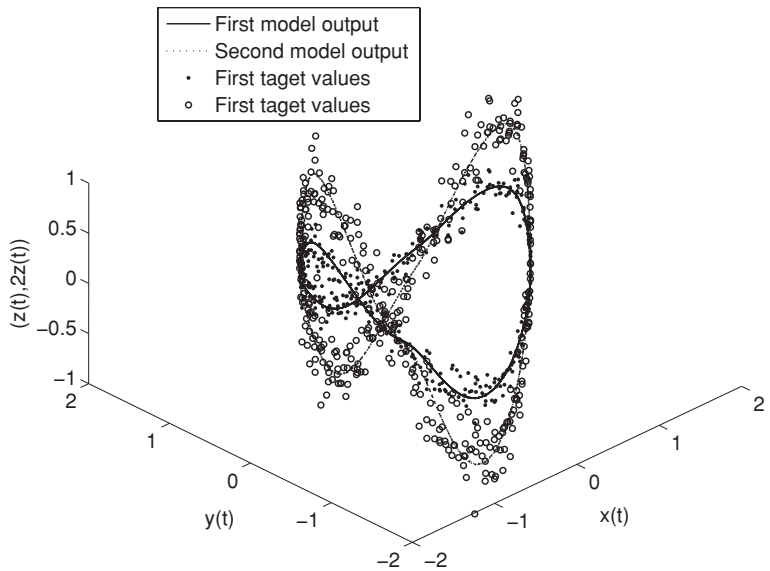
$ran\_flag = 1, K = 0$
while $ran\_flag = 1$ **do**
  evaluate the RBF on the training data set $\{f(x_n)\}_{n=1}^{L}$
  compute the model error $\{\underline{e}_n\}_{n=1}^{L}$
  compute component contributions $\alpha(h, e_k^i, e_k^j) = e_{k+|h|}^i e_k^j$ **for all** $i, j = 1, \ldots, m$
  compute correlation functions for all the m time series and all lags $0 \leq h < L$
  compute the maximum contribution to each correlation function over all lags
  apply the univariate WN test to each of the pairs
  if any of the autocorrelations does not pass the WN test **then**
    identify time series d, that has the maximum value at its autocorrelation function. Let $i = d$ and $j = d$
  else if any of the cross-correlations does not pass the WN test **then**
    identify the pair of time series d_1, d_2 that has the maximum value at their cross-correlation function, $i = d_1$ and $j = d_2$
  else
    $ran\_flag = 0$
  end if
  compute h^* via equation $h^* = \arg\max \widehat{\gamma}_{ij}(h), h > 0$ and
  compute $x^* = x_{i^*} = \psi^{-1}(e_{i^*})$ where $i^* = \arg\max_{k=1,\ldots,L-h^*} \alpha(h^*, e_k^i, e_k^j)$
  compute the CCC function, $\beta_k^{i,j} = \alpha(h^*, e_k^i, e_k^j), k = 1, \ldots, L - h^*$
  find the right and left zero crossing of the CCC function i^*, i.e., l^* and r^*
  compute $d_l = d(x_{i^*}, x_{l^*})$, $d_r = d(x_{i^*}, x_{r^*})$ and $d_c = \max\{d_l, d_r\}$
  define the local ball as $\mathcal{X}_{local} = \{x \in \mathcal{X} : \|x - x^*\| \leq d_c\}$
  add a new RBF $h(x; v)$ with initial values $v = [c_0, \sigma_0, \alpha_0, \lambda_0]^T$
  solve $E(v) = \min_v \| h(x; v) - y \|_2^2$, where $x \in \mathcal{X}_{local}$
  $K = K + 1$
  compute confidence, RMSE and $\widehat{\gamma}(h^*)$ of of the current model on the training set
end while

where λ and ω are parameters. A numerically integrated trajectory of the attracting cycle (x, y, z) is shown in Figure 1a.[3] A second data set, also displayed in Figure 1a, is constructed by amplifying the z coordinate as $(x, y, 2z)$.

[3] In this example, we are concerned only with fitting data on the limit cycle and ignore transients.



(a) Plots of the training and validation data sets for multivariate Pringle data set.



(b) The testing data set and the output of the four mode model.

Figure 1: The training, validation, and testing data sets and the output of the multivariate algorithm on the multivariate Pringle data set. (For a color version of this figure see the supplemental material, available online at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00060.)

A good projection for this problem, in the sense of minimizing the Lipschitz constant of f, has been shown to be the x, y-plane (Broomhead & Kirby, 2005). Thus, the (x, y) data values are the inputs to our function, and the range data values $(z, 2z)$ constitute the associated output of the function. The training set was selected to consist of 101 points (almost two cycles), while an additional 100 points were used for validation. In this experiment, the testing data set was selected to be the next 500 points, or approximately nine cycles. We use the Hanning-RBF in equation 3.6 in the numerical experiments in this section.

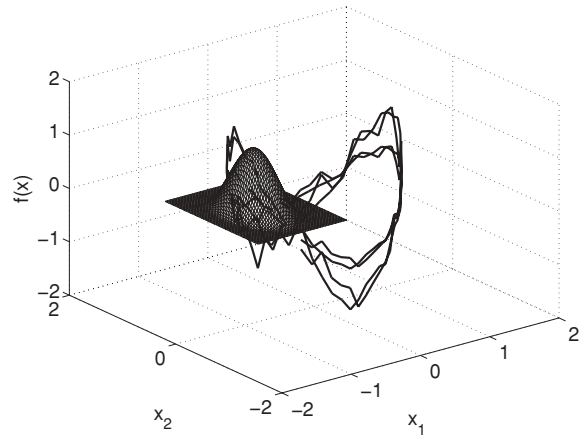
Figure 2 shows the output of the first RBF allocated by the multivariate algorithm. Figure 2a is the output associated with the first component of the multivariate range data, while Figure 2b is the output associated with the second component. The final model consists of four RBFs, and training was terminated as a result of the residuals' passing the i.i.d. test on the residual. The training, validation, and testing data sets and the output of the multivariate algorithm on multivariate Pringle data set are shown in Figure 1b. The performance of the RBF fit on the multivariate Pringle data set in the RMSE sense is shown in Figure 3a. The confidence level of the fitted model on the training and the validation set data sets as new basis functions are added to the model are shown in Figure 3b. After four RBFs have been added to the model, the confidence level of the i.i.d. hypothesis test is well above 95% for both the training and validation data, indicating a lack of structure in the residuals.

This simple example illustrates the ideal behavior of the multivariate algorithm when the output time series are highly correlated. Two univariate models in this instance would double the model complexity. Our next example, while harder to visualize, is a more challenging test for the algorithm.

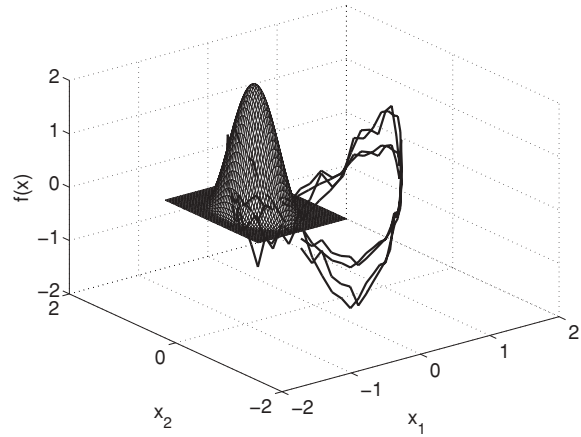
6.2 Multivariate Mackey-Glass. In this example, we apply the multivariate RBF algorithm to the Mackey-Glass delay differential equation (Mackey & Glass, 1977) which is often used in time series forecasting benchmark studies (see, e.g., Platt, 1991; Kadirkamanathan & Niranjan, 1993; Yingwei, Sundararajan, & Saratchandran, 1997). It is a classic example of the application of Takens' theorem where the optimal embedding dimension has been computed to be four. The Mackey-Glass equation is given by

$$\frac{ds(t)}{dt} = -bs(t) + a\frac{s(t-\tau)}{1+s(t-\tau)^{10}}. \tag{6.1}$$

A numerical integration of this delay differential equation produces a chaotic time series with only short-range time coherence, thus making accurate prediction difficult. The time series used in this letter was generated by integrating equation 6.1 with now-standard model parameters

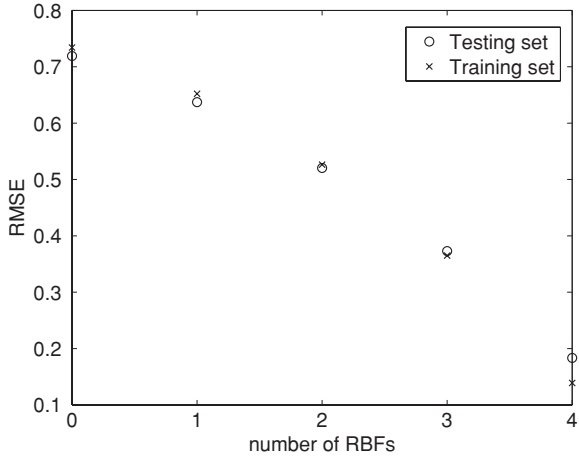


(a) The first RBF in relation with the training data set for the first time series.

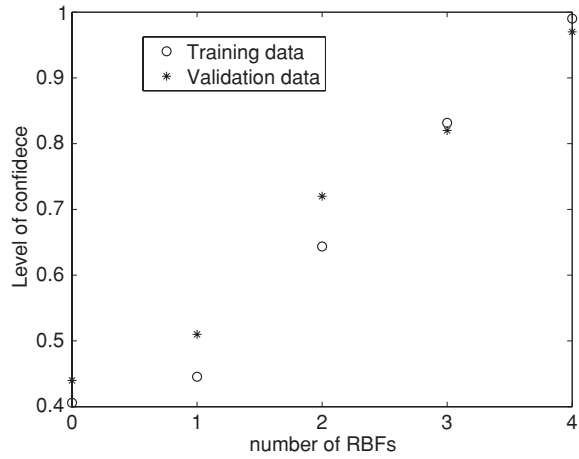


(b) The first RBF in relation with the training data set for the second time series.

Figure 2: The first RBF allocated by the algorithm for the case where $m = 2$. Hanning RBF was used in this fit. The residuals of the four-mode model pass the i.i.d. test. (For a color version of this figure see the supplemental material, available online at http://www.mitpressjournals.org/doi/suppl/10.1162/NECO_a_00060.)



(a) The RMSE plot of the model as new basis functions are added. The training stops after four basis functions have been added since the confidence level (see graph (b) below) of the hypothesis test exceeds 95%.



(b) The confidence level of the fitted model on the training (circles) and the validation (stars) data sets as the new basis functions are added to the model.

Figure 3: The performance of the multivariate RBF algorithm as basis functions are added. Both the RMSE error and the confidence level of the i.i.d. hypothesis test are shown. The raw data set used in this example is the multivariate Pringle as shown in Figure 1.

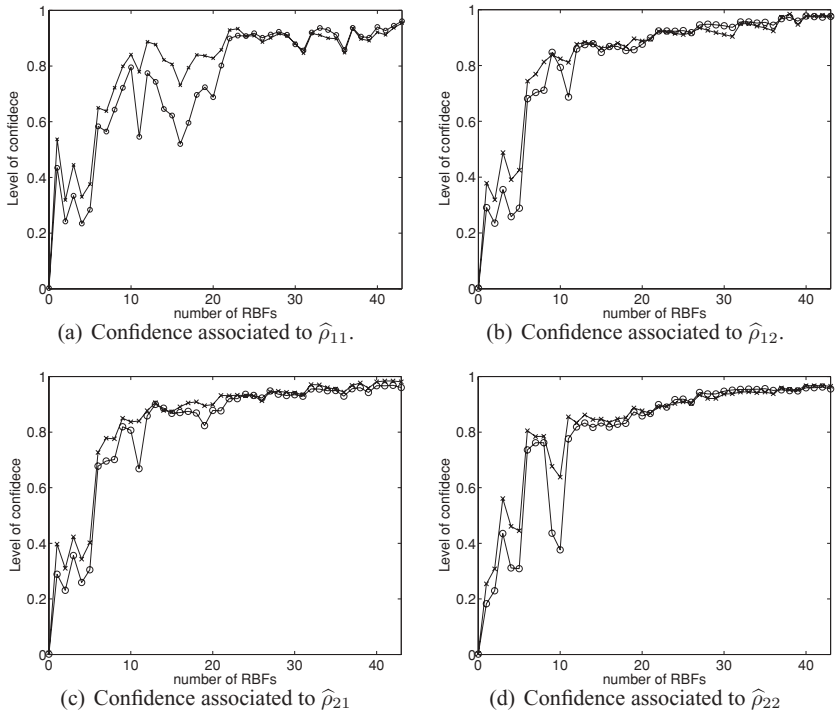
$a = 0.2$, $b = 0.1$, and $\tau = 17$ using the trapezoidal rule with $\Delta t = 1$, with initial conditions $x(t - \tau) = 0.3$ for $0 \leq t \leq \tau$ ($\tau = 17$). The initial 1000 data points corresponding to transient behavior are discarded. The next 3000 data points are reserved for the training set; points 4001 to 5000 are used as the validation set. Finally, the test set consists of 500 data points starting from point 5001. In these experiments (uniformly distributed), noise was added to the data with a standard deviation of 0.05. We use the Arctan-Hanning sRBF in equation 3.7 in the numerical experiments in this section.

At each time n, a forecast is made of the values $s_{n+\upsilon_1}$ and $s_{n+\upsilon_2}$—υ_1 and υ_2 samples ahead using the time-delay embedding consisting of the samples s_n, s_{n-6}, s_{n-12}, and s_{n-18}. Hence, for each n, the desired mapping f should take the time-delayed vector to this pair of future samples:

$$f : (s_n, s_{n-6}, s_{n-12}, s_{n-18}) \rightarrow (s_{n+\upsilon_1}, s_{n+\upsilon_2}).$$

In the experiments that follow we consider $(\upsilon_1, \upsilon_2) = (25, 50)$ and $(\upsilon_1, \upsilon_2) = (50, 75)$. These parameters are similar to those selected in Yingwei et al. (1998), where only a univariate output is considered (see also Liebert, Pawelzik, & Schuster, 1991; Farmer & Sidorowich, 1987; Kennel, Brown, & Abarbanel, 1992). The autocorrelation function of this time series goes through zero between lags 13 and 14 and is approximately -0.75 at lag 25. Given that the two time series in the range have significant correlation, we expect that fitting a two-variate model using the proposed algorithm rather than two univariate models will result in a reduced-order model. As we shall see below, this is indeed the case.

Figure 4 shows the confidence level of the multivariate model as new Arctan-Hanning sRBFs are added to the model for the case $(\upsilon_1, \upsilon_2) = (25, 50)$. We provide the confidence levels for all pairs of the time series. In almost all cases, the confidence level of 95% was achieved for the training set and validation set with approximately the same number of RBFs. In general, we advocate that this confidence level be attained for both sets unless the RMSE on the validation set increases suggesting the possibility of overfitting. The number of sRBFs required to achieve 95% of confidence on both the training and validation data sets is 43 for the two-variate problem. In this case, the RMSE of the final model is 0.0175, which is below the noise floor. The associated univariate 25 steps-ahead prediction problem requires 20 sRBFs to achieve 95% of confidence on the training and validation sets and results in a model with an RMSE of 0.0128, again below the noise floor. The univariate model for 50 steps-ahead prediction requires 28 sRBFs to reach 95% of confidence on training and validation sets with an RMSE below the noise floor of 0.0192. So, in summary, for this data set, the sRBF representation of f using two univariate models consisting



(a) Confidence associated to $\widehat{\rho}_{11}$. (b) Confidence associated to $\widehat{\rho}_{12}$.

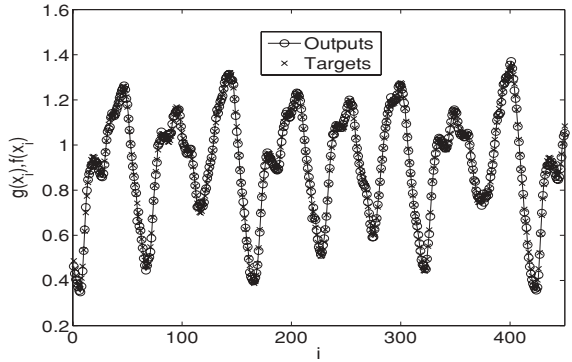
(c) Confidence associated to $\widehat{\rho}_{21}$ (d) Confidence associated to $\widehat{\rho}_{22}$

Figure 4: The confidence levels of the multivariate model as the new Arctan-Hanning sRBFs are added to the model for the case of 25- to 50-steps-ahead prediction of noisy Mackey-Glass data set.

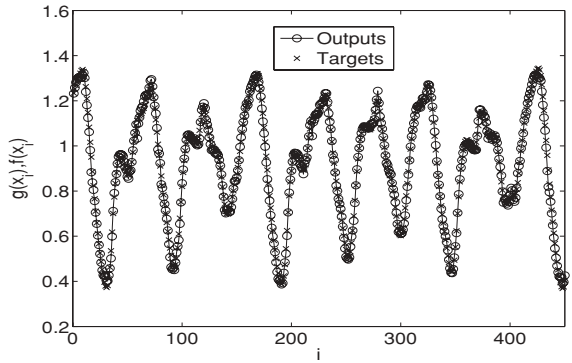
of 53 modes is over 20% larger than the two-variate model consisting of 43 sRBFs, while the errors for both representations are under the noise floor.

Figure 5 shows both outputs of the two-variate model for the Mackey-Glass experiment for the case $(\upsilon_1, \upsilon_2) = (25, 50)$. Figure 5a shows the output for the 25-steps-ahead prediction, while Figure 5b provides the model output for the 50-steps-ahead prediction. The log plot of the RMSE of the model as new sRBFs are added to the model is shown in Figure 5c.

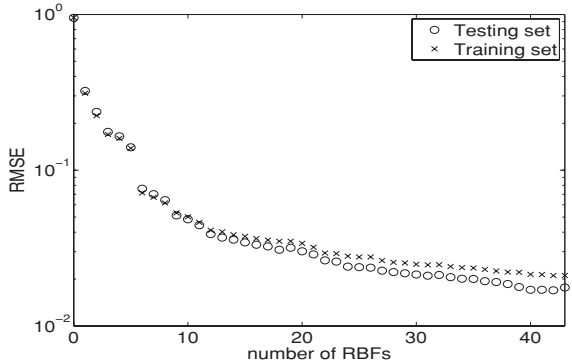
We repeated this experiment for the case $(\upsilon_1, \upsilon_2) = (50, 75)$. For this case, the two-univariate model representation was over 60% larger than the single two-variate, model consisting of 74 and 49 sRBFs, respectively, with comparable errors below the noise floor. This example provides convincing evidence of the savings one may obtain using the cross-correlation structure of the residuals of the model. Details of this second example may be found in Jamshidi (2008).



(a) Time evolution of the first component of the model output forecasting 25 steps ahead.



(b) Time evolution of the second component of the model output forecasting 50 steps ahead.



(c) The log plot of the RMSE of the model as new basis functions are added to the model.

Figure 5: The performance and the output of the two-variate sRBF fit for the case of 25- to 50-steps-ahead prediction of the noisy Mackey-Glass data set.

7 Conclusion

We observed that the extension of the ACF test for i.i.d. noise to k-variate ranges produces models of significantly smaller order than using k-univariate models while attaining comparable prediction errors. This is a consequence of the fact that the correlations as well as the cross-correlations of the residuals of the multivariate model are being exploited during the model-building process. We applied the proposed algorithm to the problem of representing data as the graph of a function as well as the benchmark problem of forecasting the chaotic time series produced by the Mackey-Glass equation. In each case, we saw a substantial reduction in the order of the k-variate model when compared to k-univariate models.

The opportunity for applications of the proposed RBFs is significant: various problems in nonlinear signal processing, optimal control, computer vision, pattern recognition, and prediction, such as the financial time-series problem. In future work, we will apply these functions for representing data on manifolds as graphs of functions (Broomhead & Kirby, 2000, 2001), as well as the low-dimensional modeling of dynamical systems (Broomhead & Kirby, 2005).

The algorithm proposed here exploits the existence of cross-correlation between the output time series. Hence, we expect the advantage of this approach to be especially significant in cases where the range data possess a related geometric or statistical structure.

Acknowledgments

This work is partially supported by NSF grants MSPA-MCS 0434351, ATM-530884, and DOD-USAF-Air Force FA-9550-08-1-0166. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the AFOSR.

References

Abarbanel, H., Gilpin, M. E., & Rotenberg, M. (2005). *Analysis of observed chaotic data*. New York: Springer.

Anderle, M., & Kirby, M. (2001). Correlation feedback resource allocation RBF. In *Proceedings of 2001 IEEE International Joint Conference on Neural Network* (pp. 1949–1953). Piscataway, NJ: IEEE.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation, 15*, 1373–1396.

Brand, M. (2003). Charting a manifold. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems, 15*. Cambridge, MA: MIT Press.

Brockwell, P. J., & Davis, R. A. (1991). *Time series: Theory and methods* (2nd ed.). Berlin: Springer.

Broomhead, D. S., & Kirby, M. (2000). A new approach for dimensionality reduction: theory and algorithms. *SIAM Journal of Applied Mathematics, 60*(6), 2114–2142.

Broomhead, D. S., & Kirby, M. (2001). The Whitney reduction network: A method for computing autoassociative graphs. *Neural Computation, 13*, 2595–2616.

Broomhead, D. S., & Kirby, M. (2005). Large dimensionality reduction using secant-based projection methods: The induced dynamics in projected systems. *Nonlinear Dynamics, 41*, 47–67.

Broomhead, D. S., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems, 2*, 321–355.

Buhmann, M. D. (2003). *Radial basis functions*. Cambridge: Cambridge University Press.

Coifman, R., & Lafon, S. (2006). Diffusion maps. *Applied Computational and Harmonic Analysis, 21*, 5–30.

Cook, D., Buja, A., Cabrera, J., & Hurley, C. (1995). Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics, 4*(3), 155–172.

Donoho, D. L., & Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *PNAS, 100*(10), 5591–5596.

Farmer, J. D., & Sidorowich, J. J. (1987). Predicting chaotic time series. *Physical Review Letters, 59*(8), 845–848.

Friedman, J. H., & Tukey, J. W. (1974). A Projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, C-23*(9), 881–890.

Harder, R. L., & Desmarais, R. N. (1972). Interpolation using surface splines. *Journal of Aircraft, 9*(2), 189–191.

Hardy, R. L. (1997). Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysical Research, 76*(8), 1905–1915.

Holmes, C. C., & Mallick, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation, 10*(5), 1217–1233.

Hundley, D. (1998). *Local nonlinear modeling via neural charts*. Unpublished doctoral dissertation, Colorado State University.

Hundley, D., Kirby, M., & Miranda, R. (1995). Spherical nodes in neural networks with applications. In S. H. Dagli, B. R. Fernandez. J. Ghosh, & R. T. Soundar Kumara (Eds.), *Intelligent engineering through artificial neural networks* (vol. 5, pp. 27–32). New York: American Society of Mechanical Engineers.

Hundley, D., Kirby, M., & Miranda, R. (1999). Empirical dynamical system reduction II: Neural charts. In K. Coughlin (Ed.), *Semi-analytic methods for the Navier–Stokes equations* (pp. 65–83). Providence, RI: American Mathematical Society.

Jamshidi, A. A. (2004). A new spatio-temporal resource allocation network (ST-RAN). Unpublished M.Sc. thesis, Colorado State University.

Jamshidi, A. A. (2008). *Modeling spatio-temporal systems with skew radial basis functions: Theory, algorithms and applications*. Unpublished doctoral dissertation, Colorado State University.

Jamshidi, A. A., & Kirby, M. J. (2006, June). Examples of compactly supported functions for radial basis approximations. In H. R. Arabnia, E. Kozerenko, & S. Shaumyan (Eds.), *Proceedings of the 2006 International Conference on Machine learning; Models, Technologies and Applications* (pp. 155–160). Las Vegas, NV: CSREA Press.

Jamshidi, A. A., & Kirby, M. J. (2007). Towards a black box algorithm for nonlinear function approximation over high-dimensional domains. *SIAM Journal of Scientific Computation, 29*(3), 941–963.

Jamshidi, A. A., & Kirby, M. J. (2008). Skew-radial basis functions for modeling edges and jumps. In *8th IMA International Conference on Mathematics in Signal Processing* (pp. 51–54). Essex, U.K.: Institute of Mathematics and Its Applications.

Jamshidi, A. A., & Kirby, M. J. (2010). Skew-radial basis function expansions for empirical modeling. *SIAM Journal of Scientific Computation, 31*(6), 4715–4743.

Kadirkamanathan, K., & Niranjan, M. (1993). A function estimation approach to sequential learning with neural networks. *Neural Computation, 5*(6), 954–975.

Karayiannis, N. B., & Mi, G. W. (1997). Growing radial basis neural networks: Merging supervised and unsupervised learning with network growth techniques. *IEEE Transactions on Neural Networks, 8*(6), 1492–1506..

Kennel, M. B., Brown, R., & Abarbanel, H. D. I. (1992). Determining embedding dimension for phase-space reconstruction using geometrical construction. *Physical Review A, 45*(6), 3403–3411.

Kirby, M., & Miranda, R. (1996). Circular nodes in neural networks. *Neural Computation, 8*(2), 390–402.

Lee, P. V., & Haykin, S. (2001). *Regularized radial basis function networks theory and applications*. New York: Wiley.

Liebert, W., Pawelzik, K., & Schuster, H. G. (1991). Optimal embeddings of chaotic attractors from topological considerations. *Europhysics Letters, 14*(6), 521–526.

Mackey, M. C., & Glass, L. (1977). Oscillation and chaos in physiological control systems. *Science, 197*, 287–289.

Mardia, K., Kent, J., & Bibby, J. (1980). *Multivariate analysis*. Orlando, FL: Academic Press.

Nabney, I. T., McLachlan, A., & Lowe, D. (1996). Practical methods of tracking nonstationary time series applied to real world data. In S. K. Rogers & D. W. Ruck (Eds.), *SPIE Proceedings Applications and Science of Artificial Neural Networks II* (pp. 152–163). Bellingham, WA: SPIE.

Packard, N. H., Crutchfield, J. P., Farmer, J. D., & Shaw, R. S. (1980). Geometry from a time series. *Physical Review Letters, 45*(9), 712–716.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation, 3*, 246–257.

Park, J., & Sandberg, I. W. (1993). Approximation and radial-basis-function networks. *Neural Computation, 5*, 305–316.

Platt, J. (1991). A resource allocating network for function interpolation. *Neural Computation, 3*, 213–225.

Powell, M. J. D. (1992) The theory of radial basis functions in 1990. In W. Light (Ed.), *Advances in numerical analysis* (pp. 105–210). New York: Oxford University Press.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(5500), 2323–2326.

Roweis, S., Saul, L., & Hinton, G. (2002). Global coordination of local linear models. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15*. Cambridge, MA: MIT Press.

Schaback, R., & Wendland, H. (2000). Characterization and construction of radial basis functions. In N. Dyn, D. Leviatan, D. Levin, & A. Pinkus (Eds.), *Multivariate*

approximation and applications: Eilat proceedings (pp. 1–24). Cambridge: Cambridge University Press.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*, 1299–1319.

Takens, F. (1980). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick; Proceedings of a symposium held at University of Warwick 1979–1980* (pp. 366–381). Berlin: Springer-Verlag.

Teh, Y. W., & Roweis, S. (2002) Automatic alignment of local representations. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15*. Cambridge, MA: MIT Press.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319–2323.

Vapnik, V. (1995). The nature of statistical learning theory. Berlin: Springer-Verlag.

Verbeek, J. J., Roweis, S. T., & Vlassis, N. (2004). Nonlinear CCA and PCA by alignment of local models. In S. Thrün, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing Systems, 16*. Cambridge, MA: MIT Press.

Wendland, H. (2005). *Scattered data approximation*. Cambridge: Cambridge University Press.

Whitney, H. (1936). Differential manifolds. *Annals of Mathematics, 37*(3), 645–680.

Yingwei, L., Sundararajan, N., & Saratchandran, P. (1997). A sequential scheme for function approximation using minimal radial basis function neural networks. *Neural Computation, 9*, 461–478.

Yingwei, L., Sundararajan, N., & Saratchandran, P. (1998). Performance evaluation of a sequential minimal radial basis function (RBF) neural network learning algorithm. *IEEE Transactions on Neural Networks, 9*(2), 308–318.

Received October 13, 2009; accepted May 22, 2010.