DISSERTATION

MODELING SPATIO-TEMPORAL SYSTEMS WITH SKEW RADIAL BASIS FUNCTIONS: THEORY, ALGORITHMS AND APPLICATIONS

Submitted by

Arthur(Arta) Amir Jamshidi

Department of Mathematics

In partial fulfillment of the requirements

for the degree of Doctorate of Philosophy

Colorado State University

Fort Collins, Colorado

Summer, 2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

COLORADO STATE UNIVERSITY

July 11, 2008

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY ARTHUR AMIR JAMSHIDI ENTITLED MODELING SPATIO-TEMPORAL SYSTEMS WITH SKEW RADIAL BASIS FUNCTIONS: THEORY, ALGORITHMS AND APPLICATIONS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Professor Eugene Allgower

Professor Rick Miranda

Professor Louis Scharf

Professor Michael Kirby, Adviser

Professor Gerhard Dangelmayr, Department Head

ABSTRACT OF DISSERTATION

MODELING SPATIO-TEMPORAL SYSTEMS WITH SKEW RADIAL BASIS FUNCTIONS: THEORY, ALGORITHMS AND APPLICATIONS

The discovery of knowledge in large data sets can often be formulated as a problem in nonlinear function approximation. The inherent challenge in such an approach is that the data is often high dimensional, scattered and sparse. Given a limited number of exemplars one would like to construct models that can generalize to new regions or events. Additionally, underlying physical processes may not be stationary and the nature of the nonlinear relationships may evolve. Ideally, a good model would be adaptive and remain valid over extended regions in space and time.

In this work we propose a new Radial Basis Function (RBF) algorithm for constructing nonlinear models from high-dimensional scattered data. The algorithm progresses iteratively adding a new function at each step to refine the model. The placement of the functions is driven by one or more statistical hypotheses tests that reveal geometric structure in the data when it fails. At each step the added function is fit to data contained in a spatio-temporally defined local region to determine the parameters, in particular, the scale of the local model. Unlike prior techniques for nonlinear function fitting over scattered data, the proposed method requires no *ad hoc* parameters and it behaves effectively like a black box. Thus, the number of basis functions required for an accurate fit is determined automatically by the algorithm. An extension of the algorithms to multivariate case, i.e., the dimension of the range of the mapping is greater

or equal to two, is also carried out. This approach produces more parsimonious models by exploiting the correlation among the various range dimensions. The convergence properties of the algorithms are shown from different prospectives.

To further enhance the order and conditioning of the models we introduce several new compactly supported RBFs for approximating functions in $L^P(\mathbb{R}^d)$ via overdetermined least squares. We also propose a skew-radial basis function expansion for the empirical model fitting problem to achieve more accuracy and lower model orders. This is accomplished by modulating or skewing, each RBF by an asymmetric shape function which increases the number of degrees of freedom available to fit the data. We show that if the original RBF interpolation problem is positive definite, then so is the skew-radial basis function when it is viewed as a bounded perturbation of the RBF.

We illustrate the utility of the theoretic and algorithmic innovations via several applications including modeling data on manifolds, prediction of financial and chaotic time-series and prediction of the maximum wind intensity of a hurricane. In addition, the skew-radial basis functions are shown to provide good approximations to data with jumps. While the algorithms presented here are in the context of RBFs, in principle they can be employed with other methods for function approximation such as multilayer perceptrons.

Arthur(Arta) Amir Jamshidi
Department of Mathematics
Colorado State University
Fort Collins, Colorado 80523
Summer, 2008

To my mother

ACKNOWLEDGEMENTS

I would like to express my sincerest thankfulness to my adviser, Prof. Michael Kirby, for his continuous advice, encouragement to perform independent research, and for his constructive criticisms throughout my Ph.D. program. I have learned a lot from his research style and enthusiasm. These, along with his friendship, made a pleasant and productive research experience for me which surely impacts my career.

I would like to sincerely thank my committee members, Prof. Eugene Allgower and Prof. Rick Miranda in the Department of Mathematics and Prof. Louis Scharf in the Department of Electrical and Computer Engineering, for their invaluable assistance, advice and support.

I would like to acknowledge partial financial support from the National Science Foundation Award DMS-9973303 and grant numbers DMS-0530884, DMS-0434351 and ATM-0715426 as well as the DOD-USAF-Office of Scientific Research under contract FA9550-04-1-0094. In addition, a kind donation by Siemens Corporate Research (SCR) has supported early parts of this work.

Special thanks is due to my sisters, Anahita and Mandana and my brother, Ali, for their love and support.

Finally, I would like to express my deep gratitude to my mother and father, who are the first teachers in my life and sources of inspiration.

Table of Contents

List of Algorithms

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Knowledge Discovery and Empirical Modeling

The process of knowledge discovery may be viewed as an interacting combination of human experience, cognition, observation and analysis. Whether knowledge is acquired by researchers trained in the application of the Scientific Method or more informally driven by everyday experiences, the process almost always involves the collection of data in some form or another. This description loosely depicts how progress is made through understanding our environment and applies equally well to knowledge discovery in, e.g., the Bronze Age, the Renaissance and for the purposes of our current interest, the Information Age. The field of information technology coupled with the digital revolution associated with the widespread availability of inexpensive data collection devices has completely transformed the landscape of knowledge discovery.

Researchers are now confronted by large sets of empirical observations of phenomena for interest often with no domain specific theory, e.g., a conservation law, to guide their investigation. The challenge is to determine relationships amongst the data that reveal phenomena, or knowledge, of interest. We will argue from a geometric perspective below that these relationships are often nonlinear and require the application of mathematical ideas to extract them. This motivates the basic topic of this dissertation, i.e., the problem of extracting nonlinear relationships in large high-dimensional and potentially scattered data sets.

1.2 Nonlinearity

One can provide a long list of examples of nonlinear phenomenon in nature. Chaos, for example, requires both nonlinearity and three dimensions to exist. Additionally, we can refer to evidence such as nonlinear optical phenomena, nonlinear wave interactions, nonlinear fluid-structure interactions and many others. However, we propose to motivate the need for nonlinear models from a perspective involving mathematical constructions that are inherently nonlinear.

For example, the modeling of a manifold or the representation of data as a graph of a function, $(x, f(x))$, with f being the nonlinear relation. Related to these problems is the implementation of Whitney's theorem, [29, 30] which proves, under certain circumstances, the existence of a nonlinear inverse mapping for (perfect) manifold reconstruction. Additionally one may envision constructing an atlas of charts from empirical data as a representation. Problems of this type arise in sampling theory on manifolds. Takens' theorem also provides evidence of a nonlinear mapping from a sampled scalar value to a representation of the data manifold up to a diffeomorphism. This motivates another large class of problems, i.e., prediction via time-delayed embeddings, i.e., $x_{n+1} = f(x_n, ..., x_1)$. Additionally, we may consider mappings between two manifolds, $\mathcal{M}$ and $\mathcal{N}$ given by $\mathcal{M} = f(\mathcal{N})$, [16]. In terms of mathematical modeling of physical systems we are concerned with estimating f in the system $x' = f(x)$, where x' represents the time evolution and the nonlinear relation f is the vector field providing instructions [31]. In addition to these mathematical constructions, there are the general problems of classification and regression, see e.g., [145, 65, 134]. These are more classically the arenas in which researchers have developed approaches for nonlinear approximation. We expand on our motivations below.

1.3 Data on Manifolds

Our geometric viewpoint for this problem arises from the fact that many apparently complicated physical phenomena exhibit *self-organization*, a tendency that reveals itself by the formation of coherent structures or patterns in data. The effect of the self-organization may be interpreted as a tendency for the data from the process to cluster in small volumes of the total space. From a state space perspective this coherency suggests the possible existence of a manifold, or a fractal set embedded in a manifold, of dimension much smaller than the ambient space. In this setting, the emphasis is on developing a practical mathematical theory and algorithms for modeling the data, i.e., constructing qualitatively equivalent representations, for example, up to a diffeomorphism [89]. Our main interest in this problem relates to representing data on a manifold as the graph of a function [29, 30] and the reduction of dynamical systems [31].

The problem of the representation of data on manifolds is further suggested by Taken's embedding theorem. Roughly speaking, Taken's theorem states that a scalar observable from data on an m dimensional manifold may be time-delay embedded into $2m+1$ dimensions effectively reconstructing the manifold up to a diffeomorphism, [135]. Interesting questions may be addressed within this framework, e.g., are two scalar observables generated by the same process? Can we find smooth mappings between their reconstructed manifolds?

1.4 Whitney's Theorem

Given an assumption data may reside on a manifold, Whitney's theorem provides a guide to data modeling which includes the construction of a nonlinear inverse mapping. Whitney's theorem states that an m dimensional manifold may be embedded in a linear subspace of dimension $2m + 1$, [146]. Further, the reduction is linear and the reconstruction is nonlinear. A new dimension reduction approach that employs *secant based projections* is developed in [29, 30]. This is a global method of dimension reduction. New

algorithms and optimality criteria for computing the good parameterizing subspaces and hence estimates for the dimension of the manifold are given in [29, 30]. It has been shown that the reduction mapping should be bilipschitz (and hence dimension preserving) with optimized Lipschitz constants. These ideas are illustrated on sample PDEs, such as the Kuramoto-Sivashinksy Equation where known manifold dimensions were computed accurately. Here the appropriate modes, and hence parameterizing scales, for representing the problem, are determined such that the data in the reduced space is (at least) a diffeomorphic copy of the raw data. To be able to go back to the space of raw data nonlinear maps are required. There is no explicit construction of such maps in Whitney's theorem. We propose to use RBFs to construct such maps. High quality nonlinear functions are needed to more fully realize the implementation of Whitney's theorem on real data.

1.5 Manifold Learning and Charts

Manifold learning is a growing area of research given the importance of geometry–not just statistics–in the analysis of data. In short, manifold learning is concerned with the local representation of data as an atlas of charts, see, e.g., [91, 62, 63]. These data models coalesce local (parametric or nonparametric) models to obtain a globally valid nonlinear embedding function. More recently, nonlinear mappings from a high dimensional sample space to a low dimensional vector space effectively recovering an internal coordinate system for the manifold from which the data is sampled, is considered in [27]. This algorithm employs a mapping that preserves local geometric relations in the manifold and is pseudo-invertible. Decomposition of the sample data into locally linear low-dimensional patches, merging these patches into a single low dimensional coordinate system via stochastic optimization processes and computation of the forward and reverse mappings between the sample and coordinate spaces is also studied in [27]. In [127], 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 which 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 [136]. In this method one local model is centered on each training point so its scaling is the same as that of local linear embedding (LLE), [126], and Isomap, [137]. Another approach for manifold learning based on aligning mixtures of linear models is proposed in [139].

In general, in this setup we expect that our data set X, although lying in the high-dimensional space $\mathbb{R}^N$, will have a much lower dimension k; and therefore we would obtain an optimal dimensionality-reducing mapping if we could construct the analogue of "charts" on X. Although we expect that our data sets will not always be manifolds, we have found this chart construction to be an excellent paradigm. We propose to use optimal *linear* charts functions ϕ_i, in the sense that ϕ_i is a linear mapping from the ambient space $\mathbb{R}^N$ containing the data set X to the model space $\mathbb{R}^k$. In this sense the mapping ϕ_i can be viewed as an optimal projection locally, and there is a large literature on the construction of such projections; for a summary see [89]. The requirement that the projection function be a chart mapping, i.e., that it have a local inverse, is made by enforcing an explicit construction of the inverse mapping ϕ_i^{-1}. The construction of ϕ_i^{-1} is achieved via a radial basis function, the training of which is effected by learning the identity function on U_i, after composing with the pre-determined projection mapping ϕ_i. This local method based on constructing an altas of charts affords the encoding dimension m [62, 63]. This prior work employs a tree-based approach to partition the

manifold and singular value decomposition is used to estimate local dimension and use MLPs for the inverse mappings.

The basic algorithm described above again has the need for the construction of a nonlinear mapping from data at its heart. We propose that the highly developed RBF algorithm presented in this dissertation will provide an important component to work in manifold learning.

1.6 A Brief History

In this section we provide a brief overview of data fitting. Details and references to current literature are provided in chapters that follow.

The beginnings of empirical data fitting may be traced to Gauss's work on using least squares to construct linear models. In general, linear models are only capable of representing phenomena where the principle of superposition holds true. Additionally, approximating nonlinear phenomena with piecewise linear models suffer from lack of smoothness. More recently A. Einstein used correlation structure of data for time series analysis, [39, 50, 150]. Other instances of data analysis in the beginning of the twentieth century can be found in, e.g., [107, 109, 59]. Over the last two decades we have seen a tremendous growth in this area motivated by new ideas for computing nonlinear models, see, e.g., [32, 121, 122]. Certainly neural networks designed for function approximation and the solution of the curse of dimensionality have attracted significant attention [145, 128]. In nonlinear models one of the most critical questions is associated with model order determination, i.e., how many nonlinear functions are needed to minimally represent the data. In general there are three main approaches to determine the complexity of a neural network for function approximation: penalized likelihood, [133, 4], predictive assessment, [55], and growing and pruning techniques, [47, 43, 38]. In this dissertation we focus on constructing a growing and pruning technique for nonlinear function approximation. We employ hypotheses tests to evaluate if any geometric structure resides in the residuals

of the data. The autocorrelation test for IID noise provides an excellent tool for this approach.

1.7 Open Problems to Be Addressed

In this dissertation a comprehensive set of questions associated with nonlinear model fitting from data are addressed, including:

- How can the need for ad hoc, or user adjusted, parameters be eliminated?
- How should algorithms differ when the data is temporal, spatial or spatio-temporal?
- What are the mathematical properties of convergence of the algorithm?
- What mathematical criteria are important to design functions with improved generalization?
- How can we better approximate data with jumps, or severely asymmetric data in general?
- What are optimal ways to extend the algorithm to ranges in higher dimensions?
- How do these algorithms work on real world data?
- How can we adapt the models in real time?

1.8 The Organization of the Dissertation

The organization of this dissertation is as follows: Chapter 2 provides the univariate black box radial basis function algorithm for nonlinear function fitting. Relevant literature is reviewed. The superior performance of the algorithm in comparison to the current state of the art is shown. Chapter 3 introduces new compactly supported RBFs for least square function fitting. Connections to the current literature of compactly supported RBFs is given. Chapter 4 introduces skew-radial basis functions and shows

their superior performance in function fitting. In Chapter 5, the theoretical and technical background of the algorithm is described and its convergence properties are investigated. A multivariate extension to the algorithm is provided in Chapter 6. Finally, Chapter 7 summarizes the contributions of this work and discusses future work.

This dissertation includes chapters that are actual papers that have appeared [72, 71], have been submitted [69] or about to be submitted [70, 68]. Since these papers are designed to be self-contained there is some minor duplication in the presentation.

Chapter 2

TOWARDS A BLACK BOX ALGORITHM FOR NONLINEAR FUNCTION APPROXIMATION OVER HIGH-DIMENSIONAL DOMAINS

Abstract We propose an algorithm for constructing nonlinear models from high-dimensional scattered data. 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 reveals geometric structure when it fails. At each step the added function is fit to data contained in a spatio-temporally defined local region to determine the parameters, in particular, the scale of the local model. Unlike currently available techniques for nonlinear function fitting over scattered data, the proposed method requires no *ad hoc* parameters. Thus, the number of basis functions required for an accurate fit is determined automatically by the algorithm. We illustrate the approach using several illustrative problems including modeling data on manifolds and the prediction of financial time-series. The algorithm is presented in the context of radial basis functions but in principle can be employed with other methods for function approximation such as multi-layer perceptrons.

2.1 Introduction

The problem of extracting nonlinear relationships in large high-dimensional scattered data sets is of central importance across fields of science, engineering and mathematics. The beginnings of empirical data fitting may be traced to Gauss's work on

using least squares to construct linear models. Over the last two decades we have seen a tremendous growth in this area motivated by new ideas for computing nonlinear models, see, e.g., [32, 145, 128, 121, 122]. Today, diverse areas such as machine learning, optimal control, and mathematical modeling of physical systems often rely significantly on the ability to construct relationships from data. Subsequently there have been a multitude of applications including financial time-series analysis, voice recognition, failure prediction and artificial intelligence all of which provide evidence for the importance of nonlinear function approximation algorithms. Our interest in this problem relates to representing data on a manifold as the graph of a function [29, 30] and the reduction of dynamical systems [31].

A common element in empirical data fitting applications is that the complexity of the required model including the number and scale of representation functions is not known *a priori* and must be determined as efficiently as possible. A variety of approaches have been proposed to determine the number of model functions, i.e., the model order problem. A generally accepted measure of quality of such data fitting algorithms is that the resulting models generalize well to testing data, i.e., data associated with the same process but that was not used to construct the model. This requirement is essentially that the data not be overfit by a model with too many parameters or underfit by a model with too few parameters.

One general approach to this problem is known as *regularization*, i.e., fitting a smooth function through the data set using a modified optimization problem that penalizes variation [138]. A standard technique for enforcing regularization constraints is via cross-validation [51, 141]. Such methods involve partitioning the data into subsets of training, validation and testing data; for details see, e.g., [55].

Additionally, a variety of model growing and pruning algorithms have been suggested, e.g., the upstart algorithm in [47], cascade correlation [43], optimal brain damage [38] and the resource allocating network (RAN) proposed by Platt [118]. Statistical

methods have also been proposed that include, e.g., Akaike information criteria (AIC), Bayesian information criteria (BIC), minimum description length (MDL) [133, 4] and Bayesian model comparison [101]. In [124, 104, 58] the issue of selecting the number of basis functions with growing and pruning algorithms from a Bayesian prospective have been studied. In [7], a hierarchical full Bayesian model for RBFs is proposed. The maximum marginal likelihood of the data has also been used to determine RBF parameters [112]. For a more complete list of references the reader is referred to [66].

In general, model order determination via both regularization and growing and pruning algorithms can be computationally intensive and data hungry. More importantly, however, is that these algorithms do not explicitly exploit the geometric and statistical structure of the residuals during the training procedure. In addition, many algorithms in the literature require that anywhere from a few to a dozen *ad hoc* parameters be tuned for each data set under consideration.

This chapter presents an approach for the model order determination problem free of *ad hoc* parameters. The algorithm is based on detecting any structure in the model residuals. As in previous work [5, 6] on which this algorithm is based, such structure is quantified via a statistical hypothesis test to determine whether the residuals are IID. The main contribution of this chapter is the implementation of the algorithm over higher dimensional domains using a spatio-temporal ball rather than a temporal window for constructing local training sets. This innovation is particularly critical if the data is periodic or quasi-periodic or, more generally, resides on a manifold. Further, a new initialization procedure is adopted that greatly accelerates the optimization algorithms. To illustrate the absence of *ad hoc* parameters, diverse data sets are fit without making any changes to the algorithm. In particular, no parameters are varied, or tuned, across data sets.

While we use radial basis functions to demonstrate the algorithms, the methodology holds for other function fitting problems. In particular, these methods may be applied

directly to multi-layer perceptrons. Also, we restrict the scope of the presentation to the case of batch data and will present extensions to the on-line algorithm elsewhere.

The organization of this chapter is as follows: Section 2.2 reviews the radial basis function approach for multiscale approximation of scattered data. Section 2.3 develops the background for the hypothesis test for IID noise. Section 2.4 introduces the algorithm including the concept of spatio-temporal windowing and appropriate stopping criteria. Section 2.5 demonstrates the performance and robustness of the system using different data sets. Finally, Section 2.6 provides some concluding remarks and discusses future work. For the purposes of keeping this chapter self-contained we present some background details of the algorithm proposed in [5, 6] and [66].

2.2 Radial Basis Functions

Radial Basis Functions (RBFs) were introduced for the function approximation problem as an alternative to multilayer perceptrons [32]. Part of their appeal is the variety of efficient algorithms available for their construction. In the extreme, the basis functions may be selected randomly (following the distribution of the data) with fixed scales. In this instance the resulting optimization problem is simply an over-determined least squares problem to determine the expansion coefficients. One may improve on this approach at modest expense by employing a clustering algorithm to determine the basis function centers [108]. Furthermore, RBFs may be adapted with rank one updates or down-dates [111, 113]. Over the years RBFs have been used successfully to solve a wide-range of function approximation and pattern classification problems [22, 66]. More recently, RBFs have been proposed as a tool for the simulation of partial differential equations, see, e.g., [1].

An RBF expansion is a linear summation of special nonlinear basis functions. In general, an RBF is a mapping $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ that is represented by

$$f(x) = Ax + \alpha_0 + \sum_{k=1}^{K} \alpha_k \phi_k(\|x - c_k\|_W), \tag{2.1}$$

where x is an input pattern, ϕ_k is the kth RBF centered at location c_k, and α_k denotes the weight for kth RBF and A is an $m \times n$ matrix. The term W denotes the parameters in the weighted inner product

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

The term $Ax + \alpha_0$ affords an affine transformation of the data and is useful so that the nonlinear terms are not attempting to fit flat regions. More general polynomials may be used for this purpose [130]. As usual, the dimensions of the input n and output m are specified by the dimensions of the input-output pairs.

The general training problem for the RBF is the determination of the unknown parameters: $\{A, \alpha_k, c_k, W, K\}$. The focus of this chapter is to determine an optimal value for K, i.e., the model order. Of course optimizing K depends on high quality values for the remaining parameters and we propose algorithms for this purpose.

The requirements of the RBF expansion are the same as standard data fitting problems, i.e., given a set of L input-output pairs $\{(x_l, y_l)\}_{l=1}^L$, $\mathcal{X} = \{x_l\}_{l=1}^L$ and $\mathcal{Y} = \{y_l\}_{l=1}^L$, the goal is to find the underlying mapping f such that $y_l = f(x_l)$. In the standard situation we have more data than equations so we can't expect to satisfy each equation exactly. Thus the problem is to minimize the cost function

$$E(A, \alpha_k, c_k, W, K) = \frac{1}{2} \sum_{l=1}^{L} \| f(x_l) - y_l \|^2,$$

where now the metric $\| \cdot \|$ is generally the Euclidean inner product and is distinct from the W-weighted inner product used to compute the contributions of the basis functions.

One of the attractive features of RBFs is the variety of basis functions available for the expansion. In particular, these functions come in both local and global flavors and include

$$\phi(r) \in \{\exp(-r^2), r, r^2 \ln r, r^3\}.$$

These functions satisfy the criteria of RBFs as described in [122] and the associated approximation theorem states that if $\mathcal{D}$ is a compact subset of $\mathbb{R}^d$, then every continuous

real valued function on $\mathcal{D}$ can be approximated uniformly by linear combinations of RBFs with centers in $\mathcal{D}$. In this study we restrict our attention to (local) Gaussian RBFs, i.e., $\phi(r) = \exp(-r^2)$ but the algorithm works in principle for any admissible RBF. We have also recently proposed several candidates for compact RBFs that have excellent conditioning properties associated with the overdetermined least squares problem [71] or Chapter 3.

We note that in function approximation problems one must distinguish whether the data, for which a model is desired, is available at the outset or whether the data becomes available as the model is being built. In keeping with standard terminology we refer to these problems as *batch* and *on-line* training, respectively. In this chapter we are specifically interested in learning input/output relationships from batch data.

2.3 Testing for Structure in Model Residuals

As indicated earlier, one can infer essential information about an empirical model by examining its residuals. Following [5, 6], the proposed algorithm functions on the premise that if there is structure remaining in the residuals, then one or more additional basis functions should be added. On the other hand, if there is no discernible structure then this is tantamount to a stopping criterion.

We seek to model data that may be viewed as the superposition of a signal and IID noise observing that if data is noise free then one may simply add IID noise to employ this algorithm. In view of this, we may expect the model residuals to be IID while the model itself represents the geometric structure of the data. Hence, an indication that an RBF model is unsatisfactory is the failure of the residuals to satisfy an hypothesis test for IID noise. This observation forms the basic idea for the stopping criterion used in this research work. The primary advantage of such a criterion is that the hypothesis test from which it stems does not involve any *ad hoc* parameters that require adjustment.

Now we outline the IID test for determining structure in the residuals and its associated algorithm.

2.3.1 Statistical Background of Test for IID Noise

The model residual for the nth data point is defined as

$$e_n = y_n - f(x_n).$$

Following [6], we denote the set of residuals for a model of order K, as

$$R^K = \{e_n\}_{n=1}^{L}, \tag{2.2}$$

where L is the cardinality of the training set. The standard definition for the sample autocorrelation function, $\widehat{\rho}(h)$, (ACF) for a set of residuals $e_1, e_2, e_3, ..., e_L$ with sample mean $\bar{e}$ and lag h is defined as

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}, \tag{2.3}$$

where $-L < h < L$, and

$$\widehat{\gamma}(h) = \frac{1}{L} \sum_{i=1}^{L-|h|} \alpha(h, e_i). \tag{2.4}$$

As proposed in [5, 6] we decompose the ACF into its component terms

$$\alpha(h, e_i) = (e_{i+|h|} - \bar{e})(e_i - \bar{e}). \tag{2.5}$$

For a fixed lag h the quantity $\alpha(h, e_i)$ is the contribution of the ith residual to the autocorrelation function. Later we focus on this quantity α and will illustrate that it reveals critical information concerning where new basis functions should be placed. Given its importance, we refer to this term as the autocorrelation contribution, or ACC [5, 6].

2.3.2 IID Hypothesis Test

As indicated above, 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 IID. The relevant theorem from statistics states that for large L, the sample autocorrelations of an IID sequence $U_1, U_2, ..., U_L$ with finite variance are

approximately IID with normal distribution with mean zero and variance $\frac{1}{L}$, $N(0, \frac{1}{L})$, [28] p. 222. Hence, if $u_1, u_2, ..., u_n$ is a realization of such an IID sequence, then 95% of the sample autocorrelations should fall between the bounds

$$\frac{-1.96}{\sqrt{L}} < \widehat{\rho}(h) < \frac{1.96}{\sqrt{L}}. \tag{2.6}$$

Note that if the underlying model is known, there is a more accurate bound for this test, described in Section 9.4 of [28]. Therefore, if one computes the sample autocorrelation 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 IID hypothesis is rejected. This test can equivalently be written in terms of χ^2 distribution. Given

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

it has been shown in [28] 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

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

2.3.3 Additional Testing Possibilities

To show rigorously that a sequence of random variables is truly IID, higher moments (if they exist) also need to be considered. In particular, if a sequence of random variables is IID, then any function of the random variables is also IID. Thus, the autocorrrelation function (ACF) of not only the sequence of random variables, but also, e.g., squares, cubes and the absolute values must also pass the above given test. For simplicity in this chapter we only consider the ACF of the raw residuals. Although this limits our ability to conclude the residuals are strictly IID, the results suggest this test is already quite powerful. Note that the test as implemented does indeed provide a necessary and sufficient condition for a sequence to be white noise.

Algorithm 1 RBF Algorithm using Spatio-Temporal Ball

$ran\_flag = 1$, $K = 0$
while $ran\_flag = 1$ **do**
  evaluate the RBF on the training data set $\{f(x_n)\}_{n=1}^{L}$
  compute the model error $\{e_n\}_{n=1}^{L}$
  compute component contributions $\alpha(h, e_i) = (e_{i+|h|} - \bar{e})(e_i - \bar{e})$
  compute ACF for all $0 < h < L$
  if the autocorrelation test is rejected **then**
    compute h^* via equation $h^* = \arg\max \widehat{\gamma}(h), h > 0$ and
    compute $x^* = x_{i^*} = e^{-1}(e_{i^*})$ where $i^* = \arg\max_{i=1,...,n-h^*} \alpha(h^*, e_i)$
    compute the ACC function, $\beta_i = \alpha(h^*, e_i)$, $i = 1, ..., n - h^*$
    optional: denoise the ACC function
    find the right and left local minimizers of 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 $v = [c_0, \sigma_0, \alpha_0]^T$
    solve $E(v) = \min_{v} \| h(x; v) - y \|_2^2$, where $x \in \mathcal{X}_{local}$
    $K = K + 1$
  else
    $ran\_flag = 0$
  end if
  compute confidence, RMSE and $\widehat{\gamma}(h^*)$ of the current model on the training set
end while

Lastly, we remark that there are alternatives to the test described above based on the autocorrelation function for testing IID or white noise. Other such tests include the difference-sign test, the rank test, and a test based on turning point [28]. These tests might also be applied as stopping criteria individually, or in conjunction with the current test based on the autocorrelation function. Another route for improvement may be possible by considering the relationship between the data inputs and the residuals of the outputs [99].

2.4 RBF Algorithm using Spatio-Temporal Ball

In this section we present the details of the proposed batch algorithm. The question of whether a new basis function should be added is answered by the IID test. We shall see

that this test also indicates where the new basis function should be added. We introduce the concept of a space-time ball for defining local regions.

Initially, the residuals of the model are equal to the original data. A step in this algorithm consists of evaluating the residuals and determining whether they indicate that a new basis function should be added. Pseudocode is provided in Algorithm 1 and we now proceed to describe the details.

2.4.1 Determining Optimal Locations of New RBFs

If the autocorrelation test indicates that the data is not IID, then the next requirement is to determine where the new basis function should be located to optimally reduce the structure in the model residuals. Following, [5, 6], we look for the point in the domain that makes the largest contribution to the ACF. This is accomplished by observing that the residuals are associated with the data in the domain in a one-to-one manner, i.e., there is a mapping, say ψ, from a data point to its residual of the form

$$e_i = \psi(x_i).$$

Thus, by identifying the residual associated with the largest contribution to the ACF we may identify the location in the domain where the basis function should be added. Note that if the maximum contribution value to the ACF function is shared by more than one residual, a single basis function would be added, selected from the residuals at random. To actually find this point first we determine the exact lag for which the autocorrelation function, $\widehat{\gamma}(h)$ reaches its maximum value h^*, i.e.,

$$h^* = \arg\max_{h>0} \widehat{\gamma}(h). \tag{2.7}$$

Note that we only consider lags $h > 0$. If we allowed $h = 0$, then the term $\alpha(0, e_i)$ would always have the maximum contribution (for some i) to the ACF and the method would be similar to those that use maximum magnitude of residuals as a criterion to allocate new basis functions critically losing the spatial component of the diagnostic.

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

$$i^* = \arg \max_{i=1,\ldots,n-h} \alpha(h^*, e_i). \tag{2.8}$$

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 will refer to this center location as x^*.

2.4.2 Spatio-Temporal Windowing

Now that the center of the new basis function has been found as described above it is necessary to determine what data should be used to determine the scale and weight of the new RBF. Consider the function $\beta_i = \alpha(h^*, e_i)$. The index i is inherited from the data labels and in the case of a time-series corresponds to a time ordering. In practice, if we plot β_i as a function of i we observe that the values of β_i decrease away from i^* which is what one expects since this index was selected given it corresponded to a local maximum in Equation (2.8). How quickly the values of β_i decrease for both $i > i^*$ and $i < i^*$ is a property of the scales of the data and the model.

For simplicity, we assume that β_i decreases monotonically for both increasing and decreasing values of i until local minima are reached 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^*. For time-series data it is apparent that spatial neighbors may not be temporal neighbors. Hence, this spatial-temporal windowing has the potential to capture substantial training data that would otherwise be ignored. For periodic, or quasi-periodic data we have found that space-time windowing is essential. As observed in [5, 6], if β_i does not decrease monotonically away from the peak then a small amount of filtering can be employed to recover monotonicity. Note that while no smoothing of β_i was required to accurately determine $\mathcal{X}_{local}$ for the examples in this chapter, it may be necessary in practice; see [5, 6, 66] for details.

2.4.3 Updating the Model

We may write the RBF expansion as consisting of K adapted terms and one new term, i.e.,

$$f^{(K+1)}(x) = f^{(K)}(x) + h(x; v),$$

where $v = [c, \sigma, \alpha]^T$ is the vector of parameters to be optimized. The new term $h(x; v)$ is initialized as

$$h(x; v) = \alpha_0 \phi(\|x - c_0\|_W).$$

The initialization of the center c_0 is at the point of most structure according to our test, i.e., $c_0 = x^*$. The vector of widths σ is very effectively initialized using the diagonal

elements of the covariance matrix of the local data,

$$\sigma_0 = \sqrt{diag(cov(\mathcal{X}_{local}))}. \tag{2.9}$$

Note here that $W = diag(\sigma_0)$, where the diagonal matrix W is the weighting term in the inner product and has its diagonal elements equal to the components in σ. This initialization of the weights has proven to be extremely valuable in accelerating the convergence of the conjugate gradient iteration. The initial value for the weight, α_0, is calculated via least squares using the initial values for center location and widths.

Once the data in $\mathcal{X}_{local}$ has been determined, the scale (width), weight and the center location of the new basis function are optimized using the conjugate gradient method with cost function

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

We only optimize the parameters associated with the new RBF and keep the others fixed.

2.4.4 Stopping Criteria

One of the most critical components of any growing algorithm is the stopping criterion. Once the optimization of the new term $h(x; v)$ is complete, a new set of residuals is computed over the entire training set. For an algorithm based on the IID test of the residuals the most natural stopping criterion to use is the 95% confidence level. Note that it is possible to interpret the level of confidence as a parameter but we do not vary it. We present several examples in the section on numerical experiments that illustrate the effectiveness of this stopping criterion.

For all applications we have computed the Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{T} \sum_{i=1}^{T} e_i^2}, \tag{2.10}$$

where T is the number of test points. To compare our results with other work we also compute two forms of the Normalized Prediction Error (NPE), namely,

$$NPE_1 = \frac{\sum_{i=1}^{T} | e_i |}{\sum_{i=1}^{T} | y_i - \bar{y} |} \tag{2.11}$$

and

$$NPE_2 = \frac{\sum_{i=1}^{T} e_i^2}{\sum_{i=1}^{T} (y_i - \overline{y})^2}. \tag{2.12}$$

It is also interesting to note that the quantity $\widehat{\gamma}(h^*)$, i.e., the total contribution to the autocorrelation function at lag h^*, monotonically decreases with the number of basis functions and becomes very flat when the algorithm has converged. Hence, this quantity could be used as a sort of statistical no progress criterion. We have compared the usual *no progress* criterion on the root mean square error with the idea of a no progress criterion on $\widehat{\gamma}(h^*)$ and have found the latter to be more robust. Although we did not need to use either as stopping criteria for our applications it is possible they could be useful for this purpose with other data sets. With this in mind, we will illustrate their behavior in the numerical experiments that follow.

2.5 Numerical Examples

Here we present several applications to demonstrate the performance of the algorithm in higher dimensional domains. The successful extension of this algorithm from one to higher dimensional domains required the introduction of the notion of a space-time window. Here we illustrate the impact of this idea on several applications. Note that throughout all the following examples the same code was employed, in particular, there were no parameters that were adjusted or tuned to the data set. Applications to one-dimensional domains for a variety of noise levels and types have been explored in [66].

2.5.1 A Simple Manifold

In this example we illustrate the representation of data on a manifold as the graph of a function. We employ the *Pringle* data set, shown in Figure 2.1, named as such given its similarity to the boundary of a potato chip; see also [30, 29]. The task is to construct a mapping from an (x, y) value in the plane to its corresponding z value on the

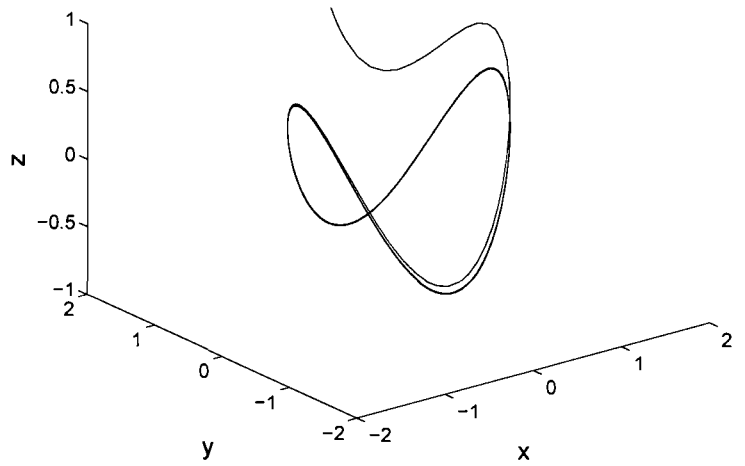


Figure 2.1: Plot of a typical Pringle set with $\lambda = 1$ and $\omega = 0.5$.

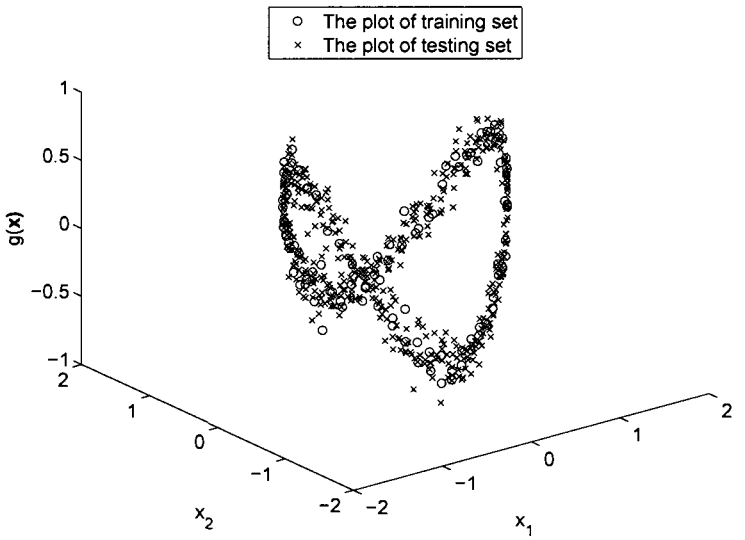
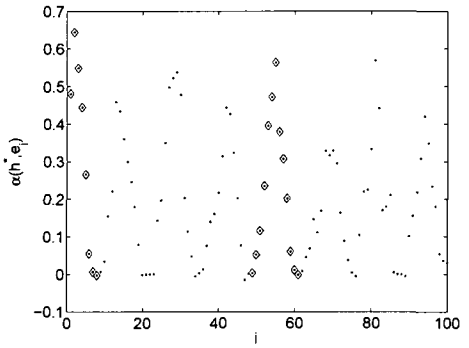
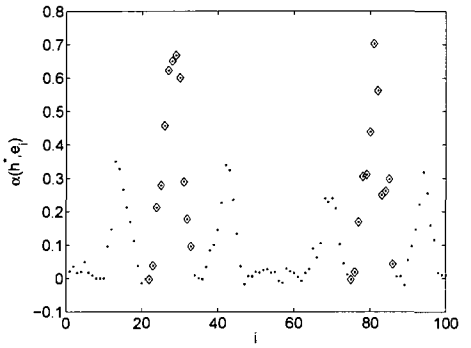


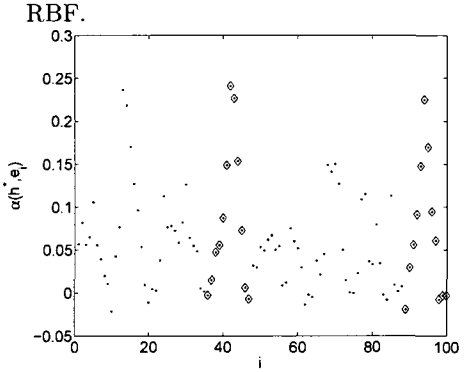
Figure 2.2: Plots of the training and testing data sets. The solution to the dynamical system is corrupted with Gaussian noise with STD of 0.1. There are 54 data points in one cycle.



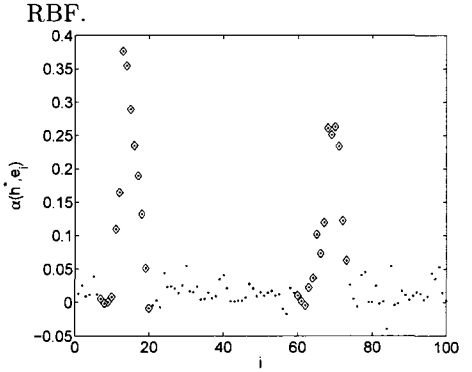
(a) The ACC function and the selected points to construct the first RBF.



(b) The ACC function and the selected points to construct the second RBF.

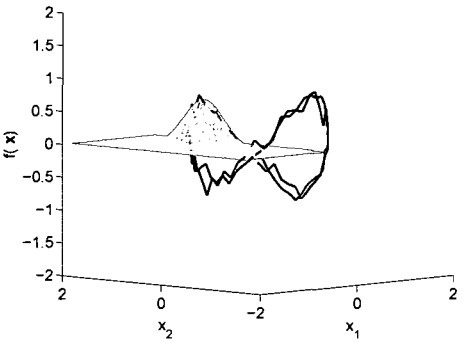


(c) The ACC function and the selected points to construct the third RBF.

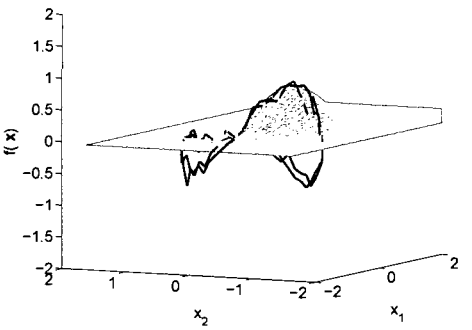


(d) The ACC function and the selected points to construct the fourth RBF.

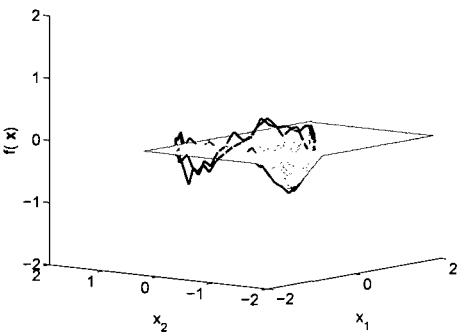
Figure 2.3: The plots of ACC functions for the four major basis functions.



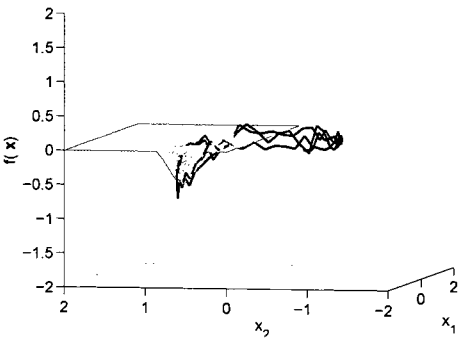
(a) The first RBF in relation with the training data set.



(b) The second RBF in relation with the training data set.



(c) The third RBF in relation with the training data set.



(d) The fourth RBF in relation with the training data set.

Figure 2.4: The primary four radial basis functions allocated by the algorithm. The residuals of the four mode model pass the IID test.

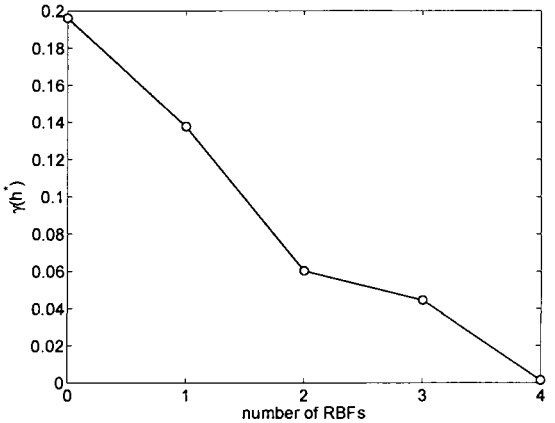
Pringle. Thus, we are fitting the graph of a function from $\mathbb{R}^2$ to $\mathbb{R}$. Such graph fitting problems are at the center of the Whitney's manifold embedding theorem where $2m+1$ dimensional domains suffice (in general) to write m dimensional manifolds as graphs; see [30, 29] for a discussion.

This data set, as proposed in [31], can be generated as the solution to the following systems of ordinary differential equations

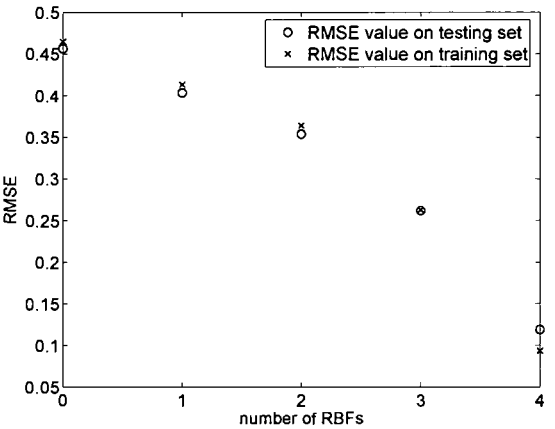
$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -x - (x^2 + y^2 - 1)y \\ \frac{dz}{dt} &= -\lambda z + 2(\lambda xy + \omega(x^2 - y^2)), \end{aligned}$$

where λ and ω are parameters. In Figure 2.1 a numerically integrated trajectory of an attracting cycle is shown. In this example, we are only concerned with fitting data on the limit cycle and ignore transients. Figure 2.2 shows the training set consisting of 101 points (almost two cycles) and testing data set consisting of 500 points, or almost 9 cycles. The fact that the solution is periodic will clearly illustrate the need for spatial as well as temporal windowing of the data. The system is capable of learning a specific part of the trajectory with a small amount of data and generalizes well to the data that resides in the same region.

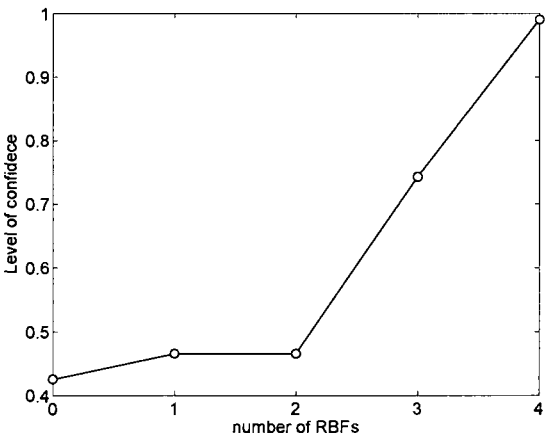
Figure 2.3 shows the ACC functions for the four major RBFs that capture the underlying structure of this data set. Again, the diamonds indicate the points in the ACC function that contribute to the RBF at that stage, i.e., they belong to $\mathcal{X}_{local}$. In Figure 2.3 (a) we see that the spatio-temporal window collects data from two peaks indicating that we have cycled through the data twice in that region. This example clearly illustrates the difference between spatio-temporal windowing and temporal windowing: *a time window would only use data from one cycle.* We see the same effect in 2.3 (b), 2.3 (c) and 2.3 (d).



(a) The plot of $\widehat{\gamma}(h^*)$ as new basis functions are added to the model.



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



(c) The confidence level of the fitted model as the new basis functions are added to the model.

Figure 2.5: The performance of the RBF fit on the Pringle data set. NOTE: The confidence level at the end of the process is 99%.

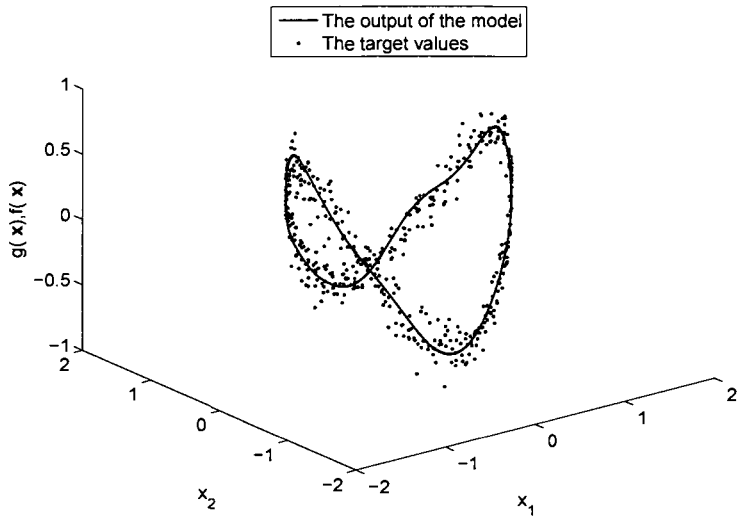
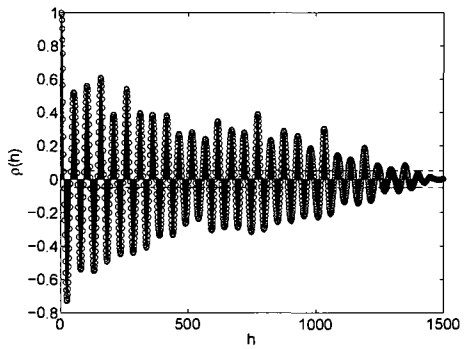
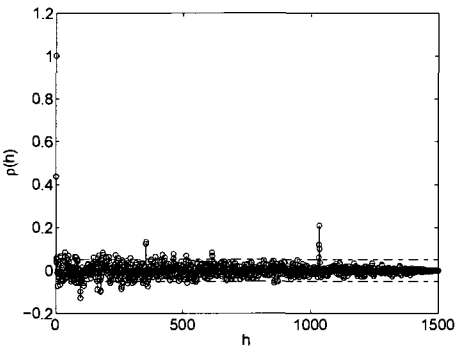


Figure 2.6: The testing data set and the output of the four mode model.

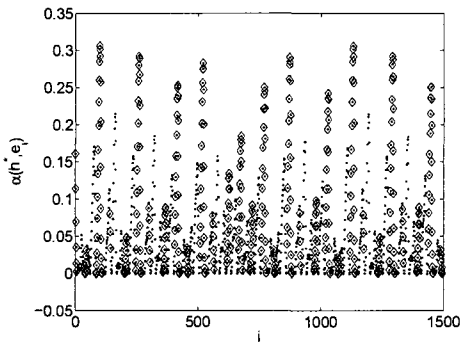
Figure 2.4 shows the location and shape of the four RBFs that are generated by the algorithm to model the data before the IID stopping criteria is satisfied. The training data and the RBFs are displayed together to illustrate how the algorithm has fit the RBFs to the data. Figure 2.5 (a) shows the maximum value of the ACC function for each step in the training process. Figure 2.5 (b) shows the performance of the model in the RMSE sense as the number of assigned RBFs increase while Figure 2.5 (c) shows the confidence level at each stage of training. We see that the first four basis functions are clearly the most significant. The four major RBFs model the data with RMSE of 0.1187 and 99% of points in the autocorrelation function resides in the 95% confidence bands. Note that neither the RMSE nor the values of $\widehat{\gamma}(h^*)$ provide reliable stopping criteria in this example. A plot of the output of the model and target values of the testing set are shown in Figure 2.6. We note that a similar experiment has been carried out for the noise-free Pringle data set; see [66] for details.



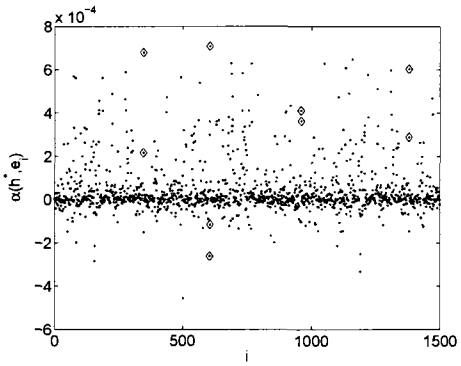
(a) Initial autocorrelation function.



(b) Final autocorrelation function.



(c) The ACC function associated with (a).



(d) The ACC function associated with (b).

Figure 2.7: The first and last autocorrelation functions and the associated ACC functions that are used to determine the local balls.

2.5.2 Mackey-Glass Time Series

The Mackey-Glass time-delay equation

$$\frac{ds(t)}{dt} = -bs(t) + a\frac{s(t-\tau)}{1+s(t-\tau)^{10}}. \tag{2.13}$$

generates a chaotic time series with short-range time coherence, where long time prediction is very difficult; it has become a standard benchmark for testing model fitting algorithms [118, 79, 152].

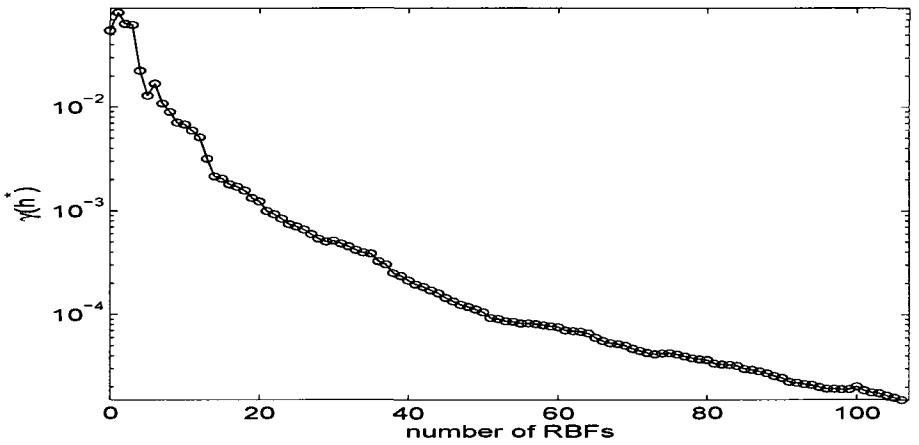
The time series is generated by integrating the equation with model parameters $a = 0.2$, $b = 0.1$ and $\tau = 17$ using the trapezoidal rule with $\Delta t = 1$, with initial conditions $s(t-\tau) = 0.3$ for $0 \leq t \leq \tau$ ($\tau = 17$). The initial 1000 data points corresponding to transient behavior are discarded. Then 4000 data points are reserved for the training set. The test set consists of 500 data points starting from point 5001. Note that not all 4000 training points collected were actually used for training the model. (These conditions are very similar to those in Platt [118].)

For purposes of comparison [153], the series is predicted with $\upsilon = 50$ samples ahead using four past samples: s_n, s_{n-6}, s_{n-12} and s_{n-18}. Hence, the nth input output data for the network to learn are

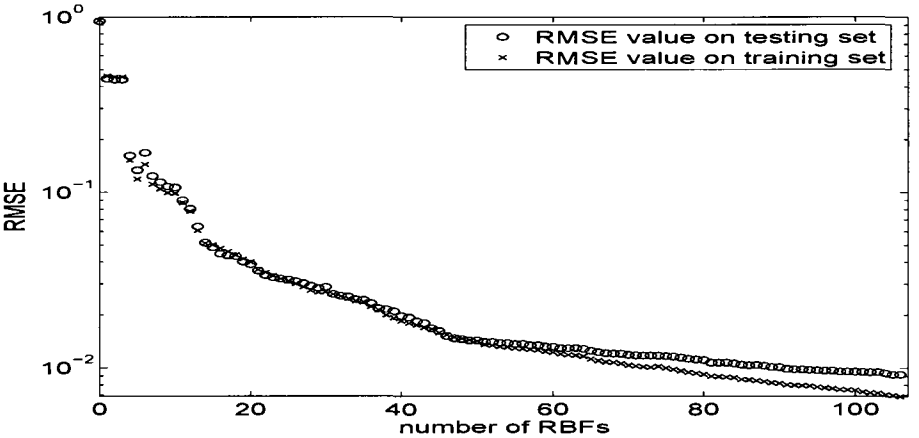
$$x_{n+\upsilon} = [s_n, s_{n-6}, s_{n-12}, s_{n-18}]^T$$

with $y_n = s_n$, whereas the υ step-ahead predicted value at time n is given by $z_{n+\upsilon} = f(x_{n+\upsilon})$, where $f(x_{n+\upsilon})$ is the network output at time n. The υ step-ahead prediction error is $\epsilon = s_{n+\upsilon} - z_{n+\upsilon}$. As such, this time series provides a good example for illustrating the construction of a nontrivial mapping from $\mathbb{R}^4$ to $\mathbb{R}$.

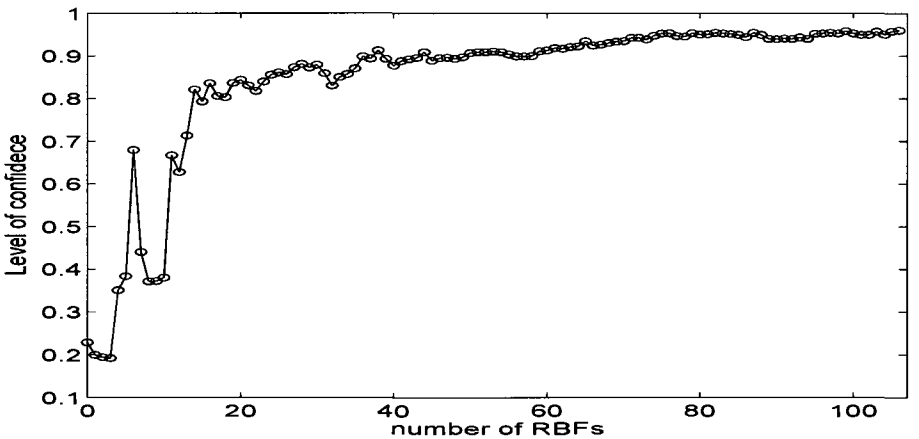
Figure 2.7 (a) shows the initial ACF (computed on the training data) while Figure 2.7 (b) shows the ACF of the residuals that indicates that the model fitting process should be terminated given 95% confidence has been achieved. Figures 2.7 (c) and (d) show the associated ACC functions, i.e., the point-wise values β_i, corresponding to the



(a) The log plot of $\widehat{\gamma}(h^*)$ as new basis functions are added to the model.



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



(c) The confidence level of the fitted model as the new basis functions are added to the model.

Figure 2.8: The performance of the RBF fit on the Mackey-Glass data set.

maximum value of the ACF in (a) and (b), respectively. From Figure 2.8 we see it is sufficient to use only 76 centers to get the 95% confidence fit for the Mackey-Glass data set with a resulting RMSE of 0.0116 (See Figure 2.9.) The output of the 76 mode model for the testing data set appears to fit the target values very well. This example is interesting in that a large number of modes is required to attain the stopping criterion.

Our algorithm based on space-time balls provides a result similar to MRAN [152] (RMSE of 0.035) using 1500 data points with 21 centers. However, at this level of RMSE, both our algorithm (21 modes) and MRAN (24 modes and 4000 data points), produce sporadic but significant overshoots and undershoots of the function in regions of high gradient. These large pointwise errors are hidden to some degree by a relatively small RMSE. The IID test is of course point-wise and reveals local un-modelled structure in the data and prevents the algorithm from terminating prematurely.

Yet, one might argue that stopping at 95% confidence and 76 modes is still premature stopping as a slightly improved final RMSE value of 0.0090 on the test data is achieved with 109 modes (but then does not improve with more). However, this example is for the special case of noise-free data. In such instances we recommend that the IID test be coupled with the RMSE test to draw optimal conclusions, unless, of course, one chooses to add noise artificially to the data. Given how close the RMSE errors are at 76 and 109 modes one must seriously consider that even in this case the 95% confidence level is arguably superior.

2.5.3 Time Series Prediction Using Exchange Rate Data Set[1]

This data set consists of daily values of the Deutsche Mark/French Franc exchange rate over 701 days; see Figure 2.10. As mentioned in [103], this data set has irregular non-stationary components due to government intervention in the Europe exchange rate

[1]We would like to thank Dr. D. Lowe at Aston University for providing us with this data set.

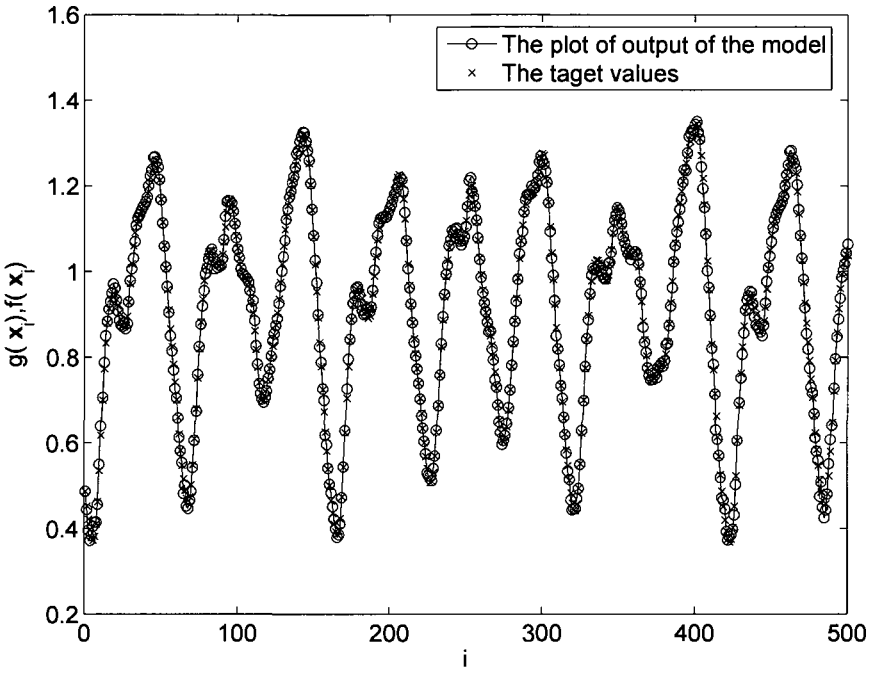


Figure 2.9: The output of the 76 mode model for the testing set compared to the target values. For this model an RMSE value of 0.0116 was obtained and the 95% of confidence stopping criteria was satisfied.

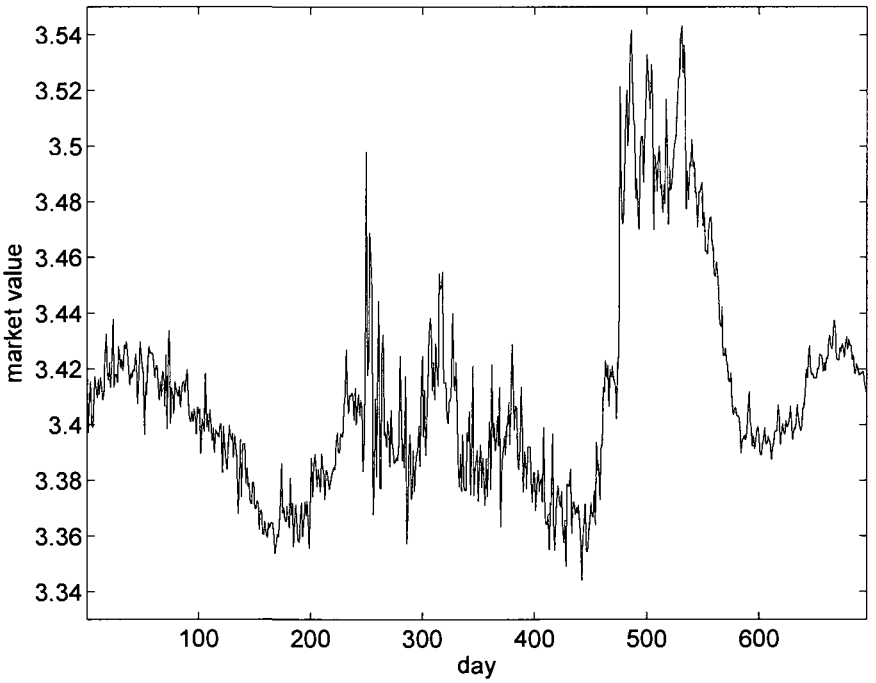


Figure 2.10: The plot of Exchange Rate data set.

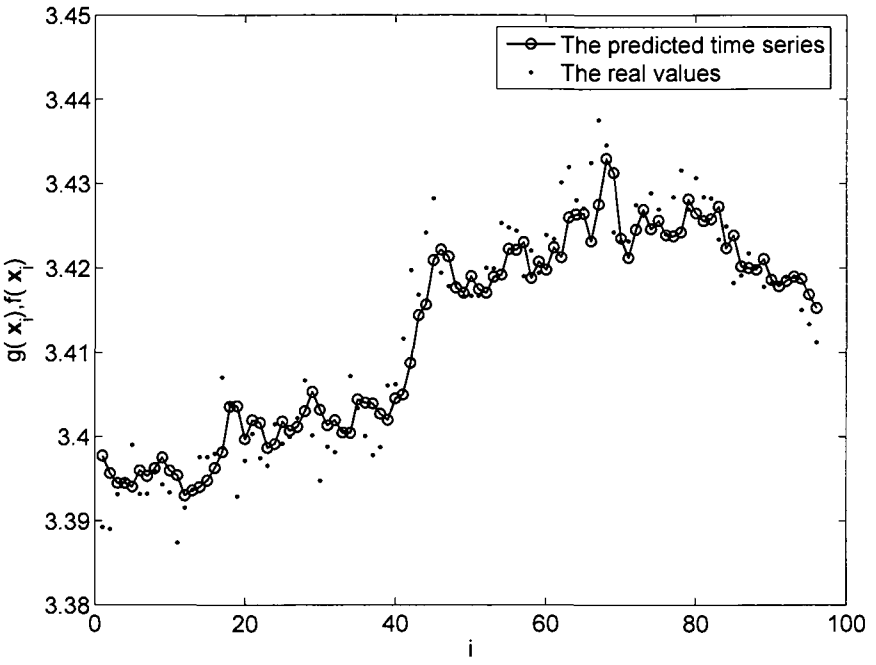
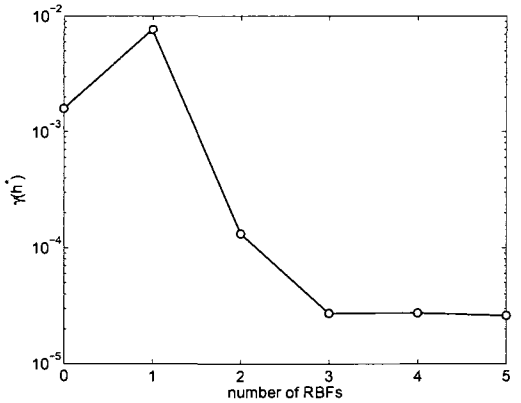


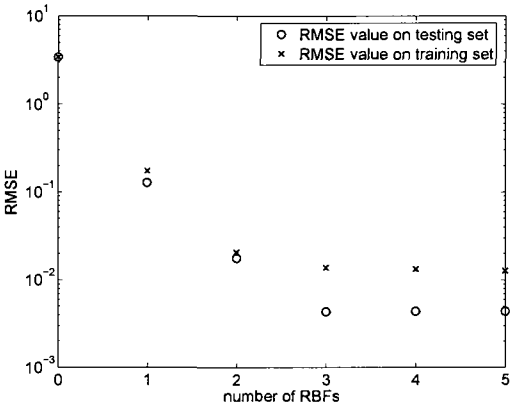
Figure 2.11: The 1-step prediction of the exchange rate data using a three mode RBF constructed using spatio-temporal balls. The associated errors for this model are: $RMSE = 0.0043$, $NPE_1 = 0.2760$, $NPE_2 = 0.1033$.

mechanism. Following [103], as there can be "day of week" effects in such data, a window of 5 previous values can be used as input, giving a data set of 696 patterns. Hence, this data set forms an interesting example of a mapping from $\mathbb{R}^5$ to $\mathbb{R}$.

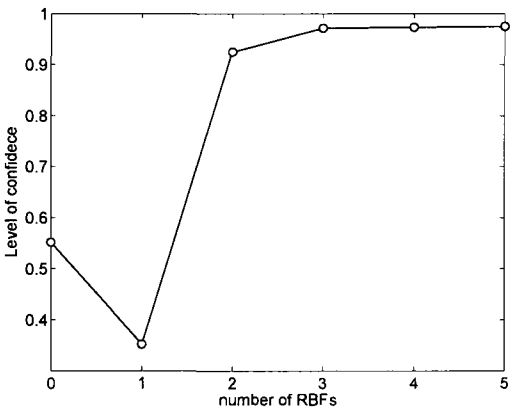
The training data for this model was taken to be the first 600 data points. The test data set was taken to be the last 96 data points. Figure 2.11 shows the output of the resulting model (1-step prediction values) and the target (market) values for the test data. The modeling process terminated with a model of order three when the 95% confidence threshold was attained (actually 97.17%). The ACC and RMSE criteria are also in agreement with the model order of three; see Figure 2.12. The three mode model produces the RMSE value of 0.0043, $NPE_1 = 0.2760$ and $NPE_2 = 0.1033$. The model has centers at $(3.4933, 3.9292, 3.2870, 3.8574, 4.0983)$, $(3.2793, 3.3475, 3.3337, 3.18433.2718)$ and $(3.3666, 3.4187, 3.6620, 3.2056, 3.6457)$ with widths $(0.4501, 2.7037, 2.2175, 2.5672, 2.9234)$,



(a) The log plot of $\widehat{\gamma}(h^*)$ as new basis functions are added to the model.

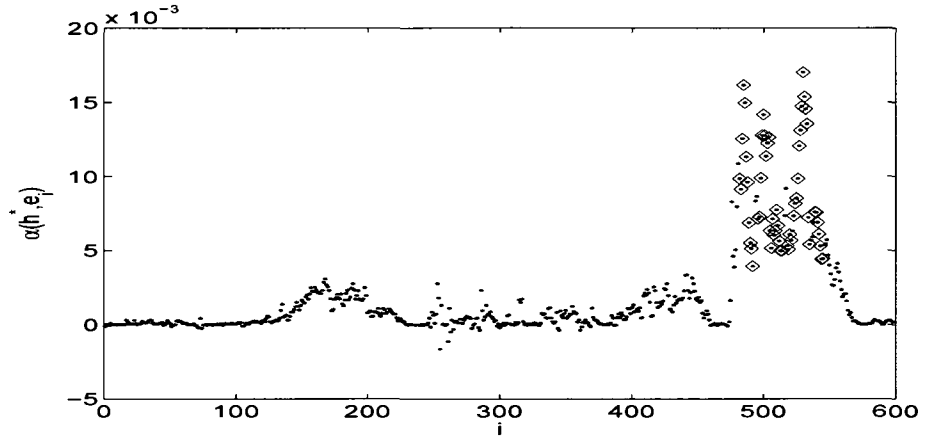


(b) The RMSE log plot of the model as new basis functions are added to the model.

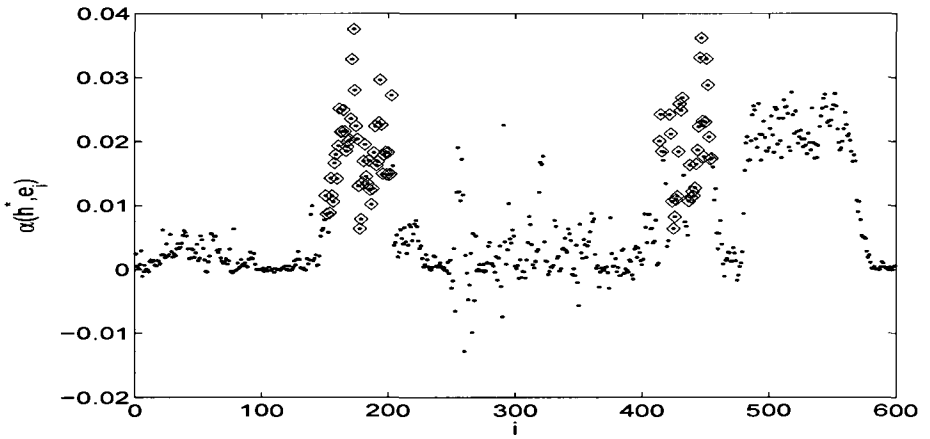


(c) The confidence level of the fitted model as the new basis functions are added to the model.

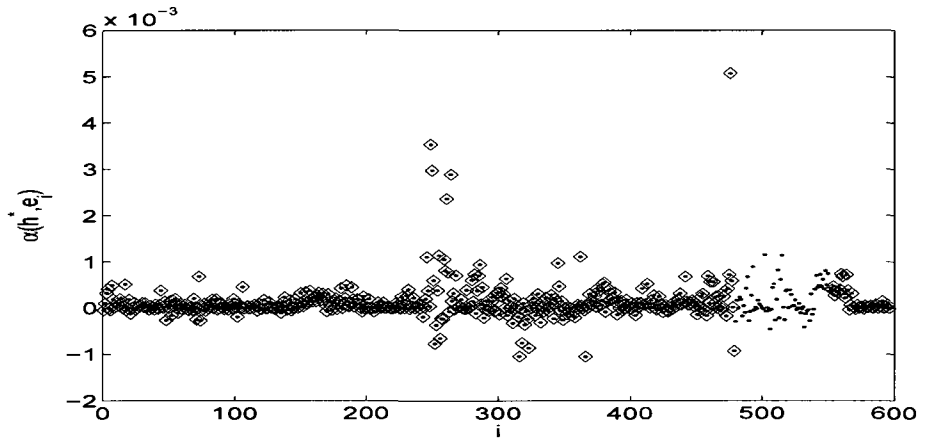
Figure 2.12: The performance of the RBF fit on the Exchange Rate data set. NOTE: The confidence level is set at 97.5% so one can observe the behavior of the system beyond the 95% of confidence. Also one might note that after adding three basis functions the 97.17% of the residuals fall within the confidence bounds.



(a) The selected data points for the first center.



(b) The selected data points for the second center.



(c) The selected data points for the third center.

Figure 2.13: The process of extracting three basis functions from the training (Exchange Rate) data set demonstrated by the ACC functions.

(0.1136, 0.1336, 8.5380, 0.6561, 0.5541) and (0.0555, 0.0358, 0.1740, 0.1939, 0.4015), and weights 3.8595, 0.5751 and 1.3805 respectively.

The results for this exchange rate data reported in [103] show a model fit with 11 RBFs and the NPE_1 of 0.336. We note that in that study the data was assumed to be on-line so the results presented here do not compare directly. However, we have observed that the on-line performance of our algorithm was comparable to the batch mode for the Exchange Rate data.

Figure 2.13 highlights the patterns in the ACC functions associated to the maximum contributions of the ACFs in the process of adding the first 3 main RBFs. Note again the need for the spatio-temporal window as evidenced in Figure 2.13 (b) and (c). Figure 2.13 (b) shows two distinct time regions contributing to the local data indicating either a periodic or quasi-periodic behavior. Figure 2.13 (b) suggests the remaining structure in the data is distributed across a preponderance of the data. A time-local windowing procedure would not capture this global structure in the data.

2.5.4 Overview of Related Work

In this section we review the work that is most similar in spirit to the algorithm presented here. It appears that the first paper to propose a growing algorithm for RBFs is [118]. In this paper, a new RBF is added to the model when the algorithm detects novelty in the input data. Briefly, the criteria for novelty include model error tolerance and the distance threshold between the input pattern and the nearest center. The new input pattern serves as the center of the new RBF, while the width of the RBF relates to a constant multiple of the distance between the new input and its closest center. If any of the criteria is not satisfied the current model parameters are adjusted using the Widrow-Hoff LMS algorithm [147]. In attempt to achieve more compact networks, a new algorithm referred to as RAN-EKF is introduced which uses the Extended Kalman Filter (EKF) instead of LMS adaptation procedure [79]. Although RAN-EKF can produce more parsimonious models it requires initialization of additional *ad hoc* parameters.

A minimal RBF neural network (MRAN) was proposed in [152]. MRAN adds a pruning strategy to RAN-EKF [79] that identifies RBFs that have negligible contribution to the overall model output over a number of consecutive inputs with respect to a threshold parameter. Also, MRAN includes an additional criterion that restrains the premature addition of new modes by smoothing the output error over a sliding window. A modified version of MRAN, called EMRAN, which utilizes an additional winner neuron strategy is proposed in [97]. Similarly, an algorithm that uses accumulated error information as a criterion for adding new RBFs is proposed in [48] and [49]. The diameter of the localized units is chosen based on the mutual distances of the RBFs. This method is able to generate small and well generalizing networks with comparably fewer epochs through training data set.

In an effort to reduce the number of required *ad-hoc* parameters a statistical test for adding new units to the network is proposed in [78]. This method, which uses EKF for training, is called Incremental Network (IncNet). In this model, if the prediction error does not fall within a certain level of confidence for the Z-statistic hypothesis test, then a new RBF will be added to the model.

More recently, a new RAN-EKF algorithm has been proposed that is applicable to both stationary and slowly varying non-stationary problems [103]. Here the novelty criterion involves testing whether the prediction error sequence corresponds to a zero mean Gaussian sequence at the 95% confidence level. A t-statistic is used to determine if the sequence has zero mean while the Weighted Sum Squared Residual (WSSR) statistic, [35], tests the normality of the sequence.

The algorithms described above that employ the results of hypothesis tests to add modes are only valid if the residuals have Gaussian distribution. In addition, perhaps more critically, all the algorithms mentioned in this section require careful adjustment of the *ad hoc* parameters for each data set to which they are applied. See Table 5.1 for a summary of these methods.

Algorithm	Noise Type	ad-hoc parameters
Platt's RAN	-	7
RAN-EKF	normal	9
MRAN	normal	12
IncNet	normal	5
RAN-WSSR	normal	6
Full-Bayesian	normal	12
New Algorithm	IID	None

Table 2.1: This table presents a comparison of related RBF algorithms. Note that the proposed algorithm is the only one that applies to the more general case of IID noise.

To name some specific practical application of RBFs, one could name its impact on DC motors [57],communications channel equalization, [95, 74, 93, 73], pattern retrieval, [151], robotics manipulations, [154], estimation of the noise density, [2], object recognition, [114], discriminating the EEG patterns, [42], estimation of ground rainfall, [149], adaptive classification, [67], sensor failure, [3], finance, [64], and other applications in astronomy and other fields. For further details about the described algorithms and their ad-hoc parameters please see [66].

2.6 Conclusions

We propose an algorithm for approximating functions from scattered data and compare its performance to the leading algorithms in the literature. To illustrate the absence of *ad hoc* parameters, all the data sets presented in this chapter were fit by exactly the same code. No adjustments were made based on the data set being fit. Hence, we claim the proposed algorithm is approaching a black-box methodology for nonlinear function approximation. This feature will permit the advancement of a variety of other algorithms, e.g., the representation of data on manifolds as graphs of functions [29, 30], pattern classification [67, 90], as well as the low-dimensional modeling of dynamical systems [31].

It is assumed that the available data represents a functional relationship, or signal, with IID additive noise. Note that if the signal contains multiplicitive noise we can take the natural logarithm of the signal to make it additive. An hypothesis test is applied to the residuals at each step in the algorithm to determine whether a new basis function should be added, and if so, where it should be added. When it has been determined that this test has been passed using the 95% confidence criterion one may infer there is no geometric structure left in the residuals and thus the model order has been found.

We extended previous work by employing a spatio-temporal window, i.e., space-time balls, for determining the local data to be used in updating the model. The examples suggest that this novelty is critical for approximating data over high-dimensional domains and in particular for data generated by dynamical systems. We have observed that significantly more data is located in the space-time balls than the temporal windows previously considered resulting in the construction of significantly improved models.

In this chapter we have assumed that the received signal is composed with additive IID noise while prior algorithms based on statistical hypotheses are restricted to Gaussian noise. Despite the fact that the IID test provides only a necessary condition, it appears to generate low order models with small RMSE.

We have presented this algorithm for batch data. It is possible to extend this approach for data that arrives in a stream, i.e., on-line data. Our preliminary results show that this approach does not require the data to be seen repeatedly as some "on-line" algorithms require. We will present these results in a companion study. Also, although the algorithm was presented here in the context of growing RBFs, in principle it can be employed with other architectures for fitting nonlinear functions such as feed-forward neural networks.

Chapter 3

EXAMPLES OF COMPACTLY SUPPORTED FUNCTIONS FOR RADIAL BASIS APPROXIMATIONS

Abstract Most applications of Radial Basis Functions (RBFs) in the literature employ basis functions from a relatively small list, that includes Gaussians, multi-quadrics, inverse multi-quadrics, cubics, linear functions and thin plate splines. These functions are attractive since they satisfy invertibility conditions for the interpolation problem. In this chapter we introduce several new compactly supported RBFs for approximating functions in $L^P(\mathbb{R}^d)$ in the over-determined least squares. In this setting the requirements on the functions are weaker and many interesting examples arise. We illustrate the utility of this broader class of RBF on the benchmark Mackey-Glass time series data. We observe that these new RBFs significantly reduce the number of modes required to approximate the data and produce models that have significantly improved condition numbers.

3.1 Introduction

As described in the previous chapters, the approximation of nonlinear relationships in scattered data is now a problem of established significance in science and engineering. Often it occurs that only empirical observations of a phenomenon are available and relationships must be estimated by means of mathematical models. It is desirable for such phenomenological models to be as simple as possible. Given the nature of the RBF

approximation problem as described in Chapter 2, it is very desirable for functions in an approximation to have minimal or even zero overlap. This is only possible if the functions have compact support.

One of the main objectives in the construction of a model from known, or training, data is to optimize the quality of its performance on new data generated by the same process. Thus we require the models to have both descriptive and predictive features. While this goal can be approached from a variety of directions[1] the inherent conditioning of the model plays a critical role in its ability to generalize. In practice, if the data model is represented generally by the mapping

$$y = f(x),$$

we are concerned with how the output of the model changes as a consequence of perturbation of the input. In particular, if

$$y + \delta y = f(x + \delta x),$$

it is desirable that the magnitude of the change in the output $\|\delta y\|$ be small if $\|\delta x\|$ is small. By definition, well-conditioned models produce small variations in δy for small variations in δx.

For nonlinear mappings, such as those generated by multi-layer perceptrons, the estimation of the condition number is complicated by the fact that the Jacobian of the map must be estimated at every point of interest [88]. This is also true in general for RBFs. However, in the case of RBFs we can determine the the condition number associated with the perturbation of the parameters simply by computing the singular values of the interpolation matrix. This information provides an important measure of the sensitivity of the model.

---

[1]For example, regularization methods and cross validation.

In general the condition number of an $m \times n$ matrix is $O(mn)$ suggesting that (nonlinear) models that employ linear transformations have poor performance bounds for large data sets of sufficient complexity. We have observed however, that the nature of the condition number depends very significantly on the type of RBFs that are employed. With this motivation we considered several forms of RBFs including those with compact support. We found that the functions generally available in the literature often have poor conditioning properties, at least for some of the data sets we have considered. In this chapter we introduce several new compactly supported functions for approximating data that possess surprisingly good conditioning properties.

The organization of this chapter is as follows: Section 3.2 provides an introduction to RBFs for nonlinear data approximation. Section 3.2.1 introduces new compactly supported RBFs and reviews the current literature in this area. Section 3.3 shows the performance of the new RBFs in context of numerical examples. Finally, Section 3.4 provides some concluding remarks and discusses avenues for future work.

3.2 Radial Basis Functions for Approximating Scattered Data

We employ the same RBF expansion as in Chapter 2 which we repeat here for convenience, i.e., an RBF is a mapping $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ that is represented by

$$f(x) = Ax + \alpha_0 + \sum_{k=1}^{N_c} \alpha_k \phi(\|x - c_k\|_W), \tag{3.1}$$

where x is an input pattern, ϕ is a RBF centered at location c_k, and α_k denotes the weight for kth RBF and A is an $m \times n$ matrix. As before, the matrix W denotes the parameters in the weighted inner product $\|x\|_W = \sqrt{x^T W x}$ and the term $Ax + \alpha_0$ performs an affine transformation of the data and is useful so that the nonlinear terms are not attempting to fit flat regions. More general polynomials may be used for this purpose [130]. As usual, the dimensions of the input n and output m are specified by the dimensions of the input-output pairs from data.

In this chapter the implementation of the RBF follows our black-box methodology for nonlinear function approximation as described in Chapter 2 or [66, 72], i.e., we employ Algorithm 1. It is assumed that the available data represents a functional relationship, or signal, with IID additive noise.

3.2.1 Compactly Supported RBFs

Recently several functions with compact support have been proposed as RBFs for the interpolation problem, see, e.g., [144, 143, 148]. For example, the C^2 function

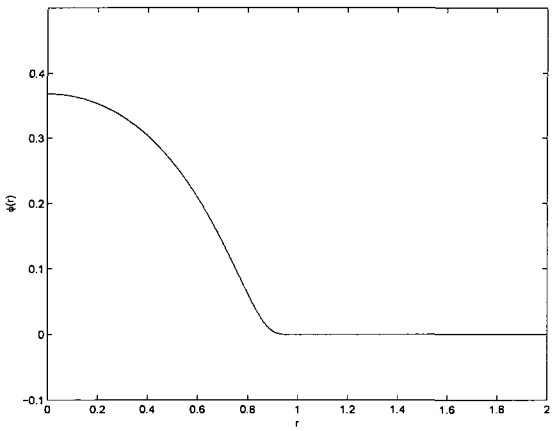
$$\phi(r) = (1-r)_+^4(1+4r), \tag{3.2}$$

has been derived as an RBF explicitly for domain dimension 4 in the sense that the resulting square interpolation matrix is a (conditional) positive definite matrix [144]. In other words, we say that this function qualifies as an RBF given the square interpolation matrix is guaranteed to be invertible. In many cases of practical interest it appears that this interpolation condition is overly restrictive. In particular, the data fitting problem is concerned with solving an over-determined least squares problem. In this setting it seems adequate to only require that the approximating basis functions be dense in an appropriate function space.

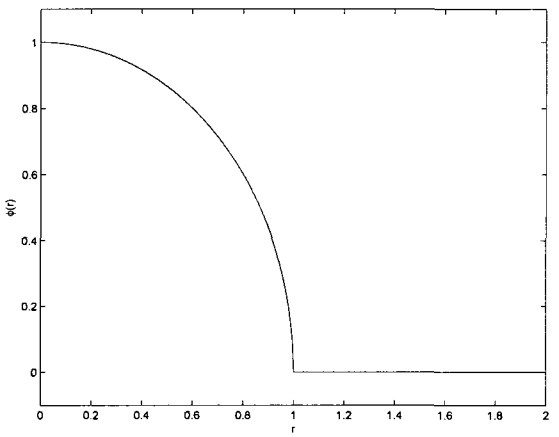
For example, as described in [115], the conditions required of basis functions to be dense in $L^P(\mathbb{R}^n)$ are very weak. For completeness, we briefly describe Park and Sandberg's theorem. Following [115], let K be a radially symmetric kernel function related to the activation function $\phi : [0, \infty) \longrightarrow \mathbb{R}$, such that, $K(\frac{x-c_i}{\sigma_i}) = \phi(\frac{\|x-c_i\|}{\sigma_i})$. The general element of the set $S_1(K)$ is expressed as

$$q(x) = \Sigma_{i=1}^{N_c} \alpha_i K(\frac{x-c_i}{\sigma_i}), \tag{3.3}$$

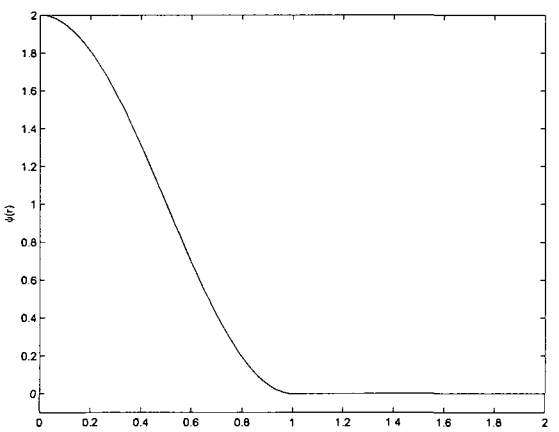
where $N_c \in \mathbb{N}$, the set of natural numbers, is the number of basis functions, $\alpha_i \in \mathbb{R}^m$ is the vector of weights, x is an input vector (an element of $\mathbb{R}^n$), c_i and σ_i are the center and



(a) The plot of the C^∞ mollifier function.



(b) The plot of C^0 quarter circle.



(c) The plot of C^∞ Hanning function.

Figure 3.1: These functions can be used as $\phi(r)$ in the radial basis function expansion.

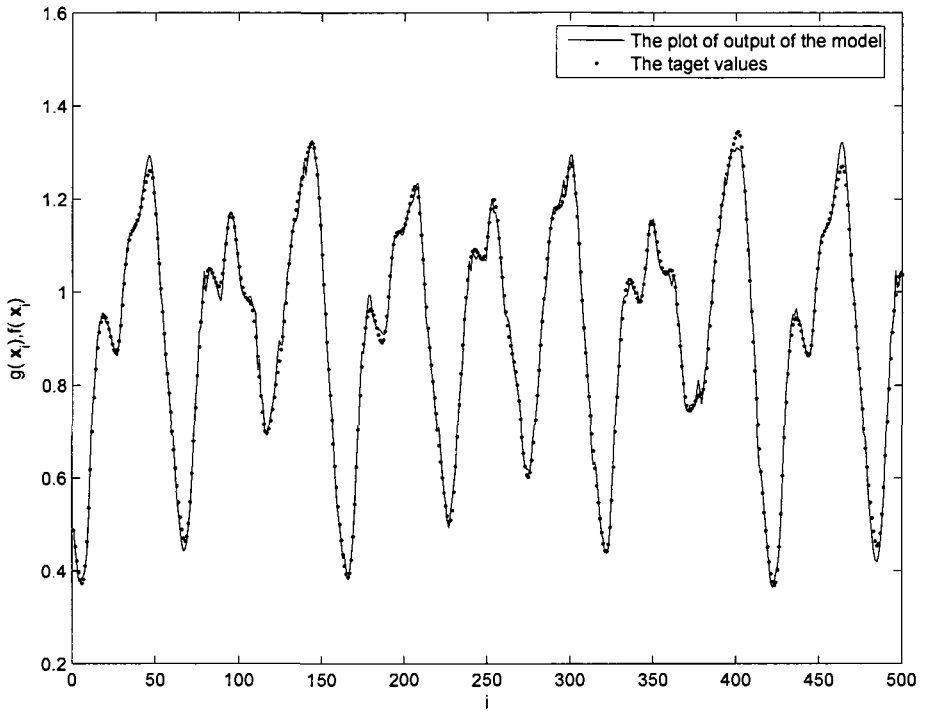


Figure 3.2: The output of the 37 mode model for the testing set compared to the target values. For this model an RMSE value of 0.0167 is obtained and the 95% of confidence stopping criteria was satisfied.

widths of the ith kernel node, respectively. If $\sigma_i = \sigma$, i.e., all the widths are constant, then this family of functions is referred to as $S_0(K)$ [115].

Park and Sandberg's $S_0(K)$ Theorem [115]:

Let $K : \mathbb{R}^n \longrightarrow \mathbb{R}$ be an integrable bounded function such that K is continuous almost everywhere and $\int_{\mathbb{R}^n} K(x)dx \neq 0$. The family $S_0(K)$ is dense in $L^p(\mathbb{R}^n)$ for every $p \in [1, \infty)$.

Park and Sandberg provide additional theorems for S_0 and S_1 with improved conditions in [116]. Motivated by these broad criteria which qualify functions as RBFs for least-squares problems, we propose several compactly supported functions that by Park and Sandberg's theorem are dense in $L^P(\mathbb{R}^n)$. In what follows we will illustrate their utility in practice in the context of over-determined least squares problem.

First, we propose the *bump* function widely used in differential geometry

$$\phi(r) = \exp(\frac{1}{r^2 - \gamma^2})H(1 - r^2),$$

for use as an RBF activation function where H is the usual Heaviside step function. This compactly supported and infinitely differential function is also widely referred to as a *mollifier*. It is shown in Figure 3.1 (a), and is qualitatively similar in nature to the widely applied non-compact Gaussian RBF, $\exp(-r^2)$. Interestingly, the failure of the Gaussian to have compact support has led some researches to arbitrarily truncate it. We observe that the Gaussian RBF satisfies the positive definiteness of the interpolation matrix for all space dimensions $d \geq 1$. Note that while the mollifier function satifies the postulates of Park and Sandberg's theorem, it has non-positive values in its Fourier transform and hence does not satisfy Bochner's *interpolation* criterion, [24], for a compact RBF [144]. Although this fact is of theoretical interest it is not of practical consequence since we are interested in the approximation (rather than interpolation) in the context of the overdetermined least squares problem.

A compact activation function with constant curvature is provided by

$$\phi(r) = \sqrt{1 - r^2}H(1 - r^2). \tag{3.4}$$

This is just the quarter circle shown in Figure 3.1 (b). Clearly this function also satisfies the postulates of Park and Sandberg's theorem. Of course this function is not differentiable where it meets the axis. While this could potentially cause problems in practice, we establish in the section on numerical experiments that the condition number of this RBF suggests it is worthy of further investigation.

Our last proposed activation function with compact support is the *Hanning* filter

$$\phi(r) = (\cos(r\pi) + 1)H(1 - r). \tag{3.5}$$

Like the bump function, this function is also infinitely differentiable; see Figure 3.1 (c). It has advantages over the mollifier function in the manner in which the function approaches

Table 3.1: This tables shows the performance of different RBFs under using an identical strategy of fit.

	Wendland RBF	Circle RBF	Mollifier
$ConditionNumber$	$3.0057e+003$	12.5845	284.3114
$RMSE$	0.0109	0.0344	0.0167
$NumberofRBFs$	51	26	37
$Confidence\%$	95	95.27	95.53

zero, i.e., there is no vanishing term in a denominator. We do not present empirical results for this case in this current chapter, however, the Hanning filter performs very well in the examples of Chapters 5 and 6.

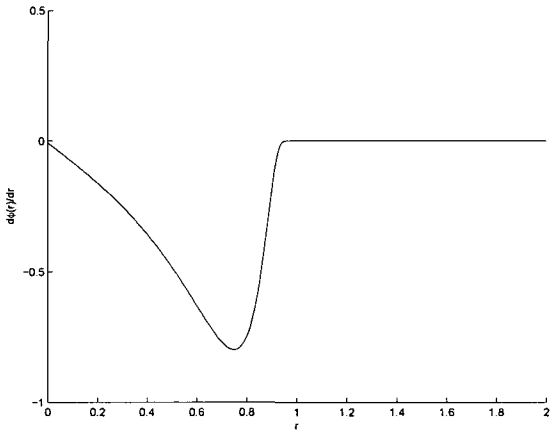
3.3 Numerical Examples

Here we employ a parsimonious growing algorithm with automatic mode determination as described in Chapter 2 or [66, 72], i.e., Algorithm 1. As described in Chapter 2, this algorithm is very attractive for comparing the qualities of various RBFs as it does not require any tuning of *ad hoc* parameters. Algorithm 1 works on high dimensional domains and employs an spatio-temporal window to identify the data points that contribute to each RBF, see Chapter 2. Recall that the placement of the functions is driven by a statistical hypothesis test that reveals geometric structure when it fails. At each step the added function is fit to data contained in a spatio-temporally defined local region to determine the parameters, in particular, the scale of the local model.

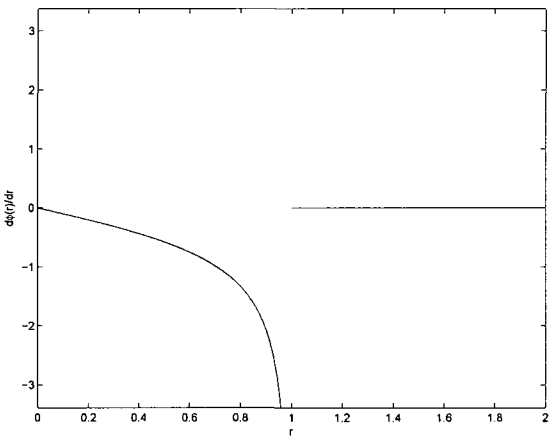
3.3.1 Mackey-Glass Time Series

In this example we again use a mapping from a time-delay embedding of the univariate time-series to a future value to illustrate the performance of the new compact functions. The data set is exactly the same as in Chapter 2, Section 2.5.2 as is the form of the modeling problem.

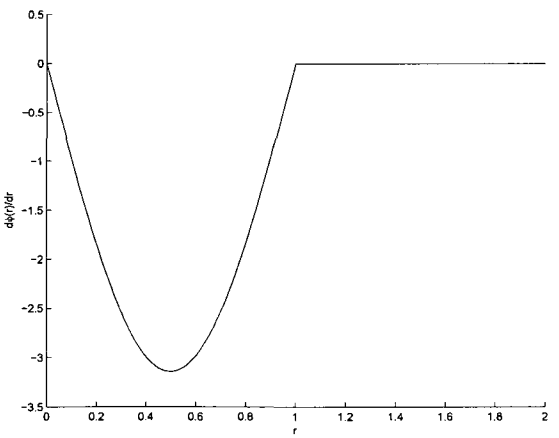
As in Chapter 2, we compare performance of the various RBFs via RMSE and the number of basis functions required. However, here we also measure the sensitivity of the



(a) The plot of the C^{∞} mollifier function.



(b) The plot of C^{0} quarter circle.



(c) The plot of C^{∞} Hanning function.

Figure 3.3: The derivatives of the compact RBFs. Small values near $r = 0$ can lead to improved conditioning of the model.

models via the condition number of the interpolation matrix of the full model. We present the final result of the fit using the mollifier in Figure 3.2. In this figure the output of the model and the associated target values are shown. A comparison of condition numbers as well as other data associated with the model fit is summarized in Table 3.1. The circle RBF has a surprisingly low condition number, three orders of magnitude lower than the polynomial RBF. Note also that, as described in Chapter 2, all the results achieve 95% confidence in the statistical test applied to the residuals.

We make no claim that the RBFs proposed here are superior to other RBFs in the literature or that they can be used for the interpolation problem. Clearly there are many factors that influence the performance of these fits. In particular, the nature of the data set will dictate to some degree which RBF is most appropriate. However it is significant that, on this data set at least, the condition numbers of the least-squares matrix are dramatically lower for the new compactly supported RBFs. This could be due in part to the profile of the derivative of each of the RBFs. We see in Figure 3.3 (a) that the derivative of the mollifier is very small near the origin. The slope rises faster than that of the derivative of the quarter circle but is more well behaved for larger values. Obviously the circle suffers from the fact that it is not differentiable at $r = 1$ as shown in Figure 3.3 (b). Clearly this blow-up is potentially problematic but in our simulations the data for each RBF was never in this region. The symmetry of the Hanning derivative shown in Figure 3.3 (c) might have advantages but we also observe the steeper slope near the origin.

3.4 Conclusions

We have proposed several new compactly supported RBFs and have illustrated some of their enhanced performance properties on the benchmark Mackey-Glass problem. Both the number of required modes and the conditioning of the final model are substantially improved over results using RBFs from the standard list. This suggests that the compactly supported RBFs proposed here provide additional options of interest in the data

fitting problem. In particular, we advocate the use of either the mollifier function or Hanning RBF as an alternative to the truncated Gaussian RBF.

We note that the condition number of the interpolation matrix depends directly on the choice of RBF and suggested an explanation of the good conditioning properties of the proposed RBFs in terms of the behavior of the derivatives. It is interesting to speculate that new RBFs may be designed by optimizing the behavior of the derivative of the RBF for purposes of numerical conditioning. Such an approach should lead to improved function generalization.

In later chapters we will consider innovations that result in Algorithm 2 and extensions to higher dimensional ranges, i.e., Algorithm 3. We employ the compact functions presented here in those settings as well and observe significantly improved performance.

Chapter 4

SKEW-RADIAL BASIS FUNCTION EXPANSIONS FOR EMPIRICAL MODELING

Abstract We propose a skew-radial basis function expansion for the empirical model fitting problem. This is accomplished by modulating or skewing, each radial basis function by an asymmetric shape function which increases the number of degrees of freedom available to fit the data. We show that if the original radial basis function interpolation problem is positive definite, then so is the skew-radial basis function when it is viewed as a bounded perturbation of the radial basis function. We illustrate the effectiveness of skewing radial basis functions via several example problems including fitting data with jumps and prediction of the maximum wind intensity of a hurricane. Further, we show this approach leads to models with both improved accuracy and reduced order.

4.1 Introduction

The importance of quantifying nonlinear relationships between sets of variables has driven researchers to devise an array of techniques for empirical modeling from data. If the underlying relationship between the domain and range variables is nonlinear and the dimension of domain variables is greater than two or three then one must resort to special techniques that overcome Bellman's curse of dimensionality [20]. The multilayer-perceptron and the associated back-propagation algorithm have attracted considerable attention for constructing such mappings [145, 128]. Alternatively, radial basis functions

(RBFs) have attracted substantial interest given that the resulting optimization problem can be broken efficiently into linear and nonlinear subproblems [121, 32, 120, 119]; see also the more recent monographs [96, 34, 144].

Empirical modeling is essentially a data fitting problem. The data may be generated, e.g., by a dynamical system that is either numerically simulated or observed. For example, a physical system such as a hurricane may be observed in nature or approximated via a numerical simulation. In either case it is of interest to model a relationship between domain variables and scalars of interest such as maximum wind intensity. Other dynamical systems, such as financial markets, behave in such a manner that equations derived from first principles do not adequately describe the behavior of the actual phenomenon. In this setting as well it is useful to be able to discover relationships by directly modeling the data.

A geometric approach to data analysis involves describing data given as a set of points on a manifold embedded in an Euclidean space as the graph of a function [29, 30]. Given data samples from a manifold, the first step is to identify an appropriate domain x for the function. Then a nonlinear data fitting technique may be used to represent the nonlinear elements $f(x)$ resulting in the representation of the data as the graph $(x, f(x))$. These ideas belong to a field that has been more generally referred to as *manifold learning*, see, e.g., [137, 126].

These examples all share the common feature of the data fitting problem in that there exists a set of domain values $\{x^{(k)} \in \mathbb{R}^n\}$ and range values $\{y^{(k)} \in \mathbb{R}^m\}$. Further, it is implicit that there exists a mapping $f(x)$ such that

$$y^{(k)} = f(x^{(k)}), \tag{4.1}$$

which we seek to determine. In general terms we can express this as a function approximation problem, i.e.,

$$f(x) = \sum_{i=1}^{n} \alpha_i \phi(x; \eta_i). \tag{4.2}$$

In addition to the nature of the function ϕ, the parameters $\{\eta_i\}$ are critical for encoding the information associated with the data. In particular, these parameters determine the location of the basis functions over the domain as well as shape parameters to govern the functions ability to match, or fit, the data.

Given the volume of the domain increases exponentially with dimension, it is simply not practical to cover a high-dimensional domain with a uniform lattice. Assuming that the auxiliary parameters η_i have been selected, then the weights α_i in Equation (4.2) are determined by satisfying the interpolation conditions given in Equation (4.1). Note that if the interpolation matrix is square then these conditions are satisfied exactly. However, if there are more data points than centers, then these interpolation conditions are approximated in the least squares sense.

In general, data fitting methods, such as those based on RBF expansions, are very limited in the nature of the parameters that can be used to adapt the shape of the functions ϕ, see Appendix for a Gallery of most prominent types of these functions.

In this chapter we investigate *skew-radial* function expansions of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i \psi(x; \varphi_i), \tag{4.3}$$

where

$$\psi(x; \vartheta_i) = z(x; \nu_i)\phi(x; \eta_i).$$

ϑ_i consists of parameters required for both the radial function $\phi(x; \eta_i)$ and the symmetry breaking function $z(x; \nu_i)$. Note that z is actually a data adapting function since it provides additional flexibility to the function to match the shape of the data. Both functions are mappings from the data domain to the real line, i.e.,

$$z, \phi : \mathbb{R}^n \to \mathbb{R}.$$

We demonstrate via several examples how this modified expansion can be used to improve a given data fitting approach by both improving the accuracy of the fit and reducing the

order of the model required. We focus on illustrating these ideas in the context of RBFs, one of the most popular approaches for nonlinear data fitting over high dimensional domains, see e.g., [52]. Park and Sandberg's theorems, [115, 116], provide theoretical background for the universal approximation properties of radial basis functions which are also inherited by skew-radial basis functions.

The organization of this chapter is as follows: Section 4.2 provides a motivating example for why skew-radial basis functions are useful. Section 4.3 reviews the connection to skew-symmetric. Section 4.4 introduces the skew-radial basis functions for data fitting and provides several examples. The positive definiteness properties of the new skew-radial basis functions are proved in Section 4.5. Section 4.6 demonstrates the added value of the developed basis functions via numerical examples. Section 4.7 provides an overview to related work. Finally, Section 4.8 provides some concluding remarks and discusses future work.

4.2 A Motivating Example

We now consider an illustrative example which, while artificial in nature, clearly indicates the *potential* need for the shape adaptation function $z(x, \nu_i)$. Consider the function $f(x)$ defined as

$$f(x) = e^{-(x-2)^2} \int_{-\infty}^{\lambda(x-2)} e^{-y^2} dy. \tag{4.4}$$

Note that if the skew parameter $\lambda = 0$, then the function $f(x)$ is symmetric. For this example we select $\lambda = -7$ and generate random points on the graph $(x, f(x))$ as shown in Figure 4.1. A data set of 450 data points is generated. A modest amount of noise is added to the data, normally distributed with mean zero and standard deviation 0.01 (these leads to the requirement of an overdetermined system). These points are partitioned into a training subset which is used to compute fitting parameters and a validation subset which is used to indicate when the training is completed. The training

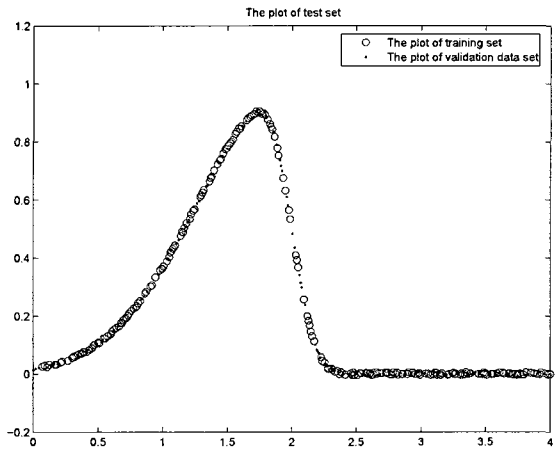


Figure 4.1: The testing and training data sets for the skew-radial data set generated by randomly sampling Equation 4.4 with the parameter $\lambda = -7$.

and validation data sets consist of 244 and 107 uniformly sampled data points from the original data set, respectively. The remaining 99 data points are used as testing.

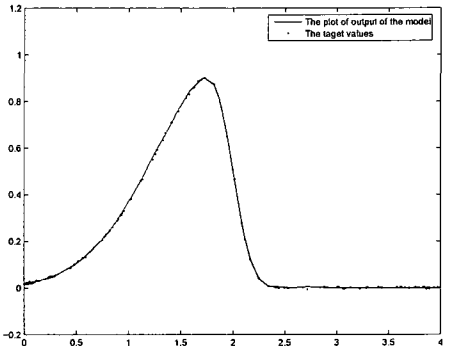
A standard RBF algorithm would seek to represent $f(x)$ as defined by Equation (4.4), e.g., as

$$f(x) = \sum_{i=1}^{n} \alpha_i \phi(\|x - c_i\|_{W_i}), \tag{4.5}$$

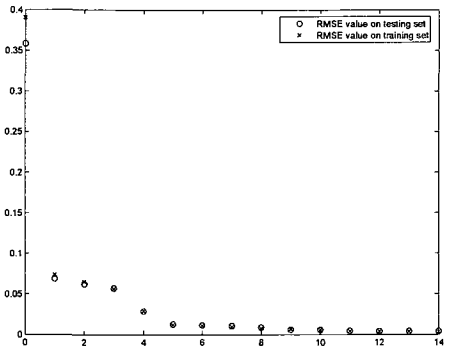
where x is an input point, ϕ is the basis function that is centered at location c_i, α_i denotes the weight for the ith basis function. The term W denotes the parameters in the weighted inner product $\|x\|_W = \sqrt{x^T W x}$.

Note the dilemma as the functions ϕ are, by construction, symmetric. In our experiment we employ Gaussian RBFs to approximate $f(x)$ and we require $n = 13$ terms to achieve an RMSE [1] of 0.0035 and a 96.80% confidence level that the residuals of the model are IID noise [66, 72] or Chapter 2. The resulting fit is shown in Figure 4.2 (a) and the RMSE performance of the model is shown in Figure 4.2 (b).

[1]Root Mean Square Error, RMSE= $\sqrt{(1/T)\sum_{i=1}^{T} e_i^2}$ where e_i's are the residuals.



(a) The the output of the model.



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

Figure 4.2: The output of the model when Gaussian's are used as RBFs and the performance of the model as new basis functions are added to the model.

It is revealing to view the manner in which the symmetric functions manage to fit the skew-radial function as shown in Figure 4.3. While it is tempting to conclude that only $n = 5$ functions are required to fit $f(x)$ we note that the residuals are not IID for that model. We return to this example in Section 4.6.1 where a model is constructed in terms of skew-radial basis functions.

4.3 Skew Statistical Distributions

The motivation for skew-radial basis functions (sRBF) stems from the fact that the shape of data to be fit is in general not radially symmetric. Our research in model order reduction, Chapter 2, indicated that the shortcomings of the fit could be directly related to the lack of flexibility in the approximating functions. The need for such asymmetric expansion functions suggested that recent developments in the literature of multivariate skew statistics could be of particular interest. In [8] a general representation of the density of an arbitrary skew distribution is given as

$$f(z|Q_m) = K_m^{-1} f_k(z) Q_m(z), z \in \mathbb{R}^k,$$

where $K_m = P(X > 0)$ and $Q_m(z) = P(X > 0|Z = z)$ for some random vectors X and Z with dimensions m and k, respectively, and with joint distribution such that Z has

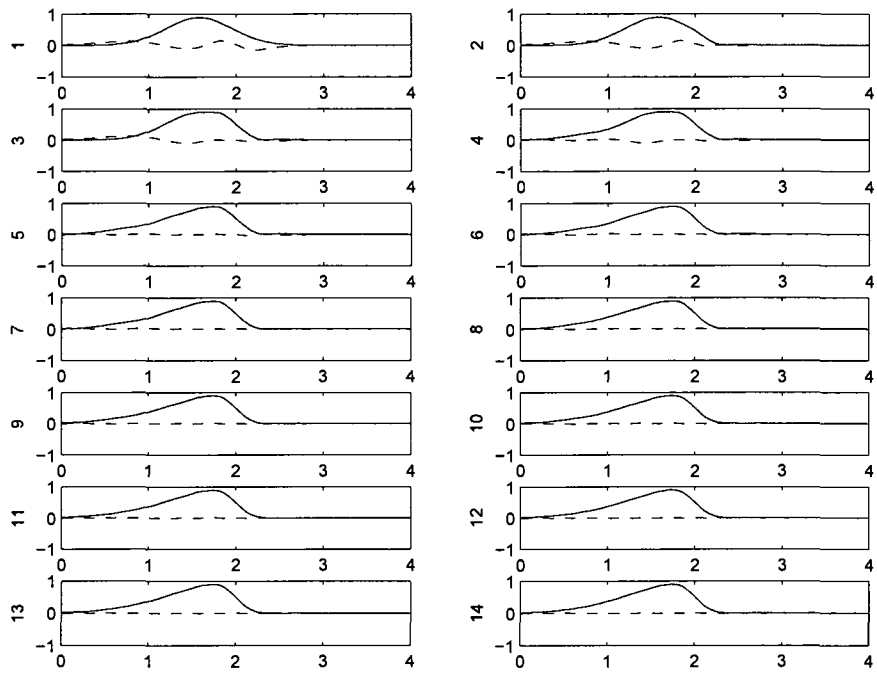


Figure 4.3: The 14 steps to fit the Skew data set. The residuals of the model are IID after 13 terms.

marginal density f_k. K_m is a normalizing constant and the term Q_m may be interpreted as a skewing function. The most general class of skew distributions is defined in terms of this density and are referred to as fundamental skew-symmetric distributions (FUSS) [8].

One could obtain different families of skew distributions by specifying a symmetric pdf f for Z and conditional distribution $X|Z = z$. For example, if we assume that $Z \sim N_k(\mu, \Sigma)$, then the FUSN class of distributions is defined as follows: Let $Z^* = [Z|X > 0]$, where $Z \sim N_k(\mu, \Sigma)$ and X is a $m \times 1$ random vector. Then Z^* has a k-variate fundamental skew-normal (FUSN) distribution, which is denoted by $Z^* \sim FUSN_{k,m}(\mu, \Sigma, Q_m)$ and its density is given by

$$f_{Z^*}(z) = K_m^{-1} \phi(z|\mu, \Sigma) Q_m(z),$$

where $Q_m(z) = P(X > 0|Z = z)$ and $K_m = E[Q_m(Z)] = P(X > 0)$. Note that one could generate further nonsymmetric distribution by specifying another function in the argument of Q_m [8].

The idea of modeling skewness by means of the construction of a mathematically tractable family including the normal distribution can be traced back to 1908, [56], where perturbation of the normal density via a uniform distribution function leads to a form of skew-normal density. For specific references on skew-Cauchy distributions, see [9, 19, 61, 53], for skew t distributions [75, 14], skew-logistic distributions [142], and skew-elliptical distributions [26].

For the purposes of this chapter we focus on the specific formulation of the skew-normal distribution. The formal definition of the univariate skew-normal (SN) family is due to Azzalini [11]. A random variable Z has an SN distribution with skewness parameter λ, and is denoted by $Z \sim SN(\lambda)$ if its density is $f(z|\lambda) = 2\phi(z)\Phi(\lambda z)$, with $z \in \mathbb{R}$ and $\lambda \in \mathbb{R}$. Here ϕ and Φ are pdf and cdf of $N(0,1)$, respectively. The case where $\lambda = 0$ reduces to $N(0,1)$. Further probabilistic properties of this distribution are studied in [11, 12]. The multivariate SN family of densities is introduced in [15] and is given by $f(z|\lambda) = 2\phi_k(z)\Phi_1(\lambda^T z)$, $z \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}^k$. Again, ϕ_k is the probability density function of a k-dimensional normal distribution, $N_k(0,1)$. Similar to what is mentioned above, the case where $\lambda = 0$ corresponds to $N_k(0, I_k)$. Further properties of the multivariate SN distribution are studied in [13].

So, in general we are motivated by the broad class of skew multivariate distributions of the form

$$f_p(y; \mu, \Sigma, D) = \frac{1}{\Phi_p(0; I + D\Sigma D')} \phi_p(y; \mu, \Sigma)\Phi_p[D(y - \mu)],$$

where $\mu \in \mathbb{R}^p$, $\Sigma > 0$, $D(p \times p)$, $\phi_p(.; \mu, \Sigma)$ and $\Phi_p(.; \Sigma)$ denote the pdf and the cdf of a p-dimensional symmetric distribution with mean μ and covariance matrix $\Sigma > 0$, respectively. Note that in this case we have a p-integral with upper bounds of $D(y - \mu)$ [54].

In this chapter we consider the simplified representation

$$f_p(y;\mu,\Sigma,\lambda) = 2\phi_p(y;\mu,\Sigma)\Phi_1\left(\lambda^T(y-\mu)\right),$$

where λ is a vector of length p and Φ_1 is the one dimensional cdf of the given distribution. In other words

$$\Phi_1[\lambda^T(y-\mu)] = \int_{-\infty}^{\lambda^T(y-\mu)} \phi_1(x;\mu,\Sigma)dx,$$

as provided in [15]. However, we emphasize that the sRBFs proposed in this chapter need not be skew distributions and as we shall see in the examples, generally they are not.

Nonetheless, a blue-print for constructing sRBFs is motivated by the literature in skew multivariate distributions. The product of a cumulative distribution function with its associated probability density function is a clear candidate. However, given our view to construct functions that efficiently adapt to data we are not concerned whether the representation is strictly speaking a true skew distribution. So, for computational efficiency, we can use the closed form Cauchy cdf (an Arctan function) with an array of different RBFs to generate a family of skew-radial functions. For examples, Erf-Cauchy RBFs (Erf is the Gaussian cdf) , Cosine-Sine RBFs, Cosine-Cauchy RBFs, and many others. An analytic representation of the cdf makes the Cauchy distributions attractive. All we need is a nonlinear modulator that can produce flexibility in the shape of a symmetric RBF. We present several concrete examples of skew-radial functions in the next section.

4.4 Skew-Radial Basis Functions

In this chapter we investigate skew-radial basis function (sRBF) expansions of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i z(x,\nu_i)\phi(\|x-c_i\|_{W_i}), \tag{4.6}$$

where the modulating term $z(x, \nu_i)$ serves to break the radial symmetry of the function $\phi(\|x - c_i\|_{W_i})$. Furthermore, in this work we focus on the special case of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i z(\lambda_i^T (x - c_i)) \phi(\|x - c_i\|_{W_i}), \tag{4.7}$$

where the vector of parameters λ_i determines the shape of the skew-radial function.

For compactness we represent the set of parameters associated with the symmetric portion of an RBF as $\eta_i = (W_i, c_i)$, and the non-symmetric portion as $\nu_i = (\lambda_i, c_i)$ and when we are not concerned about which basis function simply as η and ν. In what follows we consider specific examples of the impact of modulating a symmetric RBF $\phi(x; \eta)$ with an asymmetric function $z(x; \nu)$. Examples of compactly supported sRBFs arise by modulating compactly supported RBFs proposed in Chapter 3.

4.4.1 Erf $z(x; \nu)$; Gaussian $\phi(x; \eta)$

In our first example we employ the symmetric RBF

$$\phi(x; \eta_i) = \exp(-\frac{(x - c_i)^2}{\sigma_i^2}),$$

with an asymmetric modulating function

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

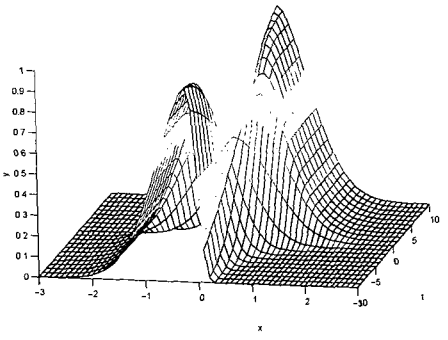
Thus, for the special case where both the domain and range of the data are one-dimensional the Gaussian-Gaussian sRBF is given by

$$\psi(x, \vartheta_i) = \exp(-\frac{(x - c_i)^2}{\sigma_i^2}) \int_{-\infty}^{\lambda (x - c_i)} \exp(-y^2) dy. \tag{4.8}$$

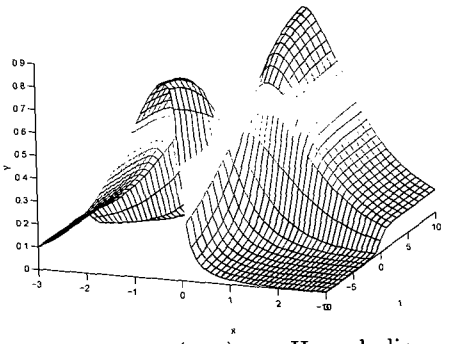
For domains in higher dimensions we have

$$\psi(x, \vartheta_i) = \exp(-\|x - c_i\|_{W_i}^2) \int_{-\infty}^{\lambda_i^T (x - c_i)} \exp(-y^2) dy. \tag{4.9}$$

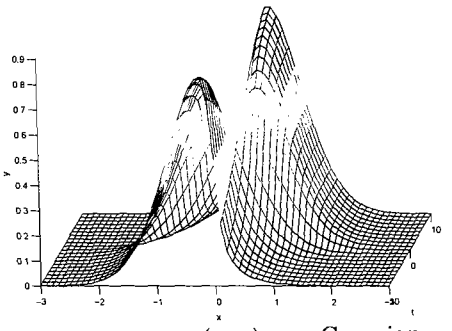
See Figure 4.4 (a) for a plot of the Erf-Gaussian case.



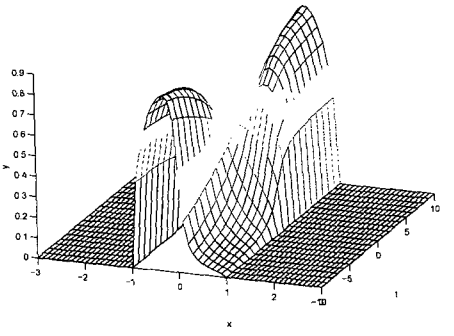
(a) Erf $z(x;\nu)$ - Gaussian $\phi(x;\eta)$.



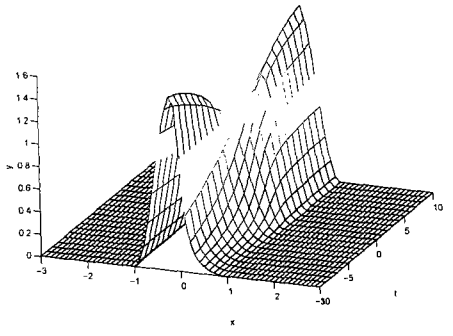
(b) Arctan$z(x;\nu)$ - Hyperbolic-Secant $\phi(x;\eta)$.



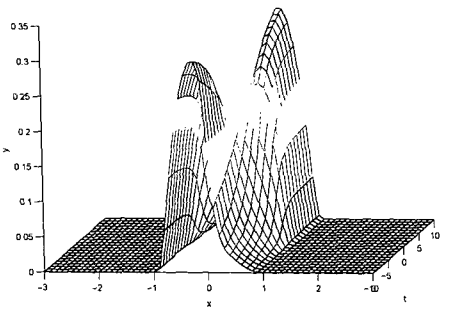
(c) Arctan $z(x;\nu)$ - Gaussian $\phi(x;\eta)$.



(d) Arctan $z(x;\nu)$ - Circle $\phi(x;\eta)$.



(e) Arctan $z(x;\nu)$ - Hanning $\phi(x;\eta)$



(f) Arctan $z(x;\nu)$ - Mollifier $\phi(x;\eta)$.

Figure 4.4: Plots of the one-dimensional (domain) skew-radial basis functions $f(x) = z(x;\nu)\phi(x;\eta)$ where the asymmetry parameter λ is varied from -10 to +10. The product functions are comprised of the following: (a) Erf-Gaussian (b) Arctan-Cauchy (c) Arctan-Gaussian as well as the compactly supported functions (d) Arctan-Circle (e) Arctan-Cosine and (f) Arctan-Mollifier.

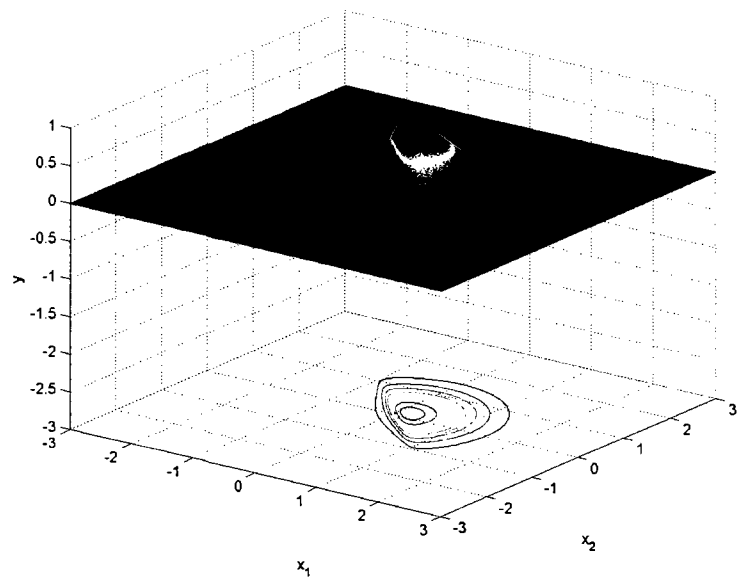


Figure 4.5: Plot of the two dimensional sRBF using Arctan for $z(x;\nu)$ and the Gaussian for $\phi(x;\eta)$ for the specific case $\lambda_1 = \lambda_2 = -10$.

4.4.2 Arctan $z(x;\nu)$; Hyperbolic-Secant $\phi(x;\eta)$

For the symmetry breaking term we may use the arctan function

$$z(x;\nu_i) = \frac{1}{\pi}\arctan\left(\lambda_i^T(x-c_i)\right) + \frac{1}{2}, \tag{4.10}$$

which is also the cumulative distribution function for the Cauchy probability density function. Thus the skew-radial Arctan-Hyperbolic Secant is given by

$$\psi(x,\vartheta_i) = \left(\frac{1}{\pi}\arctan\left(\lambda_i^T(x-c_i)\right) + \frac{1}{2}\right)\mathrm{sech}(\|x-c_i\|_{W_i}). \tag{4.11}$$

It is important to note that $\phi(r) = \cosh^{-\gamma}(r)$ for $\gamma > 0$ is positive definite, [21], so it could be used in both the least square and interpolation senses. Also note that varying γ changes the curvature of the function. This way one could get an RBF that is adaptable to the local curvature of the data. This feature could be used to further reduce the model order. See Figure 4.4 (b) for a plot of the Arctan-Hyperbolic Secant.

For further functions that are positive definite and could be used as RBFs see [125]. In particular all symmetric stable probability distribution functions are positive definite.

4.4.3 Arctan $z(x;\nu)$; Gaussian $\phi(x;\eta)$

For the symmetry breaking term we may use the arctan function given in Equation 4.10. Combined with the Gaussian RBF $\phi(x;\eta_i) = \exp(-\|x - c_i\|^2_{W_i})$, we get

$$\psi(x,\vartheta_i) = \left(\frac{1}{\pi}\arctan\left(\lambda_i^T(x - c_i)\right) + \frac{1}{2}\right)\exp(-\|x - c_i\|^2_{W_i}). \tag{4.12}$$

See Figure 4.4 (c) for a plot of the Arctan-Cauchy case.

This function also coincides with what is provided in [110]. In this reference the pdf is taken to be normal and the cdf comes from variety of distributions such as as normal, t, Cauchy, Laplace and logistic distributions.

4.4.4 Arctan $z(x;\nu)$; Circle $\phi(x;\eta)$

Now consider the Quarter Circle compact function with radius one, i.e.,

$$\phi(r) = \begin{cases} \sqrt{1 - r^2} & \text{if } r < 1, \\ 0 & \text{if } r \geq 1. \end{cases}$$

Using the Heaviside step function notation and letting $r = \|x - c_i\|_{W_i}$, we have

$$\phi(x,\eta_i) = \sqrt{1 - \|x - c_i\|^2_{W_i}}H(1 - \|x - c_i\|_{W_i}).$$

Despite the fact that this RBF is not smooth it has performed well on some data sets, see Chapter 3. Combining this with the arctan symmetry breaking term z we have the sRBF

$$\psi(x,\vartheta_i) = \left(\frac{1}{\pi}\arctan\left(\lambda_i^T(x - c_i)\right) + \frac{1}{2}\right)\sqrt{1 - \|x - c_i\|^2_{W_i}}H(1 - \|x - c_i\|_{W_i}). \tag{4.13}$$

See Figure 4.4 (e) for a plot of the Arctan-Hanning case.

4.4.5 Arctan $z(x;\nu)$; Hanning $\phi(x;\eta)$

In this example the we employ a cosine function in the same fashion as a Hanning filter to produce a RBF with compact support, [71] or Chapter 3, i.e.,

$$\phi(x;\eta_i) = \left(\cos(\|x - c_i\|_{W_i}\pi) + 1\right)H(1 - \|x - c_i\|_{W_i}).$$

To create a sRBF here we employ the Arctan function. These functions, taken together, result in the Arctan-Hanning sRBF

$$\psi(x,\vartheta_i) = \big(\frac{1}{\pi}\arctan\big(\lambda_i^T(x-c_i)\big)+\frac{1}{2}\big)\big(\cos(\|x-c_i\|_{W_i}\pi)+1\big)H(1-\|x-c_i\|_{W_i}). \quad (4.14)$$

See Figure 4.4 (e) for a plot of the Arctan-Hanning case.

4.4.6 Arctan $z(x;\nu)$; Mollifier $\phi(x;\eta)$

The Mollifier, or bump function, also has an attractive form for an RBF given its decay rate and compact support, Chapter 3. It is expressed as

$$\phi(r) = \exp(\frac{-1}{1-r^2})H(1-r).$$

So the sRBF of interest is then

$$\psi(x,\vartheta_i) = \big(\frac{1}{\pi}\arctan\big(\lambda_i^T(x-c_i)\big)+\frac{1}{2}\big)\exp\big(\frac{-1}{1-\|x-c_i\|^2_{W_i}}\big)H(1-\|x-c_i\|_{W_i}). \quad (4.15)$$

See Figure 4.4 (f) for a plot of the Arctan-Mollifier.

4.5 Interpolation with Skew-Radial Basis Functions

In the previous examples we have considered the over-determined data fitting problem which results in an overdetermined least squares system. Alternatively, one may employ these data adapted basis functions, like standard RBFs, to solve the interpolation problem. Such interpolation problems arise from explicit or implicit surface reconstruction, (a compact, orientable manifold), e.g., in image processing, [144], as well as in numerical analysis, [121], and the numerical solution of partial differential equations, see, e.g., a special journal issue devoted to this topic [1]. Given the utility of RBFs in these domains we address the issue of the suitability of asymmetric, or sRBFs, for the interpolation problem. Following the work of [25], we show that under certain conditions we are guaranteed that the sRBFs also can be used to solve the interpolation problem. In [25], it is shown that an interpolation matrix remains positive definite for a bounded

perturbation in scale and shape. We adapt their approach here to show that perturbing the symmetry of the RBFs also results in an interpolation problem with a unique solution.

Our framework is the same as in [25] and we follow the basic ideas and notation presented there. In contrast to their dilation problem, in our application we are concerned with skew-radially, perturbing the conditionally positive definite radial functions $\phi : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ in the multivariate interpolation problem. Our data is spatially finite in extent and so we assume that it consists of the set of points that is a subset of a compact set $\Omega \subset \mathbb{R}^d$, i.e., $X = \{x_1, x_2, x_3, ..., x_N\} \subset \Omega$. By assumption, if

$$A_{jk} = \phi(\|x_j - x_k\|),$$

then the quadratic form $x^T Ax > 0$ is positive definite on the space

$$V := \{\alpha \in \mathbb{R}^N : \sum_{j=1}^{N} \alpha_j p(x_j) = 0 \,\text{for all}\, p \in \mathbb{P}_m^d\}, \tag{4.16}$$

where $\mathbb{P}_m^d$ denotes the space of d-variate polynomials of order not exceeding m [25]. For examples of such (conditional) positive definite functions see Appendix, [106, 144, 131]. If $Q = \dim \mathbb{P}_m^d$ and we require that the interpolation condition

$$y_i = f(x_i)$$

be satisfied, then we seek the unique solution to the $(N+Q) \times (N+Q)$ system

$$\begin{cases} A\alpha + P\beta = y \\ P^T\alpha + 0 = 0, \end{cases} \tag{4.17}$$

where $P_{ij} = p_i(x_j)$, $i = 1, \ldots, Q$ and $\text{rank}(p) = Q < N$ of the form

$$z(x) = \sum_{j=1}^{N} \alpha_j \phi(\|x_j - x\|) + \sum_{i=1}^{Q} \beta_i p_i(x),$$

with $\alpha \in V$.

We propose a sufficient condition for the non-singularity of the system of equations 4.17 perturbed by a radial symmetry breaking function z.

Theorem 4.5.1. *Skew-radial basis functions of the form*

$$f(x) = \sum_{i=1}^{n} \alpha_i z(\lambda_i^T(x - c_i))\phi(\|x - c_i\|) + \sum_{i=1}^{Q} \beta_i p_i(x),$$

which have positive definite interpolation matrices when $\lambda = 0$ *have positive definite interpolation matrices for values of* $\lambda > 0$ *if*

$$\|\lambda\|_\infty \leq \frac{\hat{\lambda}}{NM|z'(\eta_0)| \max\limits_{1 \leq j,k \leq N} \|x_j - x_k\|_1},$$

where $\hat{\lambda}$ *is the smallest eigenvalue of the interpolation matrix* A *in Equation (4.17) restricted to* V.

Proof. Now the skew-radially perturbed basis functions form the interpolation matrix for the perturbed problem is

$$\tilde{A}_{jk} = \phi(\|x_j - x_k\|)z(\lambda_1^T(x_j - x_k)),$$

where λ_1 is the parameter which determines how skew the basis function is. The perturbed version of the system of equations (4.17) can be written as

$$\begin{cases} \widetilde{A}\widetilde{\alpha} + P\widetilde{\beta} = f \\ P^T\widetilde{\alpha} + 0 = f. \end{cases} \tag{4.18}$$

We seek to bound the term

$$(A - \tilde{A})_{jk} = \phi(\|x_j - x_k\|)z(\lambda_1^T(x_j - x_k)) - \phi(\|x_j - x_k\|)z(\lambda_2^T(x_j - x_k)).$$

It follows that

$$(A - \tilde{A})_{jk} \leq |\phi(\|x_j - x_k\|)| \cdot |z(\lambda_1^T(x_j - x_k)) - z(\lambda_2^T(x_j - x_k))|.$$

Define

$$M = \max_{jk} |\phi(\|x_j - x_k\|)|.$$

Then, by the mean value theorem,

$$(A - \tilde{A})_{jk} \leq M|z'(\eta)|(\lambda_1^T(x_j - x_k) - \lambda_2^T(x_j - x_k)),$$

for some $\eta \in [0, \lambda_1^T(x_j - x_k)]$. Further, if

$$\eta_0 = \arg\max_{\eta} |z'(\eta)|,$$

then

$$(A - \tilde{A})_{jk} \leq M|z'(\eta_0)|\lambda_1^T(x_j - x_k),$$

where we assume that we are perturbing about $\lambda_2 = 0$. If we let

$$v = x_j - x_k,$$

then

$$\lambda^T v = \sum_i \lambda_i v_i \leq \sum_i |\lambda_i||v_i|,$$

so

$$\lambda^T v \leq \|\lambda\|_\infty \|x_j - x_k\|_1.$$

Putting it all together and considering $\|A - \tilde{A}\|_2 \leq N\|A - \tilde{A}\|_\infty = N \max_{1 \leq j,k \leq N} |A_{jk} - \tilde{A}_{jk}|$, [60], we get

$$\|A - \tilde{A}\|_2 \leq NM|z'(\eta_0)|\|\lambda\|_\infty \max_{1 \leq j,k \leq N} \|x_j - x_k\|_1.$$

We require that λ be such that

$$NM|z'(\eta_0)|\|\lambda\|_\infty \max_{1 \leq j,k \leq N} \|x_j - x_k\|_1 \leq \hat{\lambda},$$

where $\|A - \tilde{A}\|_2 \leq \hat{\lambda}$(see Appendix). In other words,

$$\|\lambda\|_\infty \leq \frac{\hat{\lambda}}{NM|z'(\eta_0)| \max_{1 \leq j,k \leq N} \|x_j - x_k\|_1}.$$

□

For details concerning values $\hat{\lambda}$ for certain conditional positive definite functions see, [129] and references therein.

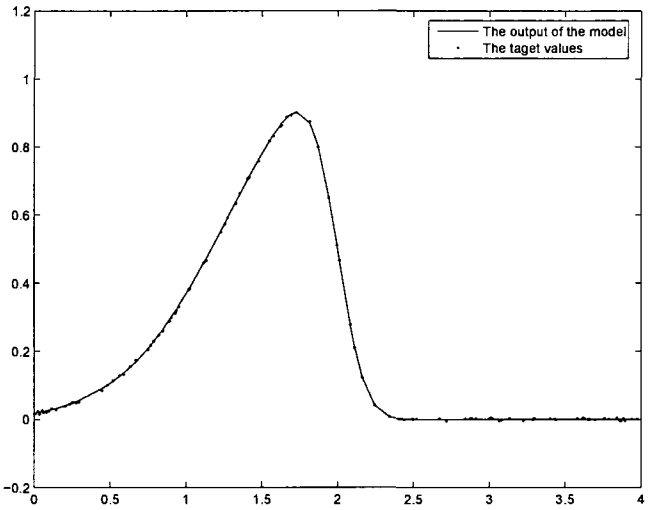


Figure 4.6: The output of the model Equation (4.19) fit to the data set shown in Figure 4.1.

4.6 Numerical Experiments

In this Section we revisit our first skew-radial example from Section 4.2 and present two additional examples to illustrate the enhanced fitting capabilities of sRBFs of the form given in Equation (4.6). To illustrate the absence of the Gibbs phenomenon using sRBFs we fit data sampled from the unit step function in Section 4.6.2. This example is followed by fitting a complex time series of maximal tangential wind velocities generated by a numerical simulation of a hurricane in Section 4.6.3.

4.6.1 Motivating Example Revisited

It was shown in Section 4.2 that attempting to fit asymmetric data by radial functions can result in models of relatively high order as the symmetric functions build up the asymmetry in the data. Now that we have seen an array of sRBFs we show how such a representation can fit the data in Figure 4.1 optimally well. To achieve this we include the skew term in the expansion,

$$f(x) = \sum_{i=1}^{n} \alpha_i \Big(\int_{-\infty}^{\lambda_i(x-c_i)} \exp(-y^2) dy \Big) \phi(\|x - c_i\|_{W_i}). \tag{4.19}$$

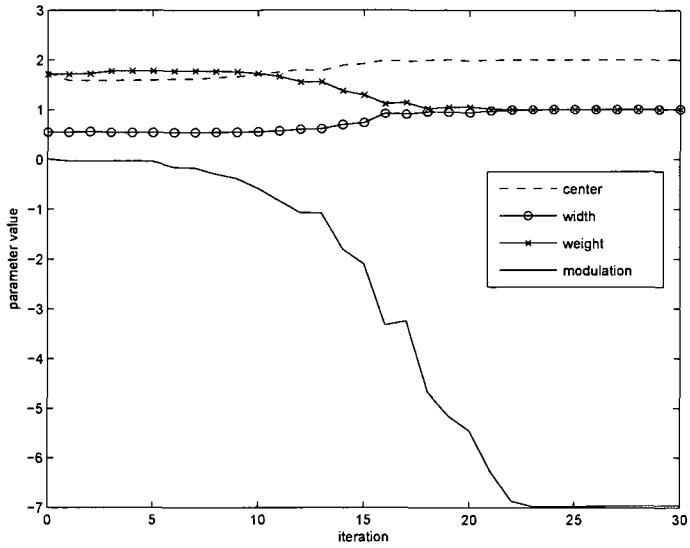


Figure 4.7: The optimization process to fit the Skew data using sRBFs.

By design, given this functional form, we expect that only one basis function is required to fit the data. This is in fact the case as is shown in Figure 4.6. Clearly this example is intended to illustrate the weaknesses of RBFs and the power of sRBFs by selecting data that is least suited to RBFs and most suited to sRBFs. In general this contrasting performance will not be so severe.

Given the simplicity of this example the numerical results are easy to interpret and provide some insight into more complicated examples. After optimization we find, that we need only $n = 1$ terms with the values converged to $\widehat{\sigma} = 1.0005$, $\widehat{\alpha} = 1.0004$, $\widehat{\lambda}$: -6.9695 and $\widehat{\mu} = 1.9999$. This modulated representation produces an $RMSE = 0.0029$ with 98.8% confidence that the residuals are IID noise.

It is informative to examine the convergence of the parameters during the optimization procedure as shown in Figure 4.7. Note that the center, width and weight of the RBF seem to be determined first and then the skewness of the modulating term refines the solution.

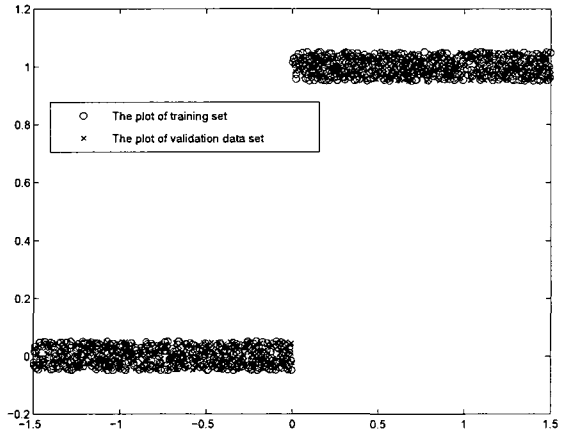


Figure 4.8: The training and validation data sets for the discontinuous step function. The training data has 926 uniformly samples data points and the validation data set consists of 701 data points.

4.6.2 The Unit Step Function

This data set is constructed to demonstrate the performance of sRBFs in fitting sharp discontinuities and edges that may arise, e.g., in the modeling of images, or time series with sudden changes or physical systems with shocks. Note that although we only consider the case of the domain being one-dimensional, the results for a two dimensional domain (discontinuous square) are similar.

In this study it is important to pay attention to the reduction in the Gibbs phenomena. The Gibbs phenomenon was first observed in the context of truncated Fourier expansions. Other variations of the Gibbs effect arise in situations such as truncated integral transforms and for different interpolation methods. The Gibbs effect for several RBFs in one dimension was first studied in [45]. For further study on Gibbs phenomenon using multiquadric RBFs see [77]. Note that the above mentioned references consider the Gibbs effect in the interpolation sense. In this work we examine the capability of our sRBFs to fit a noisy realization of step function denoted by $u(t)+n(t)$ in minimum least square sense. Where $u(t)$ denotes the step function and $n(t)$ is a uniformly distributed noise component with standard deviation of 0.1. A data set consisting of 2250 data

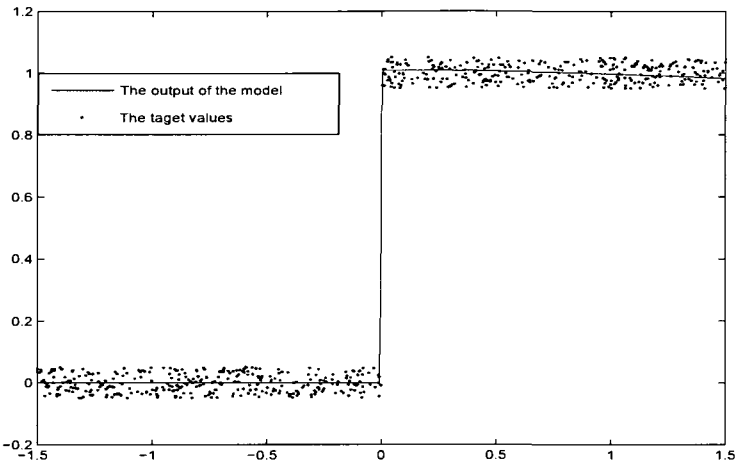


Figure 4.9: The output of the single mode sRBF model.

points is generated. Two data sets of 926 and 701 data points were randomly chosen as training and validation data sets, respectively. Figure 4.8 shows the training and testing data sets. The remaining 623 data points form the testing set.

We have employed our algorithm described in Chapter 2, to fit this time series using both radial and sRBFs.

Figure 4.9 shows the result of the fit using Erf-Gaussian RBFs. The interesting point is the sharp transition in this fit and the flexibility of the RBF to be able to achieve this without introducing the Gibbs ripples around the discontinuity. The $RMSE$ of the single RBF fit is 0.0032. The confidence level on the training set is 96.97% while the validation set reaches 99.14% of confidence. We note that the second and third order statistics on the training set are 98.05% and 97.51%, respectively. The absolute value of the residuals of the training set gets to 98.92% of confidence to be IID noise. The single RBF model has the following parameters; the skew parameter is 6377.5 the center, width and the weight are -0.00124, 9.2277, 1.0086, respectively.

In contrast to the sRBF which is able to adapt its shape to the discontinuity the RBF (Gaussian in this case) is unable to accurately adapt its shape in the neighborhood of the discontinuity as shown in Figure 4.10. Figure 4.10 (a) shows the output of the

intermediate result which employs 12 RBFs in the model. Figure 4.10 (b) shows the result of using 16 RBFs in the model. As far as the symmetric model is concerned the residual errors on the validation set are IID with 95% of confidence using 16 RBFs. The second and third order statistics and the absolute values of the residulas pass 95% of confidence to be IID at the stages of in the process of adding the 16 and 18th RBF. Note the increase in oscillatory behavior in the vicinity of the discontinuity as basis functions are added. Figure 4.10 (c) provides details about the performance of the model in terms of RMSE as RBFs are added. The final RMSE of the symmetric 21-mode model (97% of confidence of training data set) is 0.00404 (at this point higher order statistics indicate that the model residuals are IID).

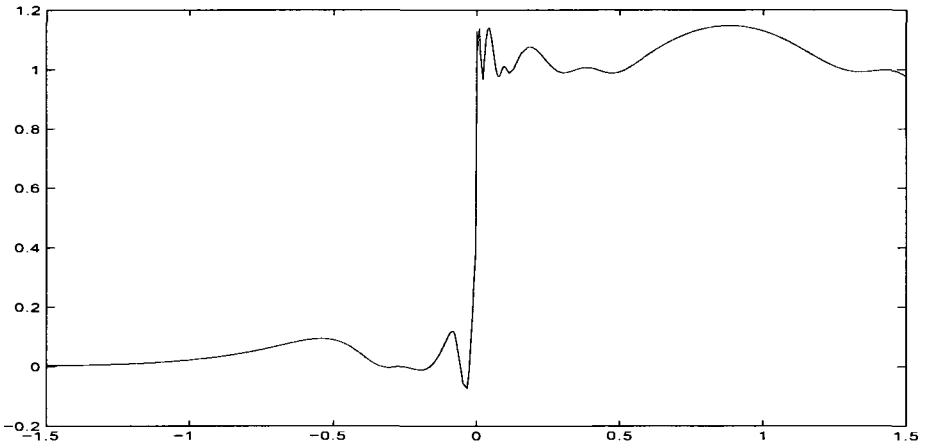
4.6.3 Hurricane Data

The maximum intensity of a hurricane is viewed as a time series in this study to compare the performance of the radial and sRBF. The data set is generated from an axisymmetric simulation of hurricane described in [117]. The main idea of axisymmetric model for a hurricane as a dynamical systems is described in [40, 41] which has similarities to [92]. We construct a prediction problem in a way that we would like to learn the behavior of the dynamical system using a part of the data set and then compare the generalization ability of the RBFs and sRBFs. We employ a time-delay embedding of the time series [36, 37], and use the radial and sRBFs to approximate f which takes four contiguous values and maps them to the next, i.e.,

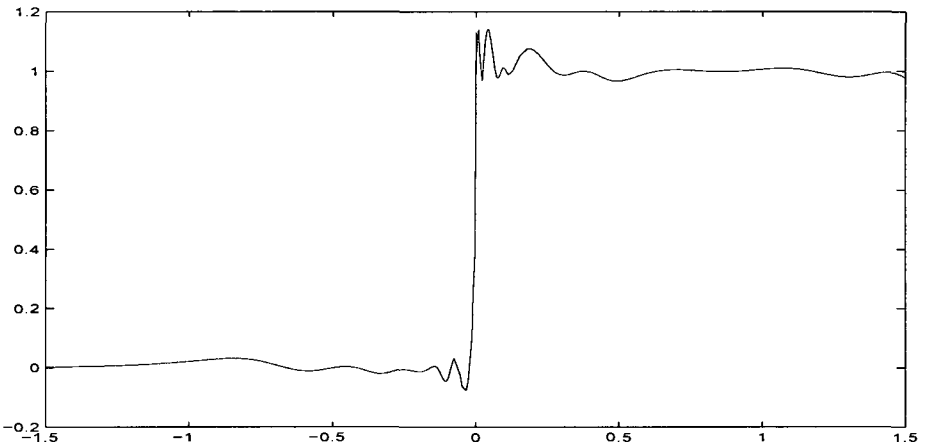
$$x_{n+1} = f(x_n, x_{n-1}, x_{n-2}, x_{n-3}).$$

One-step Prediction

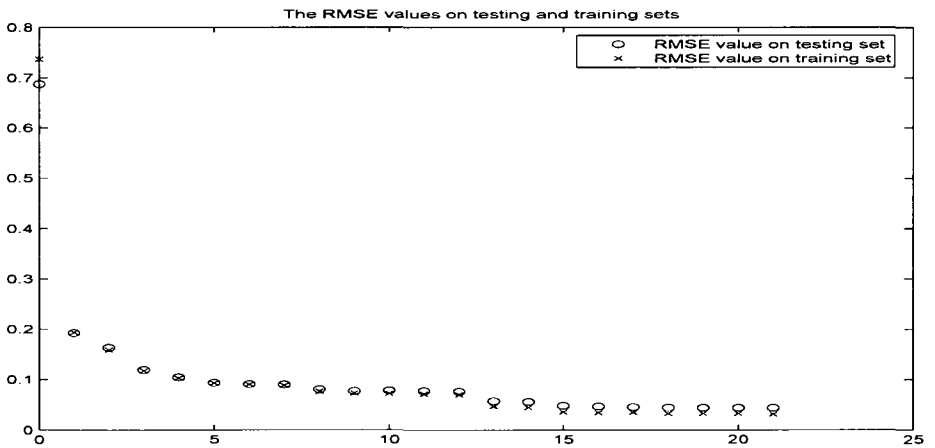
In this one-step prediction problem, we use known data in the domain of f to predict the next unknown value and then compare this with the true next value. During the data fitting stage we compare the predicted values to the true values and adjust the



(a) The output of the 12 RBF model.



(b) The output of the 16 RBF model.



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

Figure 4.10: The outcome of the final and a single mode RBFs and performance of the RBF fit.

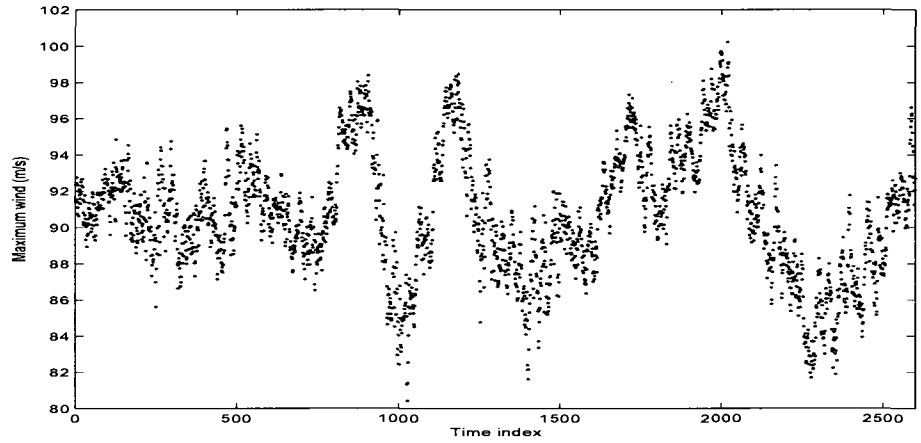
parameters in the radial or sRBFs to minimize these. Again, we employ the training algorithm presented in Chapter 2.

In this study we model the steady state behavior of the dynamics. The data set consists of 1801 training, 800 validation and 500 testing data points [2]. Figure 4.11 (a) shows the training and validation data sets. Note that after disregarding the transient part of the data we select the training data points followed by the validation and finally the testing data points. To be more realistic, there is a difference on how this experiment has been conducted with regard to the other examples in this chapter where training, testing and validation data sets were randomly (uniformly) selected from a larger data set. Figure 4.11 (b) and (c) show the predicted values using symmetric and data adapted Erf-Gaussian RBFs, respectively. Note that the asymmetric fit was complete using one RBF and the RMSE value of the testing data is 1.3036. The 96.22% and 95.62% of confidence was achieved on training and validations data sets, respectively. Where as for the symmetric RBF the training stopped based on 95.37% confidence criteria on validation set with seven RBFs. The 95% confidence on training was reached after three RBFs. When applied the model consisting of seven RBFs to the testing data set, there was a clear effect of over training. By inspection we observed that better results can be reached using only the first three RBFs, the RMSE of testing set using this pruned model is 5.7918. We observe that the sRBF is capable of fitting the space via the training and produces a model that generalizes well. Also one could note that the 95% of confidence is reached with a smaller model order and error using the sRBFs.

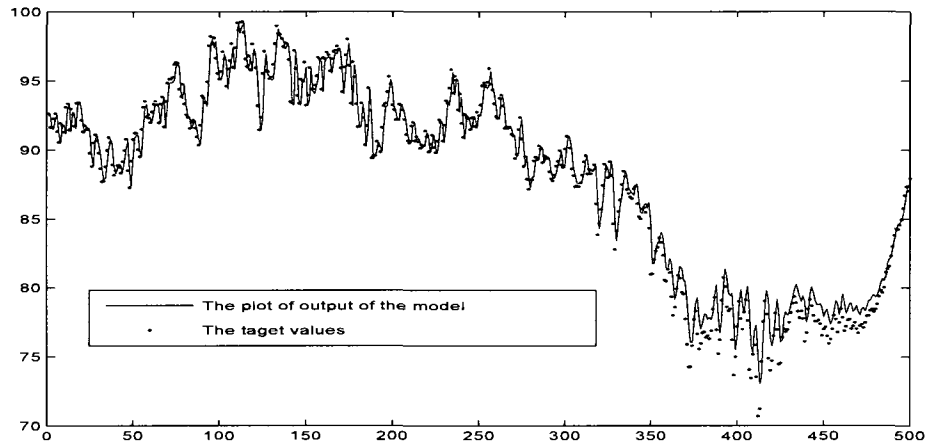
We also compared the performance of circle RBFs to the skew-radial Cauchy-circle basis functions on this data set. In both cases the final model consists of one basis function. For the symmetric RBF, the confidence on training and validation sets are

---

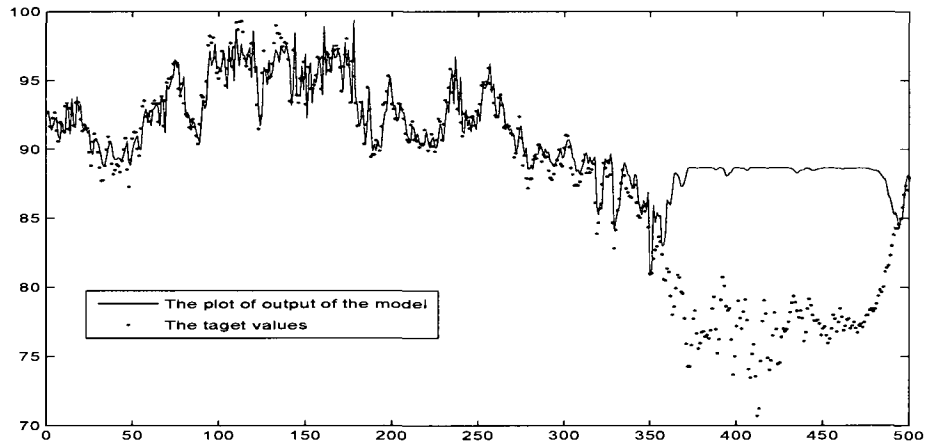
[2]We would like to thank John Persing and Mike Montgomery for providing the hurricane data used in this chapter.



(a) The training and validation data sets.



(b) The testing set and the output of the one mode sRBF.



(c) The testing set and the output of the three mode symmetric RBF.

Figure 4.11: The performance of the radial and skew-radial basis functions on Hurricane intensity prediction.

95.4% and 90.5%, respectively. The RMSE of the final model is 1.23. In the case of the skew-radial fit the confidence on training and validation data sets are 95.7% and 88.6%, respectively. The RMSE of the final model is 1.15. We note that the prediction results of Cauchy-Circle sRBF is superior to the Erf-Gaussian sRBF. The result of the Cauchy-Circle prediction is also better than for the symmetric circle RBF.

Iterated Prediction

It is interesting to look at the iterative predictions using both RBFs and sRBFs. We report the results of this experiment on the testing set and note the results are similar on the training set. In the iterated prediction problem we take the first point on the testing set (domain and the range values) and predict a future value. Then the output value of the model is used as a domain value and the last point in the domain value is disregarded. So, if x_1, x_2, x_3 and x_4 are actual (not predicted data values) we may predict the next value using the model f as

$$\tilde{x}_5 = f(x_4, x_3, x_2, x_1)$$

and at the next step

$$\tilde{x}_6 = f(\tilde{x}_5, x_4, x_3, x_2).$$

After three steps we are using only predicted data to form our new predictions, i.e.,

$$\tilde{x}_{n+1} = f(\tilde{x}_n, \tilde{x}_{n-1}, \tilde{x}_{n-2}, \tilde{x}_{n-3}).$$

We have repeated this iterated estimation 60 times. The NMSE [3] results for both symmetric and non-symmetric types are shown in Figure 4.12. We observe that the sRBF is able to produce better iterative predictions than the RBF after about 11 iterations.

---

[3]Normalized Mean Square Error, NMSE= $\frac{\sum_i e_i^2}{\sum_i (x_i - \bar{x})^2}$ where e_i's are the model residuals.

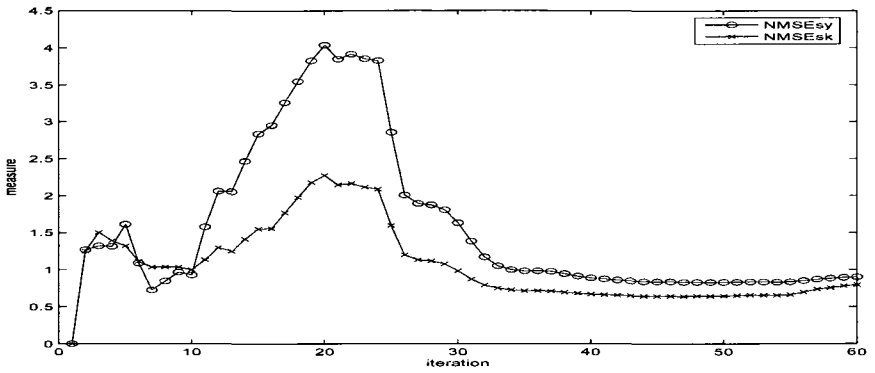


Figure 4.12: The NMSE associated with 60 steps of iterative prediction using radial (circles) and skew-radial (Erf-Gaussian) functions.

It is interesting to note that this suggests that for short spans of iterated prediction the residuals may be essentially symmetric.

Note that for the iterated prediction, the NMSEs for both the skew-radial and radial approximations exceed one for $10 < n < 30$ and $10 < n < 34$ iterations, respectively. In this region the models are performing worse than using the mean as a predictor. Nonetheless, the skew radial approximation outperforms the radial approximation in this region as well, suggesting that the data is not symmetric and should be fit with asymmetric functions. We note also that the very nature of this data, i.e., the spatial location of the point of maximum winds may vary discontinuously, makes prediction a daunting task.

To the best of our knowledge, the prediction and the modeling of hurricane intensity using RBFs has not been explored before. We anticipate that the results presented here will vary to some degree as different embedding dimensions are selected. In this study our focus was to compare the performance of the RBFs and sRBFs. Further exploration of the hurricane data set is outside the scope of this chapter and will be presented elsewhere.

4.7 Relationship to Other Work

4.7.1 Normalized Radial Basis Functions

Normalized RBFs of the form

$$f(x) = \frac{\sum_k w_k \phi(\|x - c_k\|)}{\sum_k \phi(\|x - c_k\|)}, \tag{4.20}$$

were proposed by Moody and Darken, [108]. These normalized RBFs have been compared to standard RBFs in a number of investigations and appear to have advantages, especially in the domain of pattern classification, [33, 140]. For example, it was reported in [33] that using normalized RBFs reduced the order of the model and the robustness of generalization. It has also been reported that normalized RBFs require less data when training models of dynamical systems, [76].

Functions of the form given by Equation (4.20) may also be viewed as non-radial in the sense that the normalization term is a function of the domain location and serves to break the radial symmetry. However, in contrast to the asymmetric functions proposed here there are no additional parameters in the model so this normalization term does not provide as much flexibility for the basis functions to adapt to the data. Of course the normalization of the expansion could also be employed in our context, although no attempt was made to explore that extension here.

4.7.2 Polynomial Modulation

Another approach for breaking the radial symmetry in RBFs is by modulating the RBF with a polynomial term. For example, the following expansion has been proposed

$$f(x) = \sum_k (w_k + \lambda_k^T (x - c_k)) \phi(\|x - c_k\|),$$

as well as its normalized version, [76]. Of course it is also possible to use higher order polynomials in place of the linear term indicated above. These functions do introduce

new parameters into the model via a global modulating term that will have only local impact if the RBFs employed are also local or compactly supported. Again, this term makes the representation functions non-radial but in a manner that is more restricted than the local functions we propose. For example a linear polynomial, or even low order polynomial can't bend it's shape to the data as we have seen in the examples in this chapter, e.g., the step function in Example 4.6.2.

4.7.3 Additional RBFs

An RBF based on the subtraction of two log-sigmoidal functions to generate localized robust RBFs was proposed in [94]. A composite product of log-sigmoidal functions to form localized RBFs as well as a strategy to train an RBF network based on exponentiated gradient was provided in [10]. These RBFs, have the form

$$z_i(x) = \prod_{k=1}^{n} \xi_k^{i,l} \xi_k^{i,r},$$

where $\xi^{i,l} = 1/(1 + \exp(-\beta^i \zeta^{i,l}))$ and $\xi^{i,r} = 1/(1 + \exp(\beta^i \zeta^{i,r})$ with $\zeta^{i,l} = (x - \mu^i) + \theta^i$ and $\zeta^{i,r} = (x - \mu^i) - \theta^i$. In this setting, $\beta^i > 0$ controls the shape of the RBF, x is the input, μ^i is the center of the RBF and θ^i relates to the reception field of the ith node.

Reformulated RBFs are introduced in [80] and therein prior work. These RBFs are intended to facilitate training by supervised learning based on gradient descent. The approach is based on selecting admissible generater functions that satisfy several axioms which are posed as requirements for an RBF. Here we name a few of the generating functions. Exponential generator functions,

$$g_{j0}(x) = \exp(\beta_j x), \quad \beta_j > 0.$$

For $m > 1$, the exponential generator functions correspond to $g_j(x) = \exp(\beta_j x/(1-m))$. This leads to Gaussian RBF, $\phi_j(x) = g_j(x^2) = \exp(-x^2/\sigma_j^2)$, with $\sigma_j^2 = (m-1)/\beta_j$. Linear generator functions, which generate the cosine RBFs, have the form

$$g_{j0} = a_j x + b_j, \quad a_j > 0, b_j \geq 0.$$

This type of generator produces RBFs of the form

$$\phi_j(x) = g_j(x^2) = (a_j x^2 + b_j)^{\frac{1}{1-m}}, \quad m > 1.$$

If $m = 3$ this corresponds to the inverse multiquadratic RBF,

$$\phi_j^q = g_j(x^2) = (a_j x^2 + b_j)^{\frac{1}{-2}}.$$

In [46], a certain class of oscillatory radial functions are proposed as RBFs. These RBFs lead to non-singular interpolants and become increasingly flat by scaling. This flat limit is important in that it generalizes traditional spectral methods to completely general node layouts. The RBF is given in the from

$$\phi_d(r) = \frac{J_{\frac{d}{2}-1}(\varepsilon(r))}{(\varepsilon(r))^{\frac{d}{2}-1}}, d = 1, 2, ...,$$

where $J_\alpha(r)$ denotes the Bessel function of the first kind with order α. These RBFs will give nonsingular interpolation up to d dimensions when $d \geq 2$.

4.8 Conclusions

We have proposed a class of skew-radial basis functions (sRBFs) which is formed by modulating an RBF with a symmetry breaking term. Several forms of these are suggested by multivariate skew distributions from statistics. Additional sRBFs, some of which are compactly supported, are also proposed which do not arise from the skew distribution literature.

The numerical experiments provided suggest that the proposed sRBFs adapt more flexibly to asymmetric data than their radial counterparts. The improved fit is not only the result of additional model parameters but rather the ability of the sRBFs to better match the shape of the data. In our motivating example we saw how ill-suited RBFs are at approximating functions outside of their native space, e.g., the space of sRBFs. In this case the skew-radial model required four parameters while the radial model required

26 parameters such that the model residuals were statistically IID. Similarly, we saw that sRBFs are especially well-suited for fitting data with jumps in it, such as we might find in physical flows with singularities, or in images with sharp edges arising from, e.g., shadows.

In the modeling of the hurricane maximum wind speed time-series we observed that the sRBFs produced a fit that had a significantly lower error than the RBF approach and that increasing the complexity of the RBF model did not diminish this discrepancy. In this problem we examined the models for their *relative* predictive capabilities only. Currently we are collaborating with atmospheric scientists to apply sRBFs to the problem of understanding of nonlinear relationships in the hurricane intensification process.

We anticipate that these new developments could impact other function fitting paradigms such as support vector machines, [132], and mixture models, [105].

Chapter 5

CONVERGENCE ANALYSIS

Abstract In this chapter we present a detailed analysis of an approach for constructing nonlinear empirical mappings from high-dimensional domains. We employ skew Radial Basis Functions (sRBFs) for constructing a model using data that may be scattered and 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 on training and validation data that reveals geometric structure when it fails. At each step the added function is fit to data contained in a spatio-temporally defined local region to determine the parameters, in particular, the scale of the local model. The scale of the function is determined via 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 parameter. Compactly supported skew-radial basis functions are employed to improve model accuracy, order and convergence properties. A detailed analysis of the convergence of the algorithm is presented in the context of several hypotheses tests. We illustrate the new methodologies using several illustrative problems including modeling data on manifolds and prediction of a chaotic time series.

5.1 Introduction

The discovery of knowledge in large data sets can often be formulated as a problem in nonlinear function approximation. The inherent challenge in such an approach is

that the data is often high dimensional, scattered and sparse. Given a limited number of exemplars one would like to construct models that can generalize to new regions or events. Additionally, underlying physical processes may not be stationary and the nature of the nonlinear relationships may evolve. Ideally, a good model will also be able to adapt and remain valid over extended regions in space and time.

In this chapter we consider an approach for constructing mappings of the form

$$f : U \in \mathbb{R}^n \rightarrow V \in \mathbb{R} \tag{5.1}$$

from empirical data. Again, we assume that we have samples $x^{(k)} \in U$ and $y^{(k)} \in V$ that are indexed by μ and related via

$$y^{(k)} = f(x^{(k)}). \tag{5.2}$$

In practice this may be, e.g., a mapping from a manifold U to a vector field V or, alternatively, a mapping of multiple time series $(x_1(t), \ldots, x_n(t))$ to a value at some future time $t + T$, i.e., $x(t + T)$.

In modeling with nonlinear functions we are confronted with several critical issues that either do not arise in the linear setting or where the nonlinear setting produces unique challenges. Our prescription for a useful algorithm includes two main features: the algorithm

- can be applied to a wide range of data sets where little or no detailed knowledge is available
- requires few or no user adjusted parameters

Effectively we seek an essentially *black box* algorithm that requires little expertise on the part of the user to successfully build a model.

A central assumption for our approach is that the mapping consists of the superposition of signal plus noise, i.e.,

$$f(t) = s(t) + n(t).$$

As we shall see, the algorithm exploits this assumption and iteratively fits the signal portion $s(t)$ of $f(t)$ until the residuals of model pass a Null Hypothesis for noise, i.e., they behave as $n(t)$.

A comprehensive survey of the literature related to RBFs is provided in Chapter 2, see also [66, 72], and we will not attempt to reproduce that here. While the previous chapters form a starting point for this work but as will be described below, the essential ingredients of the algorithm are substantially changed. Further, we provide a solid theoretical foundation for the convergence of the new algorithm. Serendipitously, this investigation into convergence led us to a new accelerated form of the algorithm which we show is asymptotically equivalent to the previous version.

The organization of this chapter is as follows: Section 5.2 provides introduces the Black Box RBF algorithm and highlights the features that provide enhanced modeling capabilities which include compact and skew shaped functions. Section 5.3 provides several convergence results for the algorithm and illustrates their behavior with several examples. Finally, Section 5.4 provides some concluding remarks and discusses avenues for future work.

5.2 A Blackbox Algorithm

As in the previous chapters, at the center of the *black box* algorithm for modeling mappings from scattered data is the objective that the residuals of the model should contain no geometric structure, i.e., the data should pass one or more hypothesis tests indicating that they are some form of statistical noise. This test for structure is done iteratively is applied globally at each iteration. When the data is deemed to have persistent structure, i.e., there is some location that does not appear to be noise, a new basis function is added to the point in the domain where the structure is deemed to be greatest. In contrast to our prior work which considered only the autocorrelation function test for IID noise here we implement the following suite of indicators:

- Turning point test
- Difference sign test
- Autocorrelation function test

This collection of tests permits the robust identification of structure in the residuals of the model. Further, as will be shown below, the algorithm described in this section is guaranteed to satisfy the Null Hypotheses for these tests.

Before describing details of the proposed algorithm we first characterize the general setting of the data fitting problem. We assume that we have collected L input-output pairs $\{(x_l, y_l)\}_{l=1}^{L}$, $\mathcal{X} = \{x_l\}_{l=1}^{L}$ and $\mathcal{Y} = \{y_l\}_{l=1}^{L}$, is given the goal is to find the underlying mapping f such that $y_l = f(x_l)$ for each $l = 1, \ldots, L$.[1] We would like to approximate f using an RBF expansion of the form

$$f^K(x) = Ax + \alpha_0 + \sum_{k=1}^{K} \alpha_k \phi_k(\|x - c_k\|_{W_k}), \tag{5.3}$$

where x is an input pattern, ϕ_k is the kth RBF centered at location c_k, and α_k denotes the weight for kth RBF and A is an $m \times n$ matrix. The term W denotes the parameters in the weighted inner product $\|x\|_W = \sqrt{x^T W x}$.

Thus, at iteration K of the algorithm, the problem is to minimize the cost function

$$E_K(v_K) = \frac{1}{2} \sum_{l=1}^{L} \| f^K(x_l) - y_l \|^2, \tag{5.4}$$

where the parameters associated with the newly added function ϕ are $v_K = [\alpha_K, c_K, W_K]$ and f^K is an optimal approximation for f. Note that in general not all the data is used here but only data near the point where the new center is being added; see Section 5.2.3 for details. The model residual for the lth data point is defined as

$$e_l^K = y_l - f^K(x_l).$$

---

[1] In practice the L elements are actually divided into training, validation and testing sets as will be described in Section 5.3 in the context of specific examples.

The set of residuals for a model of order K, is defined as

$$R^K = \{e_l^K\}_{l=1}^L. \tag{5.5}$$

At each iteration an hypothesis test is applied to this global set of residuals to determine whether they are representative of IID noise. If so, the algorithm terminates and if not, the point of maximum departure from IID noise is determined. This point in the range may be traced back to a point in the domain where a new basis function is to be placed. The details of how this is done are describe in Section 5.2.2. Then, the fitting parameters v_K are determined by minimizing the cost function in Equation (5.4) and the process is repeated. If no location in the data appears to possess structure, i.e., a global test on the residuals passes an hypothesis test for noise, then the algorithm stops. If the hypothesis test fails, then another iteration of the algorithm is implemented.

In what follows we present new features in this algorithm that arose through theoretical considerations (as will be described in Section 5.3) yet actually enhanced the performance via both producing smaller models and by producing lower errors. These may be viewed as basic modifications to the algorithm that was originally introduced in Chapter 2 which in turn extended the basic algorithm proposed in [6, 5].

5.2.1 Highlights of Algorithm

Since the papers [71, 66, 72, 69] appeared our extensive experience with this algorithm has led us to discover several modifications which can result in improvements in the resulting model, i.e., smaller models that require fewer iterations of the algorithm. In the following subsections we highlight the features of the new algorithm.

5.2.2 Enhancement of the Autocorrelation Function Test

Here we describe two important aspects of our algorithm. First, whether the residuals of the model are indeed IID and, second, if the residuals are not IID, how this failure may be used as a guide to place the new RBF. We present the standard Autocorrelation

Function Test followed by a simple enhancement that we have found useful in practice. For the sake of completeness we first describe the ACF test for IID noise.

The standard definition for the sample autocorrelation function, $\widehat{\rho}(h)$, (ACF) for a set of residuals $e_1, e_2, e_3, ..., e_L$ with sample mean $\bar{e}$ and lag h is defined as

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}, \tag{5.6}$$

where $-L < h < L$, and

$$\widehat{\gamma}(h) = \frac{1}{L} \sum_{i=1}^{L-|h|} \alpha(h, e_i) \tag{5.7}$$

where

$$\alpha(h, e_i) = (e_{i+|h|} - \bar{e})(e_i - \bar{e}). \tag{5.8}$$

For a fixed lag h the quantity $\alpha(h, e_i)$ is the contribution of the ith residual to the autocorrelation function.

For large L, the sample autocorrelations of an IID sequence $x_1, x_2, ..., x_L$ with finite variance are approximately IID with normal distribution with mean zero and variance $1/L$, $N(0, 1/L)$, [28] p. 222. Hence, if our set of residuals $e_1, e_2, ..., e_n$ is a realization of such an IID sequence, then 95% of the sample autocorrelations should fall between the bounds

$$\frac{-1.96}{\sqrt{L}} < \widehat{\rho}(h) < \frac{1.96}{\sqrt{L}}. \tag{5.9}$$

When this test fails it is due to the fact that some of its components are necessarily too large. To identify which components are contributing to the failure of the test we compute the lag h^* which contributes the most to $\widehat{\rho}(h)$, i.e.,

$$h^* = \arg \max_{h>0} \widehat{\gamma}(h). \tag{5.10}$$

Then, we find the residual that has the maximum contribution to the ACF for lag $h = h^*$ by solving

$$i^* = \arg \max_{i=1,...,n-h} \alpha(h^*, e_i). \tag{5.11}$$

Since i^* is simply the label of the residual we may place the new basis function initially at the point x_{i^*}.

In this chapter we redefine α to be

$$\alpha(h, e_i) = e_{i+|h|}e_i, \tag{5.12}$$

effectively ignoring the mean of the residuals in the calculations. Indeed, we observe that the mean of the residuals goes to zero after a small number of iterations of Algorithm 2. We prove in Section 5.3 that Algorithm 2, which employs the new mean-free definition of α in Equation (5.12), is asymptotically equivalent to the algorithm which uses Equation (5.8) to define α. Thus, we may view the new and old algorithms as accomplishing the same end but the behavior of the modified algorithm in the early iterations has been found to be superior resulting in better solutions more quickly.

The idea of placing a new RBF at the location of maximum contribution to the ACF follows [6, 5]. Note that the locations of the new basis functions are fundamentally changed when the mean of the residuals is ignored as proposed here.

5.2.3 Defining Local Regions

At every iteration of this algorithm (save for the last one) a basis function is added and parameters are fit. These parameters will be determined by minimizing the cost function, e.g., as described in Equation (5.4). This cost function is predicated on the data that is included in its evaluation. In [6, 5] the data was defined by those points that were contained in a contiguous time window that included the point x_{i^*}, i.e., the initial location of the new center, and bounded by locations where the ACF had local minima. In Chapter 2 this idea was extended to make the data ball not simply temporal but spatio-temporal by including all the data in the domain that resided in the initial temporal window. In this work the local data region to be used at each iteration to evaluate the parameters has been modified to include the data points between the two

zero crossings of the auto correlation function. This modification produces improved convergence properties and also provides a theoretical foundation for proving convergence of the algorithm. It is the basis for the Zero Crossing Theorem that will be proved in Section 5.3.1.

5.2.4 Optimization Algorithms and Cost Functions

We have examined the performance of a variety of optimization methods to construct models. Primarily the algorithm employs the BFGS quasi-Newton method [100]. We have also implemented conjugate gradient as well as steepest descent and variations of these methods using alternative direction of descent (ADD) on the parameters. Here we have focused on descent algorithms generated by cost functions based on errors in the least squares sense. Again, as described above, at each iteration the data used to compute the parameters of the new RBF is local to the placement of the new center of the RBF.

We have observed that at times the optimization routine which optimizes the error function over the local data may indeed lead to increases in error associated with the data not included in the local region. This has led to our preference of compactly supported RBFs as described in Section 5.2.6. In addition, we control the total error by implementing a verification that the error of the model as defined over the entire data set is not increasing. If this is the case then the algorithm simply exits the training of the parameters and proceeds to another test of IID residuals. In practice this happens about one in twenty functions.

One may envision adjusting the direction of the gradient by constraining the optimization function to prohibit increases in the error associated with data not being included in the fitting procedure. The performance of the algorithm on the test problems has not indicated that this additional complexity is warranted at this point.

5.2.5 Additional Considerations

As outlined in Section 5.2.2, the selection of an initial location in the domain for the center of a new RBF, we have consider the value h^* which maximizes $\widehat{\gamma}(h)$ as defined in Equation (5.10). The region in the domain is then selected as one which has the maximum value of $\alpha(h^*, i)$ over the residuals producing the optimal index $i = i^*$. Alternatively, we have considered looking for the region in this function $\alpha(h^*, i)$ that has the maximum area between zero crossings. The distinction between these two procedures amounts computing the maximum l_2 or l_∞ norms of $\alpha(h^*, i)$ between zero crossings. For the examples we have seen the two approaches generally agree in terms of model order but the actual location of the basis functions may vary. In general one might speculate that in the initial stages of model learning that the l_2 norm would fit more energy in the data while towards the latter stages of learning the l_∞ norm can capture small scale features that need to be targeted for fitting.

A summary of the algorithm employed in this chapter is provided in Algorithm 2.

5.2.6 Compactly Supported RBFs for Data Fitting

Most applications employ RBFs from a relatively small list, including Gaussians, multi-quadrics and thin plate splines, see Appendix. Such RBFs arise in the context of the interpolation problem which results in a linear system of equations for the weights. In this investigation we are primarily concerned with the data fitting problem which leads to an over-determined system of linear equations. In the latter situation, there is far more latitude in how the functions are selected. Of course, issues such as condition number of the interpolation matrix impact the relative quality of candidate RBFs. Several examples of compactly supported RBFs and a discussion on their properties is provided in Chapter 3.

Algorithm 2 A new RBF fitting Algorithm using Spatio-Temporal Ball.

$ran\_flag = 1$, $K = 0$
while $ran\_flag = 1$ **do**
  evaluate the RBF on the training data set $\{f(x_n)\}_{n=1}^{L}$
  compute the model error $\{e_n\}_{n=1}^{L}$
  compute component contributions $\alpha(h, e_i) = e_{i+|h|}e_i$
  compute ACF for all $0 \leq h < L$
  if the autocorrelation test is rejected **then**
    compute h^* via equation $h^* = \arg\max \widehat{\gamma}(h), h > 0$ and
    compute $x^* = x_{i^*} = e^{-1}(e_{i^*})$ where $i^* = \arg\max_{i=1,...,n-h} \alpha(h^*, e_i)$
    compute the ACC function, $\beta_i = \alpha(h^*, e_i)$, $i = 1, ..., n - h$
    find the right and left zero crossing of the ACC 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 $v = [c_0, \sigma_0, \alpha_0]^T$
    solve $E(v) = \min_{v} \|h(x; v) - y\|_2^2$, where $x \in \mathcal{X}_{local}$
    $K = K + 1$
  else
    $ran\_flag = 0$
  end if
  compute confidence, RMSE and $\widehat{\gamma}(h^*)$ of of the current model on the training set
end while

5.2.7 Skew RBFs

Fitting asymmetric data by the superposition of symmetric functions can lead to models which are non-optimal. In this section we briefly describe how RBFs may be skewed to better approximate asymmetric data. For further discussion on this topic we refer the reader to Chapter 4.

As described in Chapter 4, a skew-radial basis function (sRBF) expansions is of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i z(x, \nu_i) \phi(\|x - c_i\|_{W_i}), \tag{5.13}$$

where the modulating term $z(x, \nu_i)$ serves to break the radial symmetry of the function $\phi(\|x - c_i\|_{W_i})$. Furthermore, in this work we focus on the special case of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i z(\lambda_i^T (x - c_i)) \phi(\|x - c_i\|_{W_i}), \tag{5.14}$$

where the vector of parameters λ_i determines the shape of the skew-radial function.

See Chapter 4 for details concerning the numerical optimization of skew RBFs. We include mention of them here since they are used in some of the examples described in this later chapter.

5.3 Convergence Theory and Examples

In this section we would like to establish a theoretical framework for the RBF fitting algorithm. To establish some properties concerning the convergence of Algorithm 2, we formulate a theorem to show that at the completion of each iteration of the RBF fitting, the number of zero crossings of the residuals increases.

5.3.1 Zero Crossings

Here we make precise the definition of the number of zero crossings of a sequence $\{e_n\}$. First, we define a derived sequence of zeros and ones

The number of zero crossings of a signal sampled in time is defined using the

$$p_n = \begin{array}{ll} 1, & \text{if } e_n \geq 0 \\ 0, & \text{if } e_n < 0 \end{array} \tag{5.15}$$

for $n = 1, ..., L$. Further define

$$d_n = (p_n - p_{n-1})^2 \tag{5.16}$$

to be the indicator function at time n. So d_n is either 0 or 1 for any n. When $d_n = 1$ we say a zero crossing occurs between n and $n-1$. The number of zero-crossings of the set of residuals $R^K = \{e_n^K\}$ will be denoted by $D_1(R^K)$ and is defined by $D_1(R^K) = d_2 + ... + d_L$[2].

Theorem 5.3.1. (Zero Crossing Theorem) *The number of zero crossings of residuals of the Kth order model $\{e_i^K\}$ is larger than the number of zero crossings of residuals of a $K-1$th order model $\{e_i^{K-1}\}$, i.e., $D_1(R^K) > D_1(R^{K-1})$.*

Proof. As described in Algorithm 2, a potential place to add a new RBF is identified using the ACF test (the place that makes the maximum contribution to the autocorrelation of the signal). At the same time, a local set of data is determined and denoted $\mathcal{X}_{local}$. Since, as described above, this data is centered at x^* we may define a ball of radius d_c around $\mathcal{X}_{local}$ as $\Omega = B(x^*, d_c)$.

Let v^+ (containing parameters of center, width, weight and skew) denote the optimized parameters for the Kth, RBF ϕ_K, i.e.,

$$v^+ = \underset{v}{\operatorname{argmin}} \sum_{x \in \mathcal{X}_{local}} (\phi_K(x; v) - e^{K-1}(x))^2.$$

We would like to show that $\exists \gamma \in \Omega \supset \mathcal{X}_{local}$, s.t.,

$$\phi_K(\gamma, v^+) - e^{K-1}(\gamma) = 0.$$

---

[2]Note that D_1 denotes the number of the first order zero crossings. For kth order zero crossings D_k is defined to be the number of zero crossings of $k-1$th difference of R^K, i.e., $\nabla^{k-1} R^K$. Where R^K could be any time series of length L. More precisely, let B denote the shift operator, with $Be_n = e_{n-1}$. Therefor $\nabla e_n = (1-B)e_n = e_n - e_{n-1}$ and $\nabla^k = (1-B)^k e_n = \sum_{j=0}^{k} \binom{k}{j} (-1)^j e_{n-j}$.

This indicates that there is a new sign change in the residuals R^K.

If $\phi_K(\gamma, v^+) - e^{K-1}(\gamma) \neq 0$, then without lose of generality one could assume $\phi_K(\gamma, v^+) - e^{K-1}(\gamma) > 0$, $\forall \gamma \in \Omega$.

Let

$$\alpha_* = \min_{\gamma}(\phi_K(\gamma, v^+) - e^{K-1}(\gamma)).$$

Then $\exists v^*$, s.t., $\phi_K(\gamma, v^*) = \phi_K(\gamma, v^+) - \alpha_*$ and

$$\sum_{x \in \mathcal{X}_{local}} (\phi_K(x; v^*) - e^{K-1}(x))^2 < \sum_{x \in \mathcal{X}_{local}} (\phi_K(x; v^+) - e^{K-1}(x))^2.$$

This contradicts the fact that v^+ is the minimizer.

If $\phi_K(\gamma, v^+) - e^{K-1}(\gamma) < 0$, then the same argument as above applies.

Thus $\exists \gamma \in \Omega$, s.t., $\phi_K(\gamma, v^+) - e^{K-1}(\gamma) = 0$ which establishes the introduction of a new zero crossing to the residuals of the model of order K. By construction the RBF is contained inside the support of the zero crossings of the residuals, hence two zero crossings must be added, i.e.,

$$D_1(R^K) \geq D_1(R^{K-1}) + 1.$$

In practice we see that as a result of noise the number of zero crossings in the residual may actually increase significantly from one iteration to the next.

It is important to note that according to Algorithm 2, $\mathcal{X}_{local}$ is defined based on two consecutive zero crossings of the function $\alpha(h^*, i)$ which in turn corresponds to two zero crossings of the residuals $\{e_i^{K-1}\}$. The above analysis assumes that the RBF is compactly supported and that the support of the RBF is contained between two zero crossings of the residuals function. In the case that the RBF is not compactly supported then the tails of the RBF may exit the local region possibly resulting in the removal of some of the zero crossings in R^K. □

5.3.2 Theorems on Norms of Residuals

In what follows we assume that the signal is an additive composition of a smooth signal $s(t)$ and noise $n(t)$, i.e.,

$$x(t) = s(t) + n(t),$$

where $n(t)$ is uncorrelated with $s(t)$[3]. Here we assume that $n(t)$ is IID up to second order statistics. We will assume that $e^k(t)$ denotes the function of residuals at the kth iteration of the algorithm. Then, in one iteration of modeling the residuals become

$$e^{k+1}(t) = e^k(t) - \alpha_k \phi(t),$$

where ϕ is assumed to have compact support and is zero outside of $\mathcal{X}_{local}$.

Theorem 5.3.2. *The residuals $e(t)$ converge to the noise $n(t)$, i.e.,*

$$\lim_{k \to \infty} e^k(t) = n(t)$$

Proof. Again, in one iteration, the residuals $e^k(t)$ will be fit with the RBF $f(t) = \alpha_k \phi(t)$ to produce new residuals $e^{k+1}(t)$. We assume that the fit is taking place between two zero crossings of the residuals, i.e.,

$$e^k(a) = e^k(b) = 0$$

and

$$e^k(t) \neq 0 \ \forall t \in (a, b).$$

We will assume, without loss of generality, that the function $f(t)$ is positive on (a, b).

Observe that since $f(t) > 0$, it follows that

$$e^{k+1}(t) < e^k(t), \ \forall t \in (a, b).$$

---

[3]The connection to our previous notation is simply $e(t_n) = e_n$.

Hence we have

$$\|e^{k+1}(t)\| < \|e^{k}(t)\|$$

where the statement is true for both the 2-norm and ∞-norm.

Note that the least squares parameter is selected such that it decreases monotonically, i.e.,

$$\alpha_k < \alpha_{k-1}$$

and since it is bounded below by zero we have

$$\lim_{k\to\infty} \alpha_k = 0,$$

or, equivalently,

$$\lim_{k\to\infty} \int e^{k}(t)\phi(t)dt = 0$$

from which it follows

$$\lim_{k\to\infty} \int (s^{k}(t) + n^{k}(t))\phi(t)dt = 0$$

But since

$$\lim_{k\to\infty} \int n^{k}(t)\phi(t)dt = \int n(t)\phi(t)dt = 0,$$

where we have assumed that the noise is uncorrelated with the RBF, it follows

$$\lim_{k\to\infty} \int s^{k}(t)\phi(t)dt = 0$$

so the signal left in the residual is zero, i.e.,

$$s(t) = 0$$

and

$$e(t) = n(t)$$

in the limit.[4] □

[4]We may move the limit into the integration since the sequence of functions is Cauchy.

So, note that since by assumption the mean of $n(t) = 0$, it follows that the mean of $e(t) = 0$ in the limit. In practice, the mean converges to zero very quickly, see, Figure 5.6 for an application to the Pringle data set and Figure 5.14 for results on the Mackey-Glass data set.

As a simple corollary we have that the norm of the residuals is the same as the norm of the noise in the limit, i.e.,

Theorem 5.3.3. *The energy of the residuals $e(t)$ converges to the energy of the noise $n(t)$, i.e.,*

$$\lim_{k\to\infty} \|e^k(t)\| = \|n(t)\|.$$

Proof. This follows directly from the results above. □

Theorem 5.3.4. *The mean of the residuals $e(t)$ converges to zero, i.e.,*

$$\lim_{k\to\infty} \int e^k(t)dt = 0.$$

Proof. This follows directly from the results above. □

5.3.3 Autocorrelation Function Test

Recall the sample autocorrelation function test for white noise, [28]: For large n, the sample autocorrelation of an IID sequence $Y_1, Y_2, Y_3, ...$ with finite variance are approximately IID with distribution $N(0, \frac{1}{n})$. Hence, if $y_1, y_2, y_3, ...$ is a realization of such an iid sequence, about 95% of the absolute values of the sample autocorrelation should be smaller than $\frac{1.96}{\sqrt{n}}$. If we compute the sample autocorrelation up to lag 40 and find that more than two or three values fall outside the bounds, or that one value falls far outside the bounds, we therefore reject the IID hypothesis.

Let

$$\widehat{\rho}_k(h) = \frac{\widehat{\gamma}_k(h)}{\widehat{\gamma}_k(0)}$$

where $h > 0$.

Theorem 5.3.5. *The Null Hypothesis for the ACF test will be accepted as $k \to \infty$, i.e.,*

$$\lim_{k\to\infty} \widehat{\rho}_k(h) = 0$$

for all $h > 0$ where the sequence of residuals is generated by Algorithm 2.

Proof. By definition

$$\widehat{\rho}_k(h) = \frac{\int (e^k(t) - \overline{e}^k))(e^k(t-h) - \overline{e}^k))dt}{\int (e^k(t) - \overline{e}^k)^2 dt}.$$

In the limit we have

$$\lim_{k\to\infty} \widehat{\rho}_k(h) = \int \lim_{k\to\infty} \frac{e^k(t)e^k(t-h)}{\sigma} dt = \int \frac{e(t)e(t-h)}{\sigma} dt = 0,$$

where we have used the fact that $e_k(t) \to e(t) = n(t)$ in the limit and

$$\lim_{k\to\infty} \int (e^k(t) - \overline{e}^k)^2 dt = \sigma^2,$$

is the variance of the noise $n(t)$. □

The IID test is applied under the null hypothesis that the residuals at each step are uncorrelated. As we notice this is a necessary but not a sufficient test. We have proved that the Null Hypothesis will be satisified as $k \to \infty$. In practice, test is satisfied with 95% of confidence where k is a reasonable model order size.

We remark that one can also produce additional results for the behavior of the ACF components that are not asymptotic.

5.3.4 Spectral Analysis of Zero Crossings

In the discussion above the arguments where specific to the behavior of the algorithm and took place in the time domain. There are also some interesting and relevant results concerning the relationship between zero crossings and the autocorrelation function in the frequency domain. The results in this section are presented in [81].

Let (x, y) have a bivariate normal distribution with parameters $EX = EY = 0$, $VarX = VarY = \sigma^2$ and correlation ρ. It is reported in [81] that

$$Pr(X \geq 0, Y \geq 0) = \frac{1}{4} + \frac{1}{2\pi} \sin^{-1}(\rho). \tag{5.17}$$

Here ρ_k is the lag k autocorrelation. It is then shown in [81] that

$$EX_t X_{t-1} = \frac{1}{4} + \frac{1}{2\pi} \sin^{-1}(\rho_1).$$

Hence $ED_1 = (N-1)(\frac{1}{2} - \frac{1}{\pi} \sin^{-1}(\rho_1))$, where N is the number of data points. Further

$$\rho_1 = \cos(\frac{\pi ED_1}{N-1}). \tag{5.18}$$

This result is interesting for our analysis as it indicates that ρ_1 tends to zero as the number of zero crossing increase.

It is also shown in [82] that by the Wiener-Khintchine theorem it follows

$$\cos(\frac{\pi ED_1}{N-1}) = \frac{\int_{-\pi}^{\pi} \cos(\omega) dF(\omega)}{\int_{-\pi}^{\pi} dF(\omega)} = \frac{\gamma_k}{\gamma_0}. \tag{5.19}$$

where D_k is the number of zero crossings of the $k-1$th differenced time series. This formula shows the relation between the spectrum of the signal and the rate of zero-crossings and weighted average of the spectral mass [82]. It is important to note that the zero-crossing rate tends to admit values in the neighborhood of a dominant frequency. Note that the formula is derived for the Gaussian case. The continuous analogue of 5.19 is known as Rice's formula, [123]. Rice's formula gives the expected zero-crossing rate of a stationary Gaussian process. For further study of the the non-Gaussian case please see [17, 18]. Another analogue for a non-Gaussian case is given in [23].

For completeness we state some additional results that link ρ_k and D_k. The higher order crossings uniquely determine the spectral distribution formula for F up to a constant, [81]. The higher order crossing spectral representation is given by, [81]

$$\cos(\frac{\pi ED_{k+1}}{N-1}) = \frac{\int_{-\pi}^{\pi} \cos(\omega)(\frac{\sin\omega}{2})^{2k} dF(\omega)}{\int_{-\pi}^{\pi} (\frac{\sin\omega}{2})^{2k} dF(\omega)}, \tag{5.20}$$

from which one obtains the following formula, [84],

$$\cos(\frac{\pi ED_{k+1}}{N-1}) = \frac{-\binom{2k}{k-1} + \rho_1[\binom{2k}{k} + \binom{2k}{k-2}] - ... + (-1)^k \rho_{k+1}}{\binom{2k}{k} - 2\rho_1\binom{2k}{k-1} + ... + (-1)^k 2\rho_k}. \tag{5.21}$$

Equation (5.21) provides a recursion for obtaining $\rho_1, \rho_2, \rho_3, ...$ from $ED_1, ED_2, ...$ For example for $k = 0$ this reduces to Equation (5.18), so that ρ_1 is determined form ED_1. For $k = 1$, 5.21 gives

$$\rho_2 = 1 - 2(1 - \cos(\frac{\pi ED_1}{N-1}))(1 + \cos(\frac{\pi ED_2}{N-1})). \tag{5.22}$$

For $k = 2$, ρ_3 is determined similarly from ED_1, ED_2, ED_3. In general, ρ_k is determined from $ED_1, ..., ED_k$. Now recall that ρ_k is the kth Fourier coefficient of F.

The series of results above provides further insight into the behavior of the algorithm proposed here even if the basic theoretical result below is somewhat weaker than Theorem 5.3.5.

In what follows it will be useful to define a signal to be ρ_1-dominant if $\rho_1 > \rho_k$ for all k.

Theorem 5.3.6. *Execution of Algorithm 2 will result in the Null Hypothesis for the ACF test being accepted if the signal being fit is ρ_1-dominant.*

Proof. We have proved that the number of zero crossings of the residuals increase monotonically in Theorem 5.3.1. From Equation (5.18), $\rho_1 = \cos(\frac{\pi E(\#ZC(e_i))}{N-1})$. Thus, for this case, the autocorrelation at the maximum contributing lag decreases as a new RBF is added to the model. In other words the autocorrelation function will decrease with the addition of new RBFs. □

In actual practice, we find that ρ_1 dominates the ACF in the sense that almost always $\rho_1 > \rho_k$ for $k > 1$. Indeed, we have experimented with ignoring these components of the ACF in the modeling algorithm and have found the computational expense can be

significantly reduced. Of course, it may be necessary to include these terms if convergence criteria are not met. This subject needs more thorough investigation.

In the case where another lag becomes dominant, a higher order analysis is required. This indicates that there is a dominant periodicity in the signal and that is best removed by differencing rather than the nonlinear trend removal. However the idea is to show that the increment in the number of zero crossings also leads to a decline in the autocorrelation at this specific lag. This could be shown from Equation (5.21) and the fact that the algorithm goes through the ACC function for that specific lag and the case becomes similar to the one for lag one.

To conclude this analysis in the frquency domain, lets consider an IID hypothesis test based on higher order zero crossings. According to [84] there is no simple closed form expressions for variances D_j; various approximations are available in [83, 85]. A very useful and simple approximation can be obtained if $\rho_k \to 0$, as $k \to \infty$ sufficiently fast. Under the hypothesis of white noise D_k has an asymptotic normal distribution and also we know ED_k exactly. Thus, one could form probability limits for D_k. Approximate 95% probability limits for D_k's are given in [85] as

$$(N-1)[\frac{1}{2}+\frac{1}{\pi}\sin^{-1}(\frac{k-1}{k})] \pm 1.96\{(N-1)[\frac{1}{4}-(\frac{1}{\pi}\sin^{-1}(\frac{k-1}{k}))^2]\}^{\frac{1}{2}}. \qquad (5.23)$$

The hypothesis of white noise is rejected if at least one D_j, $j = 1, ..., K$, falls outside the bounds. When all the $D_j, j = 1, ..., K$, fall inside the limits, the initial rate of increase in the D_j resembles that of white noise and the hypothesis of white noise is accepted. For further reading about this test see [83, 85].

5.3.5 First Order Tests

Although the ACF test described in the previous section is arguably more powerful, it is also of interest to consider a convergence analysis of the proposed algorithm in

terms of first order tests on the residuals. To make the chapter self contained we include definitions of these tests in what follows. First we review additional definitions used to determine whether a time series is WN of IID.

The Turning Point Test

The Turning Point Test (see, e.g., [86]): If $y_1, y_2, y_3, ...$ is a sequence of observations, we say that there is a turning point at time i, $1 < i < n$, if $y_{i-1} < y_i$ and $y_i > y_{i+1}$ or if $y_{i-1} > y_i$ and $y_i < y_{i+1}$. If T is the number of turning points of an IID sequence of length n, then since the probability of a turning point at time i is $2/3$, the expected value of T is $\mu_T = E(T) = \frac{2(n-2)}{3}$. It can also be shown, [86], that for an IID sequence the variance of T, $\sigma_T^2 = Var(T) = \frac{16n-29}{90}$. A large value of $T - \mu_T$ indicates that the series is fluctuating more rapidly than expected for an IID sequence. On the other hand a value of $T - \mu_t$ much smaller than zero indicates a positive correlation between neighboring observations. For an IID sequence with n large, it can be shown that T is approximately $N(\mu_T, \sigma_T^2)$. This means we can carry out a test of IID hypothesis, rejecting it at level α if $\frac{|T-\mu_T|}{\sigma_T} > \Phi_{1-\frac{\alpha}{2}}$, where $\Phi_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. A commonly used value of α is 0.05 for which the corresponding value of $\Phi_{1-\frac{\alpha}{2}}$ is 1.96.

Theorem 5.3.7. *Execution of Algorithm 2 will result in the Null Hypothesis for the Turning Points test being accepted.*

Proof. Using Zero Crossing Lemma, by adding new RBFs the number of zero crossings increases. Thus the number of turning points increase. Eventually, the number of turning points enters to the acceptable bounds derived from turning point test, i.e., $\mu_T = E(T) = \frac{2(n-2)}{3}$, $\sigma_T^2 = \frac{16n-29}{90}$, $\frac{|T-\mu_T|}{\sigma_T} < \Phi_{1-\frac{\alpha}{2}}$=1.96. At this point we accept the hypothesis test and algorithm is converged. □

Difference-Sign Test

Difference-Sign Test, [86]: For this test we count the number S of values of i such that $y_i > y_{i-1}$, $i = 2, ..., n$ or equivalently the number of times the differenced series $y_i - y_{i-1}$ is positive. For an IID sequence it is clear that $\mu_s = ES = \frac{n-1}{2}$. It can also be shown, under the same assumption, that $\sigma_S^2 = Var(S) = \frac{n+1}{12}$. For large n, S is approximately $N(\mu_S, \sigma_S^2)$. A large positive (or negative) value of $S - \mu_S$ indicates the presence of an increasing (or decreasing) trend in the data. We therefore reject the assumption of no trend in the data if $\frac{|S-\mu_S|}{\sigma_S^2} > \Phi_{1-\frac{\alpha}{2}}$. The difference-sign test must be used with caution. A Set of observations exhibiting a strong cyclic component will pass the difference-sign test for randomness since roughly half of the observations will be points of increase.

Theorem 5.3.8. *Execution of Algorithm 2 will result in the Null Hypothesis for the difference sign test being accepted.*

Proof. Using Zero Crossing Lemma, by adding new RBFs the number of zero crossings increases. This causes the number of different signs form a damped oscillating behavior and will eventually enter the confidence bounds of the difference sign test, i.e., $\mu_S = E(S) = \frac{n-1}{2}$, $\sigma_S^2 = Var(S) = \frac{n+1}{12}$, $\frac{|s-\mu_s|}{\sigma_S} < \Phi_{1-\frac{\alpha}{2}}$. At this point the algorithm terminates. □

5.3.6 Subspace View of Convergence

Here we pursue a more geometric discussion of convergence of the algorithm by viewing it as a process for performing a rank-one update of the interpolation matrix. This approach can be viewed as Krylov subspace expansion of a vector space. We consider the correlation function as the cost function and attempt to find the optimal directions to minimize this cost.

At each step in the iteration a new dimension is added to the vector space which is independent of the previous dimensions as a result of the ACF test. Effectively it is

the translation and the dilation of the basis function that produce linearly independent directions. According to the structure of the algorithm the scale of the RBFs change at each level. This procedure provides a one dimensional expanding Krylov subspace at each step. We assume that the input signal is composed of signal plus noise. When the least squares solution of the weights becomes zero we detect the orthogonality of the signal space and the noise component. Note that at the stopping point the noise is orthogonal to the native space of the RBFs that are used to fit the data. Thus we have a decomposition of the signal and noise.

The current algorithm for RBFs could also be used for optimization based on other criterion. e.g., if one would likes to use the information matrix, this algorithm can be used to identify the regions in the data that are contributing the most to this matrix and try to capture the signal model in a way that is in the direction of the maximum reduction in the given sense.

5.3.7 Numerical Results

We have provided several theoretical results concerning the converge of the algorithm proposed in this chapter. Here we present numerical results that both corroborate the theory as well as provide some indication of the rates of convergence on specific data sets. We emphasize that all the results were produced without changing any parameters in the code, i.e., the algorithm was treated as a black box. We apply the algorithm to the *Pringle* and the Mackey-Glass data sets using the Cosine and the Arctan-Hanning RBFs, respectively. In these experiments the data is partitioned into training, validation and testing data sets. Note that code used here is also used, with the multivariate extensions, in Section 6.4.

We note that in these numerical experiments the support of the compact RBFs was not restricted to be contained in the local data region. Also, for these results the radius of the ball is selected to be the maximum of the radius of the left or right zero-crossing points

to the initial center location. Despite these deviations from our theoretical framework the convergence of the algorithm is not adversely affected.

Here we present several applications to demonstrate the performance of the algorithm in higher dimensional domains. Note that throughout all the following examples the same code was employed, in particular, there were no parameters that were adjusted or tuned to the data set. We present the results starting from dimension two.

Pringle data set

This data set was introduced in Chapter 2. As before, the task is to construct a mapping from an (x, y) value in the plane to its corresponding z value on the boundary of a *Pringle*. Thus, we are fitting the graph of a function from $\mathbb{R}^2$ to $\mathbb{R}$. Such graph fitting problems are at the center of the Whitney's manifold embedding theorem where $2m + 1$ dimensional domains suffice (in general) to write m dimensional manifolds as graphs; see [30, 29] for a discussion.

Recall Figure 2.1 where we showed a numerically integrated trajectory of an attracting cycle. Again, in this example, we are only concerned with fitting data on the limit cycle and ignore transients. Figure 5.1 shows the training set consisting of 101 points (almost two cycles) and 100 data points for validation. The testing data set consisting of 400 points, or almost 8 cycles. The fact that the solution is periodic will clearly illustrate the need for spatial as well as temporal windowing of the data. The system is capable of learning a specific part of the trajectory with a small amount of data and generalizes well to the data that resides in the same region.

Figure 5.2 shows the location and shape of the four RBFs that are generated by the algorithm to model the data before the IID stopping criteria is satisfied. The training data and the RBFs are displayed together to illustrate how the algorithm has fit the RBFs to the data.

Figure 5.3 (a) shows the maximum value of the ACC function for each step in the training process. We observe that it is essentially zero after four RBFs have been added

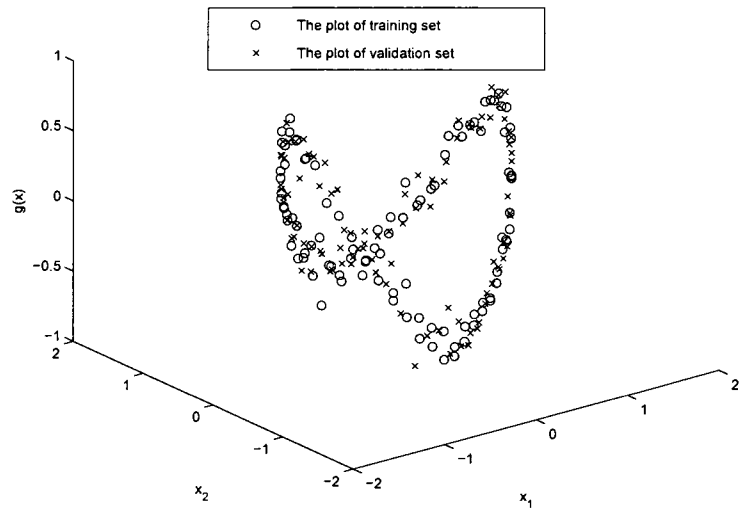


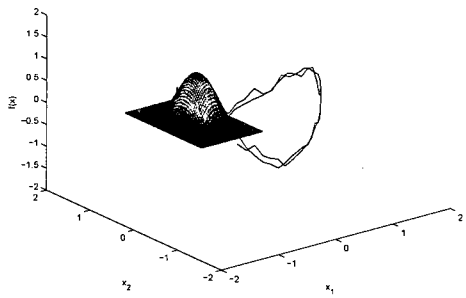
Figure 5.1: Plots of the training and validation data sets used in this numerical experiment. The solution to the dynamical system is corrupted with Gaussian noise with STD of 0.1. There are 54 data points in one cycle.

to the model. Figure 5.3 (b) shows the performance of the model in the RMSE sense as the number of assigned RBFs increase while Figure 5.3 (c) shows the confidence level at each stage of training on the training and validation data sets. Note that confidence levels are over 95% as of the fourth RBF for both the training and validation sets.

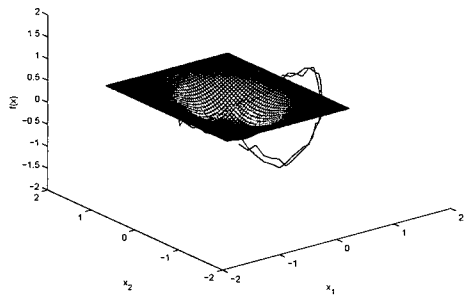
Figure 5.4 (a) shows the $\widehat{\rho}(h^*)$ for each step in the training process. Note that $h^* = 1$ for all the process. Figure 5.4 (b) shows the $\widehat{\gamma}(0)$ as the number of assigned RBFs increase. Figure 5.4 (c) shows the $\widehat{\gamma}^{k+1}(0)/\widehat{\gamma}^{k}(0)$ at each stage of training where k denotes the number of RBFs in the model.

Figure 5.5 (a) shows the number of turning points for each step in the training process. Figure 5.5 (b) shows the number of different signs as the number of assigned RBFs increase while Figure 5.5 (c) shows the number of zero crossings at each stage of training.

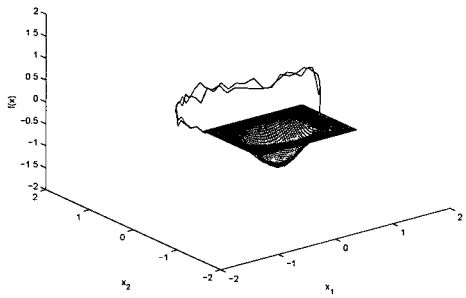
Figure 5.6 (a) shows the plot of the histogram of the final residuals. Figure 5.6 (b) shows the confidence of the model based on χ^2 as the number of assigned RBFs increase while Figure 5.6 (c) shows the mean of the residuals at each stage of training. It is



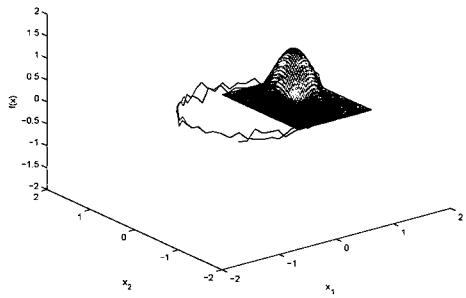
(a) The first RBF in relation with the training data set.



(b) The second RBF in relation with the training data set.

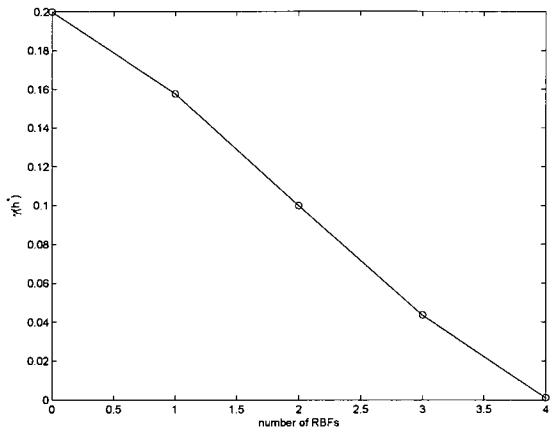


(c) The third RBF in relation with the training data set.

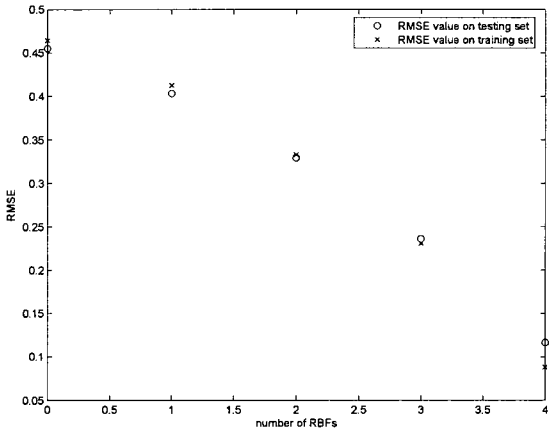


(d) The fourth RBF in relation with the training data set.

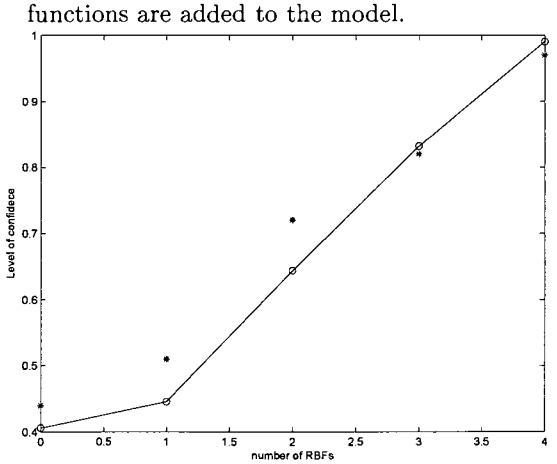
Figure 5.2: The primary four radial basis functions allocated by the algorithm. The residuals of the four mode model pass the IID test. The Hanning, or shifted cosine RBF was used in this fit.



(a) The plot of $\widehat{\gamma}(h^*)$ as new basis functions are added to the model.

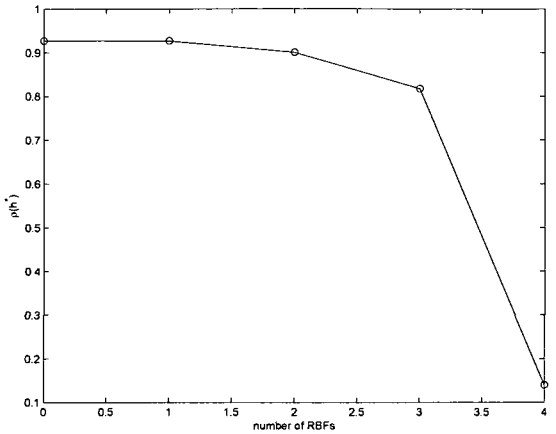


(b) The RMSE plot of the model as new basis functions are added to the model.

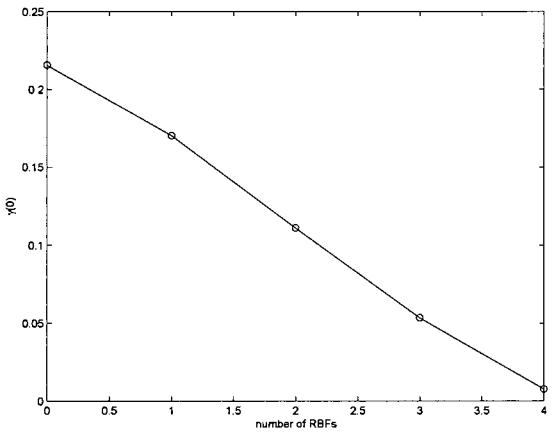


(c) The confidence level of the fitted model on the training (circles) and the validation (stars) data sets as the new basis functions are added to the model.

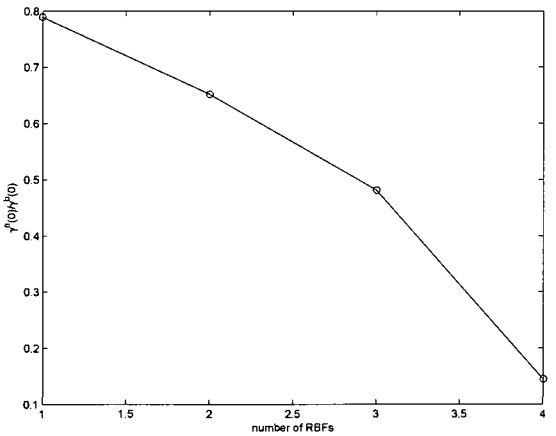
Figure 5.3: The performance of the RBF fit on the Pringle data set. NOTE: The confidence level at the end of the process is 99% on the training data set and 97% on the validation data set.



(a) The plot of $\widehat{\rho}(h^*)$ as new basis functions are added to the model.

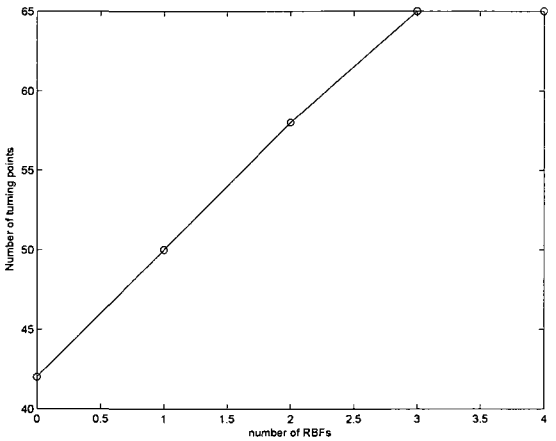


(b) The plot of $\widehat{\gamma}(0)$ as new basis functions are added to the model.

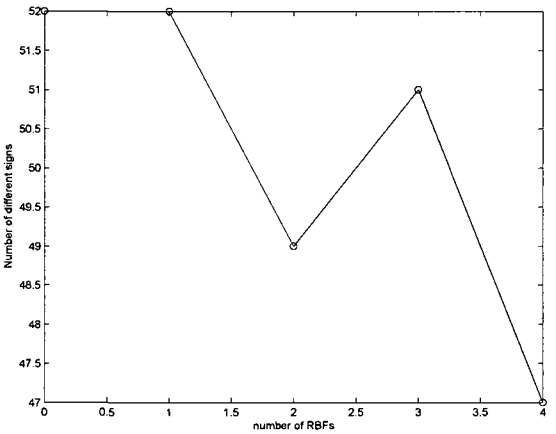


(c) The plot of $\widehat{\gamma}^{k+1}(0)/\widehat{\gamma}^{k}(0)$ as new basis functions are added to the model where k denotes the number of RBFs in the model.

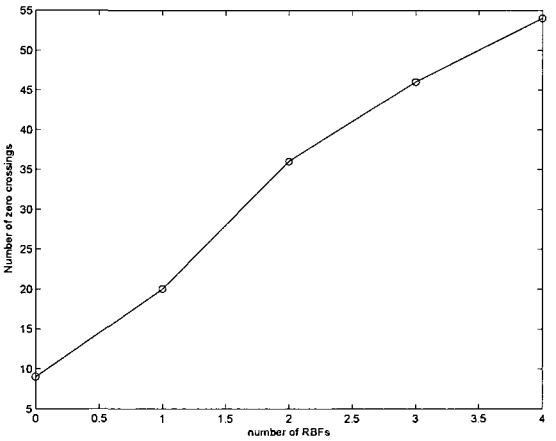
Figure 5.4: The behavior of $\widehat{\rho}(h^*)$, $\widehat{\gamma}(0)$ and $\widehat{\gamma}^{k+1}(0)/\widehat{\gamma}^{k}(0)$ plotted as functions of the number of RBFs in the model for the Pringle data set.



(a) The number of turning points as new basis functions are added to the model.

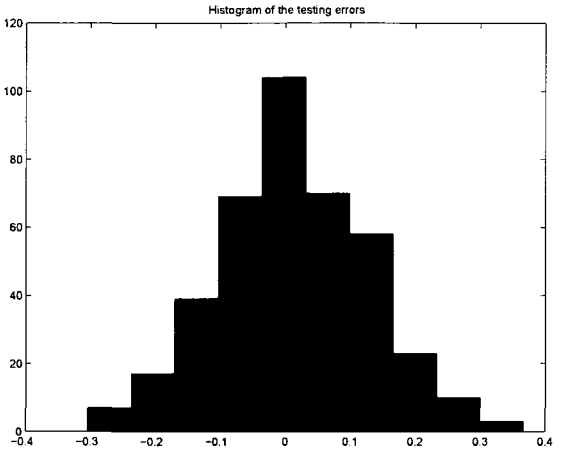


(b) The number of different signs as new basis functions are added to the model.

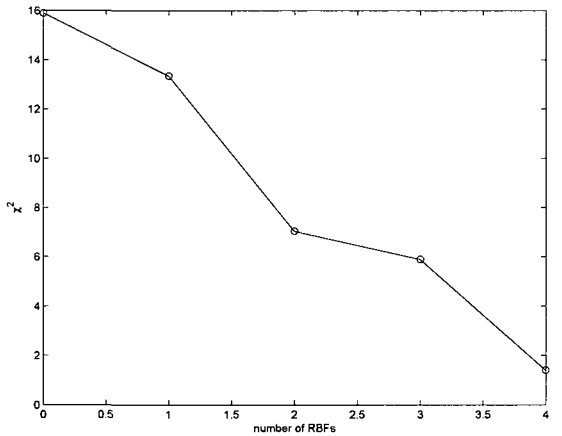


(c) The number of zero crossings as new basis functions are added to the model.

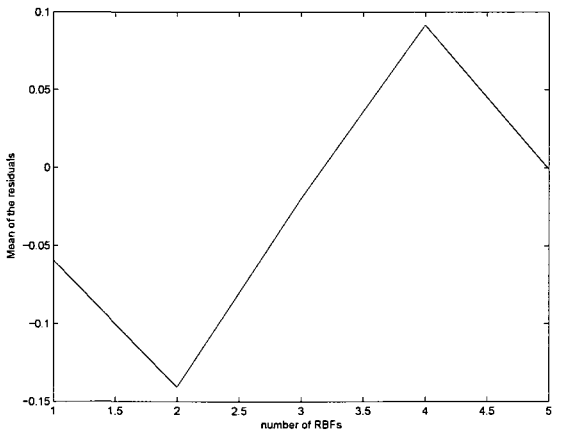
Figure 5.5: Diagnostics related to the hypothesis tests indicate that the algorithm is converging and that the model clearly requires four RBFs to fit the Pringle data set.



(a) The plot of the histogram of the final residuals.



(b) The confidence of the model based on χ^2 as new basis functions are added to the model.



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

Figure 5.6: Properties of the residuals of the Pringle model using Hanning RBFs.

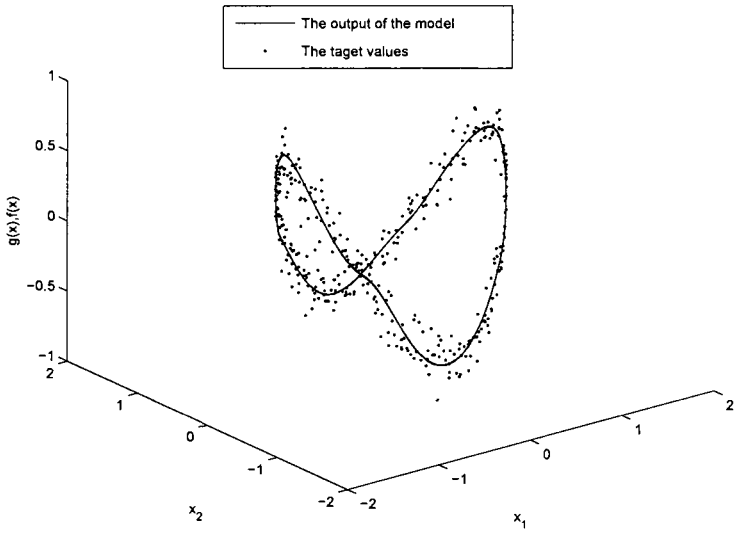


Figure 5.7: The testing data set and the output of the four mode model.

interesting that the mean of the residuals becomes essentially zero as the last function is added. This permits us to interpret the algorithm as having satisfied the full ACF test, even though the mean of the residuals was ignored during training.

A plot of the output of the model and target values of the testing set are shown in Figure 5.7.

The condition number of the four mode model is an extremely low value, namely 1.65.

Mackey-Glass Data Set

This example uses the same data, a numerical simulation of the Mackey-Glass time-delay equation, [102], as employed in the numerical experiments in Chapter 2. Here, as before, we illustrate the mapping from a time-delay embedding of the univariate time-series to a future value. In these experiments noise was added to the data with a standard deviation of 0.05.

Again, for purposes of comparison with [153], the series is predicted with $\upsilon = 50$ samples ahead using four past samples: s_n, s_{n-6}, s_{n-12} and s_{n-18}. Hence, the nth input

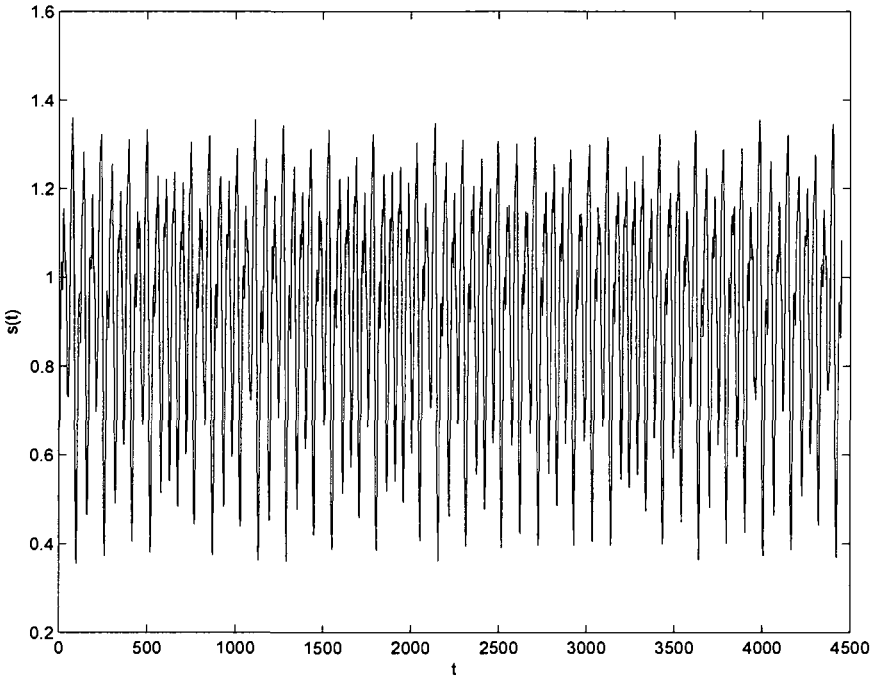


Figure 5.8: The data set used for the current study including noise with a STD of 0.05.

output data for the network to learn are

$$x_{n+\upsilon} = [s_n, s_{n-6}, s_{n-12}, s_{n-18}]^T$$

with $y_n = s_n$, whereas the υ step-ahead predicted value at time n is given by $z_{n+\upsilon} = f(x_{n+\upsilon})$, where $f(x_{n+\upsilon})$ is the network output at time n. The υ step-ahead prediction error is $\epsilon = s_{n+\upsilon} - z_{n+\upsilon}$. As such, this time series provides a good example for illustrating the construction of a nontrivial mapping from $\mathbb{R}^4$ to $\mathbb{R}$, [98, 44, 87].

Our goal here is to illustrate the convergence properties of the algorithm on Mackey-Glass data set. This data set is particularly interesting as it requires a total of 24 RBFs before the termination criterion is achieved. Figure 5.8 shows the data set with added noise that is used in this study. To give a better view of small regions of the data points 2000 to 2400 are shown in Figure 5.9.

From Figure 5.10 we see it is sufficient to use only 24 centers to get the 95% confidence fit for the Mackey-Glass data set with a resulting RMSE of 0.0186 (See Figure

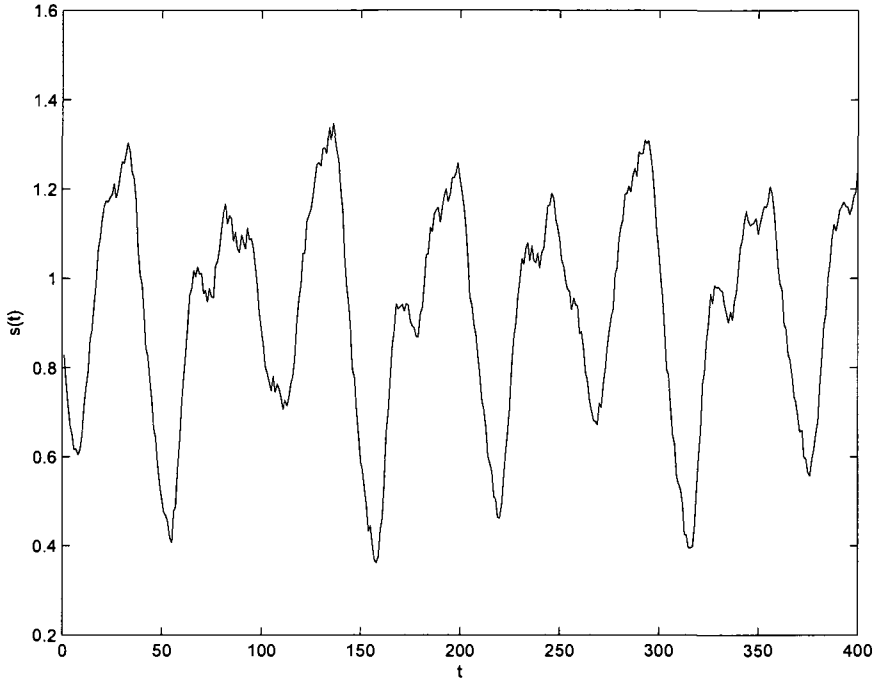


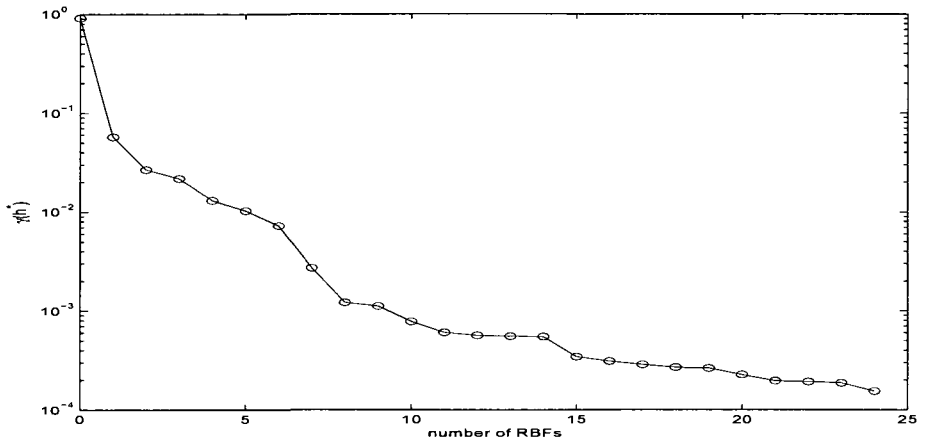
Figure 5.9: Points 2000-2400 of the data set used for the current study.

5.11 to assess the fit visually.) The output of the 24 mode model for the testing data set generated by Algorithm 2 appears to fit the target values very well. We remark that the experiment in Chapter 2 using Algorithm 1 resulted in a model consisting of 76 modes with an RMSE of 0.0168 when the %95 of confidence was satisfied.

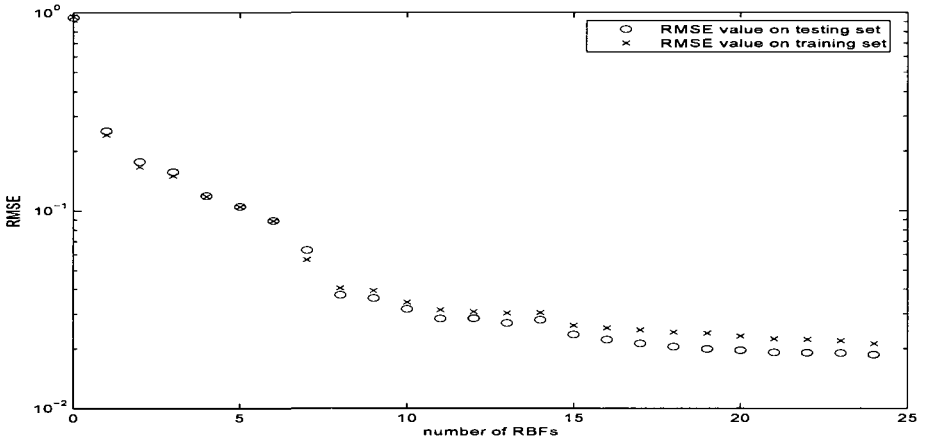
Figure 5.12 (a) shows the $\widehat{\rho}(h^*)$ for each step in the training process. Note that $h^* = 1$ for all the process. Figure 5.12 (b) shows the $\widehat{\gamma}(0)$ as the number of assigned RBFs increase while Figure 5.12 (c) shows the $\widehat{\gamma}^{k+1}(0)/\widehat{\gamma}^{k}(0)$ at each stage of training.

Figure 5.13 (a) shows the number of turning points for each step in the training process. Figure 5.13 (b) shows the number of different signs as the number of assigned RBFs increase while Figure 5.13 (c) shows the number of zero crossings at each stage of training. Note that there are 3000 points in the training set and 1200 zero crossings by the time the model has converged.

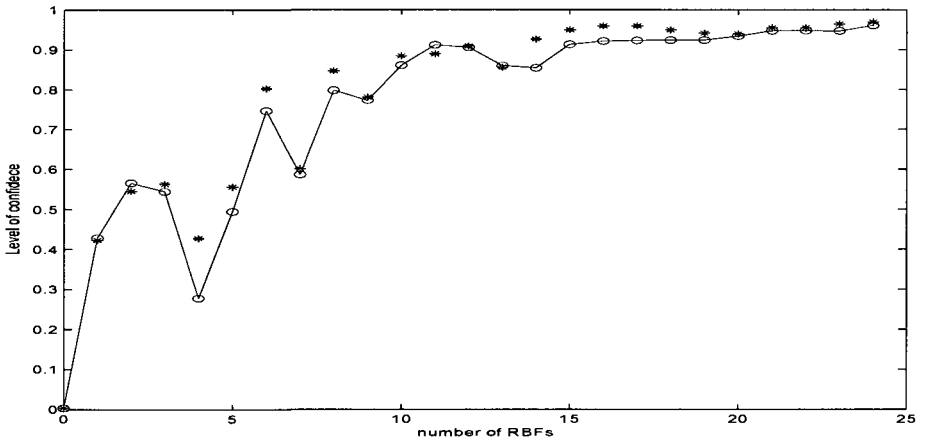
Figure 5.14 (a) shows the plot of the histogram of the final residuals. Figure 5.14 (b) shows the confidence of the model based on χ^2 as the number of assigned RBFs increase



(a) The log plot of $\widehat{\gamma}(h^*)$ as new basis functions are added to the model.



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



(c) The confidence level of the fitted model as the new basis functions are added to the model.

Figure 5.10: The performance of the RBF fit on the Mackey-Glass data set using Algorithm 2 and Arctan-Hanning skew radial basis functions. Note that over 95% confidence is achieved with 24 modes.

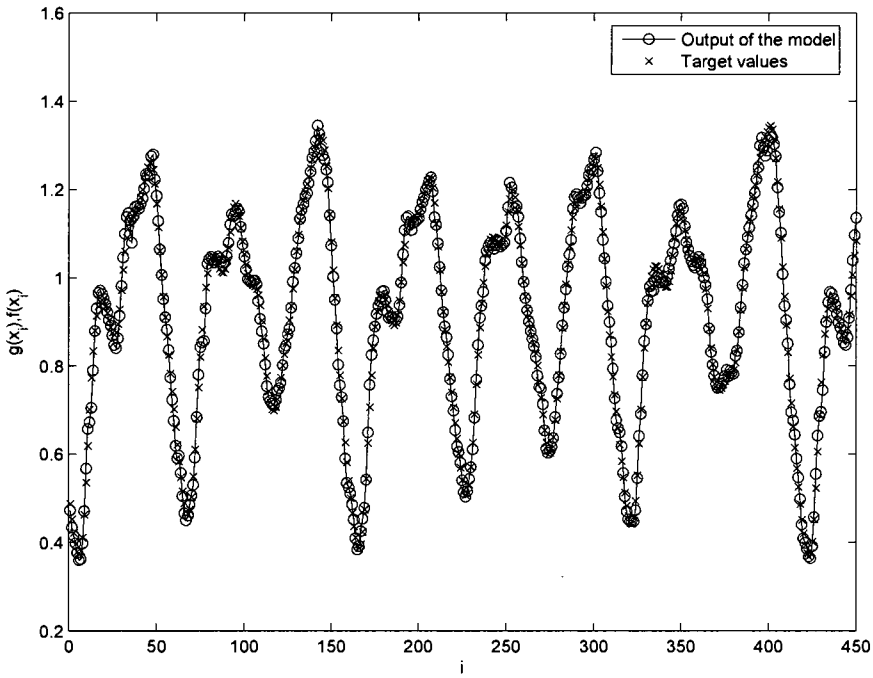
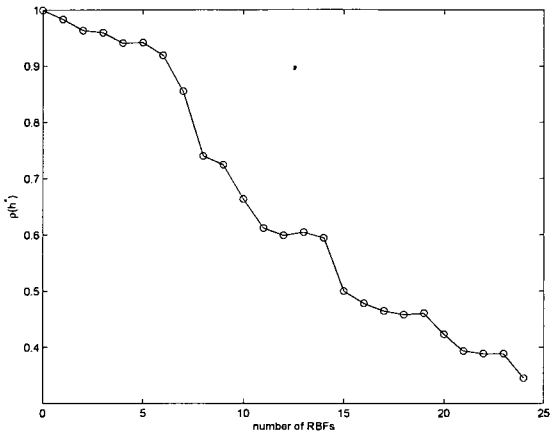
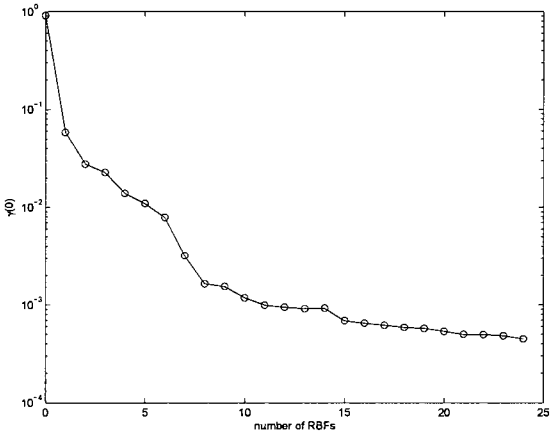


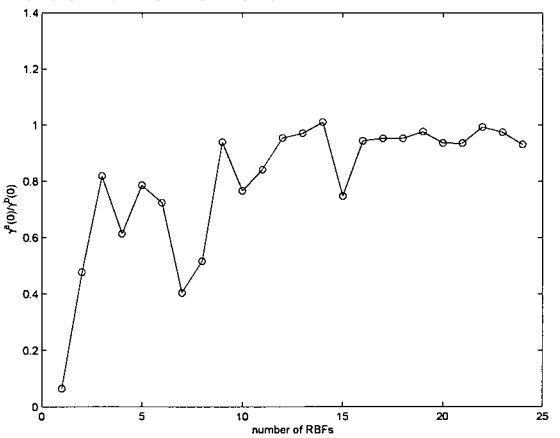
Figure 5.11: The output of the 24 mode model for the testing set compared to the target values. For this model an RMSE value of 0.0186 was obtained and the 96% of confidence stopping criteria was satisfied both of the validation and the training data sets. The model has the condition number of 1325.3



(a) The plot of $\widehat{\rho}(h^*)$ as new basis functions are added to the model.

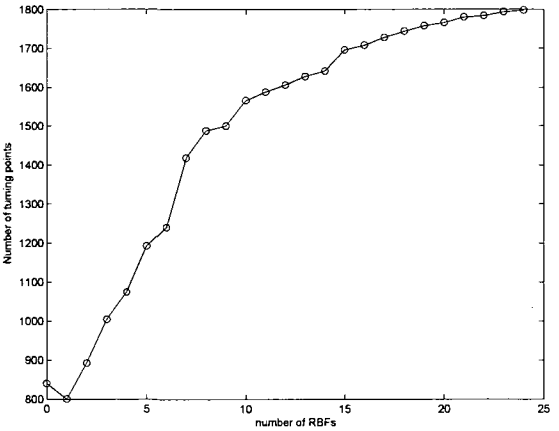


(b) The plot of $\widehat{\gamma}(0)$ as new basis functions are added to the model.

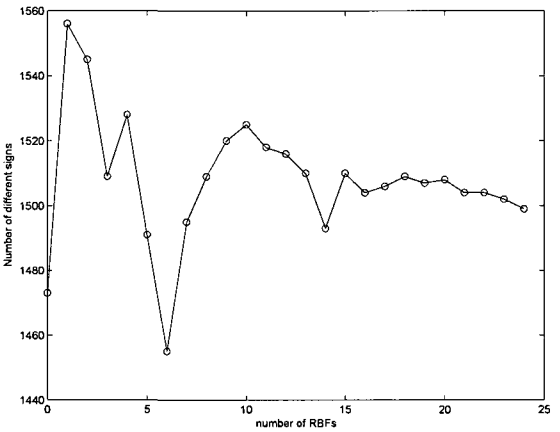


(c) The plot of $\widehat{\gamma}^{k+1}(0)/\widehat{\gamma}^{k}(0)$ as new basis functions are added to the model.

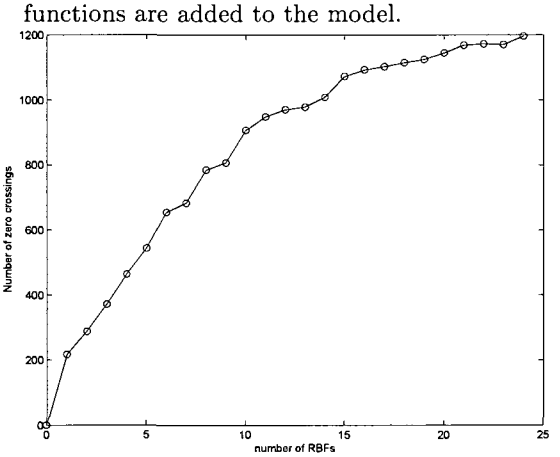
Figure 5.12: The Mackey Glass model statistical performance with Arctan-Hanning RBF.



(a) The number of turning points as new basis functions are added to the model.

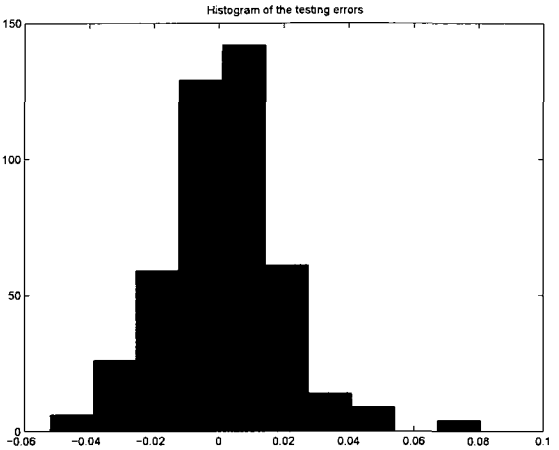


(b) The number of different signs as new basis functions are added to the model.

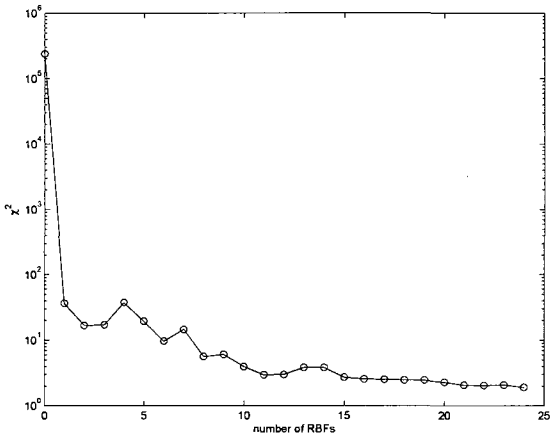


(c) The number of zero crossings as new basis functions are added to the model.

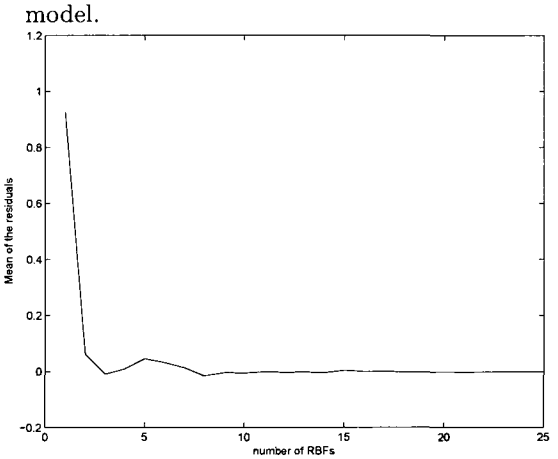
Figure 5.13: First order measures and zero crossings for the Mackey-Glass training data set.



(a) The plot of the histogram of the final residuals.



(b) The confidence of the model based on log χ^2 as new basis functions are added to the model.



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

Figure 5.14: Properties of the residuals of the Mackey-Glass model using Arctan-Hanning sRBFs.

while Figure 5.14 (c) shows the mean of the residuals at each stage of training. Although 24 basis functions are required before the algorithm may successfully terminate we see that the mean is approximately zero after 8 iterations. This provides numerical evidence that our omission of the mean in the definition of the ACF is in fact justified.

5.4 Conclusions

In this Chapter we present Algorithm 2 which has many similarities to Algorithm 1, but includes several fundamental innovations. Of primary importance is that we have redefined the manner in which we locate the places where new RBFs should be added. Like Algorithm 1, Algorithm 2 is capable of modeling a non-linear time series without adjusting any ad hoc parameters. We established a suite of convergence properties that Algorithm 2 possesses. We demonstrated via numerical experiments that the performance of Algorithm 2 is consistent with the theoretical convergence results developed in this chapter. Compactly supported and skew symmetric RBFs are employed in the simulations. We see that skew RBFs result in models that require far fewer modes when compared to the RBFs in the literature and that the condition numbers of these models are significantly improved.

Chapter 6

EXTENSION OF ALGORITHM TO RANGE DIMENSION $M \geq 2$

Abstract In this chapter we present an approach for constructing nonlinear empirical mappings from high-dimensional domains to ranges of dimension one or more. We employ RBFs and the extensions proposed in this dissertation as skew Radial Basis Functions (sRBFs) for constructing a model using data that may be scattered and sparse. The algorithm progresses iteratively adding a new function at each step to refine the model. Again, the placement of the functions is driven by a statistical hypothesis test but now in a manner that accounts for correlation in the range variables. The test is applied on training and validation data and reveals geometric structure when it fails. At each step the added function is fit to data contained in a spatio-temporally defined local region to determine the parameters, in particular, the scale of the local model. The scale of the function is determined via 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. 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.

6.1 Multivariate Extension

We propose an algorithm for constructing nonlinear models from high-dimensional domains to high-dimensional ranges from scattered data. The proposed algorithm is an extension to our previous work, [66, 72] or Chapter 2. Similar to the univariate case, the algorithm progresses iteratively adding a new function at each step to refine the model. Again, 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 spatio-temporally defined local region to determine the parameters and in particular, the scale of the local model. Unlike the available non-linear function fitting methods that leave the extension of the algorithm to higher-dimensional ranges as a trivial extension of the single-dimensional range, we provide more parsimonious models by requiring that the residuals possess no structure in each dimension as well as between pairs of dimensions. This algorithm does not require *ad hoc* parameters. Thus, the number of basis functions required for an accurate fit is determined automatically by the algorithm. These advantages extend the scope of applicability of the univariate algorithm to a much larger class of problems that arise in nature and addressed in different areas of science. A challenge in this work is to convert the multivariate statistical hypothesis test for IID noise into a practical algorithm.

Specifically, we propose an extension of the approach presented in the previous chapter for constructing mappings of the form

$$f : U \in \mathbb{R}^n \rightarrow V \in \mathbb{R}^m, \tag{6.1}$$

from empirical data where now m may be greater than 1. Again, we assume that we have samples $x^{(k)} \in U$ and $y^{(k)} \in V$ are indexed by k and related via f, a nonlinear function, as

$$y^{(k)} = f(x^{(k)}). \tag{6.2}$$

In practice this may be, e.g., a mapping from a manifold U to a vector field V or, alternatively, a mapping of multiple time series $(x_1(t), \ldots, x_n(t))$ to values at some future time $t+T$, i.e., $(x_1(t+T), \ldots, x_n(t+T))$. It is often desirable to employ a time-delay embedding of time series so that the mapping will be from points of the form $(x_1(t), x_1(t-\tau), \ldots, x_1(t-n\tau), \ldots, x_n(t), x_n(t-\tau), \ldots, x_n(t-n\tau))$.

6.2 Testing for Structure in Multivariate Model Residuals

We denote the set of residuals for a model of order K, as

$$R^K = \{e_n\}_{n=1}^{L}, \tag{6.3}$$

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

$$\widehat{\Gamma}(h) = \begin{cases} \frac{1}{L}\sum_{k=1}^{L-h} \alpha(h, e_k), & \text{if } 0 \leq h \leq n-1 \\ \widehat{\Gamma}^{'}(-h), & \text{if } -n+1 \leq h \leq 0. \end{cases} \tag{6.4}$$

Similar to the univariate case we decompose the ACVF into its components as $\alpha(h, e_k) = (e_{k+h} - \bar{e})(e_k - \bar{e})'$. Further more $\alpha(h, e_k^i, e_k^j) = (e_{k+h}^i - \bar{e}^i)(e_k^j - \bar{e}^j)$ 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 will 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

$$\widehat{R}(h) = [\widehat{\rho}_{ij}(h)]_{i,j=1}^{m} = [\widehat{\gamma}_{ij}(h)(\widehat{\gamma}_{ii}(0)\widehat{\gamma}_{jj}(0))^{\frac{-1}{2}}]_{i,j=1}^{m}, \tag{6.5}$$

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 [28].

As mentioned in the univariate case in Chapter 2, 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 IID. We need to extend our definition of white noise to the multivariate case. 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

$$\Gamma(h) = \begin{cases} \Sigma, & \text{if } h = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6.6}$$

We use the notation $\{e_t\} \sim IID(0, \Sigma)$ to indicate that the random vectors $\{e_t\}$ are independently and identically distributed 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, [28]. 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 [28]: Suppose that

$$X_{t1} = \sum_{j=-\infty}^{\infty} \alpha_j Z_{t-j,1}, \{Z_{t1}\} \sim IID(0, \sigma_1^2), \tag{6.7}$$

$$X_{t2} = \sum_{j=-\infty}^{\infty} \beta_j Z_{t-j,2}, \{Z_{t2}\} \sim IID(0, \sigma_2^2), \tag{6.8}$$

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

$$\widehat{\rho}_{12}(h) \text{ is } AN(0, n^{-1} \sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)). \tag{6.9}$$

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,

$$n^{-1} \sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j+k-h). \tag{6.10}$$

As it is reported in [28] 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 $\widehat{\rho}_{12}(h)$, i.e., 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 prewhited series $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are also independent. Under H_0, the above theorem 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 n^{-1}. An appropriate test for independence can therefore be obtained by comparing the values of $|\widehat{\rho}_{12}(h)|$ with $1.96n^{\frac{-1}{2}}$. If we prewhiten only one of the two original series, say $\{X_{t1}\}$, then under H_0 the above theorem 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

n^{-1} and covariance $n^{-1}\rho_{22}(k-h)$. Hence for any fixed h, $\widehat{\rho}_{12}(h)$ also falls (under H_0) between the bounds $\pm 1.96 n^{\frac{-1}{2}}$ 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 IID hypothesis is rejected. This test can equivalently be written in terms of χ^2 distribution. Given

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

it has been shown in [28] 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

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

6.3 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 IID 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 WN or IID test, then the cross-correlations among the time series are considered. If there is structure in the auto-correlations or cross-correlations of the time series then the IID 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 which 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, i.e., 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^*, i.e.,

$$h^* = \arg\max \widehat{\gamma}_{ij}(h), h > 0. \tag{6.11}$$

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

$$i^* = \arg \max_{k=1,\ldots,n-h^*} \alpha(h^*, e_k^i, e_k^j). \tag{6.12}$$

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 will 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. Similar to the univariate case in Chapter 2, 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, i.e., $c_0 = x^*$. The vector of widths σ is very effectively initialized using the diagonal elements of the covariance matrix of the local data,

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

Note here that $W = diag(\sigma_0)$. The initial value for the multivariate weight, α_0, is calculated via least squares using the initial values for center location and widths. Then the parameters associated with the new basis function are optimized by solving the nonlinear optimization procedure using BFGS. Note that all the multivariate range values associated to $\mathcal{X}_{local}$ contribute to the optimization procedure.

Similar to the univariate case we could use one of the statistical tests, RMSE or normalized prediction error or another measure of structure as stopping criteria. Pseudocode of this algorithm is provided in Algorithm 3.

6.4 Numerical Results

In this section we show that multivariate algorithm constructs a single model that is more parsimonious than multiple univariate models.

Algorithm 3 A multi-variate RBF 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 $i, j = 1, ..., m$
  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 **then**
    identify the pair of time series $d1, d2$ that has the maximum value at their cross-correlation function, $i = d1$ and $j = d2$
  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 $i^* = \arg\max_{k=1,...,n-h^*} \alpha(h^*, e_k^i, e_k^j)$
  compute the CCC function, $\beta_k^{i,j} = \alpha(h^*, e_k^i, e_k^j)$, $k = 1, ..., n - h^*$
  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 $v = [c_0, \sigma_0, \alpha_0]^T$
  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

6.4.1 Multivariate Pringle Data Set

To begin, in this section we use the Pringle data set, to show the usefulness of the multivariate algorithm in a case where there is full correlation between the first and the second time series. This data set was introduced in Chapter 2 in Figure 2.1. Given it is easy to visualize, this data set helps us to gain insight about the multivariate algorithm. Figure 5.1 shows the univariate training set consisting of 101 points (almost two cycles) and 100 data points for validation. In this instance the second output is the multiple of the first out put with a factor of two. This example illustrates the ideal behavior of the multivariate algorithm when the output time series are highly correlated. Two univariate models would in this instance double the model complexity.

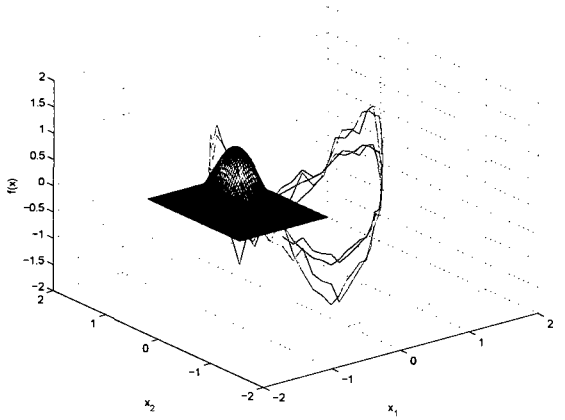
Figure 6.1 shows the first RBF allocated by the multivariate algorithm to both time series. Similar to the univariate case the final model has four RBFs. The training, validation and testing data sets and the output of the multivariate algorithm on multivariate Pringle data set are shown in Figure 6.2. The performance of the RBF fit on the multivariate Pringle data set in the RMSE sense is shown in Figure 6.3. 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 6.3.

6.4.2 Multivariate Mackey-Glass

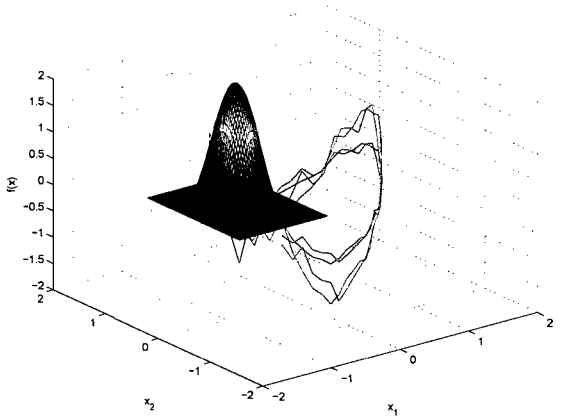
This example uses the same data, as described in Section 5.3.7. However in this section we are interested in the multivariate prediction. For this purpose the series is predicted with 25 and 50 or 50 and 75 samples ahead using four past samples: s_n, s_{n-6}, s_{n-12} and s_{n-18}. Hence, we would like to approximate f is the following relation,

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

where $(\upsilon_1, \upsilon_2) = (25, 50)$ or $(\upsilon_1, \upsilon_2) = (50, 75)$.

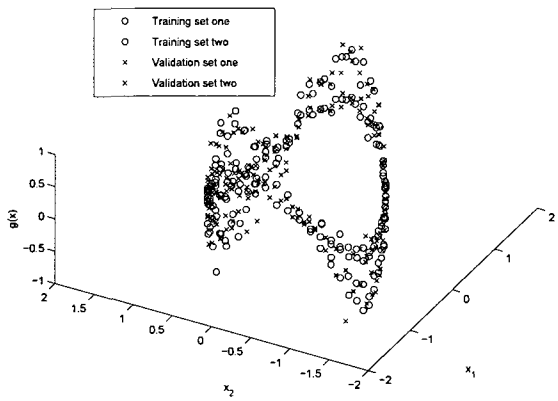


(a) The first RBF in relation with the training data set for the first time series.

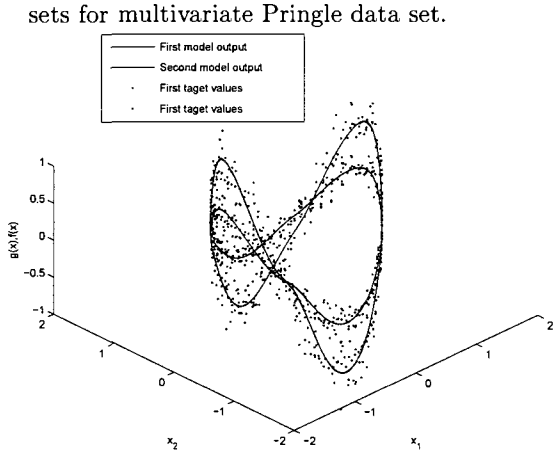


(b) The first RBF in relation with the training data set for the second time series.

Figure 6.1: 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 IID test.

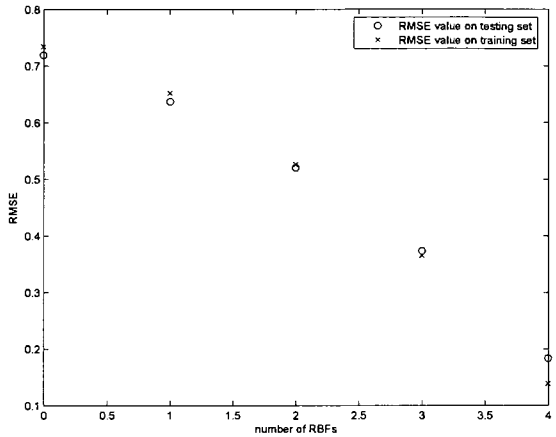


(a) Plots of the training and validation data sets for multivariate Pringle data set.

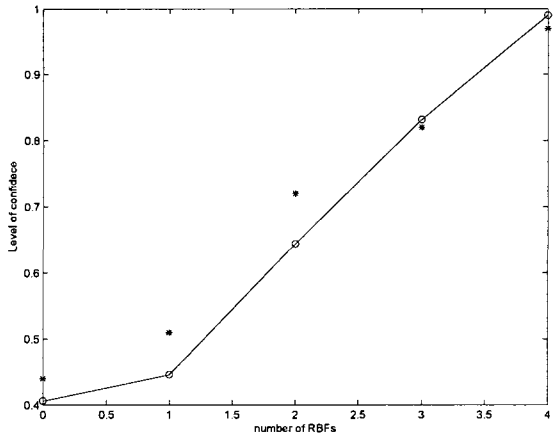


(b) The testing data set and the output of the four mode model.

Figure 6.2: The training, validation and testing data sets and the output of the multivariate algorithm on multivariate Pringle data set.



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



(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.

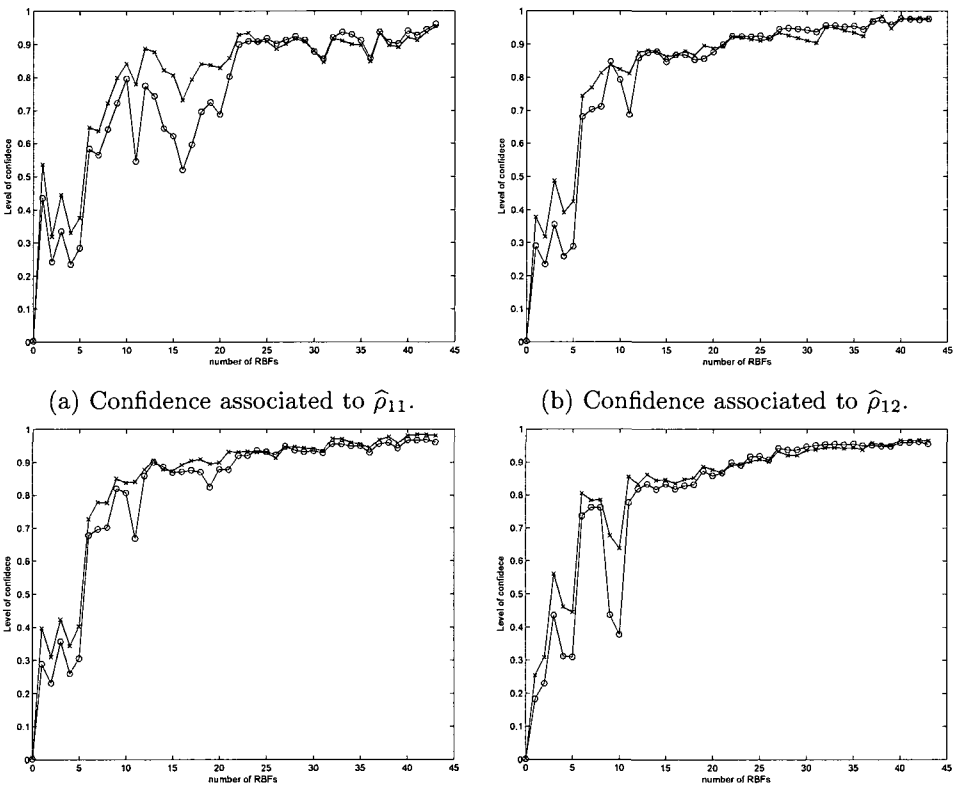
Figure 6.3: The performance of the RBF fit on the multivariate Pringle data set.

Figure 6.4 shows the confidence level of the multivariate model as new Arctan-Hanning sRBFs are added to the model for the case of 25-50 steps ahead prediction. The confidence for all the pairs of the time series are provided.

In this case the RMSE of the final model is 0.0175. The number of required sRBFs to achieve 95% of confidence on the training and validation data sets is 43. The associated univariate 25 ahead prediction problem requires 20 sRBFs to achieve 95% of confidence on the training data set and results in a model with RMSE of 0.0128. The number of sRBFs required to get to 95% of confidence on the validation data set 19. The analogue univariate 50 steps ahead predictor model has the RMSE of 0.0192 with model order 28 to reach 95% of confidence on training data set. The number of RBFs required for the confidence on the validation data set reach 95% of confidence is 26.

Figure 6.5 shows the both outputs of the multivariate model. Figure 6.5 (a) shows the output for 25 steps ahead prediction while Figure 6.5 (b) 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 6.5 (c).

In this work we would like to provide another example for predicting 50 and 75 steps ahead. Figure 6.6 shows the confidence level of the multivariate model as new Arctan-Hanning sRBFs are added to the model. In this case the RMSE of the final model is 0.0264. The number of required sRBFs to achieve 95% of confidence on the training and validation data sets is 49 and 48, respectively. The associated univariate 50 ahead prediction problem requires 25 sRBFs to achieve 95% of confidence on the training data set and results in a model with RMSE of 0.0173. The number of sRBFs required to get 95% of confidence on the validation data set is 17. The analogue univariate 75 ahead predictor model has the RMSE of 0.0265 with model order 54 to reach 95% on training data set. The number of RBFs required for the confidence on the validation data set reach 95% of confidence is 36.



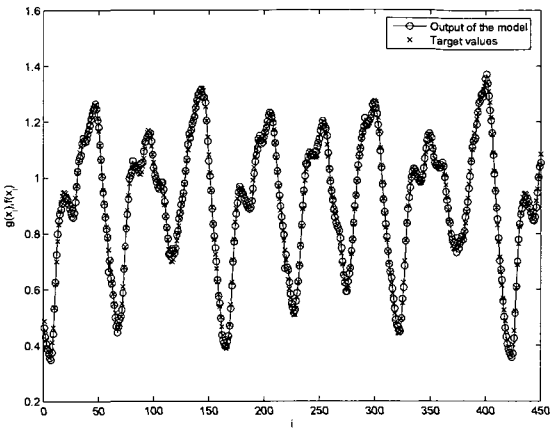
(a) Confidence associated to $\widehat{\rho}_{11}$.

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

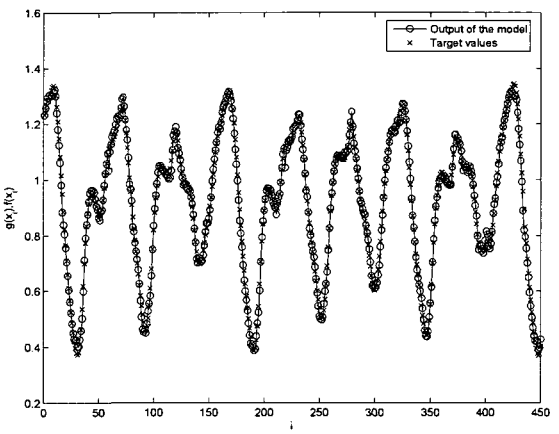
(c) Confidence associated to $\widehat{\rho}_{21}$

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

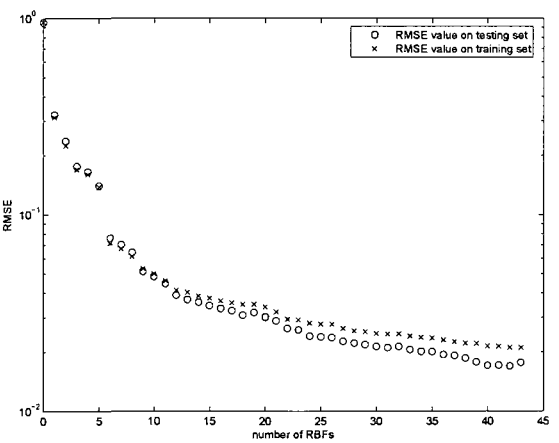
Figure 6.4: The confidence level of the multivariate model as the new Arctan-Hanning sRBFs are added to the model for the case of 25-50 steps ahead prediction of noisy Mackey-Glass data set.



(a) The first output of the model.

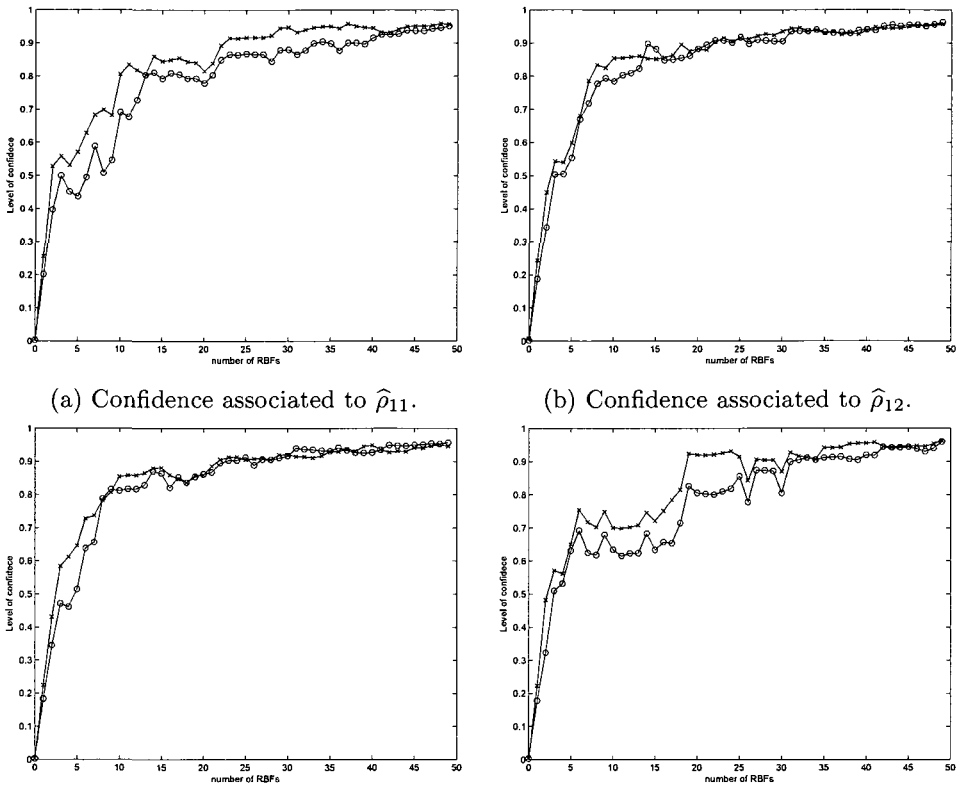


(b) The second output of the model.



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

Figure 6.5: The performance and the output of the multivariate sRBF fit for the case of 25-50 steps ahead prediction of noisy Mackey-Glass data set.



(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 6.6: The confidence level of the multivariate model as the new Arctan-Hanning sRBFs are added to the model for the case of 50-75 steps ahead prediction of noisy Mackey-Glass data set.

It is interesting to note that the multivariate study provides an sRBF model for fitting both the 50 and 75 steps ahead prediction time series with model order less than the number of sRBFs required to fit the 75 steps ahead prediction problem alone.

The log plot of the RMSE of the model as new sRBFs are added to the model is shown in Figure 6.7.

Other performance measures such as χ^2, mean of the residuals, $\widehat{\rho}(h^*)$, $\widehat{\gamma}(h^*)$, $\widehat{\gamma}(0)$ and $\widehat{\gamma}^{k+1}(0)/\widehat{\gamma}^{k}(0)$ for each time series as the new sRBFs are added to the model relieve similar facts to the study cased out in Chapter 5.

We expect that as the number of the time series in the range increase, the multivariate algorithm produces more parsimonious models.

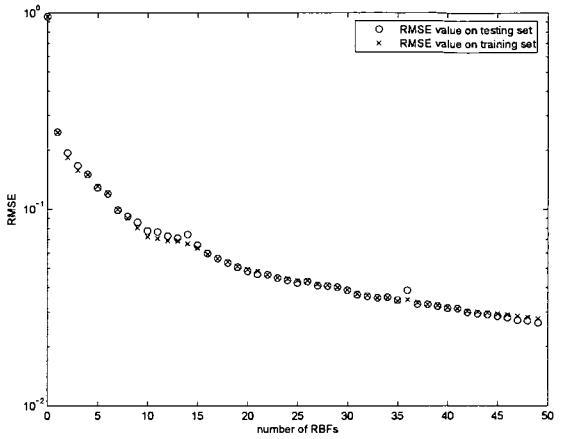


Figure 6.7: The RMSE performance of the Arctan-Hanning fit for the multivariate 50-75 steps a head prediction of noisy Mackey-Glass data set.

6.5 Conclusions

We observed that the extension of the ACF test for IID noise to multivariate ranges produces models of smaller order than using multiple univariate models. This is a consequence of the fact that the correlation of the multivariate time-series in the range is exploited during the model building process.

The opportunity for applications of the proposed RBFs is significant, e.g., 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 graph of functions [29, 30] as well as the low-dimensional modeling of dynamical systems [31].

Chapter 7

CONCLUSIONS, CONTRIBUTIONS AND FUTURE WORK

Summary of Contributions

We have proposed an algorithm for approximating functions from scattered data over high dimensional domains. We present a series of innovations of this algorithm and an approach for extending it to ranges of dimension greater than one. We have provided a detailed theoretical analysis as well as a suite of numerical results. These applications demonstrate that the proposed algorithm provides superior performance when compared to the leading algorithms in the literature.

It is assumed that the available data represents a functional relationship, or signal, with IID additive noise. An hypothesis test is applied to the residuals at each step in the algorithm to determine whether a new basis function should be added, and if so, where it should be added. When it has been determined that this test has been passed using the 95% confidence criterion one may infer there is no geometric structure left in the residuals and thus the model order has been found.

We note that the algorithm performs well on data with a low signal-to-noise ratio. This follows from the iterative behavior of the algorithm. At each step less signal is left in the residuals but the algorithm does not actually terminate until there is statistical evidence that no signal remains. Thus, even very small amounts of signal structure will lead to the addition of a new basis function.

Here we summarize the main contributions of this dissertation.

- Development of an algorithm for generating reduced order models from data requiring no ad-hoc parameters.
- Introduction of a space-time ball for model fitting.
- Detailed convergence analysis of algorithm.
- Introduction of new compactly supported basis functions and condition number analysis.
- Introduction of skew Radial Basis Functions for fitting edges and asymmetric data in general.
- Extension of algorithm to multi-variate range.
- Application to time-series prediction resulting in best models in the literature.

Development of an algorithm for generating reduced order models from data requiring no ad-hoc parameters. To illustrate the absence of *ad hoc* parameters, all the data sets presented in this dissertation were fit by exactly the same code through out each chapter. No adjustments were made based on the data sets being fit. Hence, we claim the proposed algorithm is effectively like a black box for nonlinear function approximation. This feature will permit the advancement of a variety of problems in nonlinear signal processing, optimal control, computer vision, pattern recognition and prediction or other algorithms, e.g., the representation of data on manifolds as graphs of functions [29, 30], pattern classification [67, 90], as well as the low-dimensional modeling of dynamical systems [31].

Introduction of a space-time ball for model fitting. The algorithm employs a spatio-temporal window, i.e., space-time balls, for determining the local data to be used in updating the model. Further details about the choice of local ball are also reported. Modifications to the autocorrelations test is made and shown its equivalence to the

standard ACF at the stopping stages. The examples suggest these novelties are critical for approximating data over high-dimensional domains and in particular for data generated by dynamical systems.

Detailed convergence analysis of algorithm. We also establish the convergence properties of the proposed algorithms. This revealed new features about the algorithm and provided more insight in to its behavior and the enhancement of the algorithm. The proofs brought together various developments made over different chapters. We demonstrate interesting facts about the algorithm such as the energy of the final residuals approaches the energy of the noise that could have been initially mixed with a signal. The convergence is shown for variety of statistical tests including autocorrelation, different sign and turning point tests. Further geometric elaboration on the properties of the algorithm is given in the Krylov subspace expansion sense. Optimization of the RBFs were also studied thoroughly.

Introduction of new compactly supported basis functions and condition number analysis. We note that the condition number of the interpolation matrix depends directly on the choice of RBFs and suggests an explanation for the good conditioning properties for RBFs that posses the suitable derivatives to match with data. To construct more accurate models with better conditioning we have proposed several new candidate compactly supported RBFs and have illustrated some of their positive performance properties on the benchmark Mackey-Glass problem. Both the number of required modes and the conditioning of the final model are substantially improved over previous work.

Introduction of skew Radial Basis Functions for fitting edges and asymmetric data in general. We have proposed a class of skew-radial basis functions (sRBFs) which is formed by modulating an RBF with a symmetry breaking term. These include compactly and non-compactly supported sRBFs. The main attraction of these types of RBFs is their additional flexibly to fit asymmetric data. In a motivating example we have shown that

an sRBF generated by four parameters, needs a 26 parameter RBF model to be fit with symmetric RBFs. We notice a major area of employment for this type of basis functions in fitting data with jumps in it, such as we might find in physical flows with singularities, or in images with sharp edges arising from, e.g., shadows. We also modeled the maximum wind speed of a hurricane and found out that sRBFs produced a fit that had a significantly lower error than the RBF approach and that increasing the complexity of the RBF model did not diminish this discrepancy. We anticipate that these new developments could impact other function fitting paradigms such as support vector machines, [132], and mixture models, [105].

Extension of algorithm to multi-variate range. A multivariate extension of the algorithm is also carried out. Unlike the general trend in the literature that leaves the multivariate extensions of the algorithm as a trivial extension to their univariate counter part, we observed considerable improvements in our results. In one of our experiments on Mackey-Glass we show that a multivariate fit is more parsimonious than one of its univariate counterparts.

Application to time-series prediction resulting in best models in the literature.

The algorithm provides accurate results on variety of benchmark data sets including Mackey-Glass and a financial time series.

Finally, there is significant evidence now that the exploitation of the geometry present in data will allow the construction of improved models and enhanced prediction. This dissertation details a new way to find the nonlinear relations in high dimensions that we hope will prove useful in the challenging problem of knowledge discovery in large data sets. We envision future applications to classification on manifolds, pattern recognition and time-series prediction, nonlinear signal processing, optimal control and computer vision.

Bibliography

[1] In E. J. Kansa and Y. C. Hon, editors, *Computers and Mathematics with Applications*, volume 43, pages 275–619. Elsevier, 2002.

[2] W. Ahmed, D. M. Hummels, and M. T. Musavi. Adaptive RBF Neural Network in Signal Detection. In *1994 IEEE International Symposium on Circuits and Systems, ISCAS*, volume 6, pages 265–268, May-Jun. 1994.

[3] M. Aiyar, S. Nagpal, N. Sundararajan, and P. Saratchandran. Minimal Resource Allocation Network (MRAN) for Call Admission Control (CAC) of ATM Networks. In *Proceedings of IEEE International Conference on Networks (ICON)*, page 498, Sep. 2000.

[4] H. Akaike. A New look at Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974.

[5] M. Anderle. *Modeling Geometric Structure in Noisy Data.* Ph.D. dissertation, Department of Mathematics, Colorado State University, Fort Collins, CO, 2001.

[6] M. Anderle and M. Kirby. Correlation Feedback Resource Allocation RBFs. In *Proceedings of 2001 IEEE International Joint Conference on Neural Networks (IJCNN)*, volume 3, pages 1949–1953, Washington, D.C., Jul. 2001.

[7] C. Andrieu, N. Freitas, and A. Doucet. Robust Full Bayesian Learning for Radial Basis Networks. *Neural Computation*, 13:2359–2407, 2001.

[8] R. B. Arellano-Valle and M. G. Genton. On Fundamental Skew-symmetric Distributions. *Journal of Multivariate Analsysis*, 96:93–116, 2005.

[9] R. B. Arnold and R. J. Beaver. The Skew-Cauchy Distribution. *Statistics and Probability Letters*, 49(3):285–290, Sep. 2000.

[10] F. Azam and H. F. VanLandingham. An Alternate Radial Basis Function Neural Network Model. In *2000 IEEE International Conference on Systems, Man, and Cybernetics*, volume 4, pages 2679 – 2684, 8-11 Oct. 2000.

[11] A. Azzalini. A Class of Distributions which Includes the Normal Ones. *Scand. J. Statist.*, 12:171–178, 1985.

[12] A. Azzalini. Further Results on a Class of Distributions which Includes the Normal Ones. *Statistica*, 46:199–208, 1986.

[13] A. Azzalini and A. Capitanio. Statistical Applications of the Multivariate Skew-normal Distribution. *Journal Royal Statistical Society*, 61(B):579–602, 1999.

[14] A. Azzalini and A. Capitanio. Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t Distribution. *Journal of Royal Statistical Society*, 65(B):367–389, 2003.

[15] A. Azzalini and A. Dalla Valle. The Multivariate Skew-normal Distribution. *Biometrika*, 83:715–726, 1996.

[16] D. Barden and C. Thomas. *An Introduction to Differential Manifolds.* Imperial College Press, 2003.

[17] J. T. Barnett and B. Kedem. Zero-crossing Rates of Functions of Gaussian Processes. *IEEE Transactions on Information Theory*, 37(4):1188 – 1194, Jul. 1991.

[18] J. T. Barnett and B. Kedem. Zero-crossing Rates of Mixtures and Products of Gaussian Processes. *IEEE Transactions on Information Theory*, 44(4):1672 - 1677, Jul. 1998.

[19] J. Behboodian, A. Jamalizadeh, and N. Balakrishnan. A New Class of Skew-Cauchy Distributions. *Statistics and Probability Letters*, 76(14):1488–1493, Aug. 2006.

[20] R. E. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, NJ, 1957.

[21] R. Bhatia and K. R. Parthasarathy. Positive Definite Functions and Operator Inequalities. *Bulletin of the London Mathematical Society*, 32(2):214–228, 2000.

[22] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, U.K., 1995.

[23] N. M. Blachman. Zero-crossing Rate for the Sum of Two Sinusoids or Signal Plus Noise. *IEEE Transactions on Information Theory*, 21(6):671–675, Nov. 1975.

[24] S. Bochner. *Vorlesungen über Fouriersche Integrale.* Akademische Verlagsgesellschaft, Leipzig, 1932.

[25] M. Bozzini, L. Lenarduzzi, M. Rossini, and R. Schaback. Interpolation by Basis Functions of Different Scales and Shapes. *CALCOLO*, 41:77–87, 2004.

[26] M. D. Brancoa and D. K. Dey. A General Class of Multivariate Skew-Elliptical Distributions. *Journal of Multivariate Analysis*, 79(1):99–113, Oct. 2001.

[27] M. Brand. Charting a Manifold. In *Neural Information Processing Systems 15 (NIPS'2002)*, Vancouver, Canada,, Dec. 9-14 2002.

[28] P. J. Brockwell and R. A. Davis. *Time series: Theory and Methods.* Springer Series in Statistics. Springer, 2nd edition, 1991.

[29] D. S. Broomhead and M. Kirby. A New Approach for Dimensionality Reduction: Theory and Algorithms. *SIAM Journal of Applied Mathematics*, 60(6):47–67, 2000.

[30] D. S. Broomhead and M. Kirby. The Whitney Reduction Network: a Method for Computing Autoassociative Graphs. *Neural Computation*, 13:2595–2616, 2001.

[31] D. S. Broomhead and M. Kirby. Large Dimensionality Reduction Using Secant-based Projection Methods: The Induced Dynamics in Projected Systems. *Nonlinear Dynamics*, 41:2114–2142, 2005.

[32] D. S. Broomhead and D. Lowe. Multivariable Functional Interpolation and Adaptive Networks. *Complex Systems*, 2:321–355, 1988.

[33] G. Bugmann. Normalized Gaussian Radial Basis Function Networks. *Neurocomputing*, 20(1-3):97–110, 1998.

[34] M. D. Buhmann. *Radial Basis Functions.* Cambridge University Press, 2003.

[35] J. V. Candy. *Signal processing: the Model-Based Approach.* McGraw-Hill Series in Electrical Engineering, Communications and Signal Processing. McGraw-Hill, 1986.

[36] M. Casdagli. Nonlilnear prediction of chaotic time series. *Physica D*, 35:335–356, 1989.

[37] M. Casdagli, D. D. Jardins, S. Eubank, J. D. Farmer, J. Gibson, N. Hunter, and J. Theiler. Nonlinear Modeling of Chaotic Time Series: Theory and Applications. In J.H. Kim and J. Stringer, editors, *Applied Chaos*, pages 335–380. Wiley, 1992.

[38] Y. Le Cun, J. S. Denker, and S. A. Solla. Optimal Brain Damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 598–605. Morgan Kaufmann, San Mateo, CA, 1990.

[39] A. Einstein. Method for the Determination of the Statistical Values of Observations Concerning Quantities Subject to Irregular Fluctuations. In *Archives des Sciences et Naturelles*, volume 37, pages 254–256. 1914.

[40] K. A. Emanuel. An Air-sea Interaction Theory for Tropical Cyclones. Part I: Steady State Maintenance. *Journal of Atmospheric Sciences*, 43:585–604, 1986.

[41] K. A. Emanuel. The Maximum Intensity of Hurricanes. *Journal of Atmospheric Sciences*, 45:1143–1155, 1988.

[42] A. Erfanian and M. Gerivany. EEG Signals can be Used to Detect the Voluntary Hand Movements by Using an Enhanced Resource-Allocating Neural Network. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 721–724, Oct. 2001.

[43] S. E. Fahlman and C. Lebiere. The Cascade-correlation Learning Architecture. In D. S. Touretzky, editor, *Proceedings of the Connectionist Models Summer School*, volume 2, pages 524–532. Morgan Kaufmann, San Mateo, CA, 1988.

[44] J. D. Farmer and J. J. Sidorowich. Predicting Chaotic Time Series. *Physical Review Letters*, 59(8):845–848, Aug. 1987.

[45] B. Fornberg and N. Flyer. *The Gibbs Phenomenon in Various Representations and Applications*, chapter The Gibbs Phenomenon for Radial Basis Functions.

[46] B. Fornberg, E. Larsson, and G. Wright. A New Class of Oscillatory Radial Basis Functions. *Computers and Mathematics with Applications*, 51:1209–1222, 2006.

[47] M. Frean. A Upstart Algorithm: A Method for Constructing and Training Feed-forward Neural Networks. *Neural Computation*, 2(2):198–209, 1990.

[48] B. Fritzke. Fast Learning with Incremental RBF Networks. *Neural Processing Letters*, 1(1):2–5, 1994.

[49] B. Fritzke. Supervised Learning with Growing Cell Structures. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural information Processing Systems*, volume 6, pages 255–262. Morgan Kaufmann, San Mateo, CA, 1994.

[50] W. A. Gardner. Introduction to Einstein's Contribution to Time-series Analysis. *IEEE ASSP Magazine*, 4(4):4–5, Oct. 1987.

[51] F. Girosi, M. Jones, and T. Poggio. Regularization Theory and Neural Network Architectures. *Neural Computation*, 7:219–269, 1995.

[52] F. Girosi and T. Poggio. Networks and the Best Approximation Property. *Biological Cybernetics*, 63:169–176, 1990.

[53] A. K. Gupta, F. C. Chang, and W. J. Huang. Some Skew-symmetric Models. *Random Oper. and Stoch. Equation*, 10(2):133–140, 2002.

[54] A. K. Gupta, G. Gonzlez-Faras, and J. A. Domnguez-Molina. A Multivariate Skew Normal Distribution. *Journal of Multivariate Analysis*, 89(1):181–190, Apr. 2004.

[55] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, 2nd edition, 1999.

[56] F. De Helguero. Sulla Rappresentazione Analitica Delle Curve Abnormali. In G. Castelnuovo, editor, *Atti del IV Congesso Internazionale dei Matematici*, volume III, pages 287–299, Roma, Italy, 6-11 Aprile 1909. Roma R. Accademia dei Lincei.

[57] S. L. Ho, M. Fei, W. N. Fu, H. C. Wong, and E. W. C. Lo. Integrated RBF Network Based Estimation Strategy of the Output Characteristics of Brushless DC Motors. *IEEE Transactions on Magnetics*, 38(2):1033–1036, Mar. 2002.

[58] C. C. Holmes and B. K. Mallick. Bayesian Radial Basis Functions of Variable Dimension. *Neural Computation*, 10(5):1217–1233, 1998.

[59] R. H. Hooker. Correlation of the Marriage-rate with Trade. *Journal of the Royal Statistical Society*, 64:485–492, 1901.

[60] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, England, 1985.

[61] W. Huanga and Y. Chenb. Generalized Skew-Cauchy Distribution. *Statistics and probability letters*, 77(11):1137–1147, Jun. 2007.

[62] D. Hundley, M. Kirby, and R. Miranda. Empirical Dynamical System Reduction II: Neural Charts. In K. Coughlin, editor, *Semi-analytic Methods for the Navier–Stokes Equations (Montreal, 1995)*, volume 20 of *CRM Proc. Lecture Notes*, pages 65–83, Providence, RI, 1999. American Mathathematical Society.

[63] D. R. Hundley. *Local Nonlinear Modeling via Neural Charts.* Ph.D. dissertation, Department of Mathematics, Colorado State University, Fort Collins, CO, 1998.

[64] J. M. Hutchinson, A. W. Lo, and T. Poggio. A Nonparametric Approach to Pricing and Hedging Derivative Securities via Learning Networks. *The Journal of Finance*, XLIX(3):851–889, Jul. 1994.

[65] D. R. Insua and P. Muller. *Feedforward Neural Networks for Nonparametric Regression.* Springer Verlag, New York, 1998.

[66] A. A. Jamshidi. A New Spatio-Temporal Resource Allocation Network (ST-RAN). M.Sc. thesis, Department of Mathematics, Colorado State University, Fort Collins, CO, Fall, 2004.

[67] A. A. Jamshidi. An Adaptive Underwater Target Classification System with a Decision Feedback Mechanism. M.Sc. thesis, Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, Spring, 2002.

[68] A. A. Jamshidi and M. J. Kirby. A Black Box Algorithm for Function Approximation over High Dimension Ranges. *Under preparation.*

[69] A. A. Jamshidi and M. J. Kirby. Skew-Radial Basis Function Expansions for Empirical Modeling. *Submitted to SIAM Journal of Scientific Computation.*

[70] A. A. Jamshidi and M. J. Kirby. Theoretical Aspects of an Iterated Algorithm for Non-linear Function Fitting. *Under preparation.*

[71] A. A. Jamshidi and M. J. Kirby. Examples of Compactly Supported Functions for Radial Basis Approximations. In H. R. Arabnia, E. Kozerenko, and S. Shaumyan, editors, *Proceedings of The 2006 International Conference on Machine learning; Models, Technologies and Applications*, pages 155–160, Las Vegas, Jun. 2006. CSREA Press.

[72] A. A. Jamshidi and M. J. Kirby. Towards a Black Box Algorithm for Nonlinear Function Approximation over High-Dimensional Domains. *SIAM Journal of Scientific Computation*, 29(3):941–963, May 2007.

[73] D. Jianping, N. Sundararajan, and P. Saratchandran. Nonlinear Magnetic Storage Channel Equalization using Minimal Resource Allocation Network (MRAN). *IEEE Transactions on Neural Networks*, 12(1):171–174, Jan. 2001.

[74] D. Jianping, N. Sundararajan, and P. Saratchandran. Communication Channel Equalization using Complex-valued Minimal Radial Basis Function Neural Networks. *IEEE Transactions on Neural Networks*, 13(3):687–696, May 2002.

[75] M. C. Jones and M. J. Faddy. A Skew Extension of the t Distribution, with Applications. *Journal of The Royal Statistical Society Series B*, 65(1):159–174, 2003.

[76] R. D. Jones, Y.C. Lee, C. W. Barnes, G. W. Flake, K. Lee, P. S. Lewis, and S. Qian. Function Approximation and Time Series Prediction with Neural Networks. *IJCNN International Joint Conference on Neural Networks*, 1:649–665, 1990.

[77] J. H. Jung. A Note on the Gibbs Phenomenon with Multiquadric Radial Basis Functions. *Applied Numerical Mathematics*, 57:213–229, 2007.

[78] K. Kadirkamanathan. A statistical Inference Based Growth Criterion for RBF Network. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing IV*, pages 12–21, Ermioni, Greece, 1994.

[79] K. Kadirkamanathan and M. Niranjan. A Function Estimation Approach to Sequential Learning with Neural Networks. *Neural Computation*, 5(6):954–975, 1993.

[80] N. B. Karayiannis and M. M. Randolph-Gips. On the Construction and Training of Reformulated Radial Basis Function Neural Networks. *IEEE Transactions on Neural Networks*, 14(4):835–846, Jul. 2003.

[81] B. Kedem. *Binary Time Series.* Marcel Dekker Inc, New York and Basel, Feb. 1980.

[82] B. Kedem. Detection of Hidden Periodicities by Means of Higher Order Crossings I, II. Technical Report TR84-55 and TR84-56, University of Maryland, Department of Mathematics, 1984.

[83] B. Kedem. A Graphical Similarity Measure for Time Series Models. Technical Report TR85-10, University of Maryland, Department of Mathematics, 1985.

[84] B. Kedem. Spectral Analysis and Discrimination by Zero Crossings. In *Proceedings of the IEEE*, volume 74, pages 1477 – 1493, Nov. 1986.

[85] B. Kedem and G. Reed. On the Asymptotic Variance of Higher Order Crossings with Special Reference to a Fast White Noise Test. *Biometrika*, 73:143–149, Apr. 1986.

[86] M. Kendall and A. Stuart. *The Advanced Theory of Statistics: Design and Analysis, and Time-Series*, volume 3. Hafner Publishing Co., New York, 3rd edition, 1976.

[87] M. B. Kennel, R. Brown, and H. D. I. Abarbanel. Determining Embedding Dimension for Phase-space Reconstruction using Geometrical Construction. *Physical Review A*, 45(6):3403–3411, Mar. 1992.

[88] M. Kirby. Ill-Conditioning and Gradient Based Optimization of Multi-Layer Perceptrons. In J. G. McWhirter and I. K. Proudler, editors, *Mathematics in Signal Processing IV*, The Institute of Mathematics and Its Applications Conference Series: No. 67, pages 223–237. Oxford University Press, 1998.

[89] M. Kirby. *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns.* Wiley, New York, 2001.

[90] M. Kirby and C. Anderson. Geometric Analysis for the Characterization of Nonstationary Time-Series. In E. Kaplan, J. Marsden, and K. R. Sreenivasan, editors, *Springer Applied Mathematical Sciences Series Celebratory Volume for the Occasion of the 70th Birthday of Larry Sirovich*, pages 263–292. Springer-Verlag, 2003.

[91] M. Kirby and R. Miranda. Nonlinear Parametric Neural Charts. Technical report, Colorado State University, Department of Mathematics, 1995.

[92] E. Kleinschmidt. Grundlagen einer Theorie der Tropischen Zyklonen (Basis principles for a theory of tropical cyclones). *Arch. Meteor. Geophys. Bioklimatol*, A4:53–72, 1951.

[93] C. Kumar, P. Saratchandran, and N. Sundararajan. Minimal Radial Basis Function Neural Networks for Nonlinear Channel Equalization. In *IEE Proceedings-Vision, Image and Signal Processing*, volume 147, pages 428–435, Oct. 2000.

[94] C. C. Lee, C. C. Chung, J. R. Tsai, and C. I. Chang. Robust Radial Basis Function Neural Networks. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 29(6):674–685, Dec. 1999.

[95] K. Y. Lee and S. Jung. Extended Adaptive RBF Equalizer for Overcoming Co-channel Interference. *Electronics Letters*, 34(16):1567–1568, Aug. 1998.

[96] P. V. Lee and S. Haykin. *Regularized Radial Basis Function Networks Theory and Applications.* Wiley, New York, 2001.

[97] Y. Li, N. Sundararajan, and P. Saratchandran. Analysis of Minimal Radial Basis Function Network Algorithm for Real-time Identification of Nonlinear Dynamic Systems. *IEE Proceedings of Control Theory and Applications*, 147(4):476–484, Jul. 2000.

[98] W. Liebert, K. Pawelzik, and H. G. Schuster. Optimal Embeddings of Chaotic Attractors from Topological Considerations. *Europhysics Letters*, 14(6):521–526, Mar. 1991.

[99] L. Ljung. *System Identification Theory for the User.* Prentice Hall PTR, 2nd edition, 1999.

[100] D. G. Luenberger. *Linear and Nonlinear Programming.* Addison Wesley, 2nd edition, 1984.

[101] D. J. C. Mackay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 1989.

[102] M. C. Mackey and L. Glass. Oscillation and Chaos in Physiological Control Systems. *Science*, 197:287–289, Jul. 1977.

[103] A. MacLachlan. An Improved Novelty Criterion for Resource Allocating Networks. In *Proceedings of the IEE 5th International Conference on Artificial Neural Networks*, pages 48–52, Cambridge, UK, Jul. 1997. IEE, London, UK.

[104] A. D. Marrs. An Application of Reversible-jump MCMC to Multivariate Spherical Gaussian Mixtures. In M. I. Kearns and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 577–583, Cambridge, MA, 1998. MIT press.

[105] G. Mclachlan and D. Peel. *Finite Mixture Models.* Wiley-Interscience, 2000.

[106] C. A. Micchelli. Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions. *Constructive Approximation*, 2:11–22, 1986.

[107] A. A. Michelson and S. W. Stratton. A New Harmonic Analyzer. *American Journal of Science*, pages 1–13, 1898.

[108] J. Moody and C. Darken. Fast Learning in Networks of Locally-tuned Processing Units. *Neural Computation*, 1:281–294, 1989.

[109] H. L. Moore. *Economic Cycles: Their Law and Cause.* Macmillan, New York, 1914.

[110] S. Nadarajah and S. Kotz. Skewed Distributions Generated by the Normal Kernel. *Statistics and Probability Letters*, 65:269–277, 2003.

[111] M. J. L. Orr. Introduction to Radial Basis Function Networks (1996). Technical report, Center for Cognitive Science, University of Edinburgh, Edinburgh, Scotland, Apr. 1996.

[112] M. J. L. Orr. An EM Algorithm for Regularized RBF Networks. In *1998 International Conference on Neural Networks and Brain Proceedings*, pages 251–254, Beijing, China, Oct. 27-30, 1998. Publishing House of Electronics Industry, Box 173 Beijing, 100036 China.

[113] M. J. L. Orr. Recent Advances in Radial Basis Function Networks. Technical report, Institute for Adaptive and Neural Computation, Division of Informatics, Edinburgh University, Edinburgh, Scotland, Jun. 1999.

[114] P. W. Pachowicz and S. W. Balk. Adaptive RBF Classifier for Object Recognition in Image Sequences. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN*, volume 6, pages 600–605, Jul. 2000.

[115] J. Park and I. W. Sandberg. Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, 3:246–257, 1991.

[116] J. Park and I. W. Sandberg. Approximation and Radial-Basis-Function Networks. *Neural Computation*, 5:305–316, 1993.

[117] J. Persing and M. T. Montgomery. Hurricane Superintensity. *Journal of Atmospheric Sciences*, 60:2349–2371, 2003.

[118] J. Platt. A Resource Allocating Network for Function Interpolation. *Neural Computation*, 3:213–225, 1991.

[119] T. Poggio and F. Girosi. Networks for Approximation and Learning. *Proceeding of the IEEE*, 78(9):1481–1496, 1990.

[120] T. Poggio and F. Girosi. Regularization Algorithm for Learning that Are Equivalent to Multilayer Networks. *Science*, 247:978–982, 1990.

[121] M. J. D. Powell. Radial Basis Functions for Multivariable Interpolation: A Review. In J. C. Mason and M. G. Cox, editors, *Algorithms for approximation*, pages 143–167. Clarendon Press, Oxford, 1987.

[122] M. J. D. Powell. The Theory of Radial Basis Functions in 1990. In W. Light, editor, *Advances in Numerical Analysis*, volume II of *Wavelets, Subdivision Algorithms, and Radial Basis Functions*, pages 105–210. Oxford University Press, 1992.

[123] S. O. Rice. Mathematical Analysis of Random Noise. *Bell Systems Technical Journal*, 24:46–156, Jan. 1945.

[124] D. I. Rios and P. Muller. Feedforward Neural Networks for Nonparametric Regression. In D. K. Dey, P. Muller, and D. Sinha, editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 181–191, New York, 1998. Springer Verlag.

[125] H. J. Rossberg. Positive Definite Probability Densities and Probability Distributions. *Journal of Mathematical Sciences*, 76(1):2181–2197, 1995.

[126] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, Dec. 2000.

[127] S. Roweis, L. Saul, and G. Hinton. Global Coordination of Local Linear Models. In *Neural Information Processing Systems 15 (NIPS'2002)*, Vancouver, Canada,, Dec. 9-14.

[128] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing.* MIT Press, Cambridge, MA, 1986.

[129] R. Schaback. Error Estimates and Condition Numbers for Radial Basis Function Interpolation. *Advances in Computational Mathematics*, 3:251–264, 1995.

[130] R. Schaback. *Reconstruction of Multivariate Functions from Scattered Data.* preprint, University of Göttingen, 1997.

[131] R. Schaback and H. Wendland. Characterization and Construction of Radial Basis Functions. In N. Dyn, D. Leviatan, D. Levin, and A. Pinkus, editors, *Multivariate Approximation and Applications, Eilat Proceedings*, pages 1–24. Cambridge University Press, 2000.

[132] B. Scholkopf, S. Kah-Kay, C. J. C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765, Nov. 1997.

[133] P. Smyth. On Stochastic Complexity and Admissible Models for Neural Network Classifiers. In R. Lippmann, J. Moody, and D. S. Touretzky, editors, *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*, pages 818–824. Morgan Kaufmann Publishers Inc., San Mateo, CA, 1991.

[134] H. Takeda, S. Farsiu, and P. Milanfar. Kernel Regression for Image Processing and Reconstruction. *IEEE Transactions on Image Processing*, 16(2):349 – 366, Feb. 2007.

[135] F. Takens. Detecting Strange Attractors in Turbulence. *Dynamical Systems and Turbulence, Warwick 1980; Proceeding of a symposium held at University of Warwick 1979-1980*, pages 366–381.

[136] Y. W. Teh and S. Roweis. Automatic Alignment of Local Representations. In *Neural Information Processing Systems 15 (NIPS'2002).*

[137] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, Dec. 2000.

[138] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems.* John Wiley & Sons, New York, 1977.

[139] J. J. Verbeek, S. T. Roweis, and N. Vlassis. Non-linear CCA and PCA by Alignment of Local Models. In *Neural Information Processing System 16 (NIPS'2003).*

[140] M. Visani, C. Garcia, and J. M. Jolion. Normalized Radial Basis Function Networks and Bilinear Discriminant Analysis for Face Recognition. *IEEE Conference on Advanced Video and Signal Based Surveillance*, 15(16):342–347, 2005.

[141] G. Wahba. Spline Bases, Regularization, and Generalized Cross Validation for Solving Approximation Problems with Large Quantities of Data. In W. Cheney, editor, *Approximation Theory III*, pages 905–912. Academic Press, 1980.

[142] A. Wahed and M. M. Ali. The Skew-Logistic Distribution. *J. Statist. Res.*, 35:71–80, 2001.

[143] H. Wendland. Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree. *Advances in Computational Mathematics*, 4(1):389–396, Dec. 1995.

[144] H. Wendland. *Scattered Data Approximation.* Cambridge University Press, 2005.

[145] P. J. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* Ph.D. dissertation, Division of Applied Mathematics, Harvard University, Boston, MA, Aug. 1974.

[146] H. Whitney. Differential Manifolds. *Annals of Mathematics*, 37(3):645–680, Jul. 1936.

[147] B. Widrow and M. E. Hoff. Adaptive Switching Circuits. In *1960 WESCON Convention Record Part IV, Reprinted in J.A. Anderson and E. Rosenfeld, Neurocomputing: Foundations of Research*, pages 96–104, Cambridge, MA, 1988, 1960. The MIT Press.

[148] Z. Wu. Multivariate Compactly Supported Positive Definite Radial Functions. *Advances in Computational Mathematics*, 4:283–292, 1995.

[149] G. Xu and V. Chandrasekar. Radar Rainfall Estimation from Vertical Reflectivity Profile Using Neural Network. In *IEEE 2001 International Geoscience and Remote Sensing Symposium, IGARSS*, volume 7, pages 3280 –3281, Jul. 2001.

[150] A. M. Yaglom. Einstein's 1914 Paper on the Theory of Irregularly Fluctuating Series of Observations. *IEEE ASSP Magazine*, 4(4):7–10, Oct. 1987.

[151] K. Yamauchi, N. Yamaguchi, and N. Ishii. Incremental Learning Methods with Retrieving of Interfered Patterns. *IEEE Transactions on Neural Networks*, 10(6):1351–1365, Nov. 1999.

[152] L. Yingwei, N. Sundararajan, and P. Saratchandran. A Sequential Scheme for Function Approximation Using Minimal Radial Basis Function Neural Networks. *Neural Computation*, 9:461–478, 1997.

[153] L. Yingwei, N. Sundararajan, and P. Saratchandran. Performance Evaluation of a Sequential Minimal Radial Basis Function (RBF) Neural Network Learning Algorithm. *IEEE Transactions on Neural Networks*, 9(2):308–318, Mar. 1998.

[154] M. Zhihong, X. H. Yu, K. Eshraghian, and M. Palaniswami. A Robust Adaptive Sliding Mode Tracking Control Using an RBF Neural Network for Robotic Manipulators. In *Proceedings of IEEE International Conference on Neural Networks*, volume 5, pages 2403 –2408, Nov. - Dec. 1995.

Appendix

In an effort to keep this work as self contained as possible we present the additional results that are required for the proof of theorem 4.5.1 in Chapter 4 and a gallery of RBFs.

Perturbation Bound

Following [25], the system of equations for the perturbed interpolation problem

$$\begin{cases} \widetilde{A}\widetilde{\alpha} + P\widetilde{\beta} = f \\ P^T\widetilde{\alpha} + 0 = f \end{cases} \tag{1}$$

is solvable if the rank condition $\text{rank}(P) = Q \leq N$ holds and if the perturbation $\tilde{A}$ of the interpolation matrix A is bounded by

$$\|\tilde{A} - A\|_2 \leq \hat{\lambda} \tag{2}$$

where

$$\gamma^T A\gamma \geq \hat{\lambda}\|\gamma\|_2^2$$

for all $\gamma \in V$ where V is the space defined in Equation (4.16), [129].

Gallery of RBFs

The table below page presents some of the widely used (conditional) positive definite RBFs. Note, for further properties of Wendland RBFs and their valid dimension please see, e.g., [144].

$\phi(r)$	order m	Name
$\exp(-\alpha r^2), \alpha > 0$	$m \geq 0$	Gaussian
$r^{2k}\ln r, k \in \mathbb{N}$	$m \geq k+1$	thin-plate splines
r^β, $\beta > 0$, $\beta \notin 2\mathbb{N}$	$m \geq \lceil \frac{\beta}{2} \rceil$	linear or cubic if $\beta = 1 or 3$
$(c^2 + r^2)^\beta, \beta < 0$	$m \geq 0$	inverse multiquadrics
$(c^2 + r^2)^\beta, \beta > 0$, $\beta \notin \mathbb{N}$	$m \geq \lceil \beta \rceil$	multiquadrics
$(1-r)_+^4(1+4r)$	$m \geq 0$	Wendland

Table 1: This table presents the commonly used RBFs in the literature.

Gallery of skew-RBFs

The table below summarizes the sRBFs that are introduced in this paper. These basis functions are best suited for the case that there is more data points than basis functions. According to Theorem 4.5.1, one could check if an specific sRBF is positive definite to be used for interpolation. Note that the functions shown here are only examples of many functions one could generate based on the background provided in this work.

Table 2: This table shows a collection of skew-RBFs introduced in this paper. Parameters c, λ and W denote the center, skew parameter and the inner product weight, respectively.

$\phi(x)$	Name
$\exp(-\|x-c\|_W^2)\int_{-\infty}^{-\lambda^T(x-c)}\exp(-y^2)dy$	Erf-Gaussian
$\left(\frac{1}{\pi}\arctan\left(\lambda^T(x-c)\right)+\frac{1}{2}\right)\operatorname{sech}(\|x-c\|_W)$	Atan-Hyper-Sec.
$\left(\frac{1}{\pi}\arctan\left(\lambda^T(x-c)\right)+\frac{1}{2}\right)\exp(-\|x-c\|_W^2)$	Atan-Gaussian
$\left(\frac{1}{\pi}\arctan\left(\lambda^T(x-c)\right)+\frac{1}{2}\right)\sqrt{1-\|x-c\|_W^2}H(1-\|x-c\|_W)$	Atan-Circle
$\left(\frac{1}{\pi}\arctan\left(\lambda^T(x-c)\right)+\frac{1}{2}\right)\left(\cos(\|x-c\|_W\pi)+1\right)H(1-\|x-c\|_W)$	Atan-Hanning
$\left(\frac{1}{\pi}\arctan\left(\lambda^T(x-c)\right)+\frac{1}{2}\right)\exp\left(\frac{-1}{1-\|x-c\|_W^2}\right)H(1-\|x-c\|_W)$	Atan-Mollifier