PROBABILISTIC PREDICTION USING EMBEDDED RANDOM PROJECTIONS OF HIGH DIMENSIONAL DATA

A Dissertation

by

RICHARD CABLE KURWITZ

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2009

Major Subject: Nuclear Engineering

PROBABILISTIC PREDICTION USING EMBEDDED RANDOM PROJECTIONS

OF HIGH DIMENSIONAL DATA

A Dissertation

by

RICHARD CABLE KURWITZ

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Frederick R. Best
Committee Members,	Kenneth L. Peddicord
	Dennis L. O'Neal
	Yassin A. Hassan
Head of Department,	Raymond J. Juzaitis

May 2009

Major Subject: Nuclear Engineering

ABSTRACT

Probabilistic Prediction Using Embedded Random Projections of High Dimensional Data. (May 2009)

Richard Cable Kurwitz, B.S.; M.S., Texas A&M University

Chair of Advisory Committee: Dr. Frederick R. Best

The explosive growth of digital data collection and processing demands a new approach to the historical engineering methods of data correlation and model creation. A new prediction methodology based on high dimensional data has been developed. Since most high dimensional data resides on a low dimensional manifold, the new prediction methodology is one of dimensional reduction with embedding into a diffusion space that allows optimal distribution along the manifold. The resulting data manifold space is then used to produce a probability density function which uses spatial weighting to influence predictions i.e. data nearer the query have greater importance than data further away. The methodology also allows data of differing phenomenology e.g. color, shape, temperature, etc to be handled by regression or clustering classification.

The new methodology is first developed, validated, then applied to common engineering situations, such as critical heat flux prediction and shuttle pitch angle determination. A number of illustrative examples are given with a significant focus placed on the objective identification of two-phase flow regimes. It is shown that the new methodology is robust through accurate predictions with even a small number of data points in the diffusion space as well as flexible in the ability to handle a wide range of engineering problems.

ACKNOWLEDGEMENTS

A graduate education is a luxury, more so if one has a family and is returning to school. I would not have been able to enjoy this luxury without the support of my family and an opportunity from my advisor. Pursuing a PhD is a life changing experience. The lens through which I have viewed the world has been altered and now feel that I am more capable of viewing the whole. With this in mind, I would like to thank a number of those that made it possible.

First, I am forever indebted to Dr. Frederick Best for provided an opportunity to pursue an academic career by providing a unique position working with him for the Interphase Transport Phenomena Laboratory. Dr. Best has served as an advisor, mentor, colleague, and most importantly, a friend. I look forward to future discussions in the waffle house prior to flying experiments on the reduced gravity aircraft. I would like to acknowledge Dr. Igor Carron. Dr. Carron and I shared an office when I first started graduate school and no one has done more to develop my intellectual curiosity. The direction of my research and interests beyond the range of activities within the laboratory is directly attributed to Dr. Carron. I would like to thank the members of my committee for sticking with me when progress was slow and making available their precious time as I approached completion.

Second, I thank my wife, Allison, and my two children, Aberdeen and Augustus. They sacrificed the most important facet of family life, time together. I can never replace the lost weekends and evenings but hope that seeing the effort put forth will provide a

template for their dreams. It is only Allison’s consideration for my happiness and satisfaction that she was able to set aside anniversaries, birthdays, and my presence at a number of other activities. My mother and father did a wonderful job pretending to be interested in my research and continue to support my career even though they do not have a good idea what I do.

Third, I am grateful to a number of friends and colleagues that I have interacted with the past few years. I would like to thank the current and former members of the Interphase Transport Phenomena Laboratory who have provided a stimulating work environment: many of whom were my students. However, without Ken Marsden, Jae Chang, Ryoji Oinuma, Mike Ellis, David Bean, Filip Finodeyev, Katy Hurlbert, Tom Reinarts, Charles Neil, Melissa Ghrist, Kevin Supak, Adam Shephard, and Ben Larson, this endeavor would not have been as successful or as fulfilling. I would also like to acknowledge Alex and Amy Maslowski, Teresa Bailey, Kevin Clarno, Lucille Dauffy, Josh Jarrell, and Adam Hetzler for listening or for sharing a cup of coffee, both of which were important to me finishing my PhD. While I have left many friends and colleagues unnamed, they are not forgotten and are certainly appreciated.

NOMENCLATURE

Section 2

Y	signal vector
Y_i	ith element of signal vector
A	Matrix Used for Projecting Signal Vector Onto Lower Dimensional Space
A_{ij}	ith row, jth column of projection matrix
d	size of lower dimensional space
N	Original dimensional space
D	Distance matrix – matrix of interpoint distances between a set of signal vectors
D_{ij}	Distance between signal vector i and signal vector j
RMSD	Root Mean Square Deviation between two elements
n	Total number of signals
ψ_{ij}	Signal i element j
ε	Normalized RMSD
Θ_i	ith eigenvector or embedding/diffusion coordinate
ν	Normalization coefficient
σ	Window Parameter for Gaussian kernel
k	Gaussian kernel
$\tilde{a}$	Normalized kernel matrix
a	Symmetric Conjugate of $\tilde{a}$
A	Diffusion operator
λ_j	jth eigenvalue
ζ	Diffusion distance
ξ	Known point used in constructing Experimental Probabilistic Hypersurface
x	Query on the EPH
p_i	Probability at point i
I	Entropy
γ	Range of possible outcomes on the EPH
Γ	Output of known points
τ	Discritization of Output Space
ϑ	Proportionality constant
δ	Absolute difference
c	Roots of probability function
b	Roots of probability function
q	Standard deviation of Γ
g_j	Output of EPH at point j for query x
m	Number of known measurements

Section 3

h_L	Pipe Head Loss
L	Length of Pipe
f	Friction factor
V	Velocity of fluid in pipe
D	Diameter of pipe
Re	Reynolds Number
μ	Dynamic viscosity of fluid
ν	Kinematic viscosity
ρ	Density of fluid
α	Scaling factor
β	Scaling factor
ε	Pipe roughness
A	Constant
B	Constant
C	Constant
MSC	Model Selection Criterion
CW_i	Colebrook-White Friction Factor at Point i
$\overline{CW}$	Mean of Colebrook-White Friction Factor
NP	Number of Parameters ($^{\varepsilon}/_{D}$, Re)
M_i	Model Friction Factor at Point i
n	Number of Test Points
P	Probability
CDF	Cumulative Probability Function
σ	Standard deviation of distribution

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

Page

Page

LIST OF TABLES

1. INTRODUCTION

With the increased availability of both data measurement and storage technologies as well as broader use of complex computational codes, methods that allow the user to quickly and easily interpret results allowing accurate prediction are becoming paramount. Previously, the development and validation of useful engineering models was either a theory driven process or an empirical data statistically driven process. Both are hindered by high-order dimensionality and the lack of sophisticated statistical techniques to interpret multiple dimensions. Methods exist, including neural networks, which can efficiently solve linear systems, but quickly become ineffective when non-linearity is introduced or when training data are not available. Recent developments in the area of dimension reduction allow for the collapsing of complex, high dimensional, non-linear problems commonly found in engineering into optimal low dimension embeddings. The following work presents a paradigm shift in the art of predictive modeling through the development of a new methodology to quantitatively identify the spatial arrangement of data in these low dimensional embeddings collected from common engineering problems thereby allowing the user to develop probabilistic tools

This dissertation follows the style of *Nuclear Technology.*

that can be used to either classify or predict phenomena. To predict the result of an unknown function, a probability density is constructed using known information. This process is facilitated by utilizing a low dimensional mapping of the original data to provide fast, accurate results. Contrast this with classic correlation techniques based predictions that provide a single point estimate with no information regarding the uncertainty.

Scientists and engineers strive to develop simpler, faster, more accurate models that allow one to predict the behavior of nature. From visual observation based classifications to sophisticated multiscale physics codes, we desire to understand the complex, high dimensional world around us. Dimensionality manifests itself in the number of variables required to define phenomena e.g. pixels in an image, the variables in an input deck to a sophisticated computational code, or the channel readings of a multichannel analyzer. As our understanding has grown, the complexity as defined by the number of variables or dimensions of our methods has increased. However, the complexity is ultimately bounded by our ability to effectively handle or solve these high order problems. Empirical approaches to model building require an abstraction phase where underlying patterns of information are recognized. This information is then integrated with basic physical laws or phenomenological models to produce a model that can be used for predictive purposes. The development of dimensionless (unit less) terms or Buckingham Pi theory is a common approach to dimension reduction found in engineering. The collection of these terms presents a simpler model that can be visualized or easily solved. This is due to the fact that the underlying dimensionality of the problem is small compared to the actual dimensional space in which the data reside. Unfortunately, identifying these terms or developing relationships is extremely difficult. It is not unheard of for the development and refinement of these relationships to encompass decades of scientific research. This is due to the fact that a limited set of tools were available to reduce dimensions and simplify the problem. This abstraction process is made more difficult with increasing complexity, i.e. dimension of the problem or system under investigation, and nonlinearity. Methods to assist this process are the purpose of this study.

The goal of this study is to develop an algorithmic method that detects the intrinsic dimension of the data that thus will allow prediction. Several methods have been utilized throughout the last fifty years. These include principal component analysis, self organizing maps, and neural networks.[1,2,3] Several nonlinear global methods have been proposed dealing with low dimensional mappings with some success.[4,5] Recently, a geometrically oriented algorithm that essentially produces a graph of neighboring data points that approximates the manifold has been presented by Lafon.[6] The manifold's intrinsic dimension is essentially a dimension reducing mapping from the ambient space where the local information of the data is preserved.[7] Another recent approach is the use of random projections to map high dimensional data to a low dimensional subspace. This method is related to Compressed Sensing and has been shown to be a viable dimension reduction approach.[8] Methodologies would then be needed to make predictions given a query in the original dimensional space. Recent work by Beauzamy[9,10] in the construction of an Experimental Probabilistic Hypersurface (EPH) allows one to store information and make predictions based on the propagation of information from what is known to the query point. EPH is based on the theory of maximal entropy and for each query point, this technique produces a probability density which a value can be extracted. The concentration or shape of the probability density is dependent upon how far away the query is from the measure points; thus, for queries far from given data, the density is flat. Conversely, for a query near given data, the probability density is more peaked ultimately producing a Dirac function for a query at a given data location.

1.1. Statement of Problem and Scope

The proposed project will develop and demonstrate a methodology that will allow for the dimensional parameterization of two-phase flow and other large data sets as well as the ability to make predictions based solely on available data. The complete integration of dimensional reduction, embedding, and probabilistic prediction based on the embedding of observed data have not previously existed as an integrated methodology. The application of this methodology to the area of fluid dynamics and heat transfer is new and presents an exciting new approach to modeling in this area.

The rapid rise in the ability to collect large amounts of data from multiple sources such as video and digital sensors produces a temporal stream of information that is becoming increasing difficult to analyze. For instance, recent two-phase flow testing aboard NASA’s Reduced Gravity Aircraft yielded electronic data for flow rates, pressure, acceleration, etc. collected at 100 Hz and high speed video collected at 500 frames per second.[11] How does one relate or fuse the data for analysis and ultimately to make predictions?

Gene expression or protein charge data is a classic candidate for dimension reduction. Typical mass spectrometry datasets consist of thousands of parameters/variables. Recent tests performed at Texas A&M investigated a novel perfusion system to examine the effects of radiation on model respiratory tissue.[12] The data consists of three sets of three samples of intensity readings for 10399 channels that correspond to in-vivo unirradiated, ex-vivo unirradiated, and ex-vivo irradiated treatments of analog human respitory tissue from Fischer 344 rats. The 10399 channels correspond to approximately 9028 gene probes with the difference in corresponding to bacterial transcripts and other pads. Using the new methodology, nine samples were randomly compressed and embedded into a diffusion space resulting in three distinct clusters that correspond to the different classes which are shown in Figure 1.1. Without any training data, identified channels, or special processing, clusters corresponding to the treatment classes are easily identified. This amazing result demonstrates the power of this method and allows very complex data to be analyzed objectively without an existing model.

In the past, predictive models were developed because of the high cost and difficulty of using large amounts of data. In an effort to scale from previously tested conditions to a desired state, dimensional analysis provides a method for scaling using computed sets of dimensionless parameters of the given variables, even if the form of the equation was unknown. However, new kinds of data such as video and still images are highly dimensional and are difficult to utilize in predictive models which typically requires the researcher to watch the video to discern the phenomena of interest. Automated processing is limited due to the training of pattern recognition algorithms or other computationally intensive processing of the imagery. This process of using disparate data types e.g. images, color, flow rate, etc is commonly referred to as data fusion and is as area of extreme interest in the analysis of complex problems.

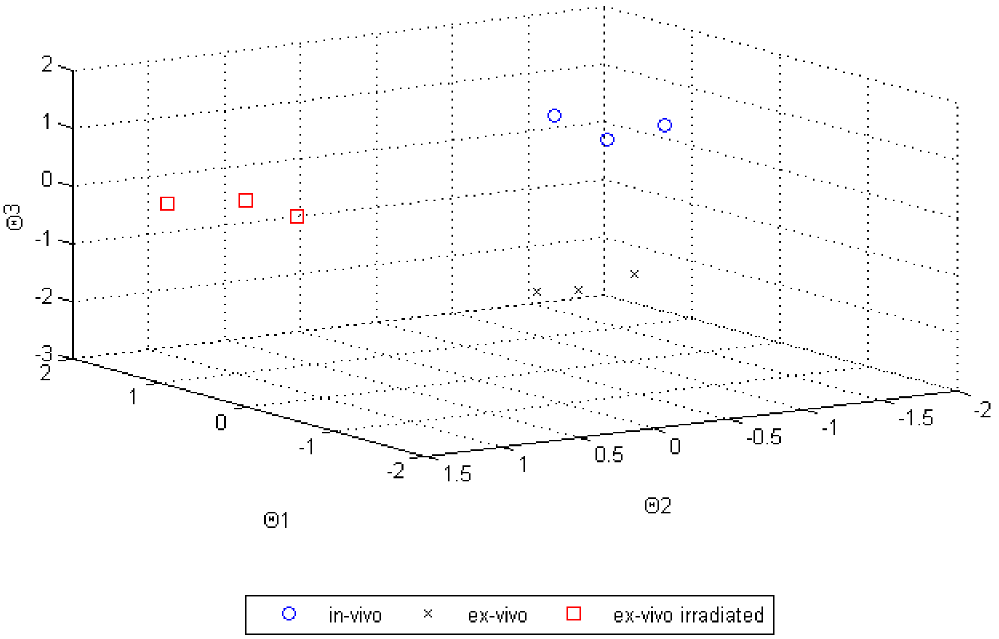


Figure 1.1 Processed Spectrum Data Plotted Using Output from the New Methodology.

An example related to fluid mechanics is shown below in Figures 1.2 and 1.3. High speed video of a flow boiling system was recorded at different heat flux levels.[13] Nine videos were used in the analysis corresponding to five values of heat flux. Figure 1.2 consists of a selection of image frames from the videos illustrating the movement of bubbles. The larger values of heat flux are shown to have a larger number of bubbles due to the higher amount of energy being transferred to the working fluid. It is apparent that the change in the number of bubbles is not linear. To demonstrate the result of the method presented in this dissertation, two videos at heat fluxes of 80 $^{kW}/m^2$, 120 $^{kW}/m^2$, 140 $^{kW}/m^2$, and 160 $^{kW}/m^2$, were used as the known inputs and the heat flux was predicted for the 100 $^{kW}/m^2$ video. Each image in the 340 frames of the video corresponds to 141 x 400 pixels resulting in 19,176,000 elements or variables to evaluate per movie. The 19 million variables are reduced to 64 variables which are embedded into a two dimensional diffusion space. These two element vectors are used to predict the heat flux. Thus, the two diffusion coordinates of the eight videos corresponding to four known heat fluxes are used to construct a probability distribution for the remaining video.

Figure 1.3 shows the diffusion coordinates (Θ_1, Θ_2) for the known and query videos. Based on Figure 1.2, it is expected that the query video at 100 $^{kW}/m^2$ should lie between 80 $^{kW}/m^2$ and 120 $^{kW}/m^2$, which is best captured by the probability distribution in Figure 1.3 that is peaked at respective values of heat flux. Taking the mean of the probability distribution results in a predicted heat flux of 106 $^{kW}/m^2$. Thus, one can predict the actual heat flux to within 10% based on the video alone. The tremendous amount of data in each video is compressed into 64 meta-variables that are embedded into a two dimensional map that accurately reflects the actual heat flux. That each observation of boiling resulted in over 19 million variables that can be accurately described through dimensional reduction by two values is an amazing feat. Thus, the utilization of visual or other higher dimensional data has been greatly enhanced.

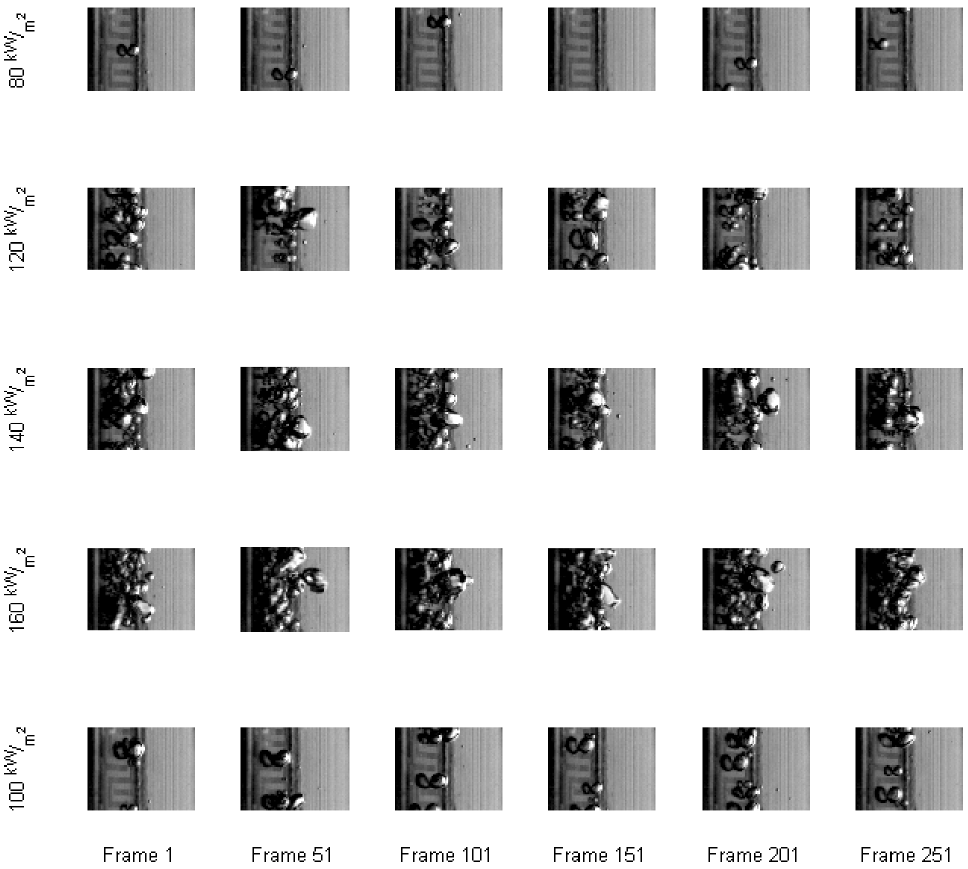


Figure 1.2 Selected High Speed Video Frames for Flow Boiling at Various Heat Fluxes.

The goal of the research is to utilize recent developments in data mining to make accurate predictions of phenomena of interest without complete understanding of the underlying physics. The research will utilize existing large data sets found in thermal-hydraulics, high speed imagery, and large scale computational models. A methodology will be developed that will allow one to discover the underlying dimension of the data and utilize existing information to make accurate predictions for a given query. Today, we are able to utilize complex computational tools that require a large number of inputs and we are able to gather voluminous amounts of data quickly and cheaply. The clear importance of this work in regard to engineering systems is the ability to utilize the considerable amount of data collected or time intensive computational output to quickly make accurate predictions.

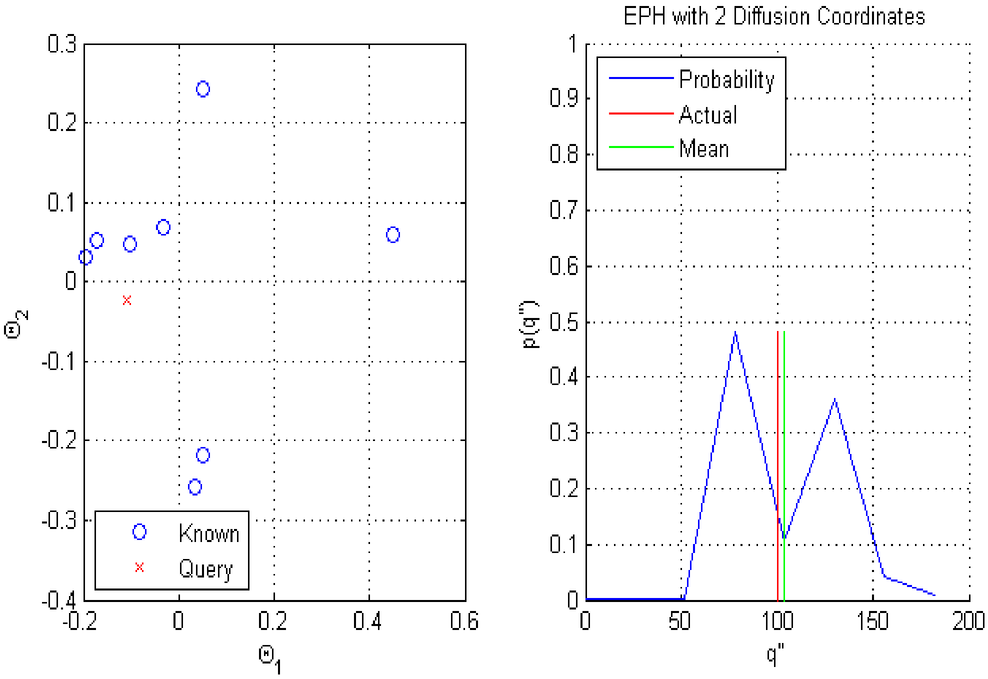


Figure 1.3 Plot of a) Embedding Coordinates and b) Probability Distribution.

1.2. Literature Review

The demand to process large amounts of data is driving the development of dimensional reduction techniques.[14] The large data sets for marketing,[15] politics,[16] and science[17] are being utilized more frequently and since the fundamental inputs to these methods are application neutral, advances in one field are quickly transferred to another. Traditionally, for dimensional reduction, principal component analysis has been utilized. However, this approach is only applicable to problems with essentially linear structures, whereas recent research has shown success for nonlinear applications when applying various linear programming techniques. The focus of this research is the application and extension of these graph based methods to higher dimensional engineering problems resulting in a low dimension model that has an accurate predictive capability.

1.2.1. Dimension Reduction Techniques

Principal components analysis (PCA) is a simplification technique whereby multidimensional datasets are reduced to lower dimensions for analysis.[18] PCA is also referred to as the (discrete) Karhunen-Loève transform or the Hotelling transform. PCA is a linear transformation that maps the high dimensional data, expressed as a covariance matrix, to a new coordinate system. The covariance matrix is a matrix made up of covariances between the elements of a vector that describes high dimensional data. This new coordinate system is ranked with the first coordinate describing the greatest variance (called the first principal component), the second greatest variance on the second coordinate, and so on. The orthogonal decomposition of the data covariance matrix can then be evaluated such that only the desired information is retained. PCA can be used for dimensionality reduction in a dataset while retaining those characteristics of the dataset that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Such low-order components often contain the "most important" aspects of the data. But this is not necessarily the case, depending on the application. PCA has the distinction of being the optimal linear transformation for keeping the subspace that has largest variance. This advantage, however, comes at the price of greater computational requirement if compared, for example, to the discrete cosine transform. Unlike other linear transforms, the PCA does not have a fixed set of basis vectors. Its basis vectors depend on the data set.

Multidimensional scaling (MDS) utilizes a number of statistical methods to reduce data to a low dimensional space as an exploratory tool,[19] This eigenvalue technique is used to visualize proximities in a low dimensional space suitable for graphing or 3D visualization allowing for interpretation of the major parameters that may provide insight into the processes underlying the perceived nearness of entities.[19] A subset of MDS is referred to as Metric Multidimensional Scaling and differs only in that the output graph is governed by a procedure called stress majorization that is essentially an optimization process on a weight function dealing with the Euclidean distance between data points. All methods use a Euclidean distance between data to form a dissimilarity matrix of inputs and produce a corresponding output matrix that optimizes some cost function. For Metric MDS, significant reduction in the stress requires only a few dimensions.[20] Therefore, only these terms are needed to accurately arrange the data for visualization. Many current methods ultimately utilize MDS but build the input matrix to extract the desired features of the manifold on which the data reside. If the dissimilarity or pair wise distances are Euclidean then the results are the same as PCA.

Since many data sets contain nonlinear structures that are invisible to PCA and MDS, new approaches are needed to handle these complex manifolds. Importantly, methods that succeed in learning nonlinear manifolds are required for use with most real world data. One of the most frequently used methods to deal with the problems of nonlinearity is kernel methods. This approach projects or maps the data into a high dimensional feature space, where each coordinate corresponds to one feature. In this feature space, a number of methods can be used to find relations in the data including PCA. Since the mapping can be quite general (not necessarily linear, for example), the relations found in this way are accordingly very general.[21]

Kernel methods owe their name to the use of kernels or radial basis functions, that enable them to operate in the feature space without ever computing the coordinates of the data in that space.[22] Kernel functions have been introduced for sequence data, text, images, as well as vectors. Kernel PCA implicitly constructs a higher dimensional space, in which there are a large number of linear relations between the dimensions. Subsequently, the low-dimensional data representation is obtained by applying traditional PCA. For good or bad, the choice of kernel adds a further parameter for the investigator to decide upon.

New methods that are able to efficiently explain hidden structure or relationships and that are less susceptible to user error are needed. Unlike classical techniques such as principal component analysis (PCA) and multidimensional scaling (MDS), a new approach needs to be capable of discovering the nonlinear degrees of freedom that underlie complex natural observations, such as human handwriting or images of a face under different viewing conditions.[4] Unlike other nonlinear approaches, Isomap is a new method that computes a globally optimal solution rather than a local solution.

Tenenbaum's Isomap algorithm extracts meaningful dimensions by measuring the distance between data points in the geometric shapes formed by items in a nonlinear data set. The technique utilizes the distances between points on the surface of a manifold or the geodesic distance. High dimensional data is typically nonlinear and is represented by complicated shapes that are difficult to resolve or project onto lower dimensions. Isomap measures the distance between any two points on the manifold, then uses these geodesic distances in combination with a classic multidimensional scaling algorithm in order to make a low dimensional representation of that data. The Isomap algorithm computes the distances between neighboring data points. For each pair of non-neighboring data points, a distance is calculated by determining the shortest path required to hop from neighbor to neighbor Finally, the classical method of multidimensional scaling is used to find a set of low-dimensional points with similar pair wise distances.

Locally Linear Embedding (LLE) is a dimensional reduction technique that utilizes an unsupervised algorithm to link local points and then maps them to a low dimension. Unlike clustering methods for local dimensionality reduction, LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima.[5] Implementing LLE starts with the calculation of how each item in a data set is related to the few items nearest to it. It then preserves these neighborhood relationships when the data is converted from its high dimensional form into a low dimensional form.[5] The method utilizes weight coefficients based on the point's neighbors and then finds a set of corresponding low-dimensional points. LLE is able to learn the local structure of nonlinear manifolds, such as those generated by images of faces or documents of text using local information.

The goal of these recent approaches is to find a system of coordinates that parameterize/describes the underlying manifold. Discovering mappings to low-dimensional representations has been a focus of much recent work on unsupervised learning.[4,5,23,24,25] A key theme of these learning algorithms is to preserve (local) topological and geometrical properties (for example, geodesics, proximity, symmetry, angle) while projecting data points to low dimensional representations.[26]

A promising recent approach is the method of Laplacian eigenmaps that map points in high dimensional space to a low dimensional representation that preserves local relationships similar to LLE. Belkin and Niyogi[23] showed a technique to construct the Laplace-Beltrami operator on a manifold from points uniformly sampled. The eigenfunctions of this operator can be used to perform dimensionality reduction. Like other methods previously mentioned, this technique uses a pair wise dissimilarity measure as input and applies a radial basis function or kernel as a weighting tool to produce a sparse distance matrix. Dimension reduction or parameterization of the manifold is then performed by choosing the n-largest eigenvalues.

The work by Lafon[6,27] and Coifman[28] describes a method of constructing what they call a diffusion kernel on the data and then employing its spectral properties to define a map that embeds the data into a lower dimension space. This approach builds upon that of Belkin and extends the method to handle general inputs that are not uniformly sampled through renormalizing the kernel matrix. The method is very similar to the construction of Laplacian eigenmaps including the use of pair wise distances and radial basis functions to produce a sparse matrix whose spectral properties can be evaluated. By combining the eigenfunctions and values, one can construct a new distance matrix that is robust to the sampling distribution of the input parameters. This method scales relatively well to high dimensional[29]; however, since high dimensional data typically lies on a low dimensional space the input is sparse in terms of the original dimension, recent work in the area of compressed sensing[24] may allow this method to be extended to very large data. Thus the signal is sparse in the high dimensional setting and can be reduced prior to being embedded. Restated, the high dimensional vector for each datum is sparse in the high dimensional setting and can be reduced prior to being embedded.

1.2.2. Cluster Analysis

Cluster analysis is used for the classification of data into groups allowing one to discover previously unseen structures or relationships. These relationships or classes allow one to provide taxonomy for phenomena, objects, or suggest statistical models with which to describe populations. This method may produce either hierarchal, with an increasing number of nested classes, or non-hierarchal clustering which is commonly used in techniques such as k-means clustering. Typically, this approach is used to divide data from a single population into a smaller number of groups. The underlying mathematics of most of these methods is relatively simple but large numbers of calculations are needed which can put a heavy demand on the computer.[30] One must determine the pair wise distance between all elements of the given population and then apply amalgamation rules to determine when two clusters are sufficiently similar to be linked together. The classification will depend upon the particular method (distance, amalgamation) used. Thus, multiple classifications are possible and it is up to the one evaluating the results to provide the expert knowledge that defines an 'optimal' classification.

1.2.3. Experimental Probabilistic Hypersurface

Recent work by Beauzamy and his student, Zeydina,[10,31,32,33] provides a mechanism to take a small number of high dimensional data and develop a probability density of possible output values for an input query. For example, one may have 300 realizations or results of a computational model that utilizes 20 inputs. If each parameter is limited to four values, there are 160,000 possible states. The method can utilize the limited amount of information (the 300 results) and produce a probabilistic density for a new set of inputs. Thus, the Experimental Probabilistic Hypersurface (EPH) allows one to store information obtained from any number of measures, in a physical experiment or in a computational code.[33] If one considers the output from given realizations to be exact for the given input, the existing information may be propagated away from known inputs to cover the entire space. If you are close to a place where the experiment has been performed, the density will be more concentrated; if you are far away, the density will be less concentrated, because you know less. This propagation requires the maximization of information entropy. The principle of maximal entropy thus governs the whole construction, which allows a construction with no artificial rules or probability laws.[31]

1.3. Contribution of Work

The present work contributes a number of new elements: first, the work presents a new, non-parametric approach to modeling high dimensional data; second, the work melds new advances in dimension reduction with a probabilistic predictive method; the reduction in dimension of video data is demonstrated; and fourth, applies the method to a number of problems in the field of nuclear engineering.

The dissertation begins with this introduction followed by the methodology section. Section 3 provides examples of the methodology as it is applied to common engineering problems found in the area of thermal-hydraulics. The method is used as a regression tool to predict a response given a new input query. Several figures are presented to demonstrate various aspects of the proposed methodology's accuracy and robustness. Section 4 applies the method to a prediction code used in nuclear engineering. Both predictions to new queries and predictions for the whole data space are made. An interesting aspect is the ability to utilize both continuous and categorical input data. A discussion of the computational time aspects is also presented. Section 5 introduces large dimensional data to demonstrate the power of the methodology. Several problems dealing with regression and classification are presented using very high dimensional data that include video and medical spectrum data. Section 6 applies the methodology to co-current two-phase flow regime classification and prediction. This section presents both microgravity and Earth gravity vertical up flow video data. The final section discusses the results and provides recommendations for further work.

2. METHODOLOGY

Engineers and scientist are increasingly using complex, high dimensional data to make important decisions. Often, the amount of data is much less than the total number of variables thus limiting the use of common statistical approaches. A common example in the nuclear industry would be large scale integrated system blow down tests for which numerous repetitions would be prohibitively expensive. A modern approach is needed to understand the underlying structure of the data and use this information to make predictions for new sets of conditions. Modern machine learning techniques coupled with recent advances in sampling theory and maximum entropy methods provide a new, unique methodology for solving these important problems.

An example of this high dimensional, low sample phenomenon can be shown in the area of facial recognition. Consider a series of facial images where the head is rotated about a fixed axis simulating a person turning their head. Even for a moderate number of pixels, the number of variables/pixels outnumbers the number of images. This issue and the fact that each variable does not change linearly, make it difficult to develop a correlation using classic approaches. The key observation is that although facial images can be regarded as points in a high-dimensional space, they often lie on a manifold (i.e., subspace) of much lower dimensionality, embedded in the high-dimensional image space. The main issue is how to properly define and determine a low-dimensional subspace of face appearance in a high-dimensional image space. Dimensionality reduction techniques using linear transformations have been very popular in determining the intrinsic dimensionality of the manifold as well as extracting its principal directions (i.e., basis vectors). The most prominent method in this category is PCA. PCA determines the basis vectors by finding the directions of maximum variance in the data and it is optimal in the sense that it minimizes the error between the original image and the one reconstructed from its low-dimensional representation. PCA has been very popular in face recognition, especially with the development of the method of "eigenfaces.[1]" Its success has triggered significant research in the area of face recognition and many powerful dimensionality reduction techniques (e.g., Probabilistic PCA, Linear Discriminant Analysis (LDA) Independent Component Analysis (ICA), Local Feature Analysis (LFA), Kernel PCA have been proposed for finding appropriate low-dimensional face representations.

2.1. Dimension Reduction Using Random Projections

In the past, most engineering problems dealt with a small number of N variables. As long as the value of N remains small, these problems can usually be handled with standard statistical techniques. As N increases, the need to reduce the number of variables from N to d becomes desirable to reduce the computational burden of analyzing data. PCA and many of the nonlinear methods work well but are computationally expensive with a computational cost of estimating PCA as $O(N^2M) + O(N^3)$ where M is the number of data points.[34]

Random projections (RP) is an emerging technique for dimensionality reduction where the computational cost is $O(dkN)$ or if the projection matrix c is sparse, $O(ckN)$

where k is a constant.[35] This technique, part of a broader technique referred to as "compressed sensing," allows a high dimensional signal vector to be projected onto a low dimensional space using a random orthonormal basis. Random projections are concerned only with the dimension reduction mapping whereas compressed sensing usually deals with both the dimension reduction and signal reconstruction. This new theory is based on the Johnson-Lindenstrauss lemma[36] that essentially states that a set of M points in a high-dimensional space of size N can be embedded with little distortion into a space of much lower dimension, d. The mapping is performed using a simple matrix of size d x N and preserves the interpoint distances between the points[37]; furthermore, it is data independent and computationally simple in that it can be applied to new data as it comes in rather than the ensemble of signals as in more classic techniques such as PCA. Researchers have demonstrated the use of random projections for capturing information about sparse or compressible signals.[24,38,39]

The Nyquist–Shannon sampling theorem states that a function, f(t), to be represented without error must be sampled at twice its highest frequency. This demand requires acquisition and storage systems to be able to handle twice the bandwidth required for the measurement, limiting the number of instruments used in sensing systems as well as the acquisition rate. Several techniques have been developed to reduce the storage requirements such as file compression but these do not address the requirements for handling the bandwidth of acquisition nor the computational power required to perform file compression. The low dimensional space or random projection of a high dimensional signal contains enough concentrated information to enable signal reconstruction with small or zero error.[40] This is due to the fact that most high dimensional signals are sparse and can be embedded into a space of much lower dimension in such a way that distances between the points are nearly preserved. The sparseness of the information in the original data space leads to dimensionality reduction and efficient modeling and has implications for the data acquisition process itself that lead to efficient data acquisition protocols.[41]

In random projections (RP), the projection is accomplished using a random matrix whose columns have unit length. Other matrices have been proposed and used to attempt to simplify the construction.[42,43,44] The usefulness of RP is that it preserves approximately pairwise distances of points in Euclidean space[45], which is desirable for embedding techniques and clustering.[35]

Starting with a signal vector, one can project Y onto a smaller space of

$$Y_i = (Y_1, Y_2, \ldots, Y_N) \qquad \text{(Eq. 2.1)}$$

dimension d using a specially constructed matrix A. The matrix A consists of dN entries that are independent and identically distributed values with a mean of zero and a standard deviation of one. The rows are essentially orthogonal which produces an

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{d1} & \cdots & A_{dN} \end{bmatrix} \qquad \text{(Eq. 2.2)}$$

incoherent basis with the signal Y. Multiplying A by the signal vector Y produces a 1 x d vector that retains the same interpoint or pairwise distance information as the original. This is shown in Figure 2.1. The N x N pairwise distance matrix calculated using

Equation 2.3 is determined for five vectors of size 1 x 10000 and compared against the corresponding N x N pairwise distance matrix random projection of size 1 x d. The average Root Mean Square Deviation is calculated for twenty different random

$$D_{ij} = \sqrt{\sum_{k=1}^{N} \left| Y_{ij} - Y_{jk} \right|^2} \qquad \text{(Eq. 2.3)}$$

projections using Equation 2.4. The averages for various values of d are shown in Figure 2.1. As the value of d increases, the RMSD decreases. For a value of approximately 100,

$$RMSD(\psi_1 - \psi_2) = \sqrt{\frac{\sum_{i=1}^{n} (\psi_{1i} - \psi_{2i})^2}{n}} \qquad \text{(Eq. 2.4)}$$

where

$$\psi_1 = \begin{bmatrix} \psi_{11} \\ \psi_{12} \\ \vdots \\ \psi_{1n} \end{bmatrix}, \psi_2 = \begin{bmatrix} \psi_{21} \\ \psi_{22} \\ \vdots \\ \psi_{2n} \end{bmatrix} \qquad \text{(Eq. 2.5)}$$

Figure 2.1 indicates that the RMSD is approximately 0.1, thus the distortion of the random projection is quite small for d << N. The predicted normalized RMSD using Equation 2.6 is based on the proof presented by and shown in Equation 2.4, which agrees with the calculated normalized RMSD.[46]

$$d \geq \frac{4\ln(N)}{\frac{\varepsilon^2}{2} - \frac{\varepsilon^2}{3}} \qquad \text{(Eq. 2.6)}$$

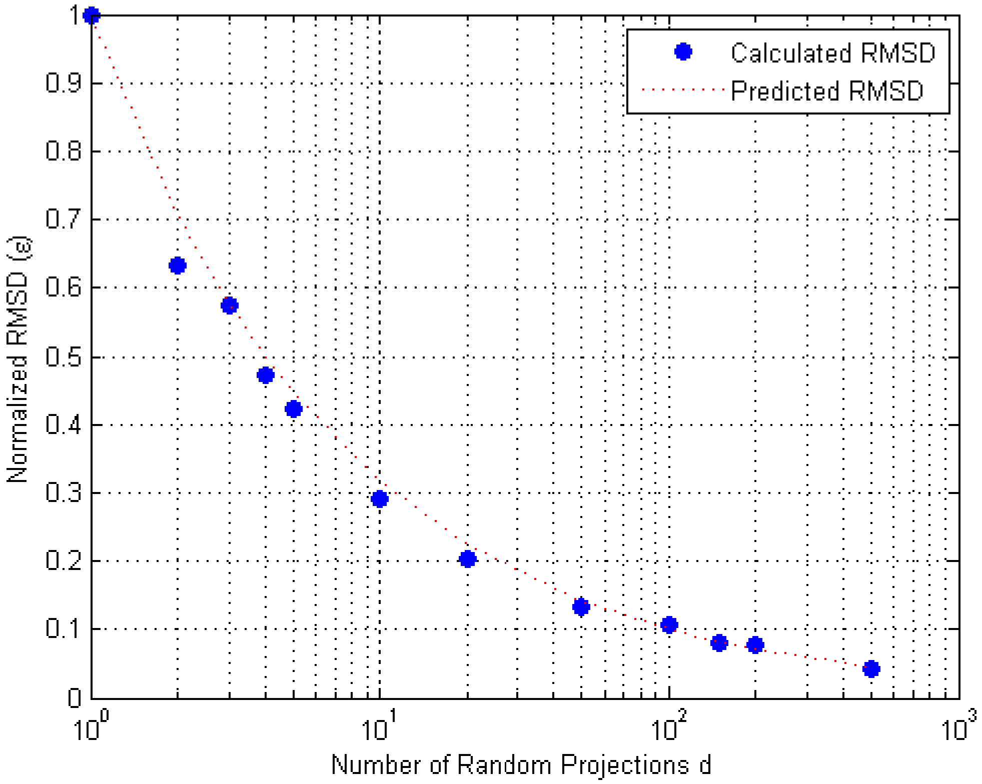


Figure 2.1 Comparison of Calculated RMSD and Predicted RMSD.

The following example illustrates the power of random projections to identify the correct facial pose out of a library of images with only a corrupted query image. The facial images are from the UMIST database[47] shown in Figure 2.2 and consist of grayscale, normalized, rotated images, with two images off axis. The query image also shown in Figure 2.2 is drawn from the database and corrupted. The images are 92 x 112 pixels that correspond to a 1 x 10304 vector or 10304 variables/dimensions. Using the figure above, it is desired to reduce the dimension of the input data with little distortion so a value of 100 is used for the random projection matrix. The projection matrix that is

10304 x 100 in size is used to reduce the size of the feature vector to a 1 x 100 vector.

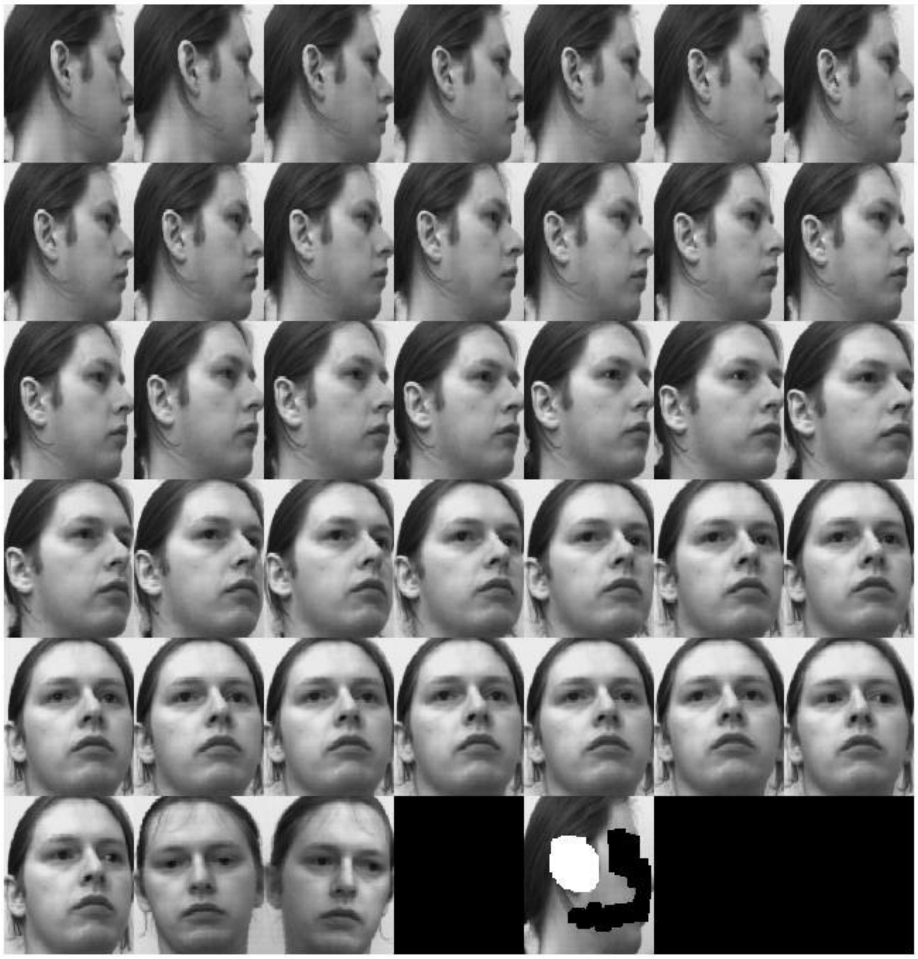


Figure 2.2 Library of Facial Poses and Corrupted Facial Pose.

This small vector is then used for comparison. The resulting identification of the correct image in the Library of Facial poses is performed using the reduced vector with the results shown in Figure 2.3. An interesting aspect is the comparison of the Euclidean distance between the query image and each image in the library for both the feature vector of the full image and the reduced vector. Figure 2.4 shows the query and corresponding image from the where the reduced feature vector accurately identifies the correct target image in the library that is the original image before corruption and is quite similar to the Euclidean distance of the full feature vector. If needed, the corrupt image could be reconstructed using a number of algorithms but is not needed for identification.[48,49]

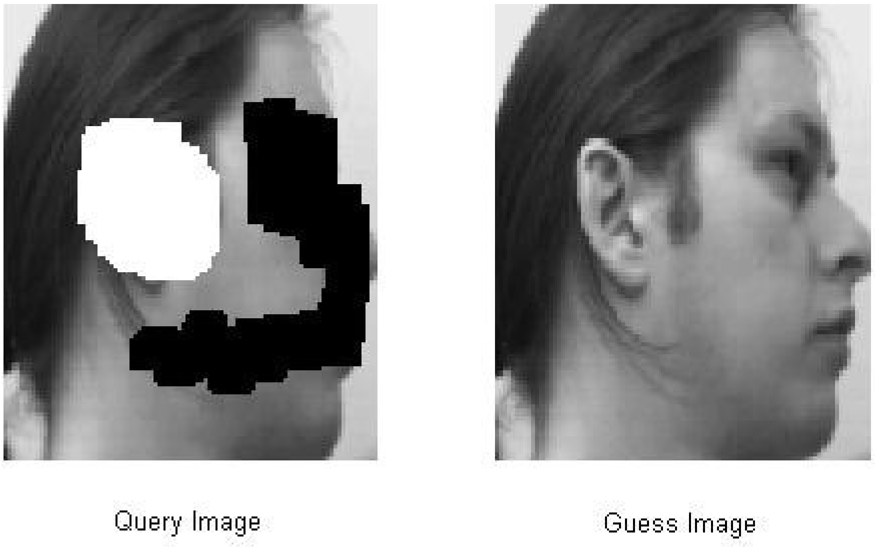


Figure 2.3 Corrupted Query Image and Identified Library Image.

The above example illustrates the use of a random matrix with mean zero and standard deviation of 1. There are several matrices that can be utilized for random projections such as discrete wavelet transforms (DWT), discrete sine and discrete cosine transforms (DST/DCT), and noiselet transforms. A study evaluating DST and DCT transforms for image compression indicated an improvement in image compression using these special matrices when the original image needed to be restored.[42] This is due to the ability to accurately capture important features in the image or data.

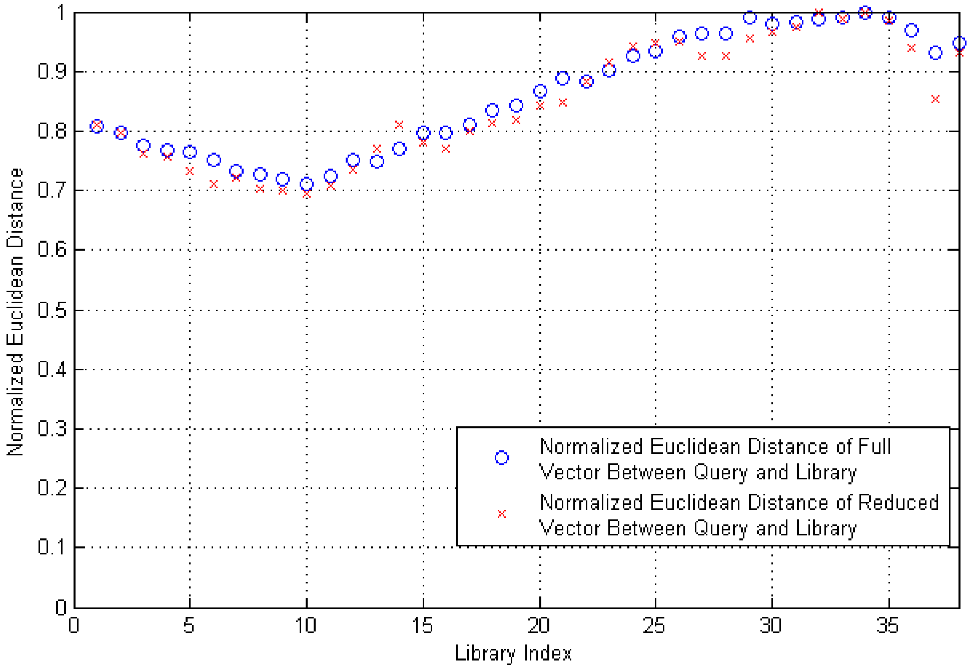


Figure 2.4 Comparison of Euclidean Distance Calculations between Query and Each Image in Library for both Full Image Vector and Reduced Image Vector.

The ability to contain all useful information in a reduced dimensional space is very desirable in the context of sorting, clustering, and analyzing high dimensional data. One can now use smaller, denser representations of high dimensional information resulting in faster, efficient, and potentially more accurate computations. The successful results of this method are applicable to a number of areas of interest to fluid mechanics and heat transfer. Sorting two-phase flow regimes is a prime example of an area where compressed sensing methods can be used to reduce the dimension of the feature space to classify regimes objectively without the need for training data. This approach would provide a simple, robust approach to regime identification by eliminating subjective labeling that is currently performed and can be applied to image, video, or sensor data such as capacitance or conductance probes.

2.2. Embedding Using Diffusion Maps

Another dimension reduction approach is the method of “Diffusion Maps.” The term, coined by Lafon,[27] describes a technique that is based on the spectrum of a normalized dissimilarity matrix. The method is based on constructing a weighted graph and then computing the first few eigenvectors and eigenvalues of the corresponding. The first few eigenvectors present a low dimensional representation of data and/or coordinates for embedding.[6,28] This approach preserves the local geometry and interpoint relationship of the data into a lower dimensional subspace.

For many high dimensional data and problems found in engineering, a low dimensional structure, a hyperplane, of the data in relation to each other is usually found. This allows one to develop models and the simplicity and accuracy of these models is determined by how small the subspace in which the data reside. The Swiss roll shown in Figure 2.5 is often utilized in machine learning due to the difficulty in projecting the points onto a lower dimensional manifold.[5,29] The Swiss roll is made up of discrete points in three dimensional space. Although the data reside in three dimensions, the data are structured and fixed to a two dimensional manifold. One could develop a mathematical model to describe this surface but this becomes exceedingly difficult with higher dimensions and fewer known points.

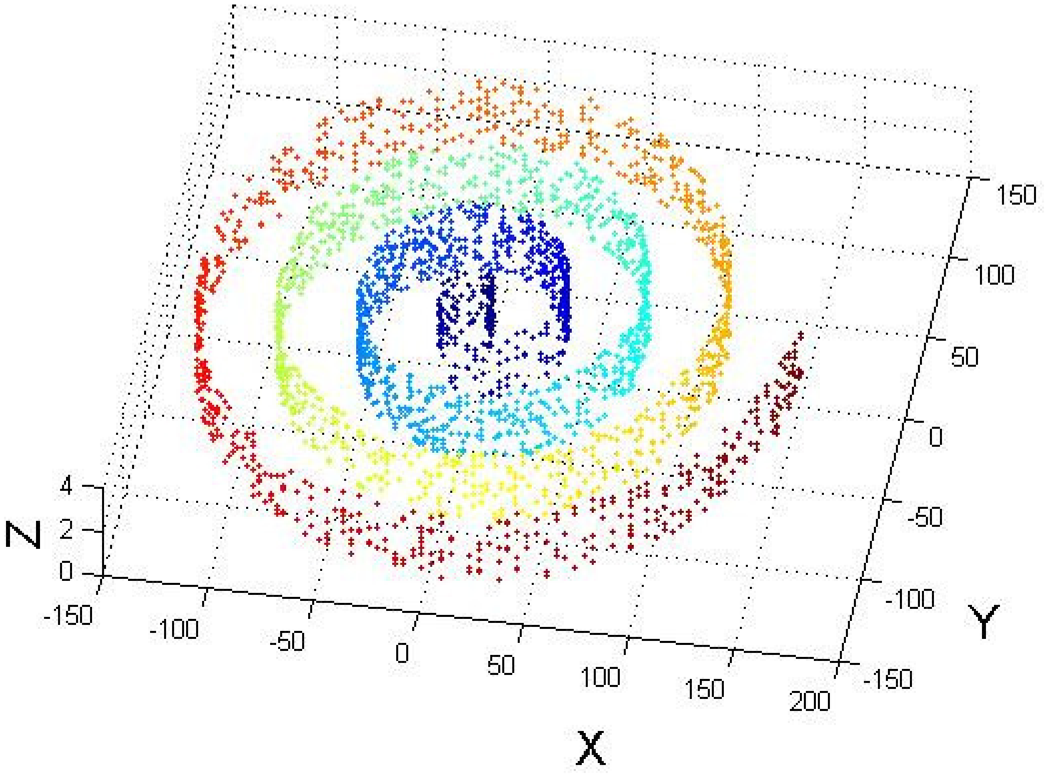


Figure 2.5 Randonly Sampled Swiss Roll

Many manifold learning techniques utilize distance measurements between points and these techniques are typically accomplished by calculation of the Euclidean distance. Unfortunately, complex structures commonly found in high dimension data introduce errors due to short circuits or paths between surfaces. An example is shown in Figure 2.6. The line segments $\overline{AB}$ and $\overline{AC}$represent the distance measurements between three points. The line segment $\overline{AB}$corresponds to points along the surface while segment

$\overline{AC}$ is for a short circuit. Points A and C are much further apart if one moves along the manifold which is shown by the dotted line segment.

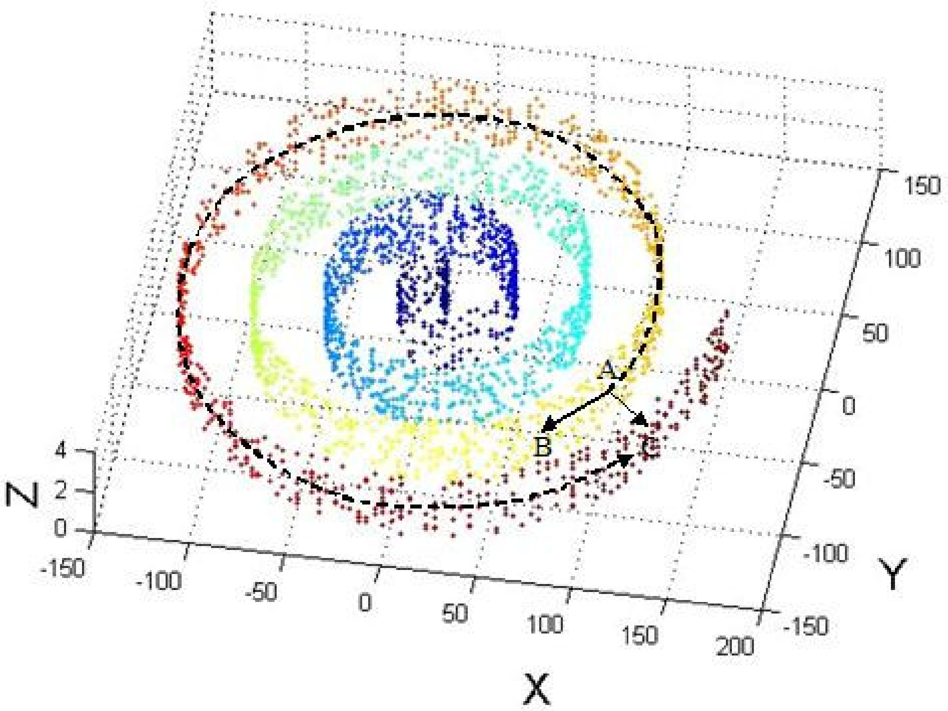


Figure 2.6 Graphical Representation of Euclidean and Geodesic Distance.

By using a radial basis function, a Gaussian kernel, the local distance information is reshaped with points far away from each other are neglected and a new weighted graph constructed. The spectrum of this matrix is calculated and results in a set of eigenvalues starting at one and monotonically decreasing. As in PCA, the first few eigenvectors are usually sufficient in describing the important features of the data. Following Lafon's method, the first eigenvector provides sufficient information to embed the swiss roll information and describe the three dimensional surface/manifold as a one dimensional structure shown in Figure 2.7. The coordinates of Figure 2.7 are the first two eigenvectors with the data from Figure 2.6 mapped into the new coordinate system. Since the swiss roll can be described as a set parametric of parametric equations with one variable, the resulting embedding being described with one eigenvector is appropriate. The line shown in Figure 2.7 closely preserves the local structure of the manifold and although the coordinate system has changed, the new space can still be used for prediction.

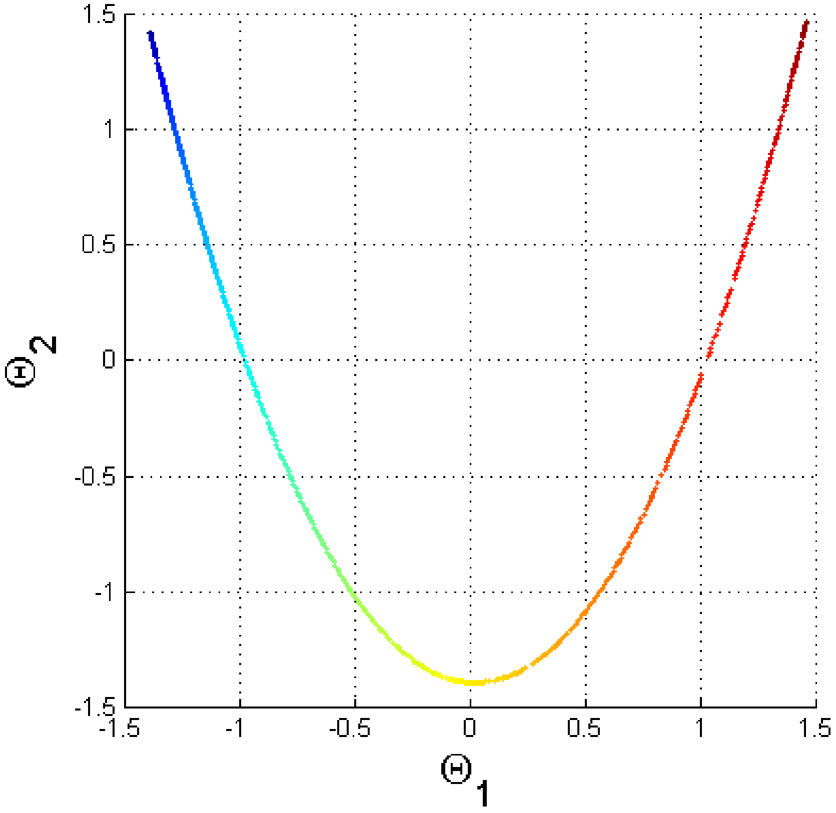


Figure 2.7 Swiss Roll Data Plotted in First Two Diffusion Coordinates.

An analog can be found in the area of Riemannian geometry where local parameters can be utilized to describe global quantities, and in the following, it will be useful to think of the data as forming a weighted oriented graph.[27] The mathematical development of these methods follows the work on functions operating on manifolds. One operator, the Laplace-Beltrami operator, is defined as the divergence or gradient of the manifold. Belkin and Niyogi[23,50] showed that the first few eigenvectors of the distance matrix are discrete approximations of the eigenfunctions of the Laplace-Beltrami operator on the manifold when data is uniformly sampled from a low dimensional manifold.

Staring with datum Ψ described by N dimensional points, the Euclidean distance is calculated for n pairwise set of vectors. A Gaussian kernel shown in Equation 2.7 is

$$k\left(\psi_i,\psi_j\right)=e^{\left(-\frac{D_{ij}}{\sigma}\right)^2} \qquad \text{(Eq. 2.7)}$$

then applied to the resulting n x n distance matrix resulting in a sparse similarity matrix. This matrix has a diagonal of ones corresponding to each datum compared to itself. Off diagonal values of the matrix correspond to a weight or distance of how “similar” two data are with each other. The similarity matrix is then normalized as shown in Equations 2.8 and 2.9:

$$v^2\left(\psi_i\right)=\int k\left(\psi_i,\psi_j\right)d\mu\left(\psi_j\right) \qquad \text{(Eq. 2.8)}$$

$$\tilde{a}\left(\psi_i,\psi_j\right)=\frac{k\left(\psi_i,\psi_j\right)}{v^2\left(\psi_i\right)} \qquad \text{(Eq. 2.9)}$$

Since we are interested in the spectral properties, eigenfunctions, of the operator it is preferable to work with the symmetric conjugate or redefining Equation 2.9 we get Equation 2.10.[51] The diffusion operator can then be defined as shown in Equation 2.11 and can be shown to be compact and self-adjoint.

$$a\left(\psi_i,\psi_j\right)=\frac{k\left(\psi_i,\psi_j\right)}{v\left(\psi_i\right)v\left(\psi_j\right)} \qquad \text{(Eq. 2.10)}$$

$$Af\left(\psi_i\right)=\int a\left(\psi_i,\psi_j\right)f\left(\psi_j\right)d\mu\left(\psi_j\right) \qquad \text{(Eq. 2.11)}$$

Thus, we can now show through Equations 2.12 through 2.14 the resulting mapping:

$$a\left(\psi_i,\psi_j\right)=\sum_{j\geq 0}\lambda_j\varphi_j\left(\psi_i\right)\varphi_j\left(\psi_j\right) \qquad \text{(Eq. 2.12)}$$

$$A\theta_j\left(\psi_i\right)=\lambda_j\theta_j\left(\psi_i\right) \qquad \text{(Eq. 2.13)}$$

$$\Phi=\left(\theta_1,\theta_2,\cdots,\theta_p\right) \qquad \text{(Eq. 2.14)}$$

The mapping shown in Equation 2.14 consists of the eigenfunctions of diffusion operator A. In his dissertation, Lafon[27] utilizes a singular value decomposition to diagonalize the kernel matrix, k. The embedding coordinates or eigenvectors are then determined from normalizing the left singular vectors which provides an orthonormal basis of k. Since each eigenfunction can be interpreted as a coordinate on the set, this mapping can be used as a diffusion metric to measure the diffusion distance between the data points which is shown in Equation 2.15.

$$\zeta_p^2\left(\psi_i,\psi_j\right)=\sum_{j\geq 0}\lambda_j^p\left(\theta_j\left(\psi_i\right)-\theta_j\left(\psi_j\right)\right)^2 \qquad \text{(Eq. 2.15)}$$

It can be shown that sufficient information is carried along in the first few eigenvectors and only these are needed for dimension reduction. Thus, for embedding high dimensional data, only the first few eigenvectors are used. Although diffusion maps are being used for a number of machine learning tasks[52,53,54] it has not found its way to modeling engineering systems. This may be due to the fact that adding new data becomes problematic since the entire embedding needs to be recomputed or engineers have usually attempted to simplify the problems in order to avoid dimensionality issues. In order to develop a methodology that utilizes these techniques, one must find a way to make use of new information or be able to locate the placement of new data.

2.3. Prediction Using the Experimental Probabilistic Hypersurface

The principle was first expounded by E.T. Jaynes when he introduced what is now known as Maximum entropy thermodynamics and suggested that thermodynamics, and in particular thermodynamic entropy, should be seen just as a particular application of a general tool of inference and information theory.[55] Essentially a Bayesian approach, the principle of maximum entropy is used to determine a unique probability distribution that makes explicit use of prior information. Thus, when the outcome is known, information entropy is zero whereas at points located far from known information, the entropy is large.

There have been several attempts to extend spectral methods for new data[27] but these are mathematical derivations that do not utilize the history of existing data and thus are biased. EPH provides a method to predict the outcome for new inputs with no bias and also provides information in how close a new query is to existing data. In fact, many of the methods mentioned previously, are compromised with new realizations or input meaning that the process or embedding must be recalculated when presented with new data. Merging dimension reduction methods with EPH provides a powerful tool that would allow one to analyze and classify data as new inputs become available. Further, EPH would allow one to quickly forecast the result prior to performing the test which is extremely useful when the cost of performing new tests becomes expensive. Comparing the forecast and new data would provide insight into how well sampled the local embedding is.

The Experimental Probabilistic Hypersurface is based on the theory of maximal entropy which is best illustrated with the example below in Figure 2.8. The example is based on a single parameter or input variable with a corresponding output. The abscissa is the range of possible inputs where two known or measured data labeled $\xi(1)$ and $\xi(2)$ are shown. The ordinate is the 'knowledge' or entropy of the system based on the information provided. Thus, at the measured points, we know the value explicitly and the entropy is zero which corresponds to a probability density in the form of a dirac; however, as we move away from our measurements, our 'knowledge' decreases until we reach maximum entropy (no information, uniform probability). This allows us to construct a density or probability distribution using the knowledge from measured data. If the queried input has already been tested and a measurement exists, the result is a certainty, and the density is a Dirac measure; otherwise, it is a true density, and this density is less and less concentrated.[9] Ultimately, the density approaches a uniform distribution far from any measured points. If a new query, x', was set between ξ(1) and ξ(2) the entropy is bounded by the information propagated by ξ(1) and ξ(2).

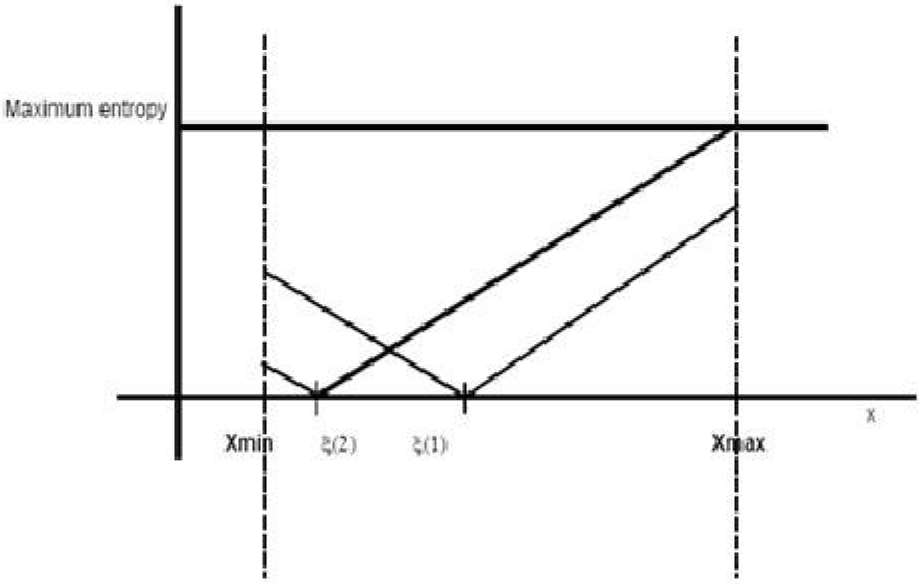


Figure 2.8 Propagation of Information from Measured Points Throughout State Space.[9]

Figure 2.9 shows the input parameters and corresponding probability density/function of the possible outputs. Providing information regarding the response of the function to the known inputs and by placing bounds on the input and output states, which is commonly done in engineering modeling, EPH allows the construction of the probability distribution for each set of inputs. For Figure 2.9, there are two known values for the given input coordinates. Thus, it is expected that the probability density will contain two peaks corresponding to the output states of ξ(1) and ξ(2) and the most likely value for x' is closer to the output of ξ(1) due to the proximity of x' to ξ(1).

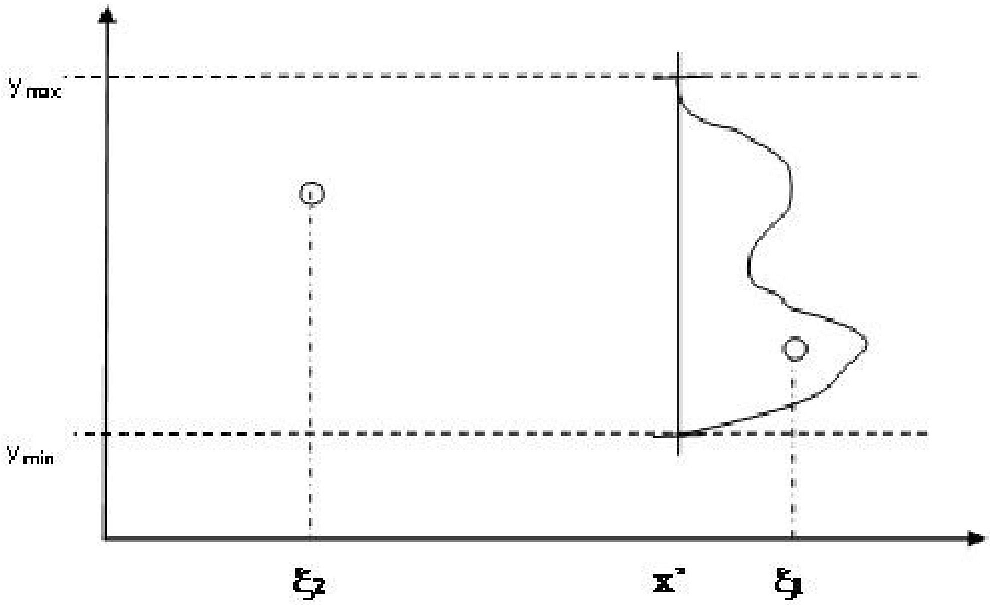


Figure 2.9 Illustration of Probability Construction for New Query X Given ε_1 and ε_2.

In Figure 2.10, the query x = 1 has a peak or near Dirac distribution due to its proximity (certainty) to a known input. Conversely, Figure 2.11 has a more open distribution due to the known points lying further away. One can observe that the known points closer to the query value exhibit more influence on the probability producing a higher probability at these locations versus the known values further away. If no points were close by, the distribution would be uniform over all possible outputs. The key point in the construction is the propagation of information from a measure point to any other point.[33]

As more points are known, the probability becomes spread amongst the known values; however, the combinations of each of these points sharpen the predictive capability focusing the peak probability toward the correct value for the query. This is shown in Figures 2.11, 2,12, and 2.13. This is due to the weighting that is proportional to the distance from the query that is applied to each known point. Thus, as one collects more data, making predictions become more accurate. Figures 2.10 through 2.12 are probabilities constructed using more known data or information about the underlying function. Although more information is quite useful, information closer to the point of query produces a more significant result as shown in Figure 2.13. The more peaked probability density is due to more information being available near the point of query.

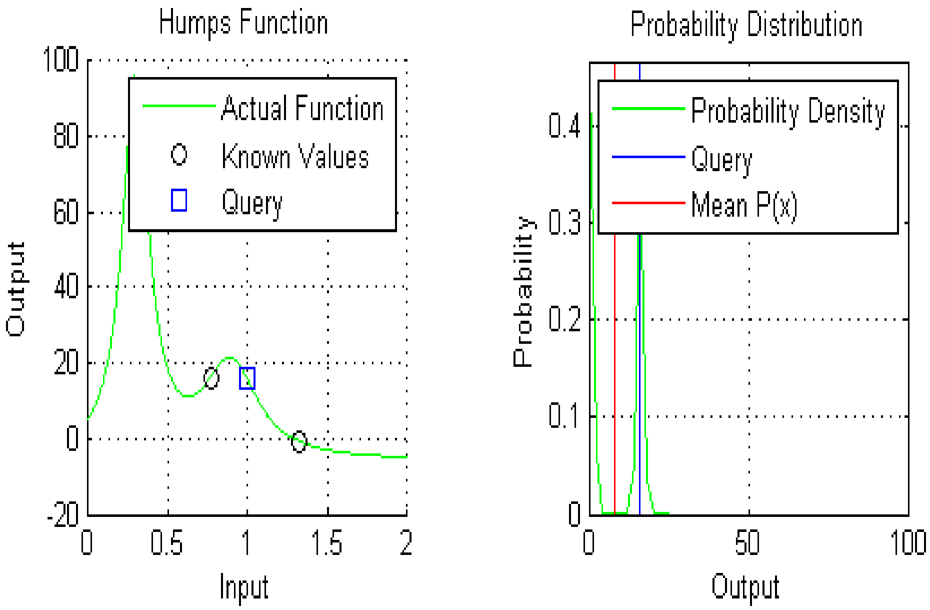


Figure 2.10 Construction of the Experimental Probabilistic Hypersurface for a Single Variable Function with Two Known Data Points. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

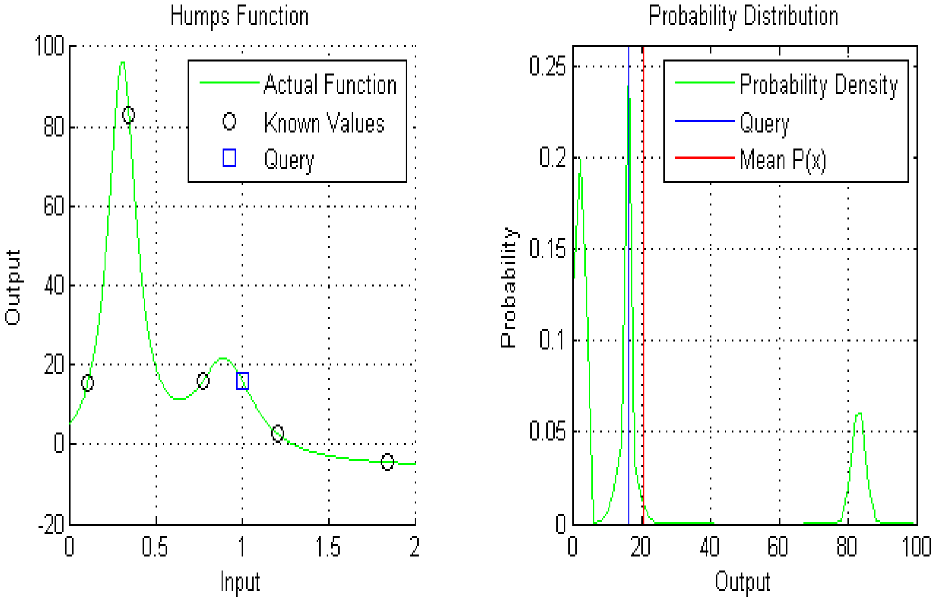


Figure 2.11 Construction of the Experimental Probabilistic Hypersurface for a Single Variable Function with Five Known Data Points. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

This is quite intuitive for the single variable functions given here but will be less so for the high dimensional functions that are presented later. Constructing the hypersurface is straightforward and will be explained in the following section. One may consult the references by Beauzamy[9] and Zeydina[10] for a detailed explanation.

If we take the above example and assume the output can be characterized as a single input parameter/dimensional function as shown in Equation 2.16.

$$\begin{aligned} output &= f\left(input\right) \\ \Gamma &= f\left(x\right) \end{aligned} \qquad \text{(Eq. 2.16)}$$

One may consider θ to represent any output or experimental result that is a function of a given set of inputs, x. Outlining the correct input parameters, the range of possible inputs, and the range of acceptable outputs is required and is considered the expert knowledge provided by the user. This information is required for one to produce a uniform law for possible outputs, values of θ. Hence, with no information provided, one assumes that the answer lies between some θ_{min} and θ_{max} with an equal probability.

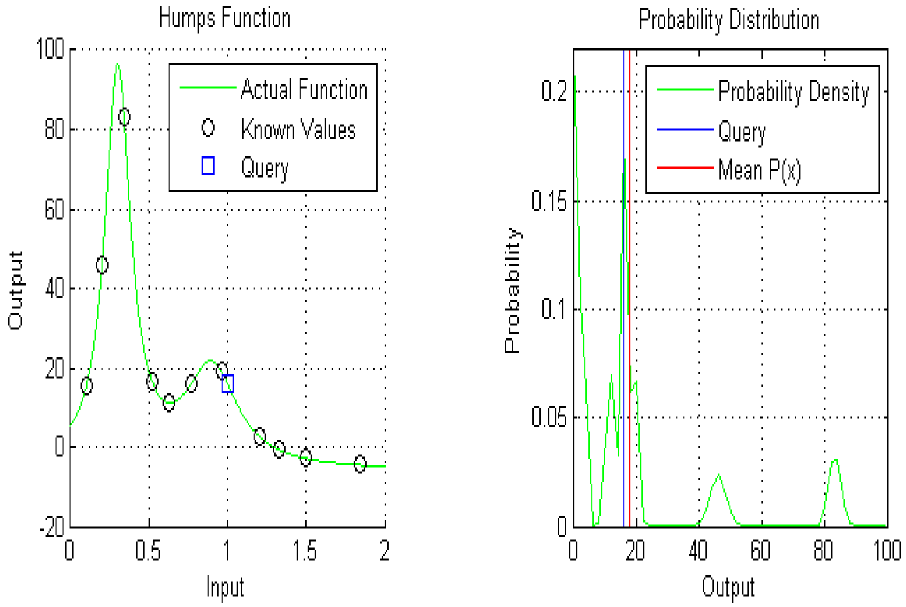


Figure 2.12 Construction of the Experimental Probabilistic Hypersurface for a Single Variable Function with Eleven Known Data Points. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

Past results such as experimental measurements or code output provide information or knowledge of the result at the specific location. This information is then conveyed throughout the output space producing a new probability density. As mentioned previously, the probability density at this specific information is a Dirac. Moving away from this point the information or influence of known values becomes less and less. This information degradation is termed entropy. Entropy or information entropy is the measure of uncertainty associated with a variable and has connections to thermodynamic entropy. Jaynes described thermodynamic entropy as being an estimate of information needed to define the detailed microscopic state of the system.[55]

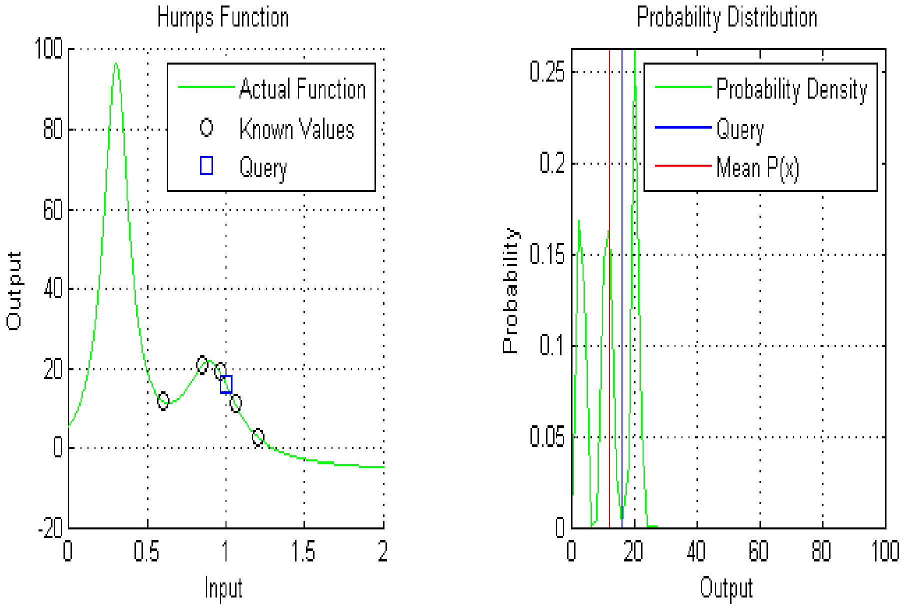


Figure 2.13 Construction of the Experimental Probabilistic Hypersurface for a Single Variable Function with Five Concentrated Known Data Points. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

The Minimal Information Lemma shows that entropy must increase linearly with the distance.[9] Beauzamy utilizes the discrete entropy as defined in Equation 2.17. Equation 2.17 defines entropy as a function of probability. The axioms of probability

$$I = \sum_j p_j \log\left(\frac{1}{p_j}\right) \qquad \text{(Eq. 2.17)}$$

apply such as $\sum p_j = 1$ and $p_j > 0$. The minimum value for entropy, *I*, is zero when the probability distribution is a Dirac (the p_j's are all zero except one). Determining the maximum value is more involved but can be directly computed realizing that the probability distribution is uniform (all p_j's are equal). Thus, we start with Equation 2.18, the maximum entropy, characterized by a uniform probability distribution over the range of possible outcomes, γ.

$$I = \log(\gamma) \qquad \text{(Eq. 2.18)}$$

The output space is divided to provide discrete coordinates that partition the range of possible outcomes. Each step is easily defined as the difference between the maximum and minimum values the result can take on divided by the range of possible coordinates.

$$\tau = \frac{\Gamma_{\max} - \Gamma_{\min}}{\gamma} \qquad \text{(Eq. 2.19)}$$

$$I(x) = \vartheta\delta(x, \varepsilon_1) \qquad \text{(Eq. 2.20)}$$

For a known point, the distribution is perturbed and the entropy is given by Equation 2.20. The entropy is proportional to the distance between x and ε_1. The proportionality constant, ϑ, is dependent upon the range, γ, and the distance between the query, ε_1, and the furthest corner of the hypersurface. Thus, Equation 2.20 can be restated to show that the value of the proportionality constant is the following:

$$\vartheta(\gamma) = \frac{I_{\max}}{\delta_{\max}} \qquad \text{(Eq. 2.21)}$$

where:

$$\delta_{\max} = |x - \varepsilon_1|$$

Now, one sets Equation 2.17 equal to Equation 2.20 using the relationship shown in Equation 2.22.

$$\sum_j p_j \log\left(\frac{1}{p_j}\right) = \vartheta\delta(x, \varepsilon_1) \qquad \text{(Eq. 2.22)}$$

Beauzamy[9,31] shows that entropy and the associated probability at a given point is a Gaussian of the form:

$$p_j(x) = e^{-cg_j^2 + b} \qquad \text{(Eq. 2.23)}$$

Equation 2.23, along with the axiom of probability indicating that the sum of discrete probabilities $\sum_j p_j = 1$ adds to unity, allows one to solve Equation 2.17 to give:

$$\log\left(\frac{1}{p_j}\right) = cg_j^2 - b \qquad \text{(Eq. 2.24)}$$

Substituting Equation 2.24 into Equation 2.22 we get Equation 2.25.

$$c\sum_t g_j^2 p_j - b = \vartheta(\gamma)\delta \qquad \text{(Eq. 2.25)}$$

Beauzamy shows that the probabilities, p_j, are symmetric and given by Equation 2.26[9].

$$q^2 = \sum_t g_j^2 p_j \qquad \text{(Eq. 2.26)}$$

Now, from Equation 2.25 and 2.26, we can derive Equation 2.26.

$$cq^2 + b = \vartheta(\gamma)\delta \qquad \text{(Eq. 2.27)}$$

Since the probabilities sum to one,we can solve for *c* and *b* using Equation 2.23.

$$e^b \sum_j e^{-cg_j^2} = 1 \qquad \text{(Eq. 2.28)}$$

So Equation 2.25 becomes:

$$c\frac{\sum_j g_j^2 e^{-cg_j^2}}{\sum_j e^{-cg_j^2}} - \ln\left(\frac{1}{\sum_j e^{-cg_j^2}}\right) = \vartheta(\gamma)\delta \qquad \text{(Eq. 2.29)}$$

A simplifying assumption can be made, with τ defined as the discritization of the output space, the summations can be replaced with the following integrations:

$$\sum_j \Gamma_j^2 e^{-cg_j^2} \approx \frac{1}{\tau}\int_{-\infty}^{\infty} \sum_j e^{-cg_j^2} d\Gamma = \frac{1}{\tau}\sqrt{\frac{\pi}{c}} \qquad \text{(Eq. 2.30)}$$

And

$$\sum_j \Gamma_j^2 e^{-cg_j^2} \approx \frac{1}{\tau}\int_{-\infty}^{\infty} \sum_j e^{-cg_j^2} d\Gamma = \frac{\sqrt{\pi}}{2\tau c^{\frac{3}{2}}} \qquad \text{(Eq. 2.31)}$$

where Γ is the output of the function we are building the probability on. Substituting into Equation 2.28 and solving for *c* and *b*.

$$c = \frac{\pi}{\tau^2} e^{(1-2\vartheta(\gamma)\delta)} \qquad \text{(Eq. 2.32)}$$

$$b = \ln\left(\tau\sqrt{\frac{c}{\pi}}\right) \qquad \text{(Eq. 2.33)}$$

$$b = \frac{1}{2} - \vartheta(\gamma)\delta \qquad \text{(Eq. 2.34)}$$

Solving for the probability, p, can be carried out by substituting Equations 2.32 and 2.34 into Equation 2.23. Equation 2.35 is the general form for the probability based on a single input parameter.

$$p_j(x) = e^{\left(-\frac{\pi\left(g_j - \Gamma_1\right)^2}{\tau^2} e^{\left(1 - 2\vartheta(\gamma)|x - \varepsilon_1|\right)} + \frac{1}{2} - \vartheta(\gamma)|x - \varepsilon_1|\right)} \qquad \text{(Eq. 2.35)}$$

Equation 2.35 provides the probability, *p*, over the range, *j*, of possible outputs, *g*. Equation 2.35 assumes one known measure or data point is provided by the input, ε, and corresponding output, Γ.

The preceding relationship can be expanded to account for more known measurements and for multiple parameters as shown in the development of Equations 2.36 through 2.41.

$$p_{ij}(x) = e^{\left(-\frac{\pi\left(g_j - \Gamma_i\right)^2}{\tau^2} e^{\left(1 - 2\vartheta_i(\gamma)\delta_i\right)} + \frac{1}{2} - \vartheta_i(\gamma)\delta_i\right)} \qquad \text{(Eq. 2.36)}$$

Now we must sum over *m* measurements, *i*, to get the probability at point *j*. This is

$$p_j(x) = \frac{1}{\sum_{i=1}^{m} \frac{1}{\delta_i}} \left(\frac{1}{\delta_1} p_{1,j} + \frac{1}{\delta_2} p_{2,j} + \ldots + \frac{1}{\delta_m} p_{m,j} \right) \qquad \text{(Eq. 2.37)}$$

extended to k measurements in Equation 2.38 and 2.39.

$$p_{ij}(x) = e^{\left(-\frac{\pi\left(g_j - \Gamma_i\right)^2}{\tau^2} e^{\left(1 - 2\vartheta_i(\gamma)\delta_i\right)} + \frac{1}{2} - \vartheta_i(\gamma)\delta_i\right)} \qquad \text{(Eq. 2.38)}$$

$$p_j\left(x\right)=\frac{1}{\sum_{i=1}^{m}\frac{1}{\delta_i^k}}\left(\frac{1}{\delta_1^k}p_{1,j}+\frac{1}{\delta_2^k}p_{2,j}+\ldots+\frac{1}{\delta_m^k}p_{m,j}\right) \quad \text{(Eq. 2.39)}$$

where the normalization coefficient, γ, is given in Equation 2.40. We can recast Equation 2.39 in a simpler form

$$\Xi_m\left(x\right)=\frac{\delta_m^k}{\sum_{i=1}^{m}\frac{1}{\delta_i^k}} \quad \text{(Eq. 2.40)}$$

Shown below in Equation 2.41 by substituting Equation 2.40.

$$p_j\left(x\right)=\sum_m \Xi_m\left(x\right)p_{m,j}\left(x\right) \quad \text{(Eq. 2.41)}$$

The EPH generates probabilities based on the available information and weights the information by the proximity of the known data to the query point. A simple example is shown below. Consider the simple, single parameter/variable function shown by Equation 2.42 and graphically in Figure 2.14. The function ranges between -7 and 4

$$y\left(x\right)=4\cos\left(\frac{\pi x}{3}\right)e^{-\sin\left(\frac{3\pi x}{4}\right)} \quad \text{(Eq. 2.42)}$$

over the range of possible inputs from 0 to 12. The periodic function is shown in green with the known points shown as black circles. The known points are all the information we have regarding the function. Thus, EPH returns a probability based on only these three points of information and their location in regard to the point of our query shown in blue. The query point, x = 6, is equidistant from two of the data and the EPH produces a higher probability at the realizations of these closer known points compared to the further third point. Now, consider Figure 2.15. There are still three known points and

EPH produces a probability density with three peaks corresponding to the three known inputs and the magnitude of the probability is based on the known data distance to the query point. As one learns more about the function, the EPH produces a more accurate density function.

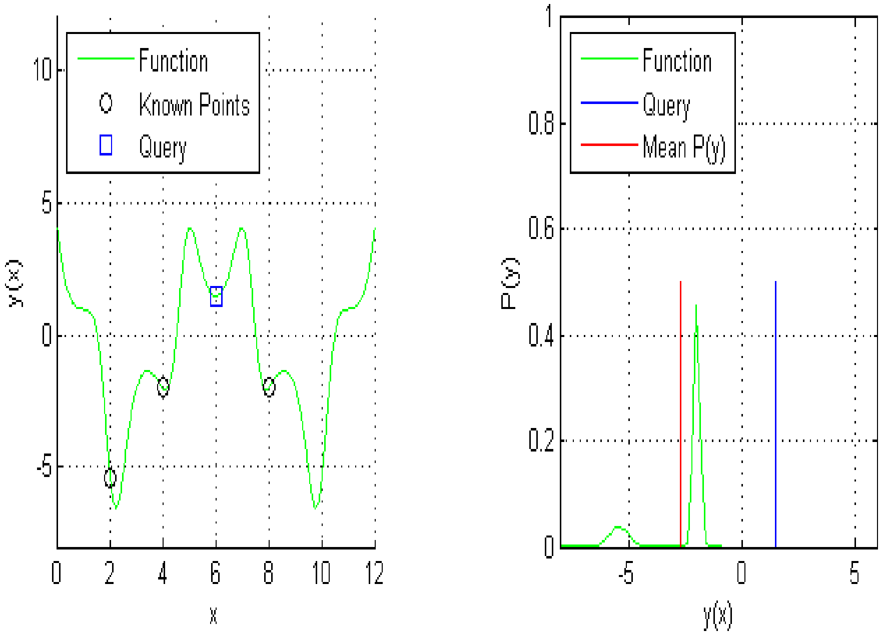


Figure 2.14 Construction of the Experimental Probabilistic Hypersurface for a Single Variable Humps Function with Three Known Data Points of Near Equal Value. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

Figure 2.16 illustrates this for eleven known points. The large probability is located near the actual value of Equation 2.41 at the query point of 6. However, the accuracy is driven by the locality of the known data.

Figures 2.14 through 2.16 shows that EPH is similar to a form of tessellation where the distance between the query and known data is of utmost importance. Expanding this to higher dimensions, it becomes clear that standard Euclidean distances could introduce errors and the manifold learning techniques shown earlier could provide more physically correct inputs to the EPH.

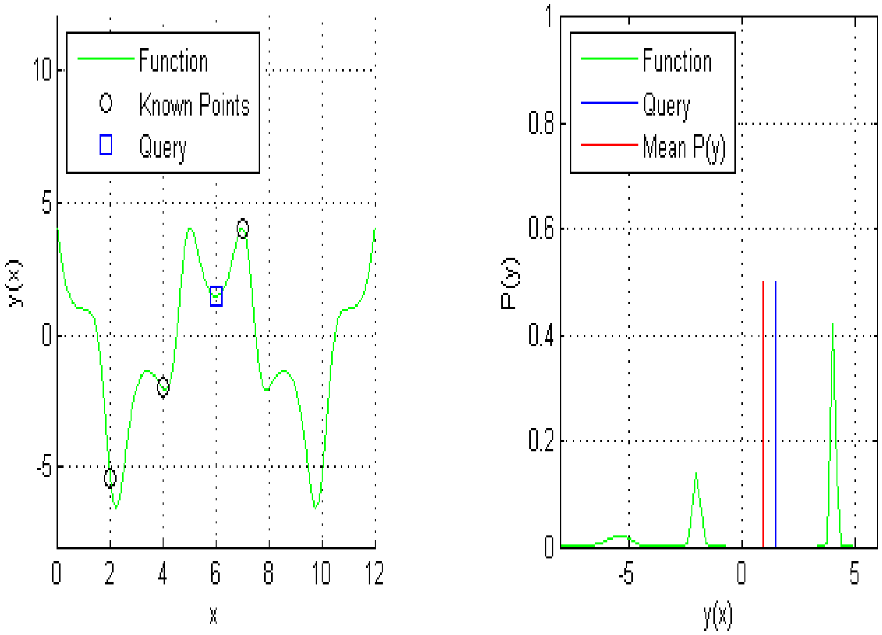


Figure 2.15 Construction of the Experimental Probabilistic Hypersurface for a Single Variable Humps Function with Three Known Data Points . a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

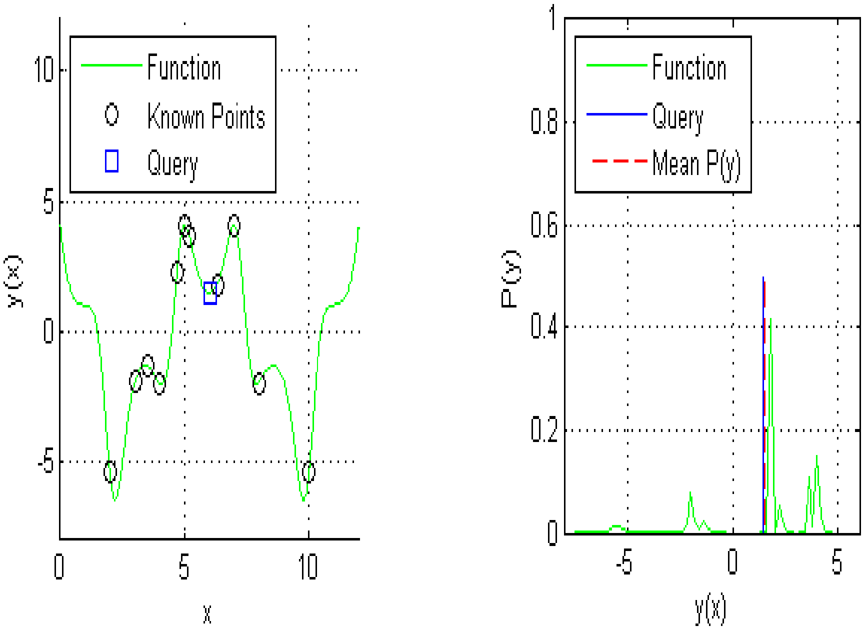


Figure 2.16 Construction of the Experimental Probabilistic Hypersurface for a Single Variable Hump Function with Eleven Known Data Points. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

2.4. Diffusion Maps and EPH

An example of the merging of diffusion space and EPH is shown below. Figure 2.17 contains three plots. The top plot shows the known data points in blue circles, the query point is indicated as a black square, and the actual function that is unknown is shown as a green line. The X and Y axis are considered the input to the function and the value of Z is the output as shown by the parametric equations in Equation 2.43.

$$\begin{aligned} X &= S\cos(S) \\ Y &= S\sin(S) \\ Z &= 0.2S \\ Z &= f(X,Y) \\ S&\Big|_0^{8\pi} \end{aligned} \qquad \text{(Eq. 2.43)}$$

The center plot is the Euclidean distance between the query point and the known data points. The smallest distance corresponds to the point closest to the query; however, all points contribute to the calculation of the probability density which is shown in the bottom plot of Figure 2.17. The mean of the probability distribution is shown for comparison with the expected actual value. This contribution places peaks in the probability density at approximately 1, 2.3 and 3.5. In fact, the most probable result is 1 and the mean of the distribution is 2 rather than the actual value of 2.2.

Figure 2.18 also contains three plots. The first replaces the values of X and Y with the first two embedding coordinates from the diffusion maps. In calculating the coordinates, the diffusion distance is calculated prior to having a Gaussian kernel applied. With the scaling factor of the kernel (σ in Equation 2.6) set to 0.1, distance values shown in Figure 2.18 greater than 0.5 are essentially infinite. This greatly affects the weighting during construction of the Experimental Probabilistic Hypersurface. A greater weighting is applied to the close points resulting in a more peaked probability distribution shown in the bottom plot of Figure 2.18. The mean value of the density

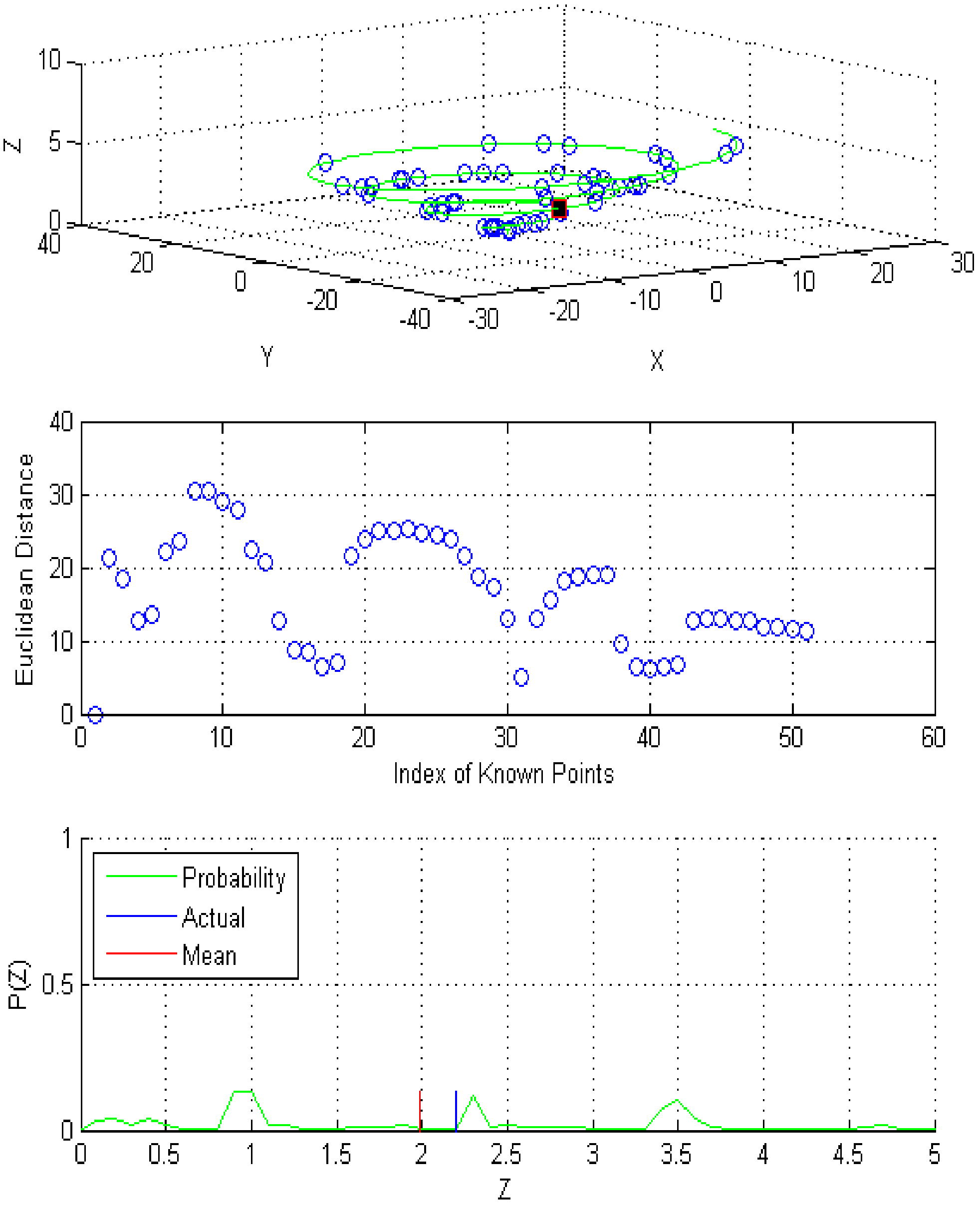


Figure 2.17 a) Helix Function (green), Known Data (blue), Query (black) b) Plot of Euclidean Distance Between Query and Index of Known Data c) Probability Density (green), Mean Probability Density (red), and Query Value (blue).

corresponds to a value of Z of 2.3. The helix example is somewhat trivial due to the fact that it only has two input dimensions. The following sections illustrate the combined methodology for problems with a larger number of prediction variables.

When dealing with data, noise is an issue that must be considered if a modeling approach is to be considered robust. An illustration can be found in the prediction of the manifold shown in Figure 2.19. The “peaks” surface is commonly used in MATLAB and is used in this example to illustrate the effects of noise on the predictive capability. Let one assume the underlying function or manifold is the grey surface. The known information regarding the surface is the 100 points, shown in blue, sampled randomly over the given domain. Thus, the known information consists of two variables and a corresponding output. The query point is shown in black and corresponds to the two input values shown in Figure 2.19. The output of the function of the query inputs is a value of approximately 15. The associated mean of the probability density indicates that the most likely value at this location is 13. One could expect better results with the proper design of experiments or expert knowledge that would allow a better sampling of the function near the query. For this smooth function, the number of random data required to accurately predict the output to a given accuracy is plotted in Figure 2.20. It is clear that not many points are needed to accurately predict the manifold. Only 100 points are needed to provide a prediction within 1%.

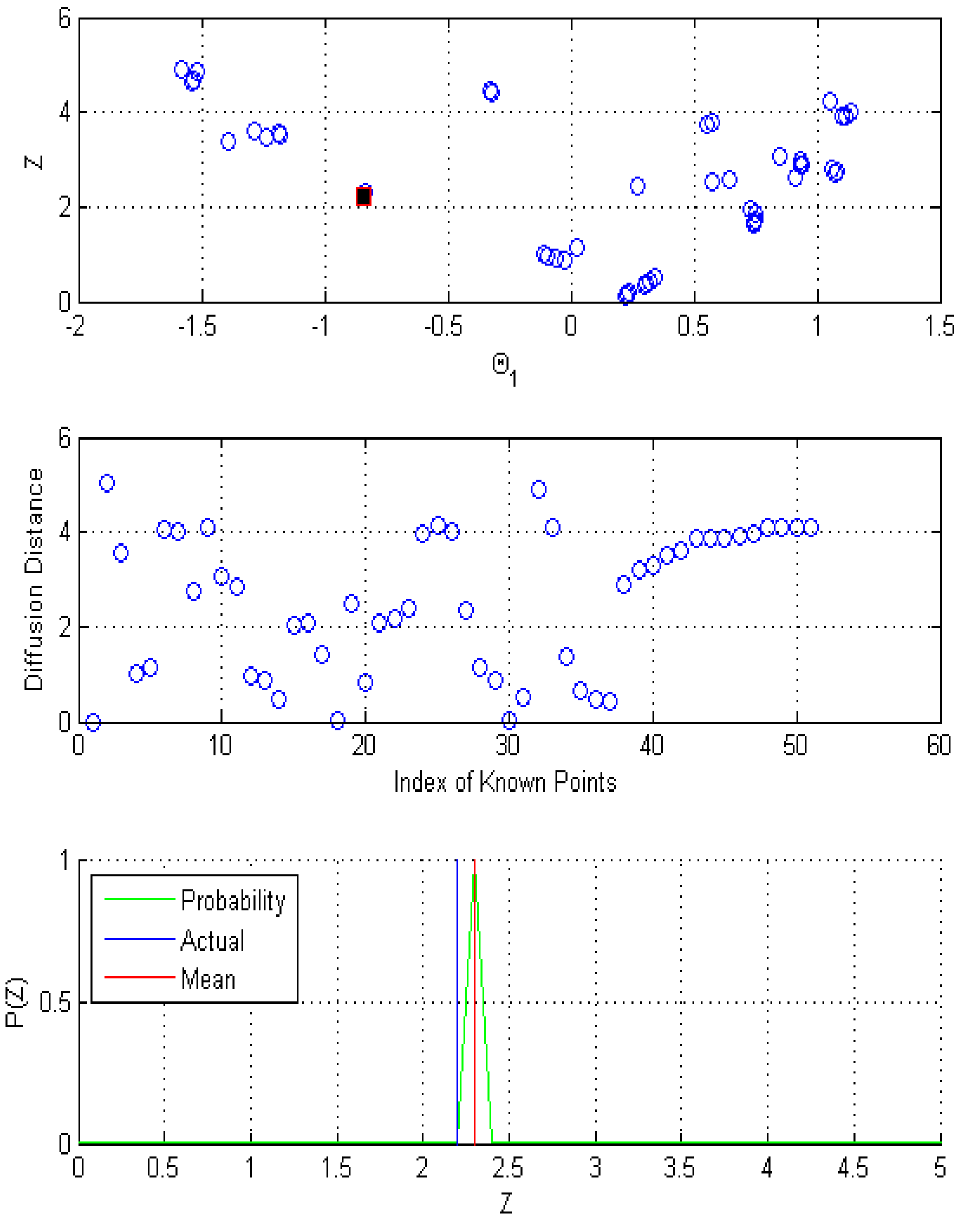


Figure 2.18 a) Helix Function (green), Known Data (blue), Query (black) b) Plot of Diffusion Distance Between Query and Index of Known Data c) Probability Density (green), Mean Probability Density (red), and Query Value (blue).

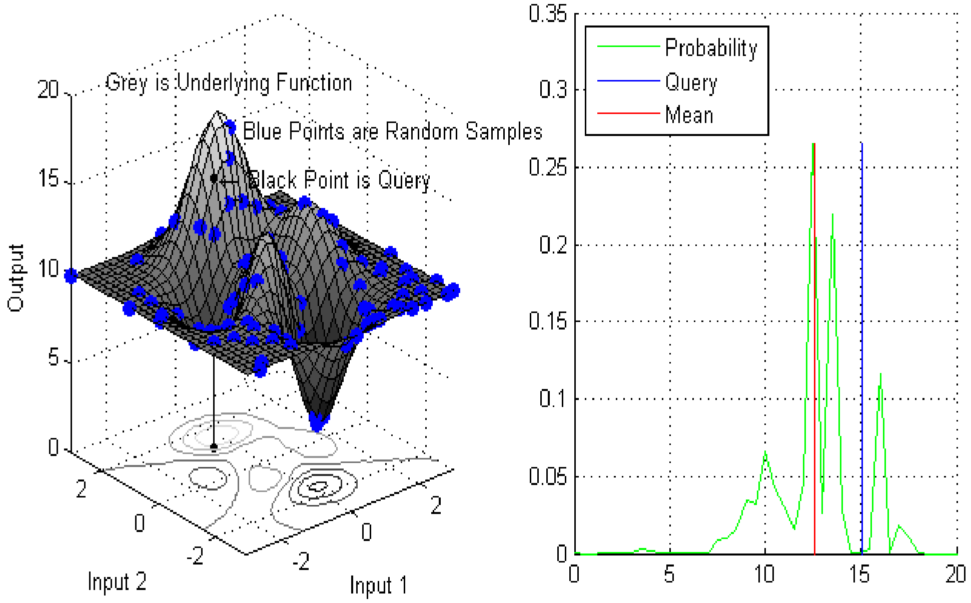


Figure 2.19 Construction of the Experimental Probabilistic Hypersurface for a Two Variable Function with 100 Known Data Points. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

Real data and many stochastic simulations produce noisy responses. The noise may manifest itself in the electronics used to measure the system, environmental factors that cannot be controlled, human error, or a myriad of other phenomena. In order to evaluate the methodology’s resilience to noise, or robustness, the function shown in Figure 2.19 was recalculated with Gaussian noise with a mean of zero and a standard deviation of 0.2. A subset of the data along with a plot of the residual between the noisy and original function is shown in Figure 2.21.

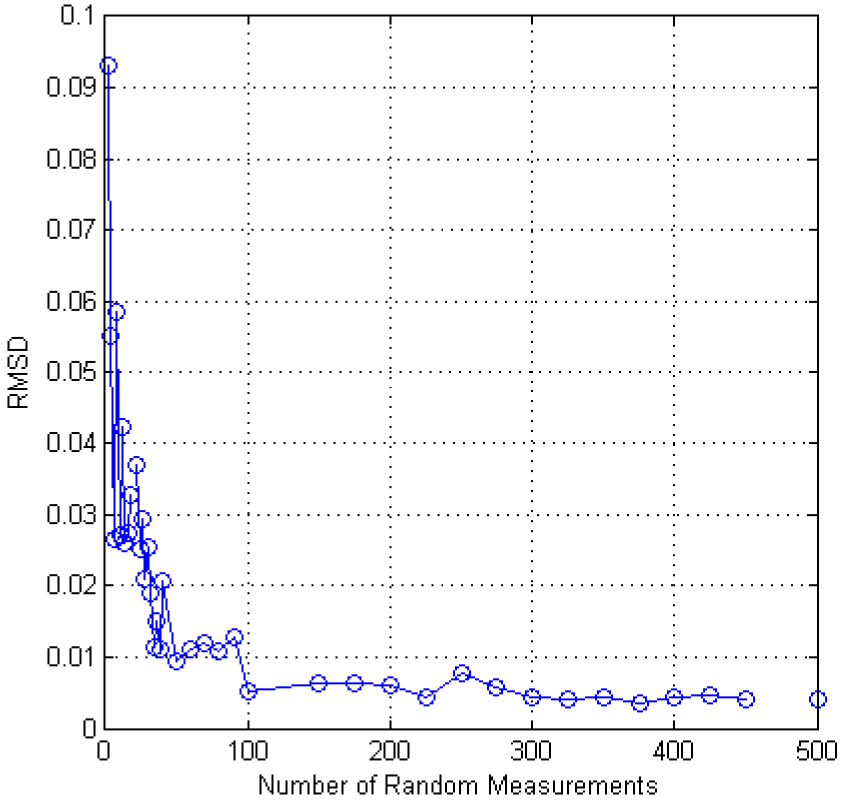


Figure 2.20 Plot of the RMSD versus Number of Random Measurements.

Random points were selected and used to construct a probability density for a particular query point as done previously. An example is shown in Figure 2.22. This process was repeated for various numbers of given points with the results used to produce Figure 2.23. This shows the percent difference between the actual value and the value with highest probability. One can see that the percent difference quickly reaches the value of the underlying noise within 100 samples. Thus, the method is robust to noise and allows one to potentially learn the inherent accuracy of the collected data as well as the form of the underlying function.

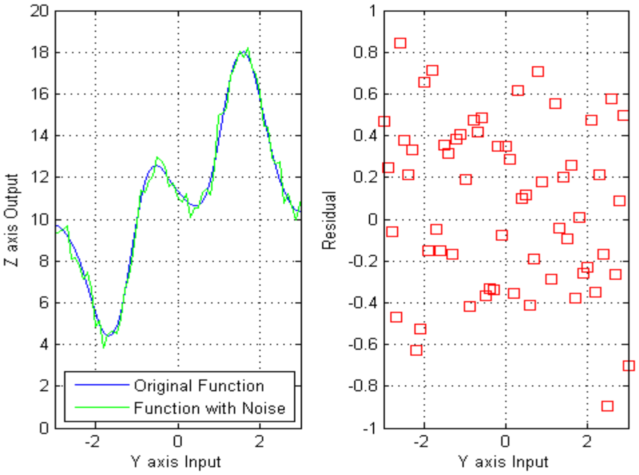


Figure 2.21 Description of White Noise Added to Function in Figure 2-10. a) Comparison of Original Function (blue) and Noisy Function (green) b) Plot of Residuals.

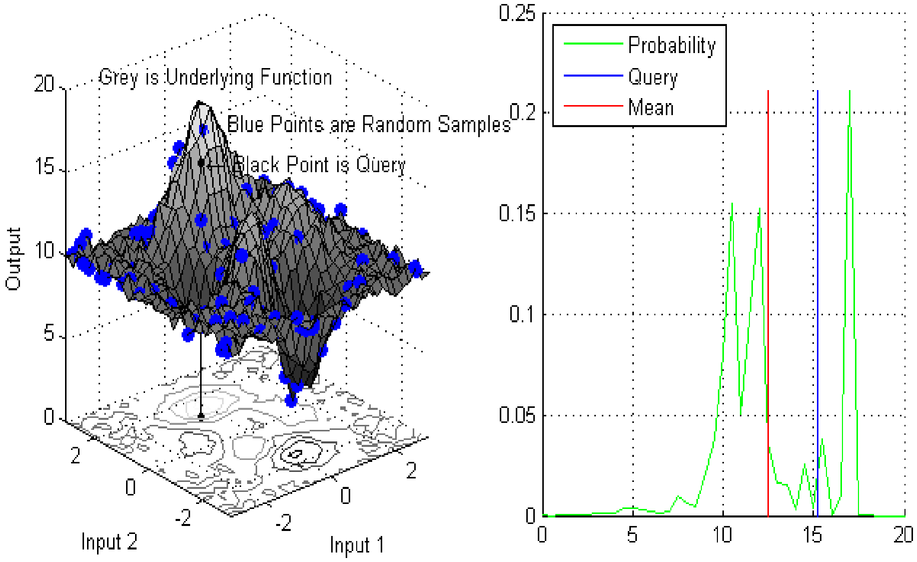


Figure 2.22 Construction of the Experimental Probabilistic Hypersurface for a Two Variable Noisy Function with 100 Known Data Points. a) Plot of Function, Known Data, and Query Point b) Probability Density Function, Mean of Probability Density, and Actual Value for Query.

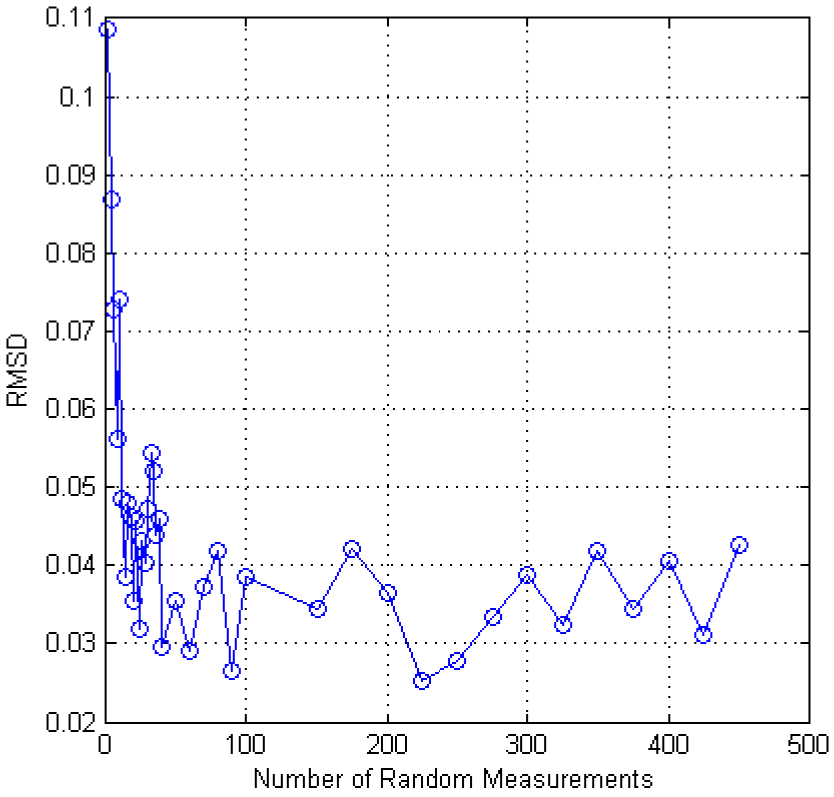


Figure 2.23 Plot of the RMSD versus Number of Random Measurements.

2.5. New Methodology

Section two has introduced the concepts of random projections, diffusion maps for manifold learning and embedding, and the experimental probabilistic hypersurface. These concepts have been synthesized into a new methodology that provides a new approach to understanding engineering phenomena. This is a fundamental change to how modeling or regression has been performed in the past. Further, the integrated use of the individual methods has never been done and provides a way of overcoming the limitations of each. Stated more clearly,

- The addition of new data produces biased results contrary to EPH which utilizes existing data/knowledge.
- The embedding of data in new coordinates from diffusion maps provides a more accurate weighting compared with the distance measurements previously used in EPH.
- Random projections provide a robust, fast technique to reduce the dimensionality of data allowing for quicker computation of both the diffusion maps and EPH.

The method to analyze high dimensional problems is shown in Figure 2.24. Both the known data and query are randomly projected onto a low dimensional manifold. The data are then embedded using the diffusion map approach discussed earlier in the section. The embedding coordinates for the known data and the query are used to construct the EPH. Based on the type of problem being solved, the probability density produced is used accordingly.

The need to reduce the dimension of the data is twofold. First, the time requirements for performing the embedding and constructing the EPH are quite sensitive to the dimensionality of the data. Second, the reduced dimension allows one to find features or trends that are difficult to find in the sparse, high dimensional space in which the data reside. Calculating the diffusion map requires the calculation of the distance matrix which is the pairwise Euclidean distance between all data, the time requirements scale as $O(\mathrm{N}^2)$ and constructing the Experimental Probabilistic Hypersurface scale as $O(\mathrm{N}^3)$.[56] Random projections require an extra operation with complexity of $O(\mathrm{dkN})$ but

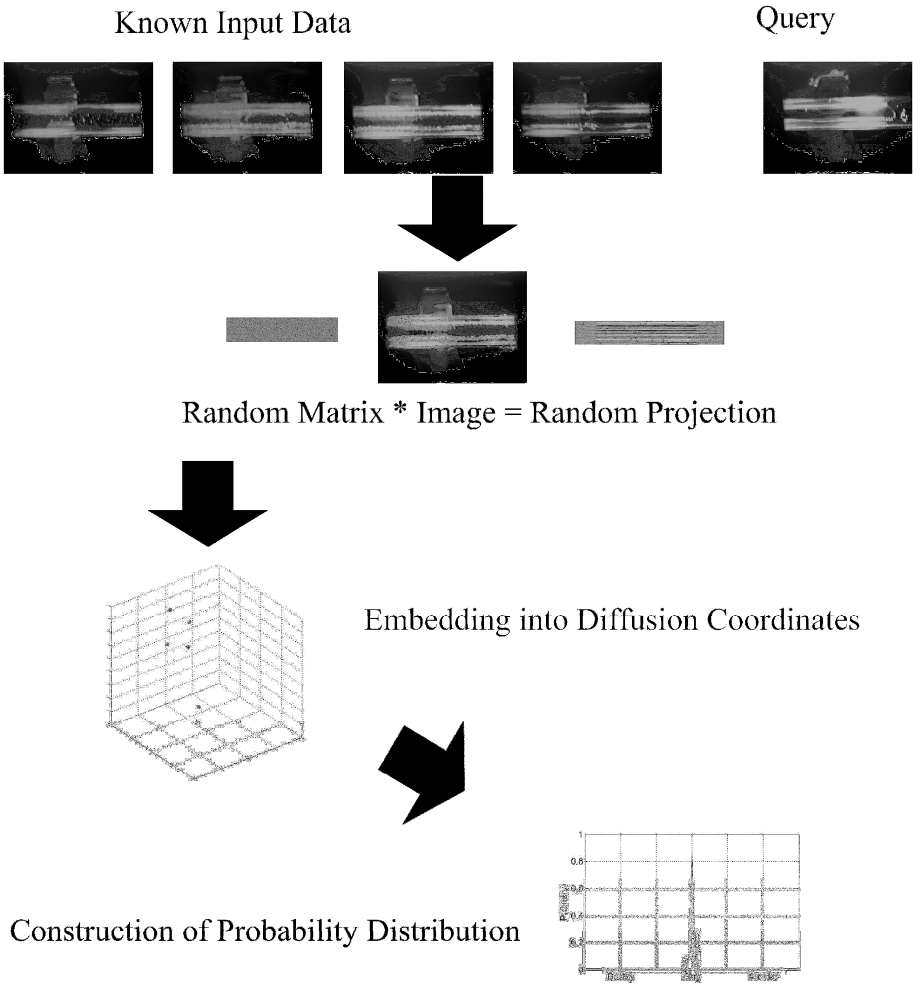

Known Input Data
Query
Random Matrix * Image = Random Projection
Embedding into Diffusion Coordinates
Construction of Probability Distribution

Figure 2.24 Methodology Graphically Represented.

this reduces the pairwise complexity to $O((\mathrm{dk})^2 + (\mathrm{dk})^3)$. Since k is a constant on the order of 4, and d is $k\log_{10}(N)$, plotting the time requirements in Figure 2.25 indicates that problems with approximately 10 or more variables/dimensions will run faster using random projections.

This approach can be used for regression and classification type problems. An example is for a given set of input data, a known output value exists. Classification type problems are a variation of regression except that the output state is not continuous but made up of discrete values. Thus, one can evaluate the cumulative distribution to determine in which category a query resides.

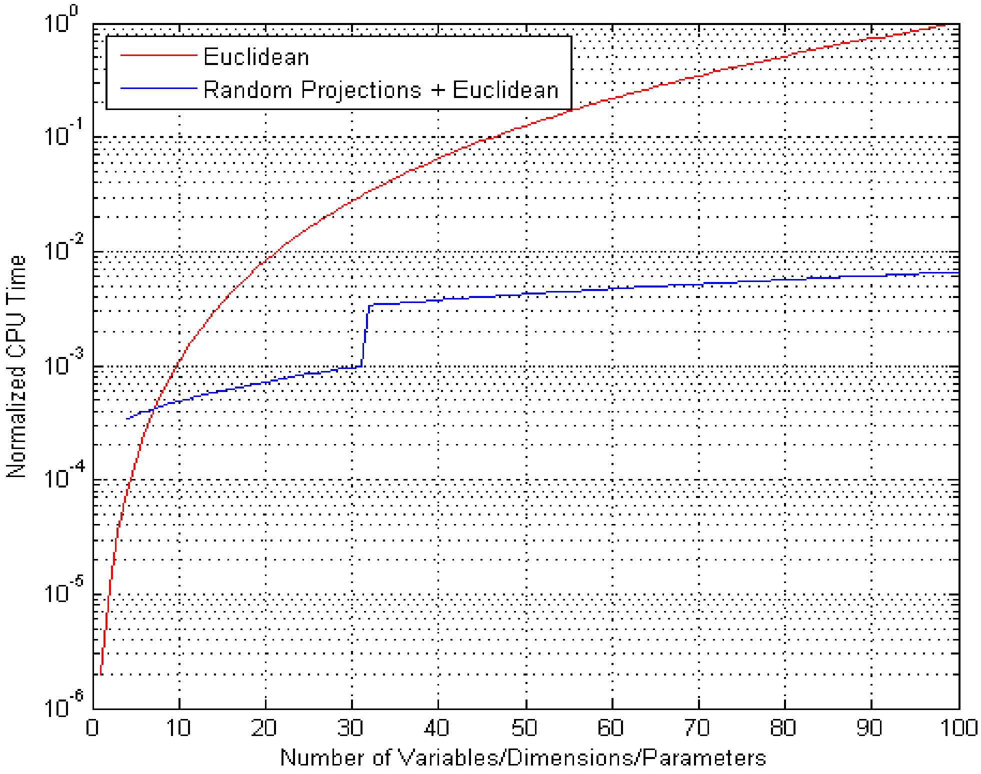


Figure 2.25 Comparison of CPU Time versus Number of Variables for Calculating Euclidean Distance for Full Vector and Reduced Vector.

This completes the development of the new methodology. The following sections apply the technique to problems of engineering interest. The problems are chosen to introduce elements with progressively higher dimensional data.

3. FRICTION FACTOR AND CRITICAL HEAT FLUX

The focus of this section is to contrast historical approaches to developing useful engineering predictive models versus the approach outlined in Section 2. Further, an important aspect to consider in this section is the development time required. This may be due to a number of factors including communication of results, the expense of performing experiments, or the difficulty of the problem. This difficulty manifests itself in the number of parameters or variables. In the past, recording and storing large amounts of information limited what problems could be tackled or required one to develop simpler models. Today, both communication of information, cost of acquiring data, and acquisition and storage limits are not limiting factors. Now, the time waiting to understand the enormous amount of data is limiting; therefore, a better method than the classic approach to modeling is needed.

A classic fluid dynamics problem is the calculation of pressure drop in a closed conduit. This elementary problem is implemented across a number of disciplines and provides an excellent history of the abstraction process. It is well established that the pressure drop or head loss is a function of the fluid properties (density, viscosity), the wall material (roughness), pipe diameter, and fluid velocity. Julius Weisbach[57] proposed Equation 3.1, which is still in use today. The head loss, h_L, per unit length of pipe, L, is a

$$\frac{h_L}{L} = f\,\frac{V^2}{2D} \qquad \text{(Eq. 3.1)}$$

function of the fluid velocity (V), pipe diameter (D), and a proportioning factor named friction (*f*) and given by Equation 3.2. The coefficients α and β depend on the wall

$$f = \alpha + \frac{\beta}{\sqrt{V}} \qquad \text{(Eq. 3.2)}$$

material and fluid. Several other relationships for pressure drop and/or friction factor have utilized different coefficients but the magnitude and meaning were debated since many didn't consider fluids other than water. These relationships were almost exclusively developed from individual experiments and the model developed statistically to varying levels of rigidity. In fact, "the dimensional rigidity of the relationship was lost on many in the fluid mechanics community, which led to the use of several irrational, dimensionally inhomogeneous, empirical formulas.[58]"

3.1. Historical Perspective

It is difficult to assign a date for the beginning of hydraulic engineering. A book that may be considered the foundation of modern hydrodynamics, *Della Misura dell'Acque Correnti*, or "On the Measurement of Running Waters," was written by Castelli.[59] Although this is not the starting point, if we consider it so, nearly 350 years elapses before the work of Moody presented a simple tool for solving the pressure drop in piping systems. Three hundred and fifty years of work to describe the simple phenomena of head loss of fluid flow through pipes. The difficulty was determining the variables of interest and how to deal with the nonlinearity of the head loss due to changes in these variables.

Osborne Reynolds observed the transition from laminar to turbulent flow and introduced a new parameter to predict the transition that would become what is known today as the Reynolds number.[60] The Reynolds number is the ratio of inertial to viscous forces with the transition for flow in pipes in the range of 2000 to 4000 and was a tremendous contribution allowing modelers to separate the two regions allowing different linear predictive models. The friction factor could then be described as a function of the Reynolds number for laminar flow based on the independent work of Hagen[61] and Poiseuille.[62] For turbulent flow, several predictions for pipe friction continue to be used including the smooth wall relationship described by Blasius[63] and the more accurate Colebrook-White[64] relationship.

Hunter Rouse incorporated these relationships into a more useful design tool, which was improved upon by Moody.[65] Of note, Rouse stated "these equations are obviously too complex to be of practical use. On the other hand, if the function which they embody is even approximately valid for commercial surfaces in general, such extremely important information could be made readily available in diagrams or tables.[66]" Stated more clearly, Rouse felt the five dimensions were too complicated for practical use and was able to reduce the needed dimensions to two where $f = F(\mathrm{Re}, \varepsilon)$. The friction factor is a function of the Reynolds number and the pipe roughness. The Moody diagram, shown below in Figure 3.1, continues to be used today because it provides a simple, accurate method to predict the friction factor given the Reynolds number and the wall roughness.

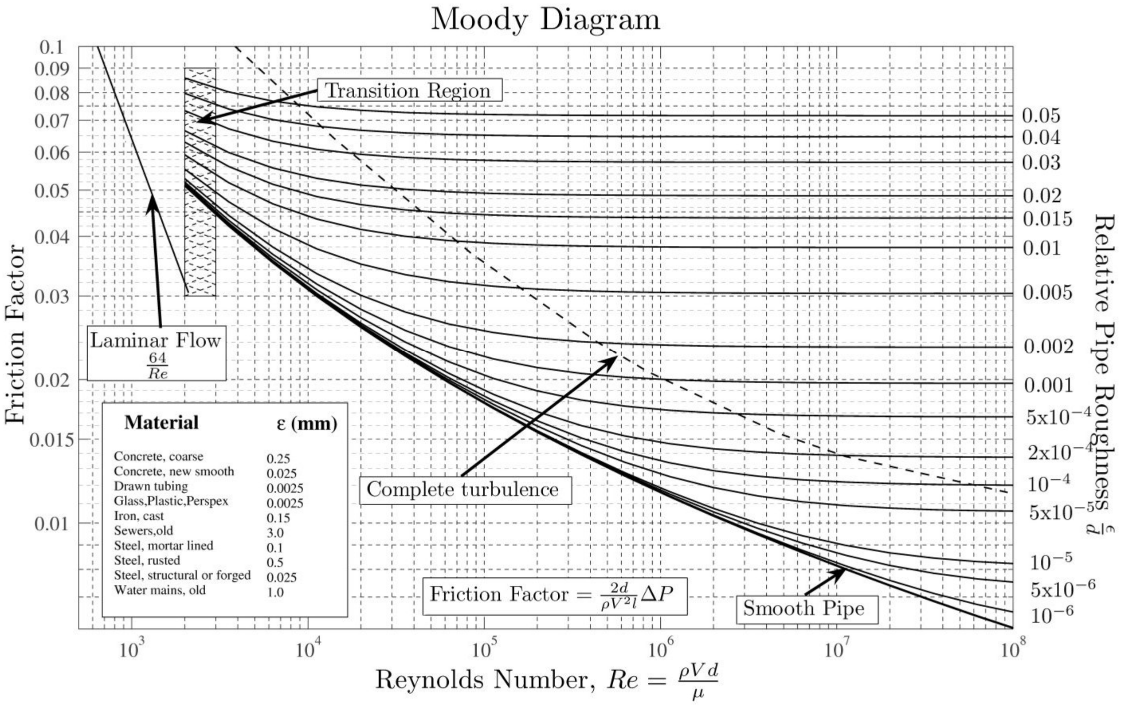


Figure 3.1 Moody Chart Used for Predicting Wall Friction.

It is well known that the shape of the velocity profile changes with increasing velocity due to changes in the boundary layer and the flow field with the move from a laminar to turbulent flow. In order to easily determine the friction factor for various flow conditions in round pipes, various authors have put forward a graphical representation, which is referred to today as the Moody Chart. Figure 3.1 is the Moody Chart, which has lines that relate the non-dimensional friction factor to the dimensionless Reynolds number and to dimensionless wall roughness. For laminar flow, Poiseuille’s law, shown in Equation 3.3, is derived relationship as a function of the Reynold’s number is used to predict pipe wall friction.

$$f = \frac{64}{\mathrm{Re}} \qquad \text{(Eq. 3.3)}$$

where

$$\mathrm{Re} = \frac{\rho V D}{\mu} \qquad \text{(Eq. 3.4)}$$

Reynolds number is typically referred to as the ratio of inertial to viscous forces and is derived from simplifications of the momentum equation. For the turbulent regime, one must consider the effects of wall roughness in order to calculate the friction factor. The effects of wall roughness must be considered and those effects are illustrated by the various lines on the Moody Chart that correspond to different roughness described by the variable ε. Roughness, ε, is a length measurement that describes the irregularities of the wall surface. Darcy and others have put forward relationships to describe friction as a function of Reynolds number and pipe roughness. Relationships of the form described by Colebrook and White, Equation 3.5, are typically used to determine the friction factor

$$\frac{1}{\sqrt{f}} = -2\log\left(\frac{\varepsilon}{D} + \frac{2.51}{\mathrm{Re}\sqrt{f}}\right) \qquad \text{(Eq. 3.5)}$$

in the turbulent range; however, Blasius[63], using similarity theory, found the friction to be well calculated by the relationship shown below. It can be seen that this relationship is for smooth pipes (i.e. ε = 0); whereas the implicit form in Equation 3.5 considers roughness. Equation 3.6 is widely used for predicting wall friction in smooth wall pipes.

$$f = \frac{0.3164}{\mathrm{Re}^{0.25}} \qquad \text{(Eq. 3.6)}$$

The Moody Chart is a mainstay of most engineers wishing to estimate the friction head loss in flow through pipes and the basis for this statistical work on the accuracy of the predicted values. There is error associated with the use of each line represented on this graph due to assumptions made in developing the model and conclusions drawn from the experimental data used in validating the models. Statements of true accuracy are rare, but White[67] has stated the Moody chart is only accurate to $^{+}/_{-}15\%$.

The starting point for this study is the survey by Drew.[68] The paper presents the results of a literature review conducted primarily to determine if the available data for turbulent flow were represented accurately by a form of Equation 3.2 as the widely accepted equation of Lees[69] (Equation 3.7) for the friction factor in isothermal flow

$$f = 0.0072 + \frac{0.612}{\mathrm{Re}^{0.35}} \qquad \text{(Eq. 3.7)}$$

through smooth pipes. The validation of the Lees model is not done statistically; however, a new relationship is derived of the same form and is compared with Lees' model qualitatively. This paper was instrumental in compiling the data used in this investigation. The writers compiled an extensive list of English, American, German, and French experimenters. Although the paper does not present any new data itself, it presents existing data in both the turbulent and laminar regions for smooth pipes as well as boundary lines of $^{+}/_{-}$ 5% of the value of the prediction model given in Equation 3.8. Again, the form is similar to Equation 3.2 and is shown in Figure 3.2 which illustrates the data compiled for the model comparison.

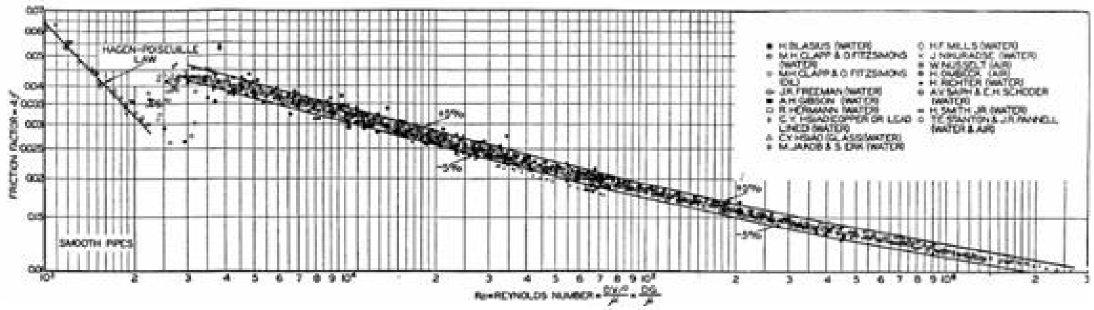


Figure 3.2 Friction Factor Data versus Reynolds Number.

$$f = 0.0056 + \frac{0.5}{\text{Re}^{0.32}} \qquad \text{(Eq. 3.8)}$$

The paper by Freeman[70] is an exhaustive study of the flow of water through rough and smooth pipes, with extreme detail placed on accuracy of the results. The data presented is mostly in the range of turbulent flow, although some data points are from the laminar flow regime. Also, great attention was placed on varying the size of pipes and the velocity. Care was taken in the design of the experiment to ensure fully developed flow in this tank emptying experiment. Differential pressure measurements were taken with a manometer with a stated accuracy of 0.001 ft H_2O. Details concerning the connection with the main pipe were discussed explaining how they were designed to minimize the disturbance to the boundary layer.

The Ombeck paper[71] contains significant data along the turbulent range. Although the paper was written in German, interpretation of the data was not difficult, but the lack of description of the experiment leaves no way of determining the accuracy of the data. Ombeck utilized air as the working fluid and drawn brass tubing. With respect to other published data, Ombeck's data fit the standard curve well but have a different asymptotic value which is shown in Equation 3.9.

$$f = \frac{0.242}{\mathrm{Re}^{0.224}} \qquad \text{(Eq. 3.9)}$$

Blasius' paper did much to further research into the turbulent and transition regions of flow in pipes.[63] The paper presents all of the data collected in the same manner as the other experimenters, so the interpretation of the data was not difficult. Blasius did his experiments on small pipe diameters made of lead and glass with both air and water and developed an equation for turbulent flow between Reynolds numbers of 4000 and 100000 given above in Equation 3.6.[63]

The Stanton and Pannell[72] paper contains data that covers the entire range of flow from laminar to turbulent. Their data, derived from both air and water experiments has high accuracy and includes a number of points in the laminar region. Of particular interest is the use of pumps or fans to produce the desired flow condition in contrast to the others use of elevated tanks.

Clapp and Fitzsimons[73] did their work with water and oil at low Reynolds numbers. The smooth pipe used was drawn copper, but their results are only for 0.494 inch diameter pipe. The thesis presented in the paper was not on the subject of friction factor versus Reynolds number, but experiments were documented and Reynolds number and Friction factor can be calculated from the printed results. The calming length of the pipe was over 100 diameters allowing the flow to fully develop.

The Poiseuille paper contains data that covers only the laminar portion of flow.[62] The data, derived from water experiments in glass with satisfactory development length, has very good accuracy, though there is a portion of data that is extremely inconsistent and which was removed from the dataset. Poiseuille used glass tubing with enough development length for good measurements.

Herman,[74] using water as his working fluid, pioneered high Reynolds number research in the turbulent range. To obtain his accurate results in the turbulent range he used very long calming lengths, from 150 to 250 length over diameter ratios, to make sure the flow was fully developed. Utilizing drawn brass and copper tubing, using the form for friction shown below, Hermann reported values of (A) 0.00135 and 0.00132 for the intercept of Brass and Copper respectively, (B) 0.099 (Brass) and 0.0998 (Copper) for the slope, and (C) -0.300 for the exponent.[74]

$$f = A + B \cdot \mathrm{Re}^{C} \qquad \text{(Eq. 3.10)}$$

3.2. Current Approaches to Predicting the Friction Factor

Currently, Poiseuille's law provides a purely theoretical approach to predicting the friction factor for laminar flow. The Colebrook and White relationship is used for turbulent flow but due to the difficulty of its implicit nature there is a desire to find explicit forms. The most common is the Blausis equation for smooth walled pipes shown in equation 3.5. More recent relationships proposed are the Swamee-Jain equation[75] and the Serghides solution[76] shown below in equations 3.11 and 3.12 respectively.

$$f = \frac{0.25}{\left(\log\left(\dfrac{\varepsilon}{3.7D} + \dfrac{5.74}{\mathrm{Re}^{0.9}}\right)\right)^{2}} \qquad \text{(Eq. 3.11)}$$

$$f = \left(A - \frac{(B-A)^2}{C-2B+A} \right)^{-2} \qquad \text{(Eq. 3.12)}$$

where

$$A = -2\log\left(\frac{\varepsilon/D}{3.7} + \frac{12}{\mathrm{Re}} \right)$$

$$B = -2\log\left(\frac{\varepsilon/D}{3.7} + \frac{2.51A}{\mathrm{Re}} \right)$$

$$C = -2\log\left(\frac{\varepsilon/D}{3.7} + \frac{2.54B}{\mathrm{Re}} \right)$$

There is an active community working on the development of friction factor predictive models.[77,78,79] The work in this area deals primarily with the accuracy of explicit formulations of the Colebrook-White formula. A number of different approaches have been used to produce an explicit form for predicting pipe wall friction that include Lambert W function,[80] the Weymouth equation,[81] and the Panhandle A and B equations.[82] A comparison of selected formulations is presented by Romeo[79] and is shown below in Table 3.1 with the addition of Sonnad's linear and continuing fraction approximations.

The Model Selection Criterion (MSC) shown in Equation 3.13, is a statistic that describes the accuracy of one model over the other. The weighted difference is modified with a term that includes the number of parameters for each model. Thus, more elaborate models with a large number of parameters are penalized versus simpler models. Larger values indicate a higher degree of similarity between the proposed model results and the

Table 3.1 Comparison of Various Predictive Models with the Colebrook-White Approximation for Friction for Turbulent Pipe Flow.

Author	Reference	Model Selection Criterion
Colebrook-White	C. F. Colebrook, Turbulent flow in pipes with particular reference to the transition region between the smooth and rough pipe laws. J. Inst. CiV. Eng. 1938-1939, 11, 133.[83]	-
Moody	L.F. Moody, Friction factors for pipe flow, Trans. ASME 66 (8) (1944) 671.[65]	4.639
Wood	D.J. Wood, An explicit friction factor relationship, Civil Engrs. ASCE 60 (December 1966).[84]	-4.019
Churchill	S.W. Churchill, Empirical expressions for the shear stress in turbulent flow in commercial pipe, AIChE J. 19 (2) (1973) 375.[85]	8.980
Jain	A.K. Jain, Accurate explicit equations for friction factor, Proc. ASCE, J. Hydraulics Div. 102 (HY5) (1976) 674.[75]	9.118
Chen	N.H. Chen, An explicit equation for friction factor in pipe, Ind. Eng. Chem. Fundam. 18 (3) (1979) 296.[86]	12.180
Round	G.F. Round, An explicit approximation for the friction-factor Reynolds number relation for rough and smooth pipes, Can. J. Chem. Eng. 58 (1) (1980) 122.[87]	3.067
Barr	D.I.H. Barr, Solutions of the Colebrook–White function for resistance to uniform turbulent flow, Proc. Inst. Civil Engrs., Part 2 71 (1981) 529.[88],[89]	12.247
Zigrang-Sylvester	D.J. Zigrang, N.D. Sylvester, Explicit approximations to the Colebrook's friction factor, AIChE J. 28 (3) (1982) 514.[90]	12.537
Haaland	S.E. Haaland, Simple and explicit formulas for the friction factor in turbulent pipe flow, Trans. ASME, JFE 105 (1983) 89.[91]	8.845
Manadili	G. Manadili, Replace implicit equations with signomial functions, Chem. Eng. 104 (8) (1997) 129.[92]	9.722
Romeo	E. Romeo, C. Royo, and A. Monzon, Improved explicit equations for estimation of the friction factor in rough and smooth pipes. Chem. Eng. J. 2002, 86, 369.[79]	22.111
Sonnad – Linear	C.T. Goudar and J.R. Sonnad, Explicit friction factor correlation for turbulent flow in smooth pipes, Industrial Engineering and Chemical Research 42 (2003), pp. 2878–2880.[78]	26.57
Sonnad - CFA		28.22

output of the Colebrook-White formula. For Table 3.1, each of the models proposed by the authors was compared with the Colebrook-White equation and the resulting MSC is shown.

$$MSC = \ln\left(\frac{\sum_{i=1}^{n}\left(CW_i - \overline{CW}\right)}{\sum_{i=1}^{n}\left(CW_i - M_i\right)}\right) - \frac{2NP}{n} \qquad \text{(Eq. 3.13)}$$

The current work is aimed at demonstrating a new approach to modeling. Although the head loss equation is based upon physics, the proportional constant known as friction is derived entirely from experimental data. This interesting aspect is similar to many other problems encountered in engineering where one can accurately describe the underlying physics but not detailed features. This is commonly due to the scale of the problem but many times the difficulty is in determining or measuring the variables or the large number of variables needed to accurately describe the phenomena. The following section will review the predictions for the wall friction in a pipe and compare the classic approach to the methodology described in this paper.

3.3. New Approach to Predicting the Friction Factor

The methodology described earlier in this paper provides an alternative approach to predicting pipe friction and demonstrates a new approach to modeling. Rather than relying on a curve fit based on the three parameters of wall roughness, pipe diameter, and Reynolds number, a pure statistical approach to sampling the manifold described by all parameters (wall roughness, pipe diameter, fluid velocity, fluid viscosity, and fluid density) would allow direct calculation and provide a measure of certainty associated with the result. The ease with which the new methodology is applied should be considered against the 400 years of development that continues to go into the prediction of head loss.

3.3.1. Comparison with Current Predictive Models

As shown in Table 3.1, explicit models have been developed that show high fidelity with the Colebrook-White formula. However, the Colebrook-White formula is a fit of data and contains no mechanistic underpinnings. The accuracy of this fit must be considered when comparing models. One may ask, “why we would desire higher fidelity explicit models when a simpler model would suffice?” Figure 3.3 contains two plots of actual data from the references listed above along with curves associated with various models. The top plot is for fluid in smooth wall pipes with $^{\varepsilon}/_{D}$ of 10^{-8} with predictions of all authors in Table 3.1. The lower plot is a close up of the range of Reynolds numbers from 10^4 to 10^5 along with selected models and includes bounding lines corresponding to $^{+}/_{-}$ 5% of the Colebrook-White formula. The root mean square deviation is the statistic used to compare the models with data. This formula, shown in Section 2, is commonly used to compare the differences between values predicted by a model or an estimator and the values actually observed. Root mean square deviation values are shown in Table 3.2. Clearly, there is little difference among the models when compared with the distribution of actual data in smooth wall piping.

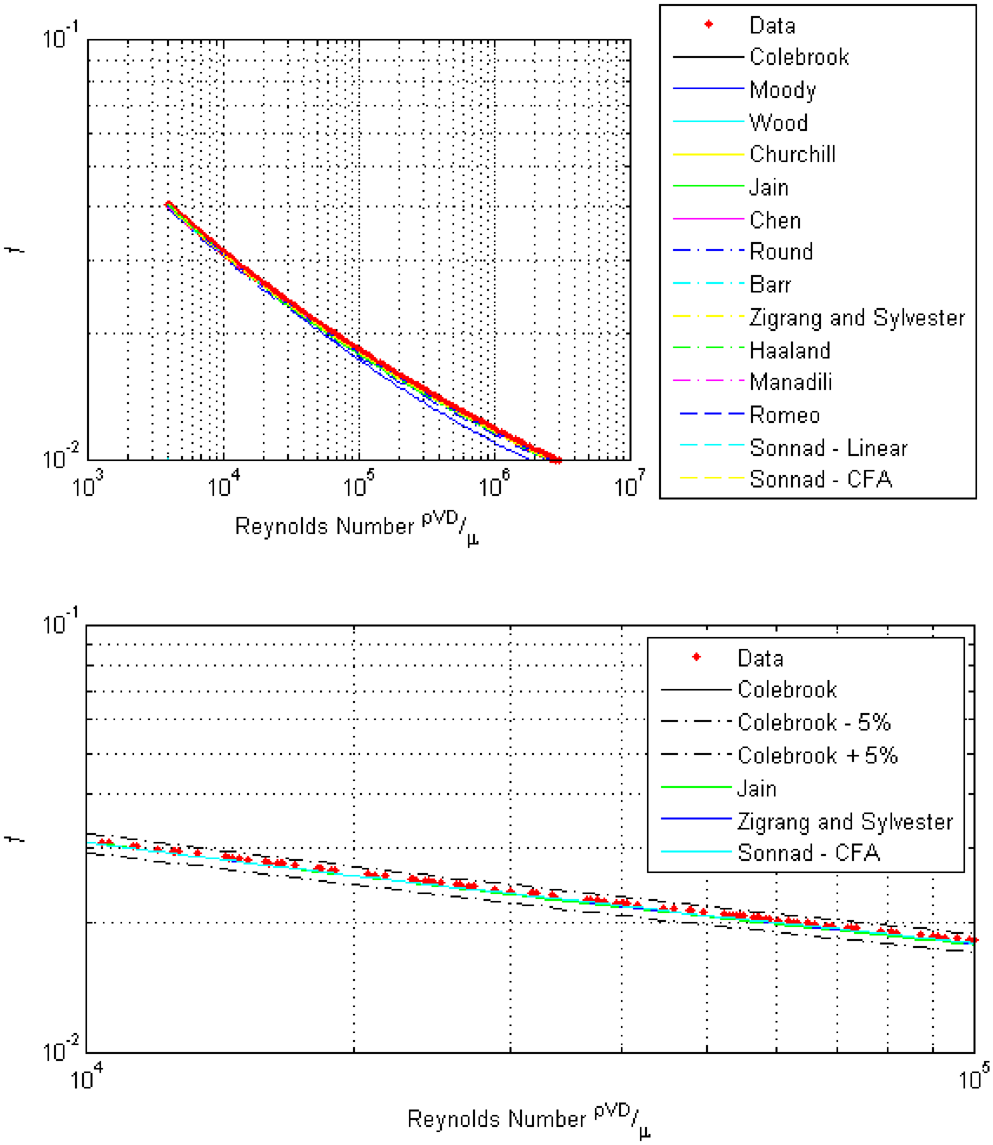


Figure 3.3 Comparison of Predictive Models for Turbulent Pipe Flow versus Data. a) $10^{3} < Re < 10^{7}$ b) $10^{4} < Re < 10^{5}$

Table 3.2 Comparison of Various Predictive Models with Data for Friction for Turbulent Pipe Flow.

Author	Root Mean Square Deviation
Colebrook-White	0.00150
Moody	0.00162
Wood	0.01608
Churchill	0.00151
Jain	0.00152
Chen	0.00150
Round	0.00167
Barr	0.00150
Zigrang-Sylvester	0.00151
Haaland	0.00153
Manadili	0.00150
Romeo	0.00149
Sonnad –Linear	0.00150
Sonnad -CFA	0.00150

To compare the new methodology with recent published methods for determining the friction factor, a synthetic dataset was created. This initial synthetic dataset consisted of 2420 points for four values of roughness, 11 pipe diameters, 11 values of velocity, and the fluid density and viscosity for three fluids. Values for Reynolds number ranged from 4000 to 9.1 x 10^9 and 31 values for $^{\varepsilon}/_{D}$ ranging from 10^{-13} to 0.0033 which is similar to the work done by previous authors.[65,78,79] The data was produced using equation 3.14. The values in the dataset correspond to the friction factor determined

$$f_{synthetic} = f_{Colebrook-White} + 3\sigma P(0,1) \qquad \text{(Eq. 3.14)}$$

from the Colebrook-White formula with Gaussian noise corresponding to a standard deviation of $0.00\overline{3}$. The root mean square deviation was used to compare the various models with the synthetic data which is shown in Table 3.3.

Table 3.3 Comparison of Various Predictive Models with Synthetic Data for Friction for Turbulent Pipe Flow.

Author	Root Mean Square Deviation (smooth wall)	Root Mean Square Deviation (All Data)
Colebrook-White	0.00017	0.00410
Moody	0.00074	0.00402
Wood	0.00951	0.01139
Churchill	0.00019	0.00408
Jain	0.00018	0.00409
Chen	0.00017	0.00410
Round	0.00028	0.00412
Barr	0.00017	0.00410
Zigrang-Sylvester	0.00017	0.00411
Haaland	0.00018	0.00410
Manadili	0.00017	0.00410
Romeo	0.00017	0.00410
Sonnad –Linear	0.00017	0.00410
Sonnad -CFA	0.00017	0.00410

3.3.2. Comparison with New Methodology

The models developed above were derived from measurements and an understanding of the Reynolds number and pipe roughness. Colebrook was able to develop a fit to data for various pipe roughness values. The authors listed above have essentially developed explicit models to closely match the results of the Colebrook-

White formula but have not considered the accuracy of the Colebrook-White formula compared with actual data. The methodology presented in the previous section is used to predict the friction factor based on available data. Rather than assuming an underlying model (Lambert W, Weymouth, or Panhandle) and adjusting the number of parameters to better fit the data or another model, this new approach uses the data to generate a probability distribution. One can then use the friction factor with the highest probability or calculate a probability for the pressure drop directly.

Twenty randomly selected data points consisting of fluid properties, velocity, pipe diameter and roughness, and the corresponding friction factor are sampled from the synthetic dataset. The five input variables, pipe roughness, pipe diameter, velocity, fluid density, and fluid viscosity are used as inputs with the corresponding friction factor used as the known outputs. A query point is also selected from the dataset and the input variables used as the query in the EPH algorithm while the friction factor is used for comparative purposes. The result is shown below in Figure 3.4. It should be noted that the data in Figures 3.4a, 3.5a, and 3.6a are plotted versus Reynold’s number for visualization purposes. The data is in five dimensions and this is what is used in the methodology irrespective of the calculation for friction factor using the Colebrook-White relationship.

The plot on the left portion of Figure 3.4 is for the random data and query point plotted on a Moody chart or Reynolds number versus friction factor. The red points are the randomly selected points whereas the blue square is the query point. The plot on the right of Figure 3.4 is a plot of friction factor versus probability produced from EPH. The green curve is the probability distribution and the blue and red lines are the actual friction factor and the mean value of the probability respectively. For the query point in Figure 3.4, the five randomly selected data points produce a probability distribution with an mean corresponding to a friction factor of 0.0155. The friction factor from the dataset is 0.0089. Figures 3.5 and 3.6 are for 50 and 500 randomly selected points respectively. As shown in Figures 3.4b, 3.5b, and 3.6b, the probability distribution becomes more peaked and the most likely friction factor predicted approaches the actual friction factor due to more information present. Friction factor is directly proportional to the pressure drop so one can evaluate the cumulative distribution function (CDF) to determine the likely range based on the desired level of accuracy.

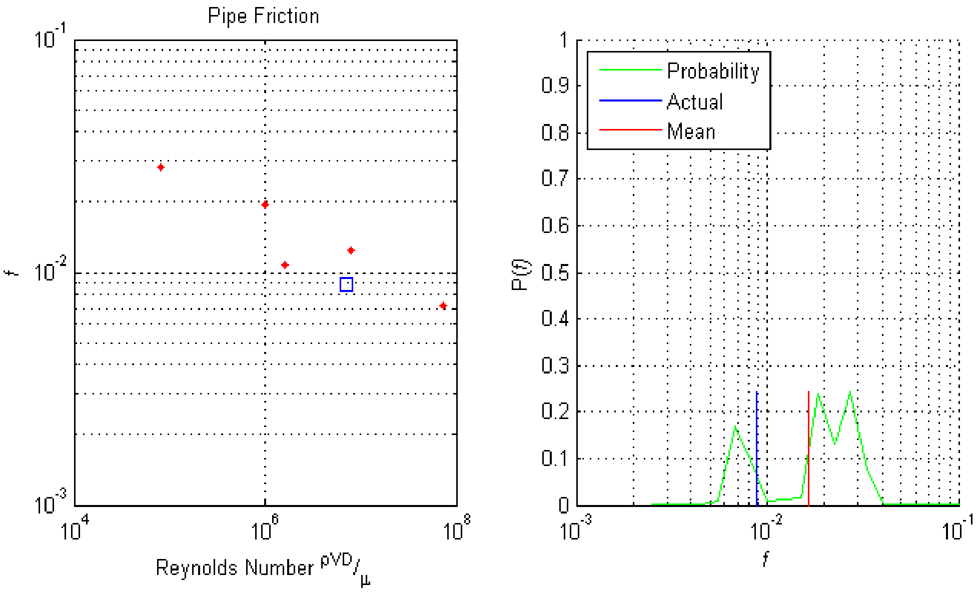


Figure 3.4 a) Plot of Five Known Data Points (red) and Query (blue) b) Plot of Probability Density (green), Mean Probability (red), and Actual Value of Query (blue).

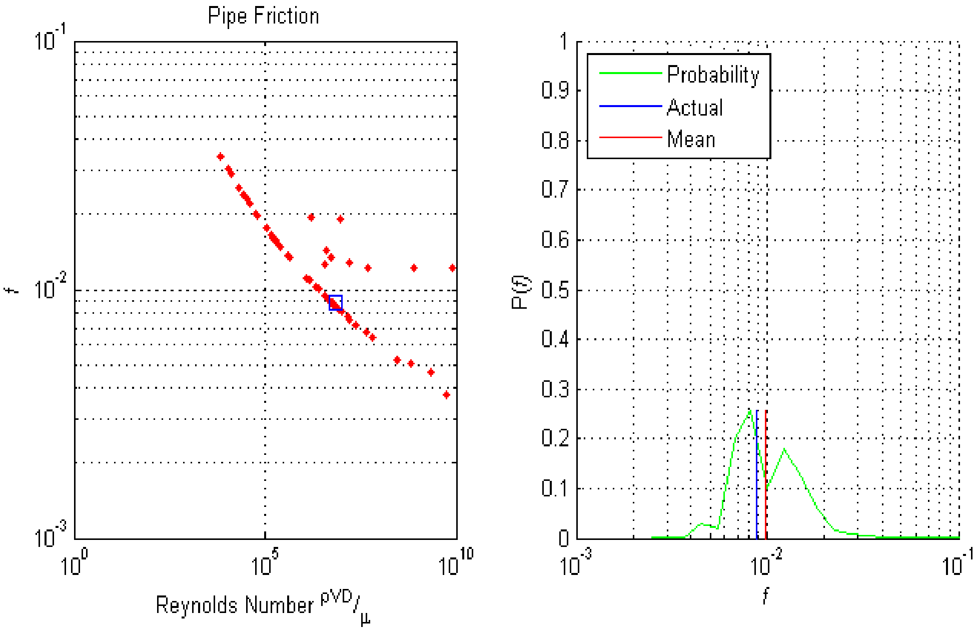


Figure 3.5 a) Plot of Fifty Known Data Points (red) and Query (blue) b) Plot of Probability Density (green), Mean Probability (red), and Actual Value of Query (blue).

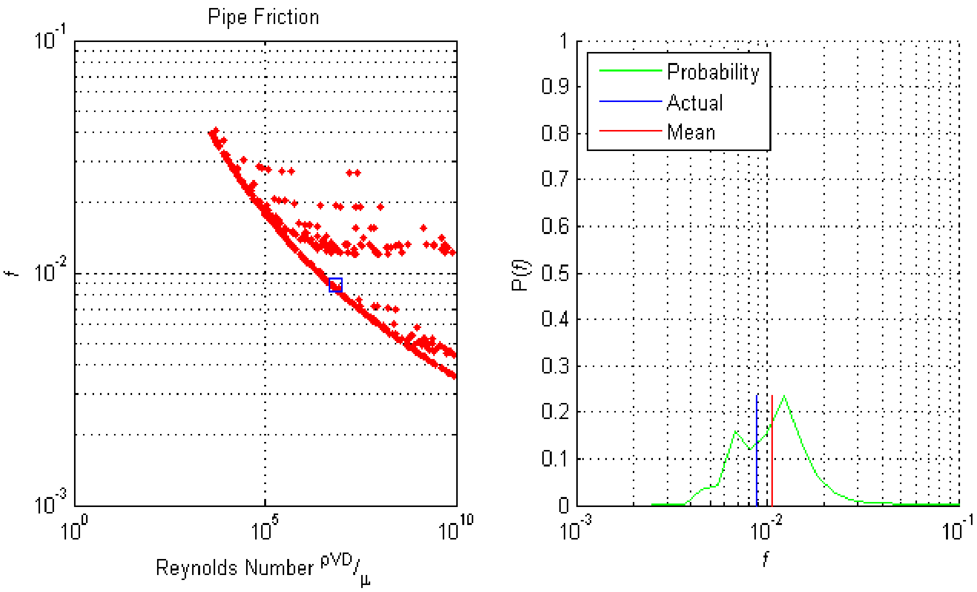


Figure 3.6 a) Plot of Five Hundred Known Data Points (red) and Query (blue) b) Plot of Probability Density (green), Mean Probability (red), and Actual Value of Query (blue).

Figure 3.7 contains three plots: one, a plot of 10 random data (red) and a query point (blue) plotted on axes of Reynolds number and friction factor; two, the corresponding probability plot; and three, the cumulative distribution plot (CDF). The CDF shows that there is less than a 18% probability that the friction factor is less than 0.01, a 78 % chance that the friction factor is between 0.01 and 0.02, and a less than 4% chance of the friction to be greater than 0.02. With more data, the probability is more peaked about the correct value of the friction factor.

Figure 3.8 illustrates this phenomenon for 50 randomly selected points. The location of the calculated mean probability is essentially equal to the friction for the query point illustrating the case that more data provide more information to construct a more accurate estimate of friction.

Now that we have seen how well EPH can predict the surface of the manifold described by several dimensions that correspond to the input variables, information is needed to provide the optimum number of data. Figure 3.9 is a plot of the Root Mean Square Deviation versus the number of known data points. Since the input data is randomly selected, each query point was predicted twenty times with different random sets of known data points. The mean RMSD between the actual data and mean probability calculated from the density produced by EPH was plotted and it can be seen that approximately 90 known points gives the same RMSD value between the data and the prediction as the models presented earlier.

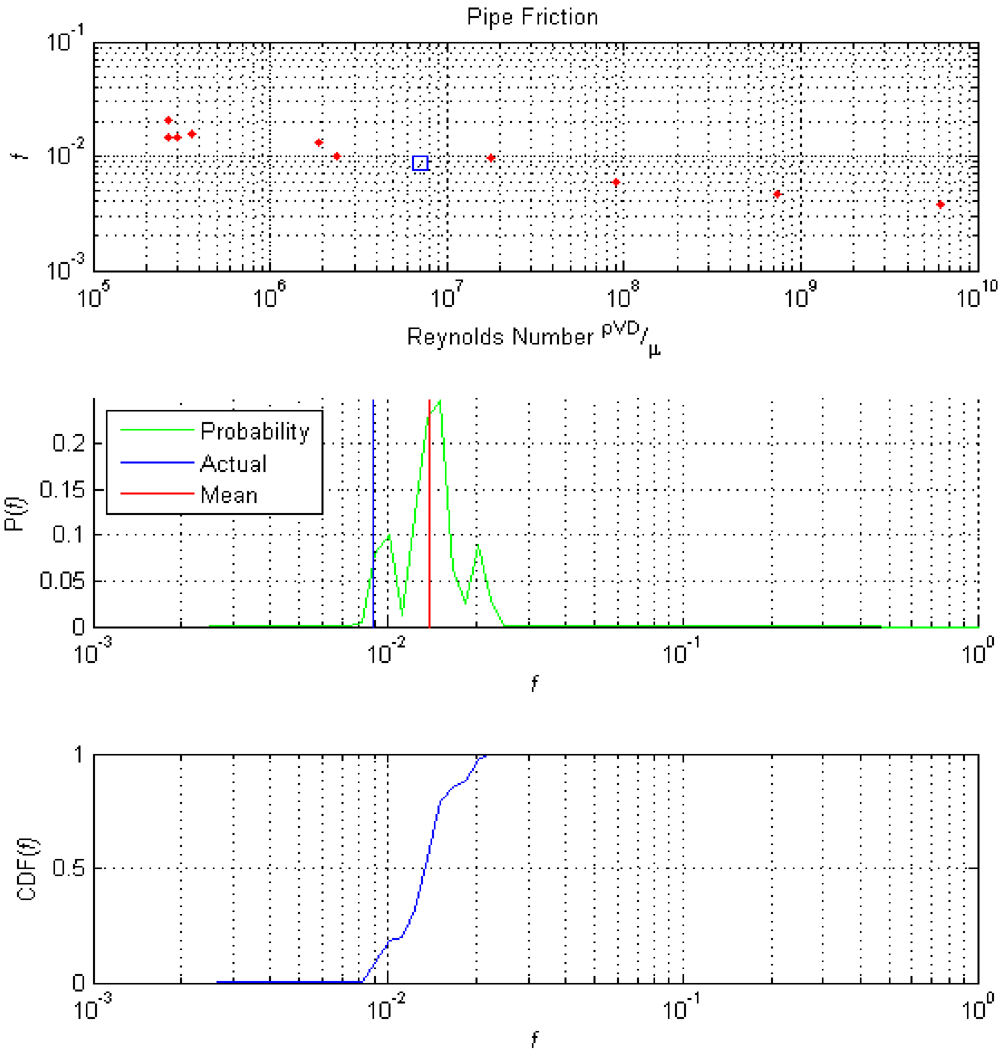


Figure 3.7 a) Plot of Ten Known Data Points (red) and Query (blue) b) Plot of Probability Density (green), Mean Probability (red), and Actual Value of Query (blue) c) Cumulative Density Function.

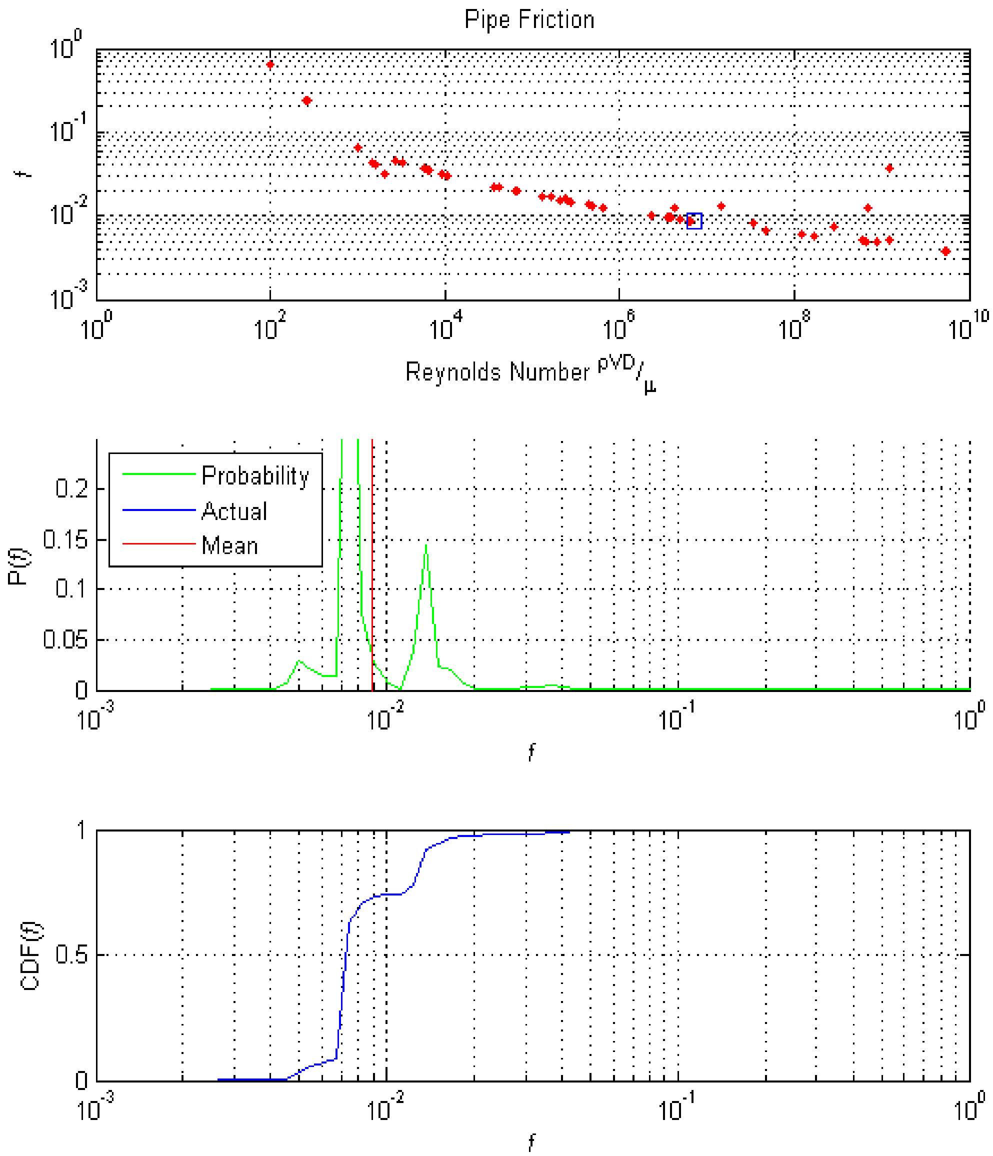


Figure 3.8 a) Plot of One Hundred Twenty Five Known Data Points (red) and Query (blue) b) Plot of Probability Density (green), Mean Probability (red), and Actual Value of Query (blue) c) Cumulative Density Function.

A more intriguing result and the focal point of the proposed method is when we combine dimension reduction techniques such as random projections or those similar to the method proposed by Lafon with the predictive power of EPH. This approach would allow for faster predictive computations and allow EPH to be utilized for much higher dimension problems. Figure 3.9 also shows the RMSD versus number of known data points for 2 reduced variables using the embedding technique of Lafon. Although it takes more known data points, 400 vs. 90, than using the actual data, the accuracy approaches that of the other methods. Thus, for high dimension problems, the decrease in computation time would warrant using the diffusion coordinates or random projections as the input to the EPH.

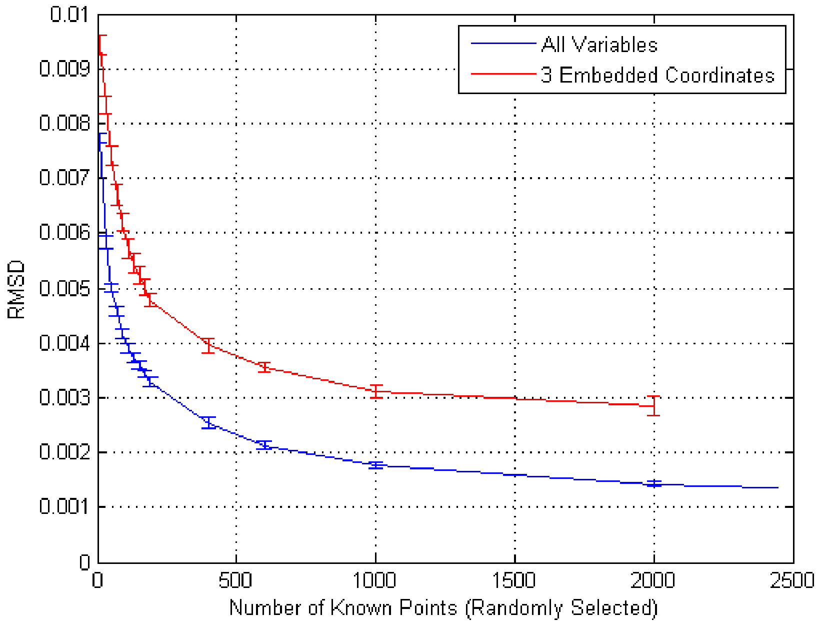


Figure 3.9 RMSD versus Number of Known Data for All Variables and for 3 Embedded Coordinates.

3.4. Critical Heat Flux

Critical Heat Flux, or CHF, is the phenomenon that describes a change in the boiling regime resulting in reduced heat transfer and serves as the upper limit of safe operation for heat transfer equipment. For heat flux controlled systems like nuclear fission and electrical heated systems, the CHF condition results in a drastic rise in wall temperature as the heat transfer coefficient is drastically reduced due to vapor blanketing whereas the CHF condition in temperature controlled systems results in a dramatic decrease in heat flux. Due to the importance of this limiting condition, a significant amount of research has been carried out to understand this phenomenon including the development of predictive correlations. An excellent overview of critical heat flux research can be found in the work of Kandlikar,[93] Thompson,[94] Hall and Mudawar,[95,96] and Groeneveld.[97] There have been several variables that affect the CHF. Six are most commonly identified,[98] which include:

- Pressure – The pressure of the working fluid has a relatively weak influence on the CHF with a gradual rise in CHF with a decrease in pressure for certain ranges.[98,99]
- Local liquid subcooling - The level of subcooling, sometimes referred to as the mixed mean temperature, requires a higher heat flux to initiate and sustain boiling leading to an increase in the critical heat flux. Again, Bergles[98] provides information regarding the effect of subcooling with similar plots from Sakurai and Shiotsu.[100]

- Mass Flux – except at high quality, mass flux is strongly correlated with critical heat flux with a near linear relationship.[101] The higher velocity leads to increased turbulence enhancing heat transfer. This effect has led researchers such as Inasaka and Nariai[102] to classify the CHF into high heat flux and low heat flux regions with mass flux as a parameter.
- Length and Diameter – These provide dimensional information and are typically combined in regression models as the ratio of the length over the diameter. The length to diameter ratio has a number of competing effects. The length affects the development of the flow regime. Nariai[103] and Inasaka and Nariai[102] conducted experiments to study the effect of tube diameter, tube length and mass flux on CHF. Others evaluating the effect of L/D ratio on CHF include Boyd,[104] Bergles,[98] Ornatskiy,[105] and Cheng.[106] Many researchers have treated the diameter effects independent of other variables.[107,108,109]

Understandably, there have been many approaches developed to predict CHF.[93,96,97] These include phenomenological and pure statistical methods. From a predictive standpoint, the current state-of-the-art approaches include the method by Celata[110] and Liu.[111] Celata and Liu utilize a liquid sublayer model with minor differences. For Celata's model, about 91% of data points are predicted within +/-30% with a root mean squared error of 17.2%. Liu's approach predicts 98% of the data within +/- 35% with a root mean squared error of 13.4%. Current Neural network approaches predict approximately 77% of the data within +/-30% with a root mean squared error of 42.3%.[112]

Since flow boiling is a very complex phenomenon, the accuracies of classical methods for CHF are usually out of the 10% error band.[113] This is further exacerbated by the difficulty in obtaining high quality data. Several approaches claim to have high accuracy but are limited to a small range of operating conditions. The approaches listed above utilize a large number of data from varied sources. Unfortunately, other than a thermal evaluation of the data such as that performed by Mudawar,[95] no statistical tests on the different populations have been performed. Thus, it is difficult to compare models that do not have a large amount of data or cover a wide range of operation due to the fact that precision of the data can be quite different than for current large data sets. This is shown clearly in the work by Deng[114] where he compares the critical heat flux measured experimentally versus his predictive model for various datasets. Based on the figures provided in the paper and the calculated root mean square error, it is quite clear that there is a distribution in the precision of measurement in the various datasets.

Determining the critical heat flux for a given set of conditions is of utmost importance to a number of industries. This is demonstrated by the myriad of models and lookup tables available.[97,115,116] This problem is well suited for the Experimental Probabilistic Hypersurface (EPH). Essentially, there are seven parameters corresponding to length, diameter, inlet quality, outlet quality, mass flux, pressure, and whether the channel is horizontal or vertical. These input variables along with the corresponding critical heat flux value are used as the database used to construct the EPH for a given set

Table 3.4 Range of Values from Critical Heat Flux Database

	Pressure bar	Mass Flux kg/m^{2}s	Diameter mm	Length mm	Inlet Quality	Exit Quality	Heat Flux MW/m^{2}s	Reference # of Data
Low	1.034	9.9	1.016	25.4	-0.64	-0.28	0.11	[94]
High	111.4	18580	37.47	3657	0.14	1.577	24.41	1746
Low	30	179	3.9	1000	-2.39	-0.38	0.5	[117]
High	200	8111	25	4996	-0.02	1.1	4.949	625
Low	70	1961	9	840	-0.26	0	0.99	[118]
High		4173	12	3657	-0.01	0.47	4.31	235
Low	140	687	10	1.83	-0.46	0.19	0.96	[119]
High		1763			-0.03	0.52	4.08	25
Low	34	637	6.2	0.6	-0.86	-0.19	1.4	[120]
High	103	18577	37.5	1.97	-0.03	0.59	8.11	400
Low	3	1440	15.8	2.44	-0.33	-0.07	0.32	[115,121]
High	10	8110			-0.11	1	4.88	50
Low	10.13	5000	2	40	-0.95	-0.93	0.2	[122]
High	162	30000			0.072	0.264	120	392
Low	1	3060	4	250	-0.4	-0.07	5.2	[123]
High	31.2	27000			0.05	0.411	61.1	149
Low	1.1	10.8	9.7	43.54	-0.21	0.09	1.495	[124]
High	12	301.4			0.023	0.807	7.572	55
Low	6.27	2417	4.4	110	-0.19	-0.05	7.37	[98,125]
High	12.84	4567	6.16	154	-0.13	0.003	15.18	16
Low	101.3	491.7	10	250	-3.52	0	0.582	[126]
High	202.6	5542		2100	6.659	0.431	4.931	292
Low	19	776	1	239	-0.68	0.66	0.285	[127]
High	72	2736		975	0	0.99	2.363	83
Low	7.403	11240	2.5	241.5	-0.54	-0.18	12.11	[128]
High	38.27	36044			-0.2	0.056	59.14	70
Low	4.38	1111	12	22000	-0.68	-0.27	0.138	[129]
High	15.73	2529			-0.1	0.718	0.464	181
Low	39.24	550	8	100	-0.11	0.012	0.19	[130]
High	98.1	6445		666	-0	1.314	9.71	66

of conditions. The dataset used in the following example consists of 4385 unique data points from various researchers listed in Table 3.4. The range of data for each parameter is also provided.

All the data provides a large amount of information for constructing probability densities for queries of CHF conditions. The data found in Table 3.4 are not uniformly or normally distributed, which is expected when the data are difficult to acquire and the database is from multiple authors. Also, many of the variables in the dataset are correlated such as the inlet and outlet thermodynamic quality. This fact demonstratesthat the hypersurface that describes the function resides in a subspace of the overall 7 dimensional space. In order to gauge the accuracy of the methodology, the root mean square deviation is calculated using a leave one out approach. All of the data except one are considered known while the one data point is used for the query. First, the EPH is constructed using only the data itself and this will be compared with the construction using the embedding coordinates.

Based on experience from the previous section on friction factor estimation, only a few hundred points were needed for the methodology. Figure 3.10 is a plot of the RMSD versus the number of known points used in the construction of EPH with and without embedding coordinates. Over 200 randomly selected points are all that is needed to accurately predict new points with an RMSD of less than 1% using data randomly selected from the database. For new points within the range of the existing variables, the entire 5055 points could be used but it would be much faster to repeatedly calculate the predicted critical heat flux using subsets of 200 points and then selecting the mean of the resulting distribution of predicted values. This approach provides a highly accurate method for predicting the critical heat flux that does not require multivariate regression.

When calculating the RMSD values shown in Figure 3.10, the known data used in constructing the EPH are randomly selected. An interesting question arises regarding the ability to predict points outside the range of input variables. For example, the data shown in Table 3.4 is partitioned into two groups. The first group is all the points with a diameter greater than 4.4 mm which will be used as our known database of critical heat flux values. The second group with diameters less than or equal to 4.4 mm will be used to test the accuracy of the methodology for predicting points outside our known range. This is somewhat akin to extrapolating to new points.

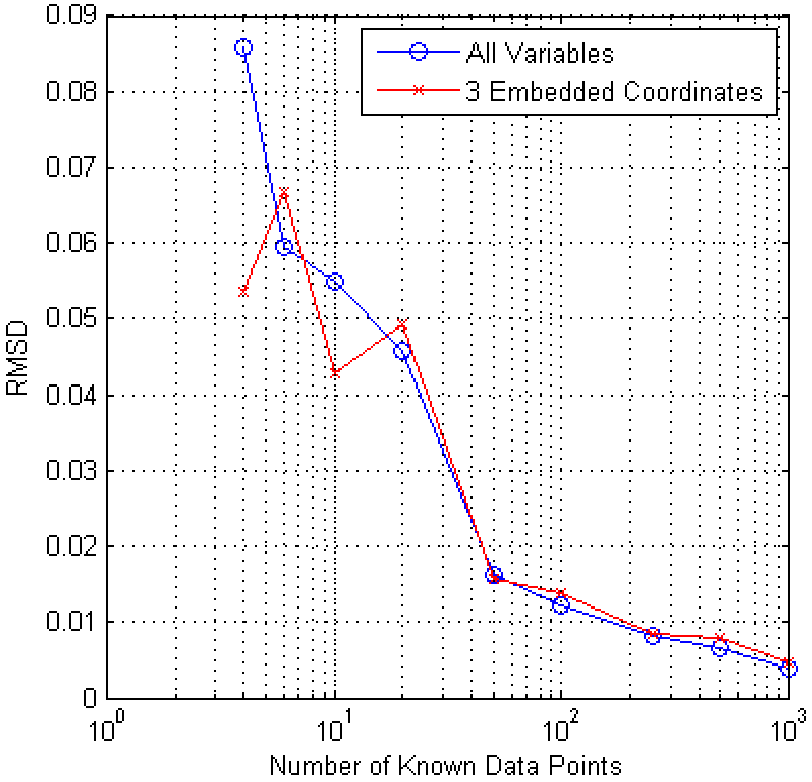


Figure 3.10 RMSD Values versus Number of Known Points for All Six Input Variables and for Three Embedding Coordinates.

4. MCNP

Computer codes are increasingly used to make predictions of complex systems. These codes may contain hundreds of input parameters that may be a mixture of continuous numeric, categorical numeric, and categorical text. The different types of variables along with the type of algorithm (deterministic or stochastic) provide a challenging problem to the proposed methodology. Further, the prediction codes are commonly used to predict that some phenomenon does not occur such as reaching the melt temperature in a nuclear fuel rod. Building on the local probability distributions shown in the previous section, the ability to create a global probability distribution is shown for real world problems. Thus, with a limited amount of computer processing one can provide important information regarding the probability of a specific phenomenon occurring.

4.1. Predicting MCNP Results with New Methodology

A Monte Carlo method is a computational algorithm that relies on repeated sampling of random inputs to compute results. This approach tends to be used when it is infeasible or impossible to compute an exact result with a deterministic algorithm (particle transport, Brownian motion, etc.). Also, the large number of computations needed to produce an accurate result requires a significant amount of computing time requiring high fidelity problems to be solved on high end computing resources. The Monte Carlo N-Particle transport (MCNP) code developed at Los Alamos National

Laboratory is an important tool in the design and evaluation of nuclear systems. Monte Carlo simulations are a useful tool to perform searches for critical configurations as well as shielding and other radiation effects analysis.[131] The interesting aspect of criticality searches is that it is typically whether a configuration or set of configurations is critical or not. The ability to determine the global probability from a small number of measurements for a complex system would greatly decrease the computational costs and allow more configurations to be quickly evaluated.

A simple MCNP model based on examples for a square lattice of plutonium nitrate filled cylinders found in the MCNP Criticality Primer was developed and is shown below in Figure 4.1.[132] The input variables correspond to the liquid level, cylinder pitch, cylinder diameter, and cylinder wall thickness, which are continuous and whether a cylinder is present in the 3 x 2 square array. Stated more clearly, each cylinder location is assigned a variable whose value is zero or one depending whether the location has a cylinder or not. The material cards for the plutonium nitrate solution and the cylinder walls were not changed for the example but could be adjusted to provide more input variables. Since some of the variables are continuous, an infinite number of input configurations are possible. The table on page 98 presents the values for the input variables used for the case presented, the multiplication factor was calculated for seven liquid levels, four cylinder diameters, five cylinder pitches, and three wall thicknesses. These were used as inputs along with varying the number and location of cylinders in the array. The number of possible configurations is infinite due to the continuous variables but considering that a finite set of input variables were used, the number of possible configurations is 6720 (7x4x5x3x2x2x2x2). The average computer time for each run was in excess of 11 minutes so calculating this drastic reduction in potential states still produces a significant computational task of over 51 days.

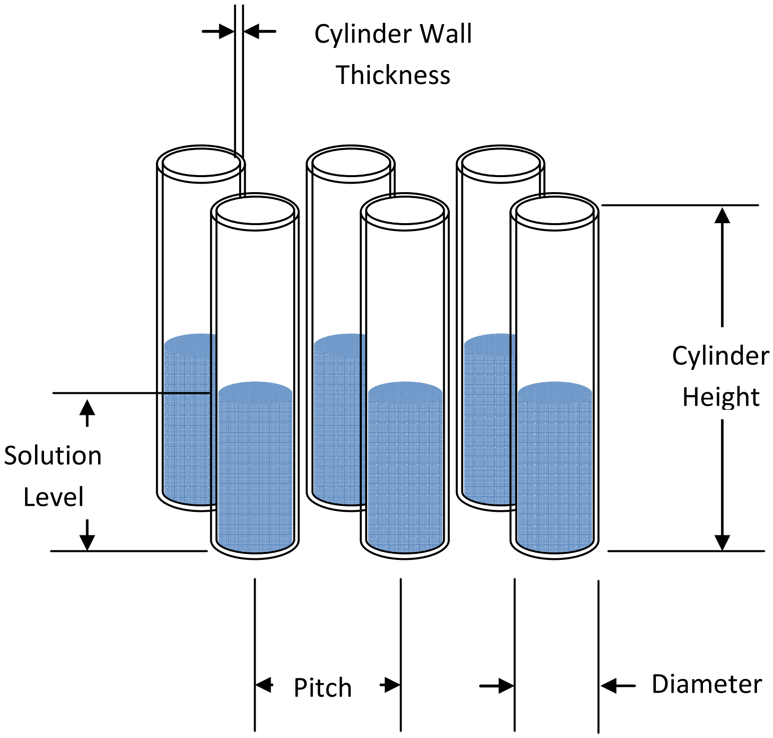


Figure 4.1 Schematic of System Modeled in MCNP.

Similar to the calculations done for the friction factor, the multiplication factor, k, was predicted using the methodology. The results are shown below in Figures 4.2 and 4.3. Figure 4.2 is a plot of the probability of possible values of k_{eff} along with the actual values of the query point and the mean of the probability distribution for 40 randomly selected known data points. The probability distribution is quite peaked and the corresponding mean of the distribution closely matches the actual value of the query. This is indicative that a number of known points are nearby; thus, we have a higher confidence in the result. Figure 4.3 is a second calculation for query value farther from known points. The probability distribution is more uniform and one has less confidence in the resulting mean probability. This is shown as a greater difference between the actual and the value based on the mean probability. The selection of 40 points was

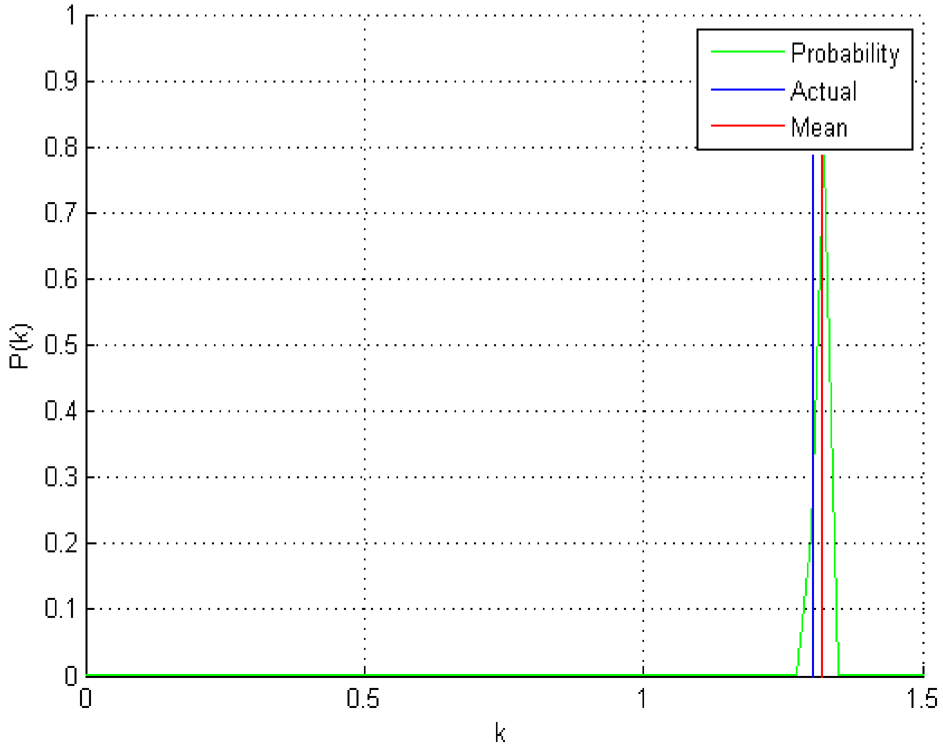


Figure 4.2 Plot of Probability Distribution (green), Mean of Probability Distribution (red), and the actual Value of Query (blue) for 40 Randomly Selected Known Data Points with Query Closer to Known Data.

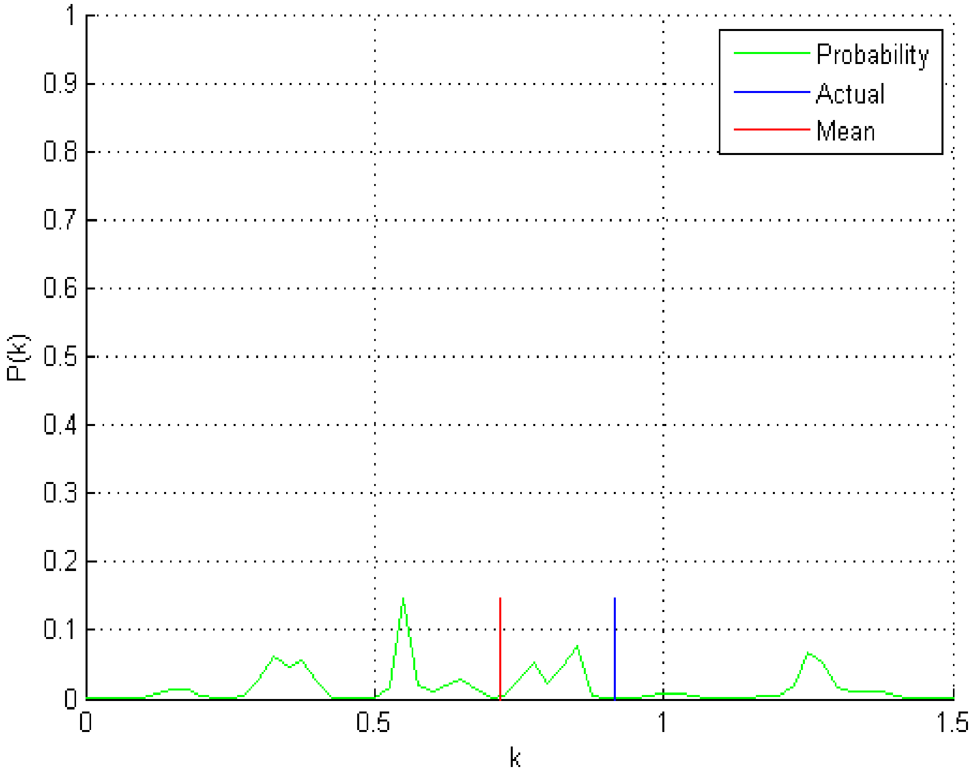


Figure 4.3 Plot of Probability Distribution (green), Mean of Probability Distribution (red), and the actual Value of Query (blue) for 40 Randomly Selected Known Data Points with Query Far from Known Points.

arbitrary. The dependence of the RMSD for all points in the database on the number of known data points is shown in Figure 4.4. Based on the RMSD error values shown in Figure 4.4, give the expected accuracy of using 40 known points to predict the multiplication factor to within 0.18 which is quite coarse for criticality work. One would need to use more points or only make queries close to known values of the multiplication factor to improve results. This is shown in Figure 4.5 where the probability distribution is constructed using 600 known points. The number of known points needed to predict the multiplication factor to within 0.05 is approximately 600 which is less than 9% of the possible combinations.

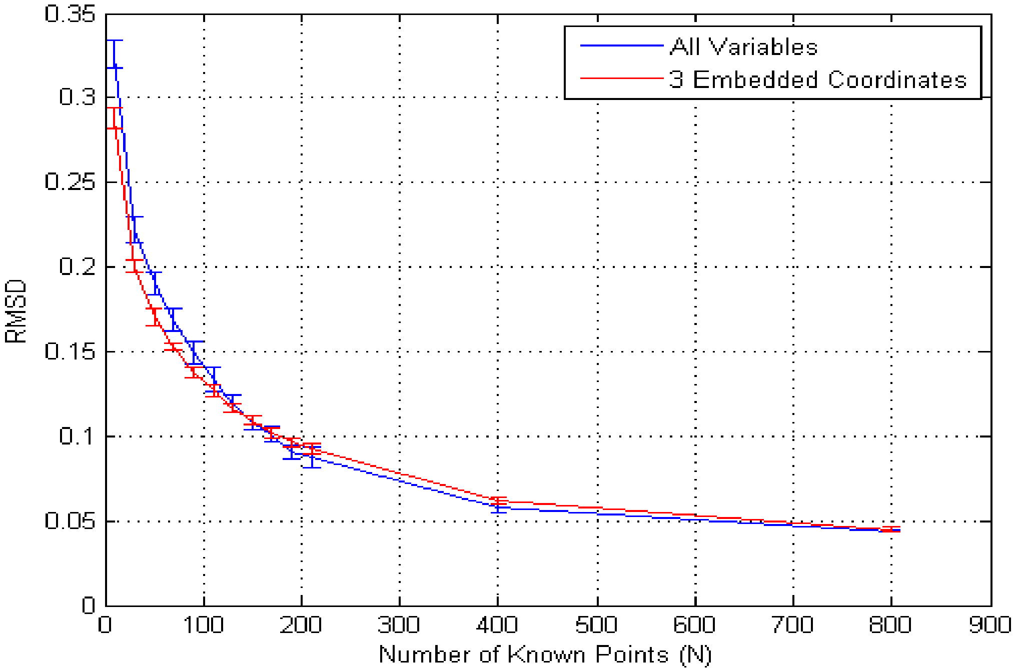


Figure 4.4 Plot of RMSD versus Number of Known Points for All Variables (blue) and for 3 Embedded Coordinates (red).

Similar to the results shown in the previous section, the methodology does quite well in predicting the actual result. The point of interest is that this methodology works well in predicting a stochastic calculation that includes categorical and continuous data. However, criticality engineers are interested in the potential for a system to achieve a critical or supercritical configuration where the multiplication factor is greater than or equal to 1. For the model presented above, a criticality analysis would determine the multiplication factor for a specific solution level, cylinder diameter, wall thickness, and cylinder pitch. Typically, an MCNP model would be developed and executed to determine the multiplication factor. The model would then be altered and executed again to evaluate the multiplication factor with different values for the input variables. An alternative analysis may look at the maximum solution level that remains subcritical. Each of these approaches provides information for a single calculation point and

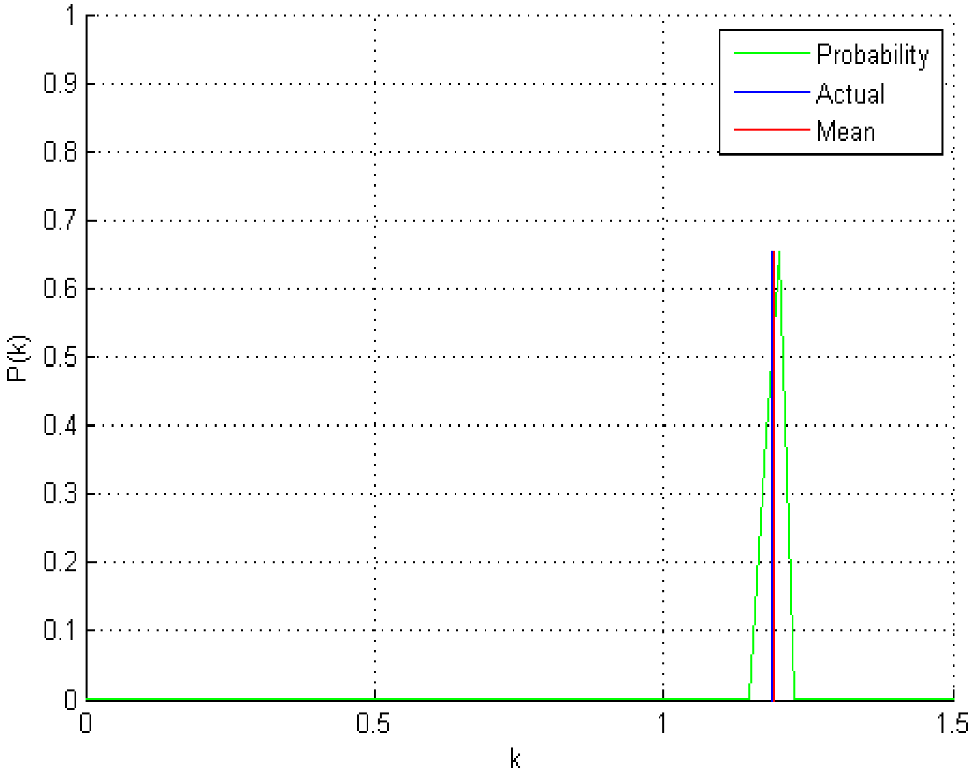


Figure 4.5 Plot of Probability Distribution (green), Mean of Probability Distribution (red), and the actual Value of Query (blue) for 600 Randomly Selected Known Data Points with Query Far from Known Points.

previous results are not directly used.The methodology outlined earlier produces information for all possible configurations using past results. Thus, one can calculate the probability of achieving criticality (or any other condition) for a range of configurations in one calculation using existing MCNP output.

Table 4.1 Values Used in the Input Deck for the MCNP Model Shown in Figure 4.1.

Parameter	Range
Pin Diameter (cm.)	5,10,20,25
Pin Pitch (cm.)	2,5,10,15,20
Pin Wall Thickness (cm)	0.01,0.1,1
Solution Level (cm)	2,10,20,30,50,60,75
Pin 1	Present 1 - Missing 2
Pin 2	Present 1 - Missing 2
Pin 3	Present 1 - Missing 2
Pin 4	Present 1 - Missing 2
Pin 5	Present 1 - Missing 2
Pin 6	Present 1 - Missing 2
Note: Always 2 pins present for any calculation	

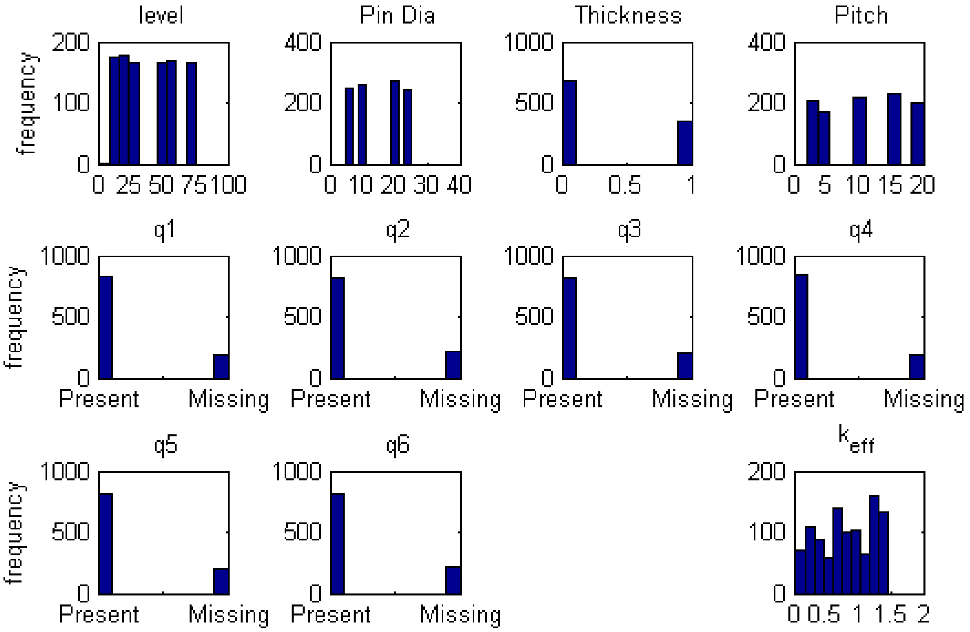


Figure 4.6 Histograms of MCNP Input Deck Variables and Resulting Values of k_{eff}.

To determine the probability that one could achieve criticality based on the range of configurations listed in Table 4.1, hundreds of MCNP calculations were carried out to produce a database of multiplication factors for random input parameters. The values of the input parameters are shown in Table 4.1 and were uniformly sampled as shown in Figure 4.6. For the 1022 conditions evaluated, there were 385 points that had a multiplication factor greater than or equal to1. Since the points are uniformly sampled, the expected probability of k being greater than 1 for all configurations should be close to385/1022 or 0.3767.The global probability was calculated for the system described above and is demonstrated below. Figure 4.7 is a plot of the probability as a function of the number of known data points for any configuration of parameters listed in Table 4.1 that would result in a) a multiplication factor greater than 1.5 or b) a multiplication factor greater than 0. To calculate this global probability, the approach used by Zeydina[33] is followed. As expected, this value is nearly zero for points above 1.5 since we have no information that any configuration can achieve a multiplication factor greater than 1.5 and since the multiplication factor must be nonnegative, the probability that keff is greater than 0 is nearly 1.

An interesting aspect of these two lines in Figure 4.7 is the relatively small amount of information (less than 150 points) needed to provide accurate information regarding the global probability. This represents a significant improvement in computational time.

As mentioned previously, criticality analysis is interested in the probability that a system could achieve a critical configuration. This is somewhat of a misnomer since most of the time an analysis involves a high fidelity calculation of the exact system that produces a specific result. How does one take advantage of the extensive modeling and computation effort that has taken place previously? Using the methodology, one can make accurate predictions using a minimum number of previous results. The following example illustrate this.

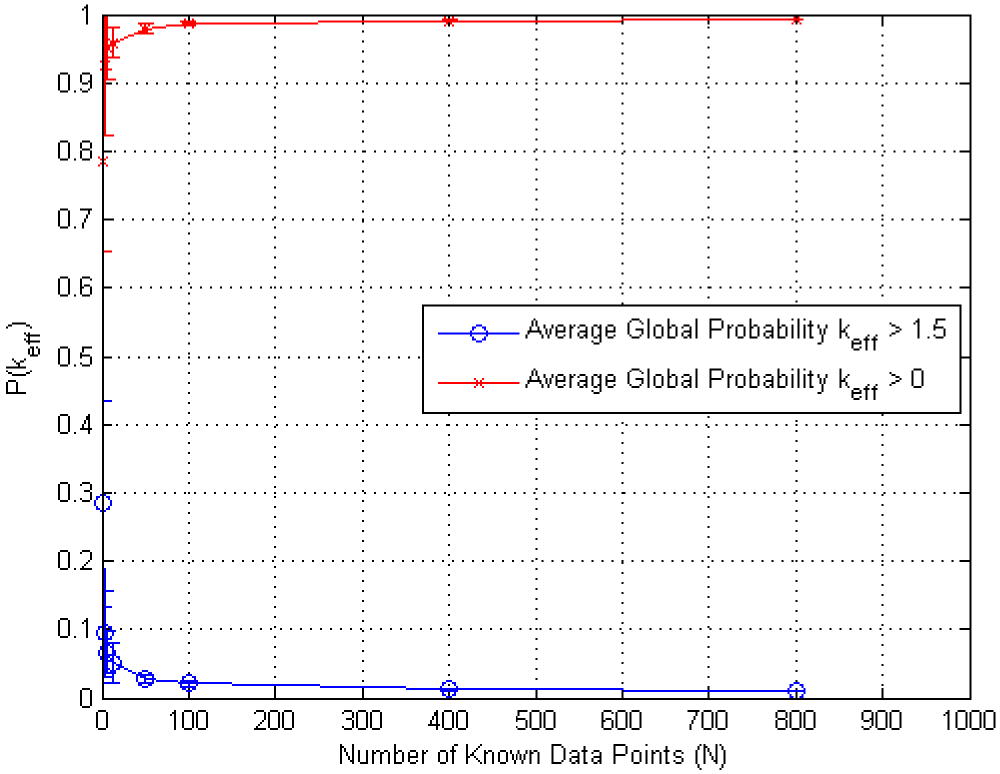


Figure 4.7 Probability Plot of Mean Global Probability versus the Number of Known Data Points for $k_{eff} > 1.5$ (blue) and for $k_{eff} > 0$ (red).

Figure 4.8 is a plot of the global probability that any configuration outlined in Table 4.1 results in a multiplication factor greater than 1. The global probability versus the number of known data points is plotted along with the ratio of critical or supercritical configurations over all calculated configurations. The global probability approaches the ratio indicating that approximately 100 known points/previous calculations are needed to describe the whole configuration space.

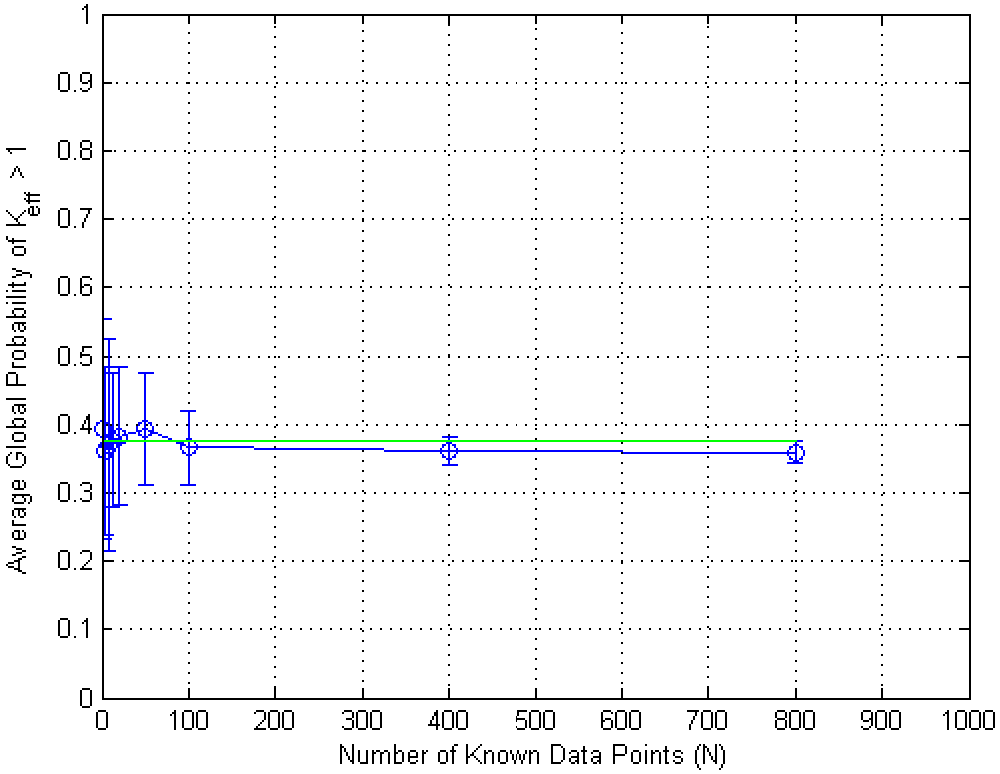


Figure 4.8 Plot of the Mean Global Probability for $k_{eff} > 1$ versus the Number of Known Data Points (blue) and the Fraction of Values of k_{eff} for All Runs Performed.

5. HIGH DIMENSIONAL DATA

The previous sections demonstrated the power of the methodology with simple, low dimensional examples. The following section now applies the methodology to high dimensional data sources. The five to ten variables needed to describe pipe friction or a criticality problem in previous sections may not seem daunting but consider the grayscale image shown in Figure 5.1 that consists of 640 x 480 pixels. These images are used in the determination of flow regime for co-current two-phase flow. Each image is a data point in $\mathbb{R}^{640x480}$ or each data point has a possible $256^{640x480}$ configurations. However, observation of the images indicate that there is some underlying structure commonly referred to as a regime that indicate the data lie on a manifold of lower dimension than the original space of $\mathbb{R}^{640x480}$. Methods to extract this manifold for use in prediction would eliminate the subjective manner flow regimes are identified. The focus of this section is to apply the methodology described in Section 2 and demonstrated on the relatively simple problems in Sections 3 and 4 to more challenging problems with thousands to millions of variables more commonly found today.

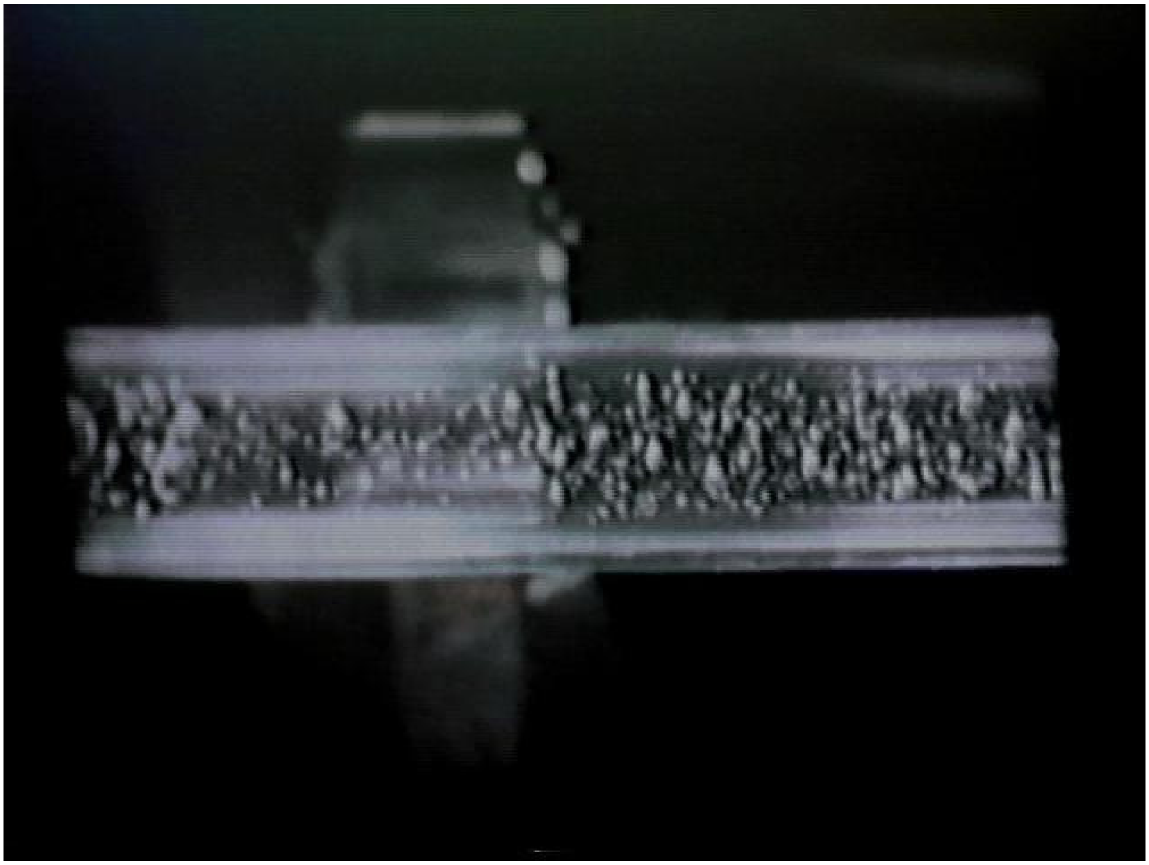
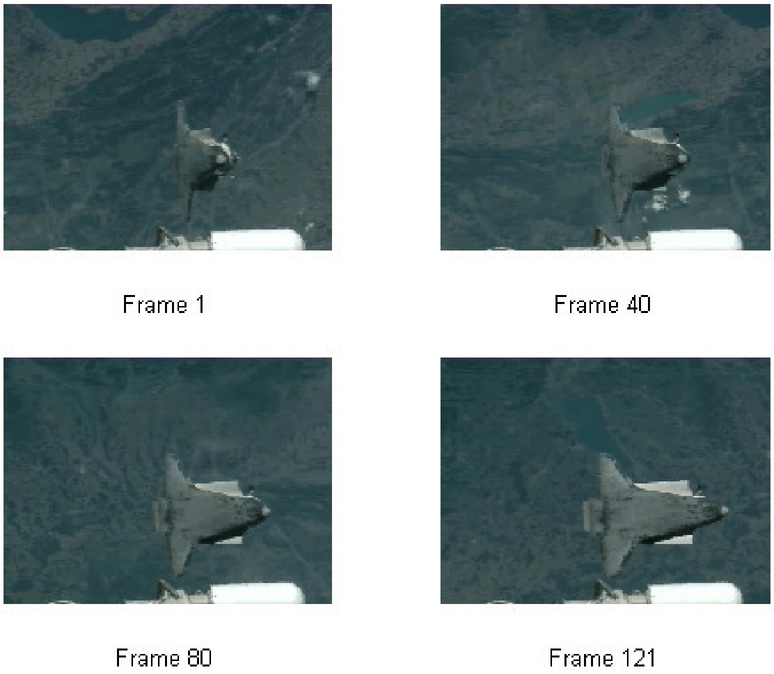


Figure 5.1 Image of Microgravity Two-Phase Flow.

5.1. Space Shuttle Rotation

A video taken from the International Space Station (ISS) of the space shuttle performing a pitch maneuver during STS-114 is used as the data source to demonstrate the ability to perform regressions using visual data. The pitch maneuver is performed to allow the visual inspection of the thermal protection tiles to determine if there has been damage during takeoff. The video consisting of 121 frames was taken as part of an inspection program to examine the protective thermal tiles and involves the orbiter performing a rotation of 60 degrees at a relative constant rate. Part of the inspection process requires knowledge of the exact angle to properly analyze the imagery.

This video provides a challenging example to feature recognition algorithms due to the changing background associated with the movement of orbiter about the earth. A sequence of images corresponding to the range of angles is shown in Figure 5.2. Identifying the pixels that best describe the orbiter pitch angle is very difficult due to the small number of test cases, the large number of total pixels, and the low signal-to-noise ratio produced from the changing background. These high dimensional, low sample size examples are increasingly found in engineering due to advances in digital technology.



Frame 1

Frame 40

Frame 80

Frame 121

Figure 5.2 A Selection of Four Frames from the Entire Range of the STS-114 Pitch Video.

The known images are dimensionally reduced using random projections with a normally distributed random matrix. The resulting vectors are much smaller in size and serve as input into the EPH. A query image is also dimensionally reduced using the same matrix. The discritization of angle is ~0.25 and the most probably angle is the discritized location closest to the mean of the probability distribution. The original images in the video are 240 x 320 pixels which correspond to 76800 parameters or variables. As a rule of thumb, this space can be collapsed to $C \cdot \log_{10}(76800)$, where C is a constant of 4 to 8.[35] Figure 5.3 is a plot of the accuracy of the technique for various sizes of random projections and number of known images in the database. The accuracy increases with more known information and 30 random projections provide enough information to make accurately predictions. This corresponds to C = 6 which corresponds to the values found in literature.

The effective size of the dimension reduction matrix in Figure 5.3 where the error between the actual angle and the predicted angle versus the number of known images in the database for several dimensions of the collapsed image is shown. Figure 5.4 is a plot of the average RMSD error of the orbiter angle versus the number of known images in the database. The plot indicates that 64 known images yields an accuracy of approximately 1.5 $^{+}$/- 1 degrees.

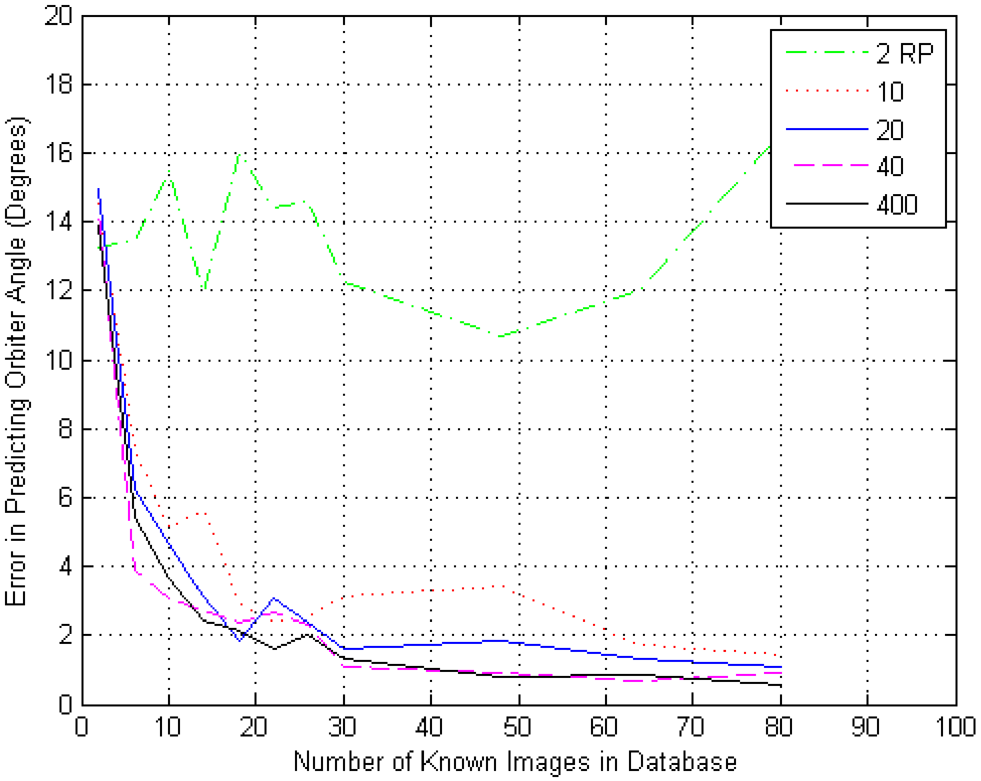


Figure 5.3 Error in Predicting Orbiter Angle versus Number of Known Images in Database for Various Size of Random Projections.

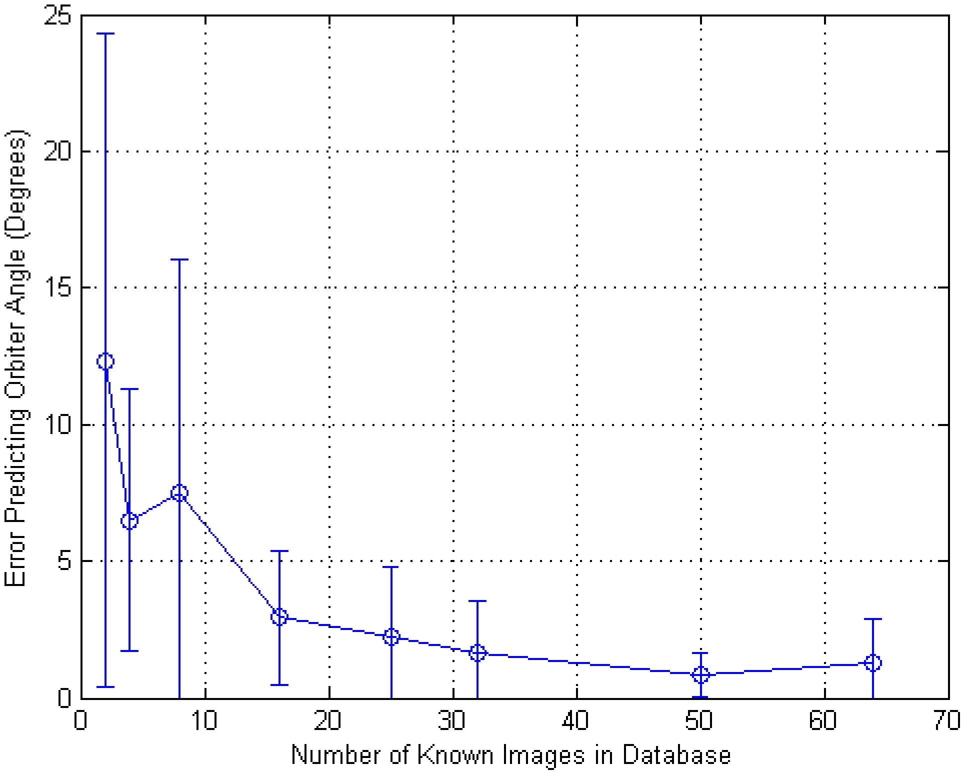


Figure 5.4 Average RMSD Error versus Number of Known Images in Database Using 30 Random Projections.

Looking further, the random projected images can be embedded into a new space using techniques described earlier. This transformation can further reduce dimension and orients the images for accurate prediction. Figure 5.5 shows the results using the first 2 embedded coordinates as inputs into the EPH. The 30 random projections are produced using the normally distributed random matrix discussed in Section 2. The accuracy is essentially the same as using only the random projections.

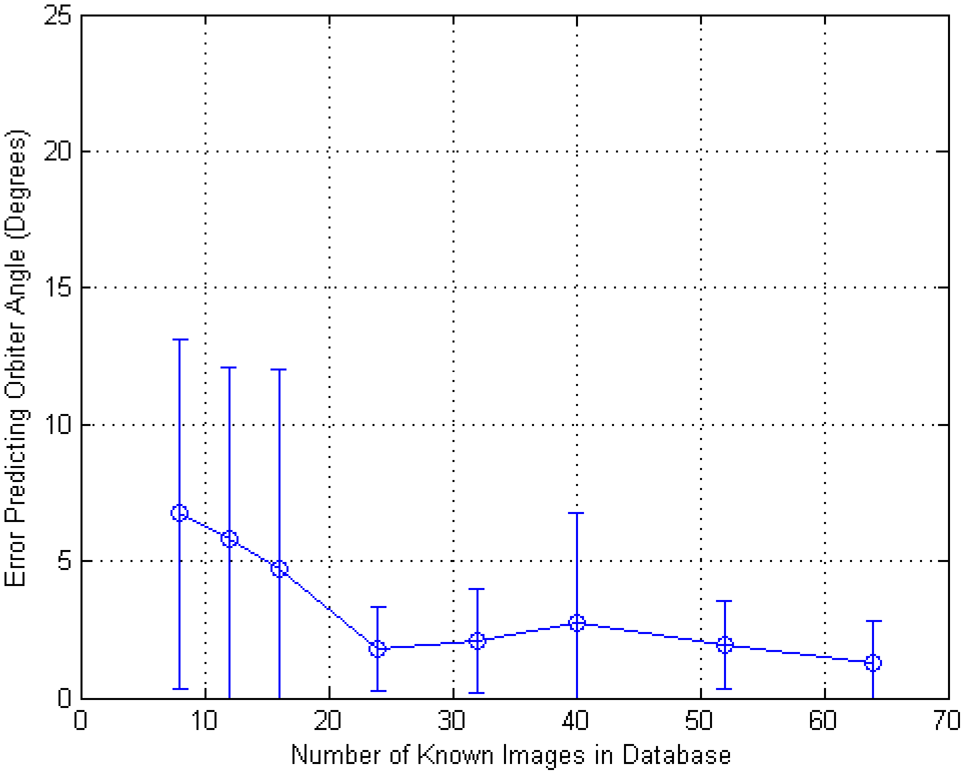


Figure 5.5 Average RMSD Error versus Number of Known Images in Database Using 2 Embedding Coordinates from 30 Random Projections.

As discussed previously, other matrices can be used to project the image onto a lower dimensional space. Using a discrete sine transform (DST) similar to the one described by Amador[42], the images were projected onto a space with size 1x30. The average error was again calculated similar to what was done for Figure 5.5. According to Amador,[42] the DST is better able to pack information into the projections and the resulting embedding more accurately reflects the rotation of the orbiter. The result is shown in Figure 5.6 where the error for 64 known images is 0.7 $^{+}$/- 0.3 degrees. The DST projections produce a better result requiring fewer known images in the database.

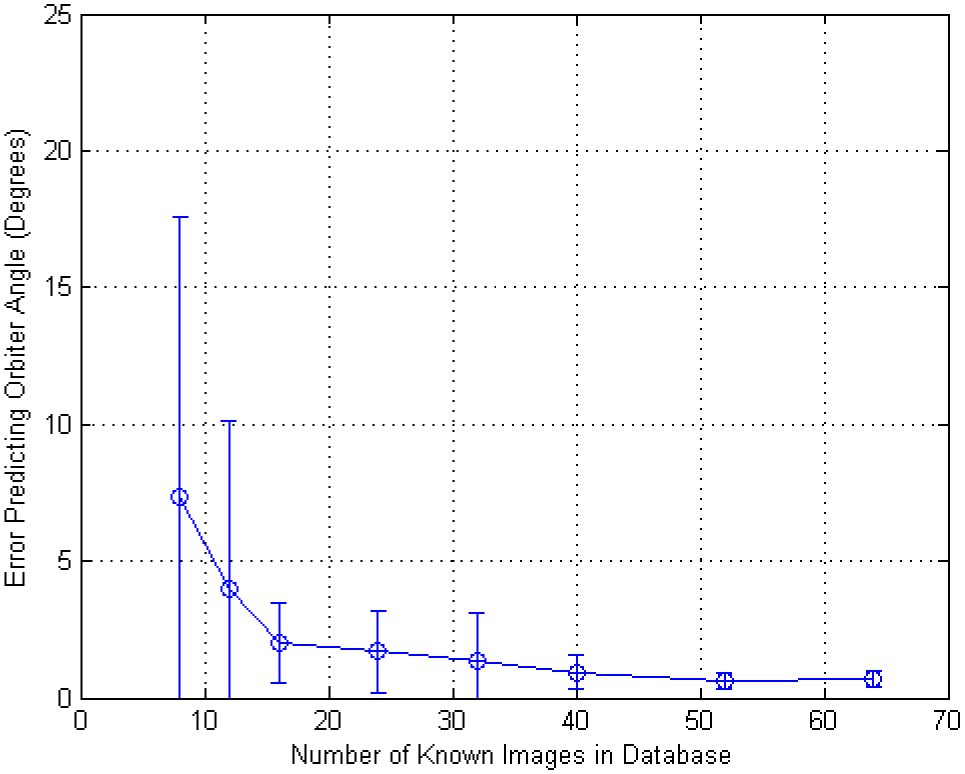


Figure 5.6 Average RMSD Error versus Number of Known Images in Database Using 2 Embedding Coordinates from 30 DST Projections.

Using the results from Figure 5.6, a prediction of orbiter angle is made using 20 known images and projecting them and the query image using a discrete sine transform matrix onto a vector of 30 elements. These vectors are then embedded using the first two diffusion coordinates. The prediction is based on the mean probability produced by the EPH using the 2 diffusion coordinates as inputs. Figure 5.7 consists of the query image, the probability plot of possible angles with the mean probability and the actual value of the query, and the resulting image based on the mean probability.

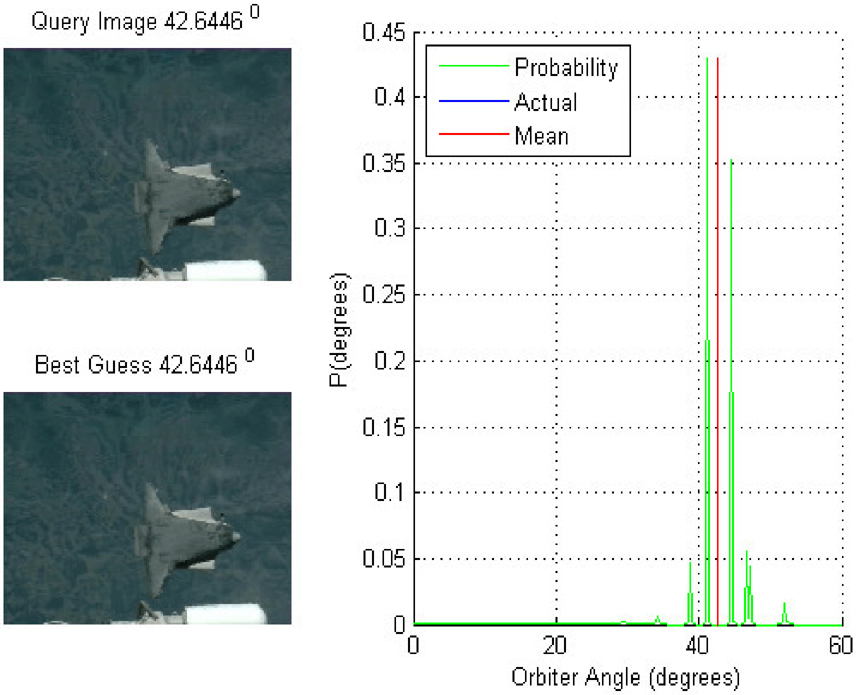


Figure 5.7 a) Query Image, b) Guess Image Based on Mean of Probability Distribution, c) Probability Distribution (green), Mean of Probability Distribution (red), and Actual Angle (blue).

Figure 5.7 indicates that the methodology can accurately predict the orbiter angle given enough information. To see what points were selected, plots of the embedding coordinates are shown in Figure 5.8 for both the known and query images. The first plot in Figure 5.8 is the two diffusion coordinates that are used as inputs to the EPH. One can see an ordering of the 20 randomly selected known images and their relation to the query image. The query is located near similar images, which is further shown in the second plot. The second plot is first coordinate, Θ_1, versus the rotation angle of the orbiter in degrees. One can see the query image is located at the correct angle nearest the similar images corresponding angles.

The shuttle rotation example illustrates the methodology's ability to reduce dimension and embed high dimensional data and use this information to make accurate predictions when provided with new data. The implications of this are threefold: 1) The reduction in dimension allows for faster calculation, 2) embedding random projections identifies key features in a noisy image, and 3) EPH provides a method for adding new data to the embedding and making accurate predictions using given information.

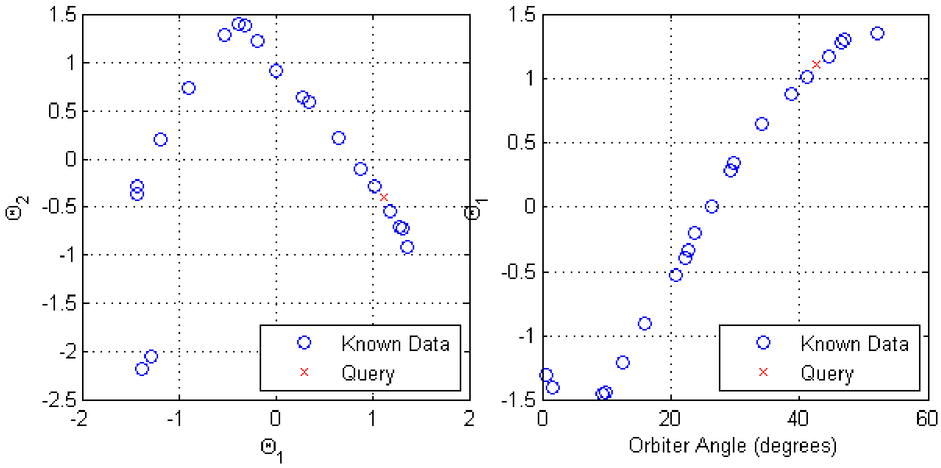


Figure 5.8 a) Plot of the First Two Diffusion Coordinates for the Known and Query Images b) Plot of the First Diffusion Coordinate versus the Orbiter Angle for the Known and Query Images.

5.2. Spectrum Data

Gene expression or protein charge data are a classic candidates for dimension reduction. Typical mass spectrometry datasets consist of thousands of parameters. Recent tests performed at Texas A&M investigated a novel perfusion system to examine the effects of radiation on model respiratory tissue.[12] The data consists of three sets of three samples of intensity readings for 10399 channels that correspond to in-vivo unirradiated, ex-vivo unirradiated, and ex-vivo irradiated classes. The 10399 channels correspond to approximately 9028 gene probes with the difference in channel number corresponding to bacterial transcripts and other pads. The nine samples were randomly compressed and embedded into a diffusion space which is shown in Figure 5.9. The result is three clusters that correspond to the different classes.

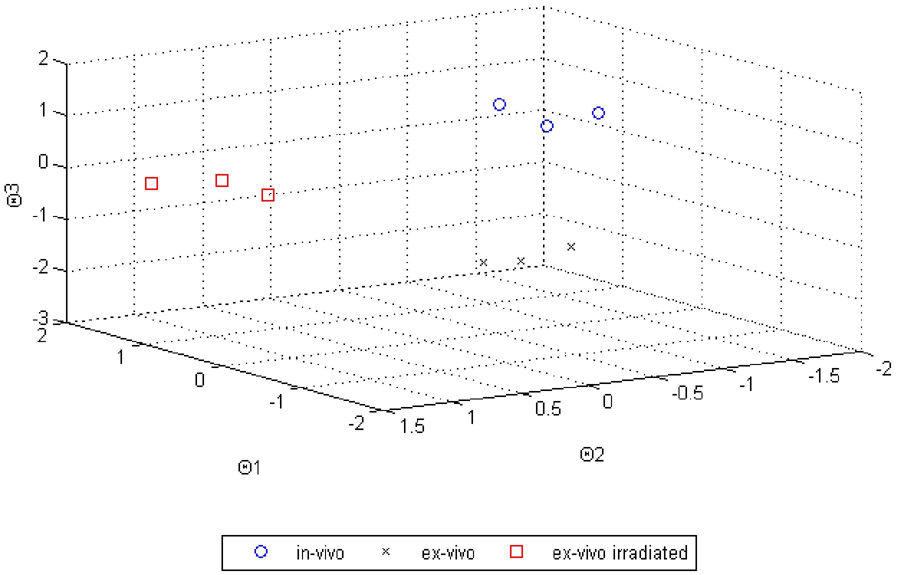


Figure 5.9 Spectrum Data Plotted in First Three Embedding Coordinates.

Based on the results above, a similar approach was used with spectrum data (SELDI-TOF data) to detect the presence of ovarian cancer. The spectrum data has 15154 channels or parameters corresponding to size and net electrical charge of proteins. A typical spectrum from the dataset by Lanclet[133] is shown in Figure 5.10. The discriminating pattern formed by a small key subset of proteins or peptides is buried among the entire ensemble of thousands of proteins represented in the sample spectrum.[133] Only specific mass/charge (M/Z) positions/channels/parameters along the spectrum horizontal axis are used for the discrimination of ovarian cancer. Identification of these parameters forms a significant research area and typically requires an extensive amount of training data.[134,135,136]

Each spectrum of the training dataset similar to the one represented in Figure 5.10 is randomly projected onto a lower dimensional space using either a matrix of normally distributed random entries, a matrix used for discrete sine transforms, or a matrix used for noiselet transforms. A small number of dimensions are then embedded into a diffusion space. An example is shown in Figure 5.11 where the original spectrum of 15154 channels is projected onto a 256 element vector using a random matrix. The resulting embedded data shows relatively nice separation between the cancer and control data. This is for the first three diffusion coordinates, all that can be shown in a 3D figure. Further separation of the two groups occurs as more diffusion coordinates are included. Using the diffusion coordinates as inputs to the EPH, one can then project whether each spectrum (patient) belongs to the cancer or non-cancerous group. The accuracy of the

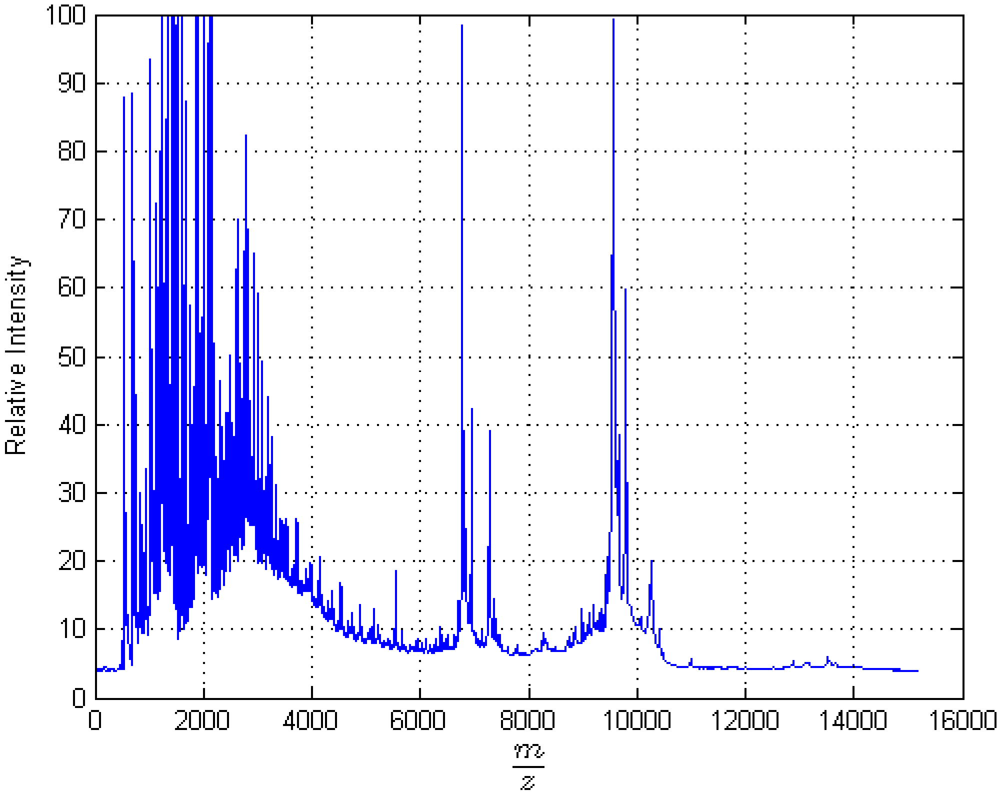


Figure 5.10 Sample Spectrum of SELDI-TOF Data.

results is shown in Figure 5.12. The plot shown in Figure 5.12 relates the accuracy of predicting cancer versus the number of diffusion coordinates for various size random projections. It is clear that the amount of information contained in the projected vector increases with increasing size and that approximately 20 diffusion coordinates are all that is needed to produce the maximum accuracy. Thus, the original 15154 element spectrum can be projected onto a space of 128 and embedded into 20 coordinates to produce an accuracy of greater than 99% for detecting cancer.

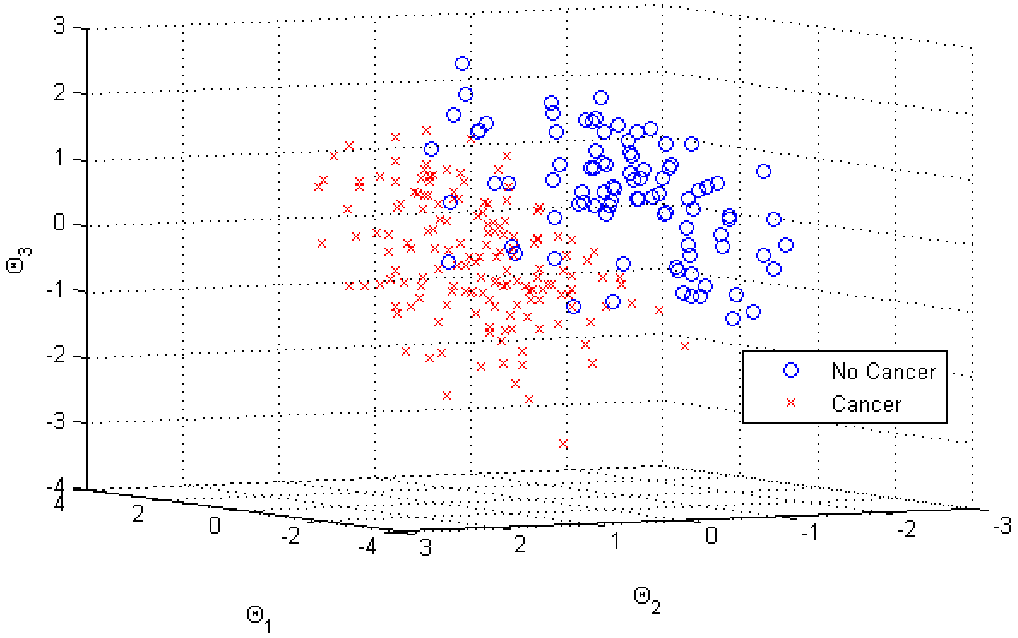


Figure 5.11 The First 3 Diffusion Coordinates of SELDI-TOF Data Randomly Projected Onto a Vector of 256 Elements.

As shown in the orbiter video in the previous section, other matrices can be used to project the original data. The spectrum data is projected using a discrete sine transform matrix resulting in a more accurate prediction with fewer number of diffusion coordinates from a smaller input vector. Figure 5.13 is a plot of the first three diffusion coordinates for data that has been projected onto a vector of 256 elements using a discrete sine transform matrix. The DST matrix is able to capture more information about the spectrum than the random matrix resulting in faster determination of the presence of Ovarian cancer which is shown in Figure 5.14 with the high prediction accuracy. The dimension reduction of the large original signal indicates that cancer information resides in a subspace of much smaller dimension that may only require a few channels compared to the total recorded.

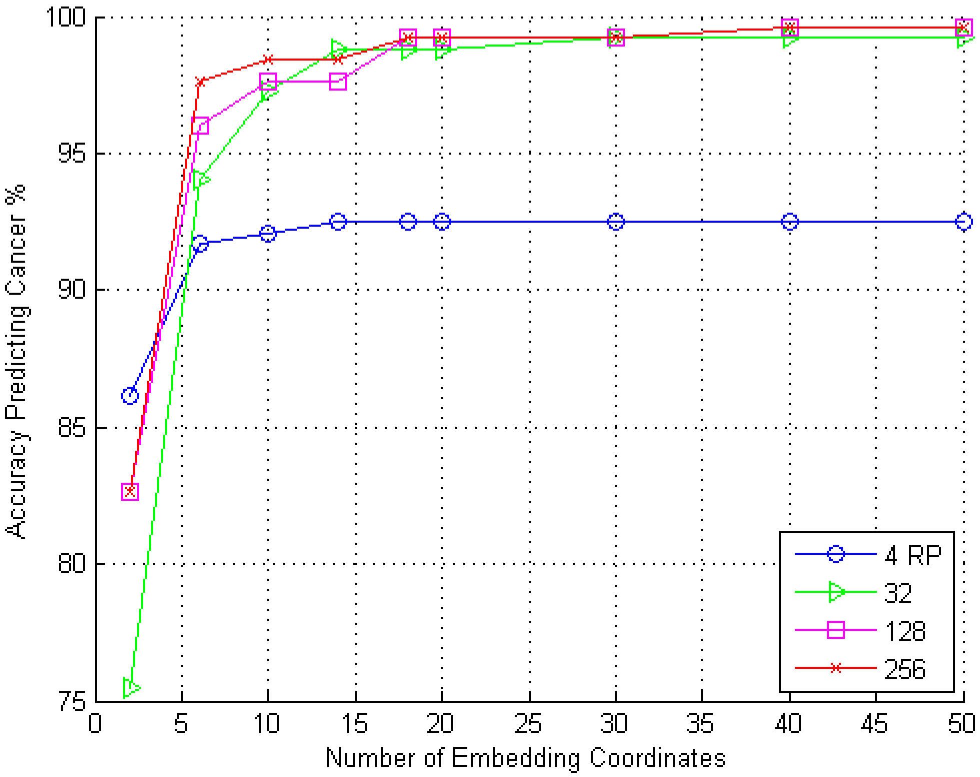


Figure 5.12 Cancer Prediction Accuracy as a Function of Number of Diffusion Coordinates for Various Size Random Projections.

A noiselet transform is another matrix of special interest due to the fact that (1) they are incoherent with systems providing sparse representations of image data and other types of data, and (2) they come with very fast algorithms; the noiselet transform runs in O(n) time, and just like the Fourier transform, the noiselet matrix does not need

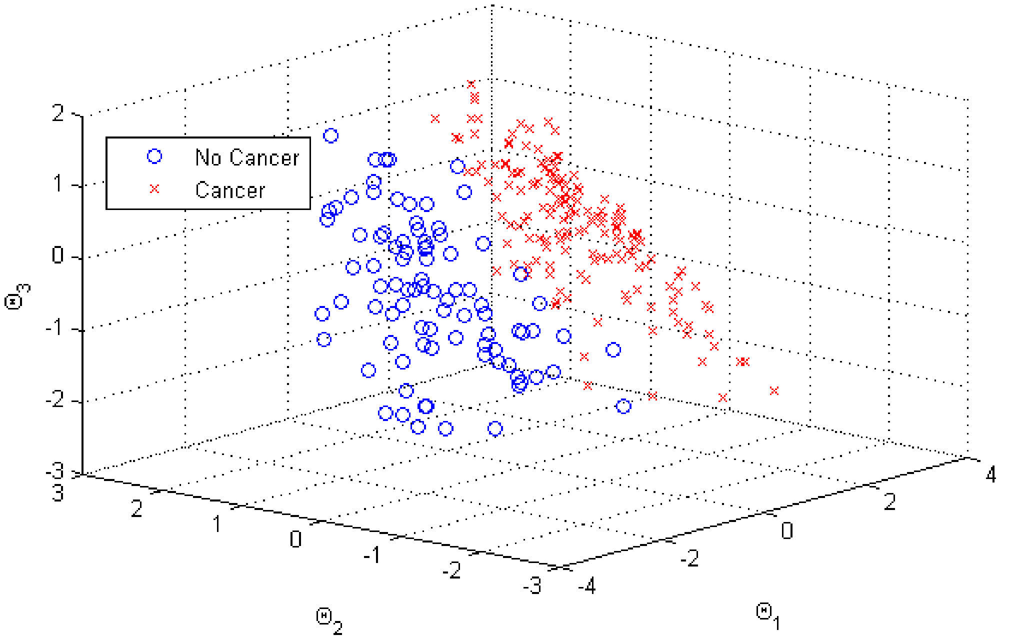


Figure 5.13 The First 3 Diffusion Coordinates of SELDI-TOF Data Projected Using a DST Matrix Onto a Vector of 256 Elements.

to be stored to be applied to a vector.[137] Figure 5.15 is a plot of the first three diffusion coordinates for both the cancer and non-cancer data. Again, as in Figures 5.11 and 5.13 a separation of cancer and non-cancer data is shown with just three diffusion coordinates. Figure 5.16 is the resulting plot of accuracy versus the number of diffusion coordinates. Only 15 diffusion coordinates are needed to predict cancer with greater than 99% accuracy. In fact, both the DST and noiselet basis produce 100% correct identification of cancer using 30 diffusion coordinates produced from projection vectors that are 256 elements long. No clustering algorithm such as k-means is used. The diffusion coordinates are used as inputs to construct the Experimental Probabilistic

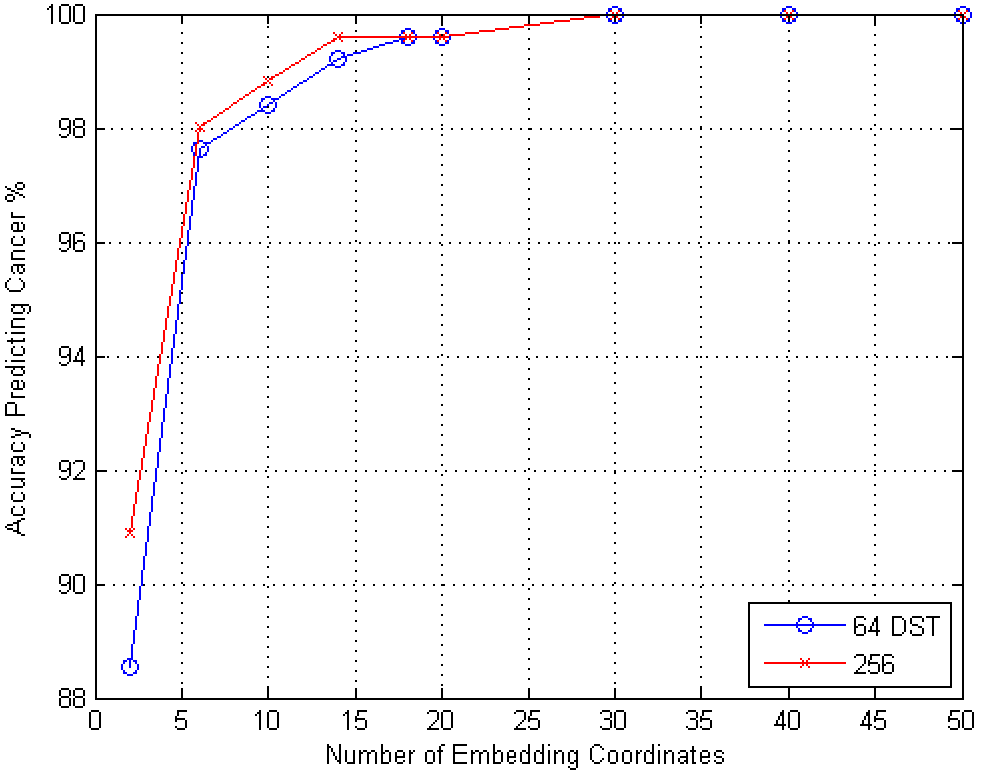


Figure 5.14 Cancer Prediction Accuracy as a Function of Number of Diffusion Coordinates for Various Size DST Projections.

Hypersurface. The total number of known data is 255 for each query with corresponding output values assigned 1 or 2 according to no cancer or cancer respectively. For each query, the resulting probability distribution was computed from the EPH and the mean taken producing an assignment to either non-cancer or cancer based on how close the mean probability is to 1 or 2 respectively. A sample probability plot is shown in Figure 5.17.

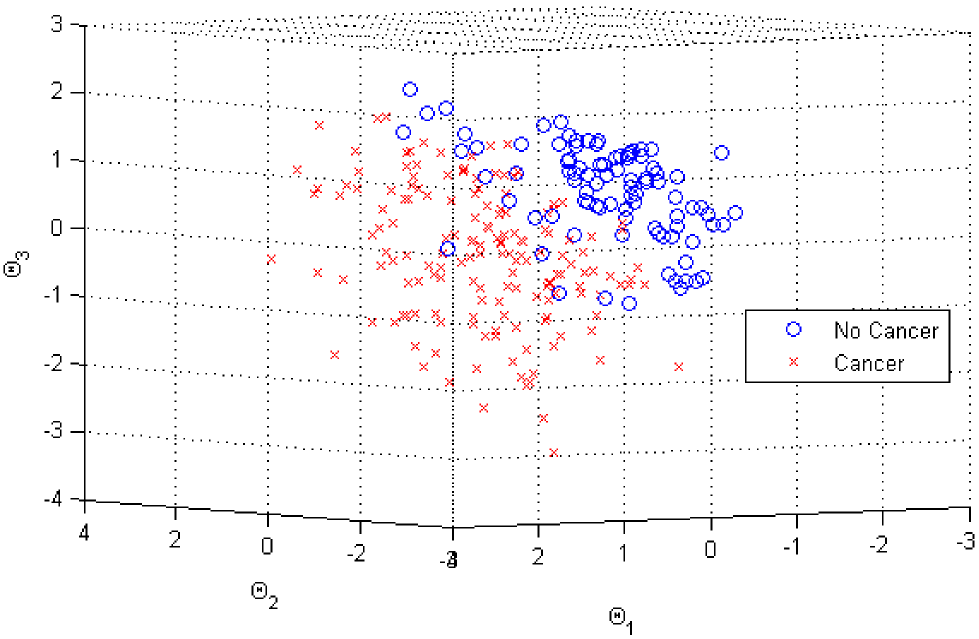


Figure 5.15 The First 3 Diffusion Coordinates of SELDI-TOF Data Projected Using a Noiselet Matrix Onto a Vector of 256 Elements.

The probability plot shown in Figure 5.17 shows the peak located over the cancer category along with the corresponding mean of the distribution which matches the actual label for the data. The discritization of the possible values is 0.33 and the range is 0.66 to 2.33. Thus, the probability is jagged since values are only calculated at 0.66, 0.99, 1.32, 1.65, 1.97, and 2.30. This results in the peak for cancer being located at 1.97 rather than 2 but this has no effect on the resulting classification. This approach is quite fascinating; not only does the accuracy rival the latest methods for prediction without false positives but it can provide an actual probability value rather than a binary yes/no result.

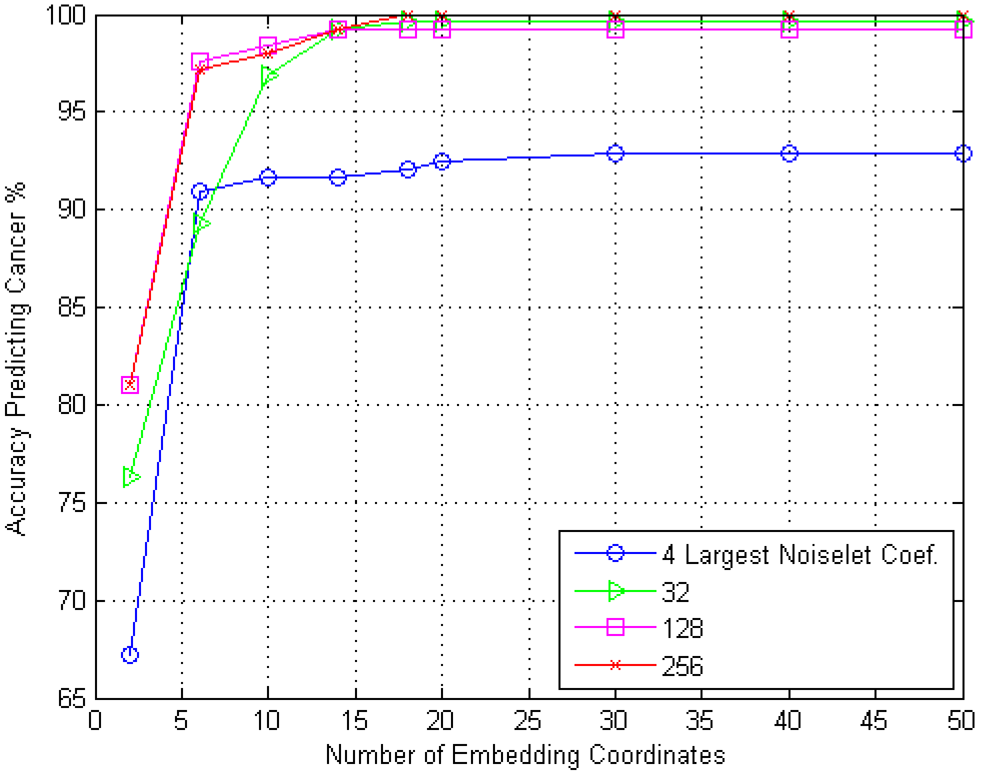


Figure 5.16 Cancer Prediction Accuracy as a Function of Number of Diffusion Coordinates for Various Size Noiselet Projections.

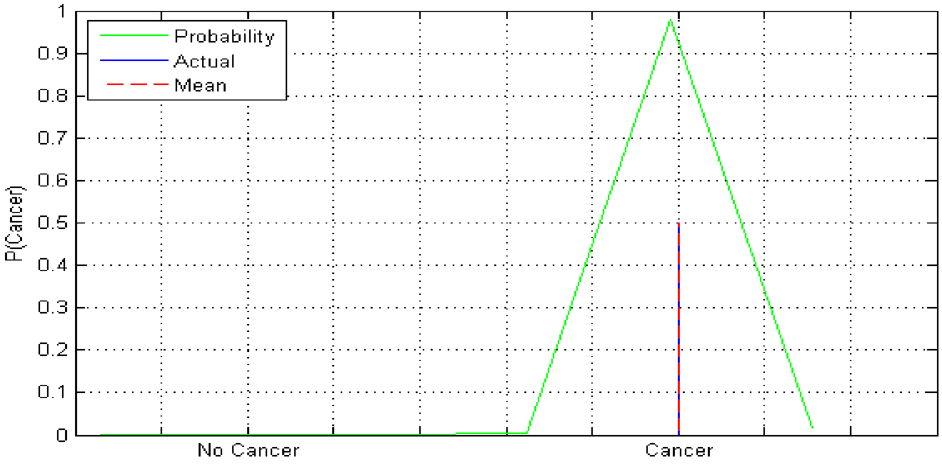


Figure 5.17 Example Output of Methodology for Ovarain Cancer Data. Probability (green), Actual Result if Cancer is Present (blue), and Mean Probability (red).

5.3. Video Analysis

Our visual capability is quite powerful; therefore, video is commonly used to record processes that can be viewed and critical phenomena identified. Analysis of video becomes more difficult; first, by sheer size of the data and second, by the complexity of the millions of pixels that describe the imagery through time. A rich community of machine learning has produced amazing results for tasks associated with feature recognition such as fingerprint identification and genomic signal clustering but these typically require specialized algorithms that are customized to the problem at hand. The ability to quickly analyze this high dimensional data using a generic methodology would be highly sought after.

The example shown in Section 1 is presented here in more detail. High Video is speed video of a flow boiling system recorded at different heat flux levels.[13] The video was recorded at 3500 frames per second at a resolution of 640 x 800. The video was cropped to 456 x 800 pixels and shrunk to 141 x 400 pixels and consisted of 340 frames or slightly less than 0.1 seconds. Even with the cropping of pixels and reduction in recording time, the total number of variables is 19,176,000. This corresponds to approximately 20MB of storage for a grayscale movie. Although the video presents a highly complex process of vapor and liquid boiling two-phase flow, observation indicates features that seem to correspond with heat flux such as the number of bubbles. These features should lie on a manifold of lower dimension that can be identified and used to identify class or magnitude such as the applied heat flux used to generate the vapor in the flowing fluid.

Nine videos were used in the analysis corresponding to five values of heat flux. Figure 5.18 consists of a selection of image frames from the videos illustrating the movement of bubbles. The larger values of heat flux are shown to have a larger number of bubbles due to the higher amount of energy being transferred to the working fluid. Due to the complexity of the boiling process the number or size of bubbles is not linear and although one can roughly estimate location by eye, an exact prediction is only a guess.

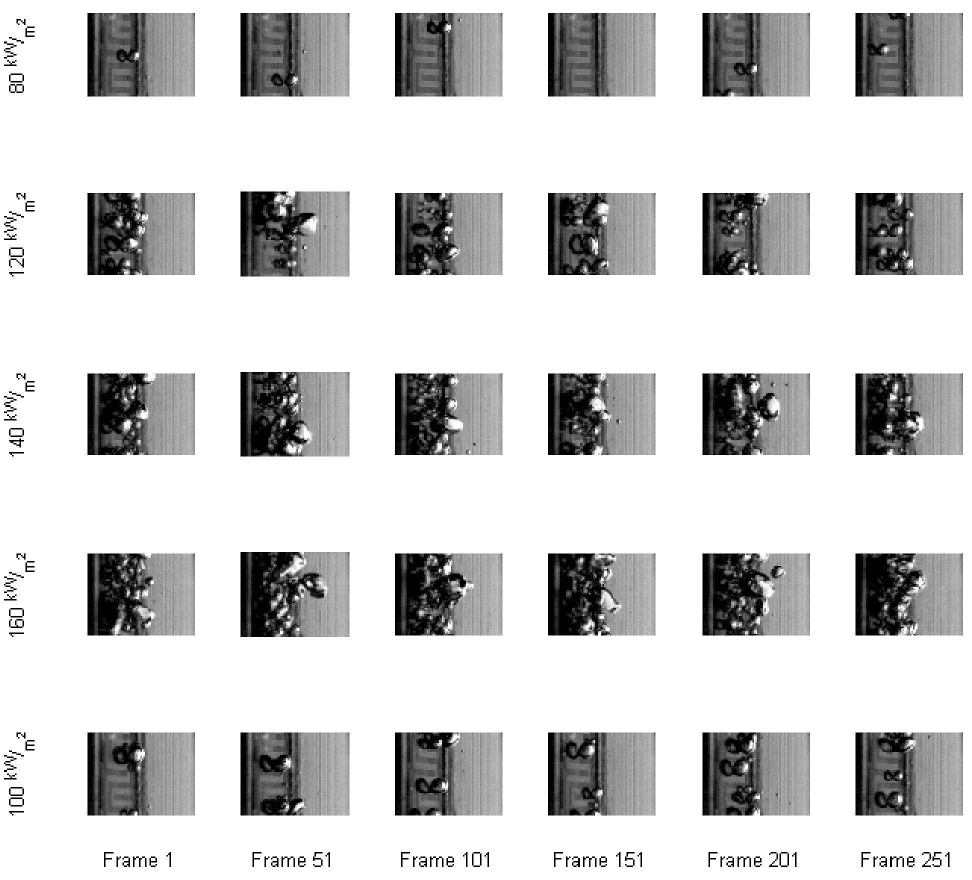


Figure 5.18 Selected Frames From High Speed Video of Flow Boiling.

Using the methodology, two videos at heat fluxes of 80 $^{kW}/m^2$, 120 $^{kW}/m^2$, 140 $^{kW}/m^2$, and 160 $^{kW}/m^2$, were used as the known inputs and the heat flux was predicted for the 100 $^{kW}/m^2$ video. Each image in the 340 frames of the video corresponds to 141 x 400 pixels resulting in 19,176,000 elements or variables to evaluate per movie. Each image is randomly compressed from 56,400 pixels to 16 variables. The 340 frames of 16 variables or 5440 are randomly projected down to 64 elements that are embedded into a two dimensional diffusion space. These two element vectors are used to predict the heat flux. The two diffusion coordinates for the eight known videos are used to construct a probability density shown in Figure 5.19 with the resulting density mean of 106 $^{kW}/m^2$. Thus, one can predict the actual heat flux to within 10%. The tremendous amount of data in each video is compressed into 64 meta-variables that are embedded into a two dimensional map that accurately reflects the actual heat flux. That each observation of boiling resulted in over 19 million variables that can be accurately described through dimensional reduction by two values is an amazing feat. Thus, the utilization of visual or other high dimensional data has been greatly enhanced and leads us to the final section of the dissertation where we will apply the methodology to the identification of two-phase flow regimes.

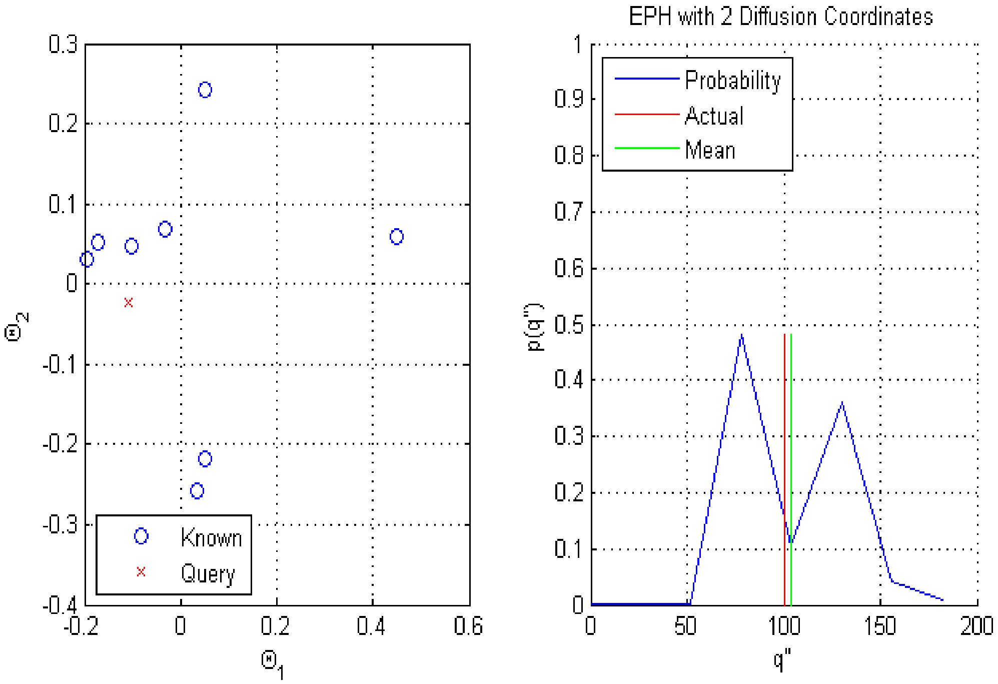
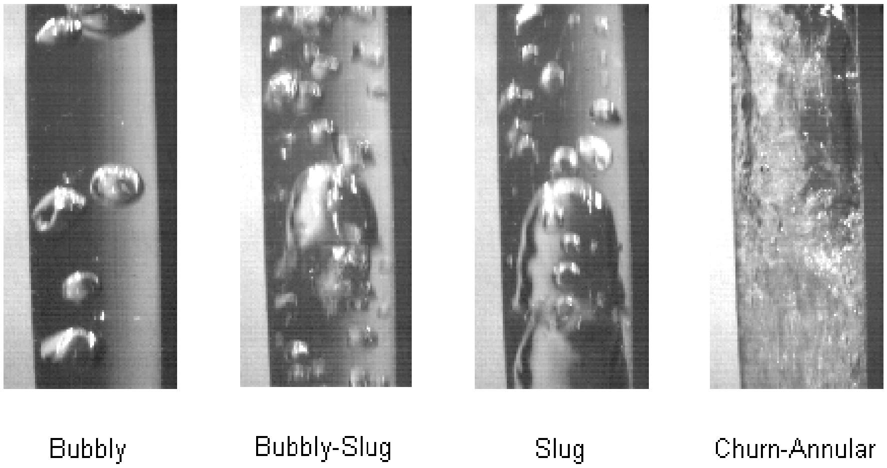


Figure 5.19 a) Plot of First 2 Diffusion Coordinates with Known Data (blue) and Query (red) and b) Plot of Probability Distribution (blue), Actual Heat Flux (green), and Mean Probability (red).

6. FLOW REGIME IDENTIFICATION

Two phase flow regime classification is typically performed by the investigator through observation of still images, high speed imagery, or statistical properties of two phase flow parameters.[138,139,140,141] The regimes are classified based on the spatial orientation of gas and liquid; however, there is not a clear set of identified regimes[142] and many times the investigator must make a decision. This leads to bias and nonstandard regime identification.

The previous sections demonstrated the power of the new modeling approach for high dimensional data including image data. There are several problems in fluid dynamics that can benefit from the analysis technique presented. One is the identification of flow regimes for multiphase flow. The spatial orientation of two fluids moving in a conduit can assume a finite number of characteristic configurations/patterns, which are termed flow regimes. Flow regimes are of prime importance to the understanding of thermal-hydraulic behavior such as pressure drop, heat transfer, critical flows, and other phenomena. The flow patterns cannot be predicted from the independent variables of the system such as the phase flow rates and their physical properties in a straightforward manner.[139]

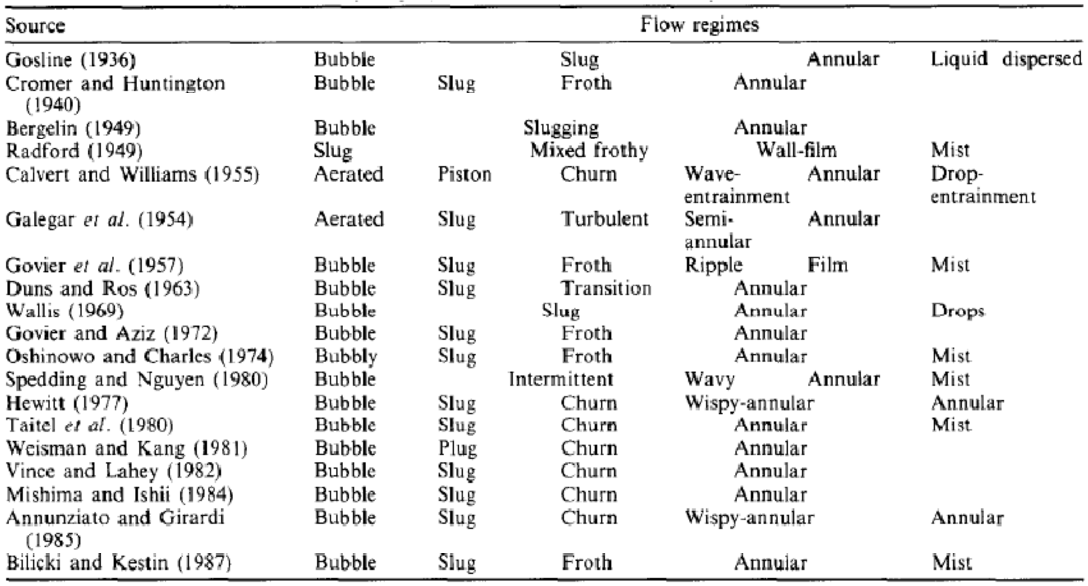


Bubbly Bubbly-Slug Slug Churn-Annular

Figure 6.1 Example Flow Regimes for 1g Vertical Flow Used in the Analysis.

A visual representation of two-phase flow regimes is shown in Figure 6.1. The four images in Figure 6.1 are taken from high speed imagery at 1000 frames per second of vertical up flow of air and water. Although several flow regimes are reported in various studies,[138,143] a consistent set of identified regimes have not been put forward.This is probably due to the fact that the classic identification method is visualization of the flows. This has led to a multitude of flow regime maps which are plots of independent system variables with regions. The use of completely visual observations for determining flow patterns has the disadvantage of being subjective. Differences in interpretation of visual observations are no doubt a major reason for experimenters having recorded different flow patterns under essentially similar flow conditions.[144] A further complication is the different proposed flow regimes used by various authors. Bi[145] presents a table, shown in Table 6.1, which illustrates the different sets of flow regimes authors' use.

Table 6.1 Table of Flow Regime Labels for 1g Horizontal and Vertical Flow.[145]



Source	Flow regimes					
Gosline (1936)	Bubble		Slug		Annular	Liquid dispersed
Cromer and Huntington (1940)	Bubble	Slug	Froth	Annular		
Bergelin (1949)	Bubble		Slugging	Annular		
Radford (1949)	Slug		Mixed frothy	Wall-film		Mist
Calvert and Williams (1955)	Aerated	Piston	Churn	Wave-entrainment	Annular	Drop-entrainment
Galegar *et al.* (1954)	Aerated	Slug	Turbulent	Semi-annular	Annular	
Govier *et al.* (1957)	Bubble	Slug	Froth	Ripple	Film	Mist
Duns and Ros (1963)	Bubble	Slug	Transition	Annular		
Wallis (1969)	Bubble		Slug	Annular		Drops
Govier and Aziz (1972)	Bubble	Slug	Froth	Annular		
Oshinowo and Charles (1974)	Bubbly	Slug	Froth	Annular		Mist
Spedding and Nguyen (1980)	Bubble		Intermittent	Wavy	Annular	Mist
Hewitt (1977)	Bubble	Slug	Churn	Wispy-annular		Annular
Taitel *et al.* (1980)	Bubble	Slug	Churn	Annular		Mist
Weisman and Kang (1981)	Bubble	Plug	Churn	Annular		
Vince and Lahey (1982)	Bubble	Slug	Churn	Annular		
Mishima and Ishii (1984)	Bubble	Slug	Churn	Annular		
Annunziato and Girardi (1985)	Bubble	Slug	Churn	Wispy-annular		Annular
Bilicki and Kestin (1987)	Bubble	Slug	Froth	Annular		Mist

Baker[138] developed one of the first two-phase flow regime maps with air-water and air-oil data in large tubes, which uses scaling factors for the horizontal and vertical axis corresponding to the superficial mass flux times a fluid property scaling factor for liquid and gas respectively. A simpler map, first popularized by Mandhane,[146] uses only the superficial velocities and a derivative is shown in Figure 6.2. Another version of the flow regime for vertical two-phase flow is shown in Figure 6.3.

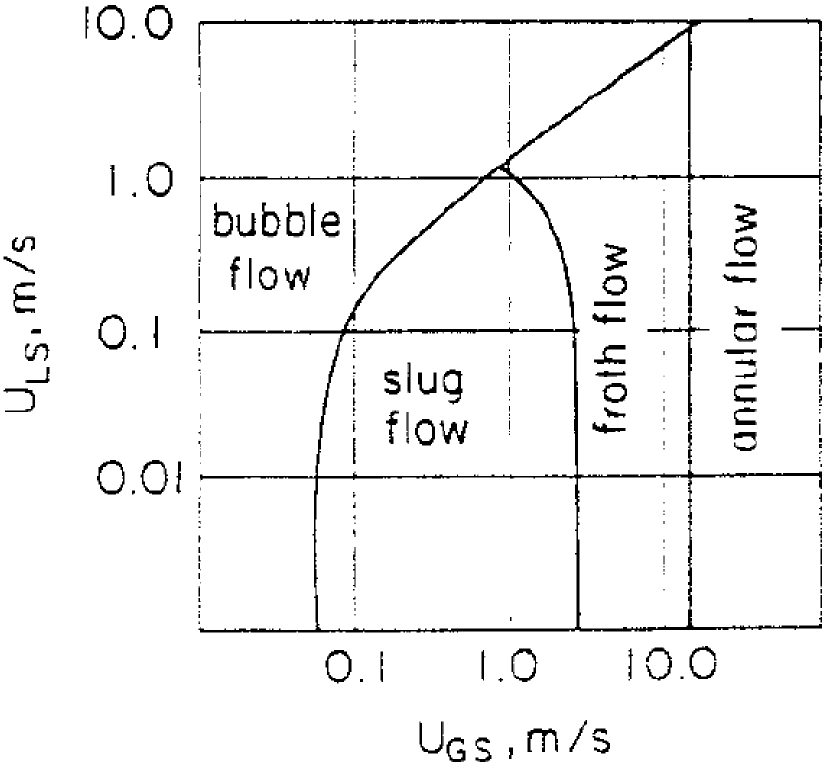


Figure 6.2 Example of a Baker Map.[138]

Most flow regime maps provide sharp transition lines between regimes. The subjective identification of regimes, the different identified regimes used to create predictions, and lack a physical basis for many of the models have led to a large number of flow regime models and maps. There have been innumerable classifications suggested.[147] Many of the published flow regime maps are shown in the following Table 6.2.[148,149] The lack of a consistent set of variables has made it difficult to identify the correct flow regime. This poses a problem when attempting to develop pressure drop, void fraction, and heat transfer models that incorporate all flow regimes without having discontinuities at the boundaries.[150] A more physical, correct prediction model would produce a probabilistic representation of the flow regimes and allow for a continuous predictive model for pressure drop and other phenomena.

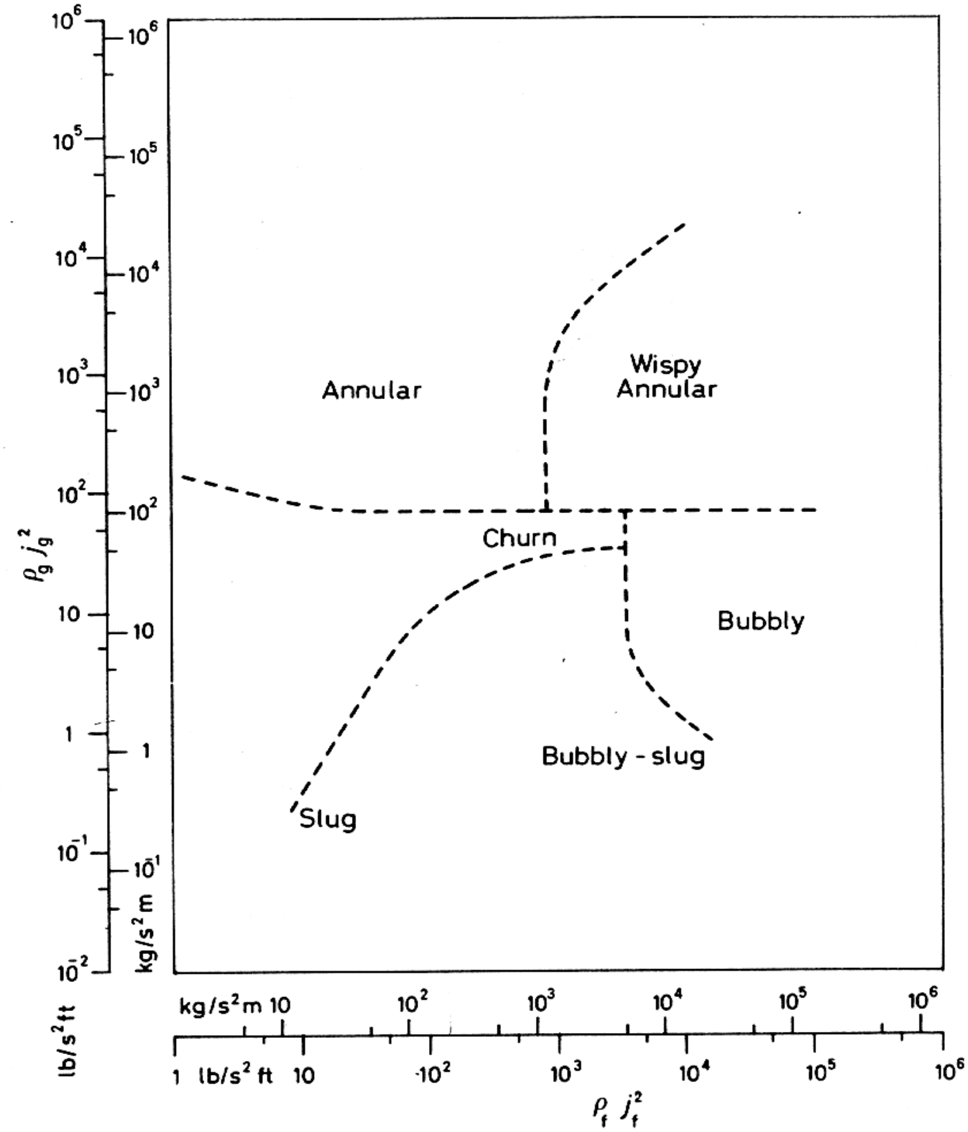
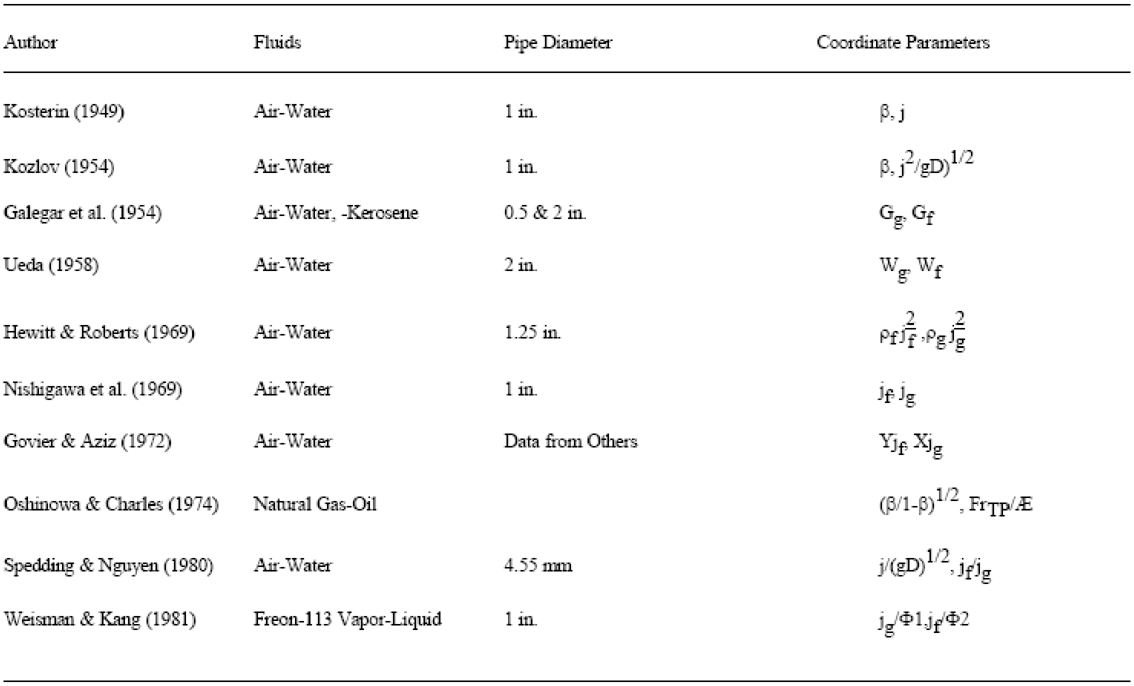


Figure 6.3 Example of Map by Hewitt from Collier.[142]

Table 6.2 Listing of Published Coordinate Parameters for Flow Regime Mapping.[148]



Author	Fluids	Pipe Diameter	Coordinate Parameters
Kosterin (1949)	Air-Water	1 in.	β, j
Kozlov (1954)	Air-Water	1 in.	$\beta, j^2/gD)^{1/2}$
Galegar et al. (1954)	Air-Water, -Kerosene	0.5 & 2 in.	G_g, G_f
Ueda (1958)	Air-Water	2 in.	W_g, W_f
Hewitt & Roberts (1969)	Air-Water	1.25 in.	$\rho_f j_f^2, \rho_g j_g^2$
Nishigawa et al. (1969)	Air-Water	1 in.	j_f, j_g
Govier & Aziz (1972)	Air-Water	Data from Others	Yj_f, Xj_g
Oshinowa & Charles (1974)	Natural Gas-Oil		$(\beta/1-\beta)^{1/2}, Fr_{TP}/Æ$
Spedding & Nguyen (1980)	Air-Water	4.55 mm	$j/(gD)^{1/2}, j_f/j_g$
Weisman & Kang (1981)	Freon-113 Vapor-Liquid	1 in.	$j_g/\Phi 1, j_f/\Phi 2$

Several objective approaches have been developed based on signals analysis of statistical moments,[151,152,153] spectral techniques,[154] and fractal dimension.[155,156] A recent approach relies on a limited number of capacitance measurements coupled with Fuzzy Set Method[139,157] utilizes some statistical moments (mean and variance) and a few wavelet transform coefficients of pixel intensity data. Sunde's visual analysis approach is quite interesting due to the fact that he has included a form of pattern recognition. Using this method, the bubbly and annular flows are correctly identified but it was more difficult to identify the slug and churn regimes. The objective approaches indicate that one can identify flow regimes with some accuracy. However, many of these approaches require specialized instrumentation and it is difficult to compare results between authors.

Since the flow regimes are described by their spatial orientation, an approach that utilizes imagery would be expected to produce accurate results.

6.1. Flow Regime Imaging

The previous section showed how images from the shuttle pitch maneuver images and flow boiling video could be embedded into a low dimensional space that could be used for organizing, clustering, or prediction. The first flow regime video is from NASA high speed archive of reduced gravity two-phase flow. An example frame capture is shown in Figure 6.4. Three flow regimes were used for identification; bubbly, slug, and annular. Ten high speed videos were used in the preliminary analysis. The videos consist of 150 frames of grayscale images that are 231 pixels wide by 191 pixels tall. Taking each individual pixel of each frame as a variable, the dimensionality of the videos is $R^{231x191x150}$or approximately 6.5 million. Since the video is highly correlated with time, several pixels would have similar values. For example, the background and tube walls are fixed and should not change significantly. Also, since the flow is in one direction, downstream pixels will be correlated with upstream pixels. Thus, the data should reside in a low dimensional subspace and the pixel intensity data can be projected onto this smaller subspace.

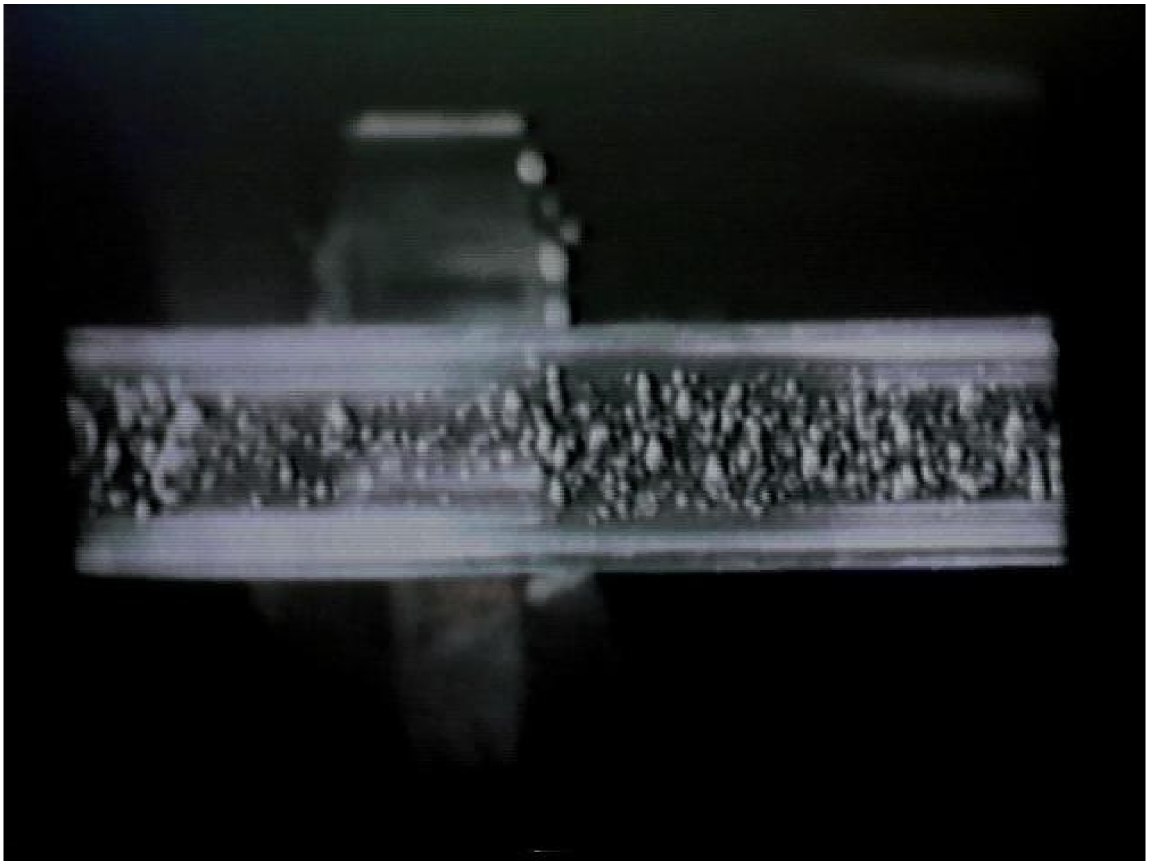


Figure 6.4 Frame Capture from High Speed Imagery of Microgravity Two-Phase Flow.

With over 44000 pixels per image and 150 images per movie, random projections are needed to provide more manageable data that allows the calculation of the Euclidean distance without a significant time or memory constraint. Each image is randomly projected using a 100 x 44121 matrix comprised of values that are normally distributed with a mean of 0 and a standard deviation of 1. Each column of the matrix is normalized prior to being multiplied by each frame of the movies. In order to compress in the time domain, a second random matrix is constructed that is 40 x 150. This matrix is then multiplied by the transpose of the random projected movie matrix. The resulting vector consists of 4000 terms that are used to produce an embedding of the 10 high speed recordings. The first two embedding coordinates from the diffusion map are shown in Figure 6.5. It is quite clear that only two coordinates are needed to correctly classify the three flow regime. Similar data could be randomly projected using the same matrices for this data and then a new embedding could be performed. However, it is difficult to produce similar test conditions for imaging. What is typically done is tests with the same flow conditions. Using the embedding coordinates as the output and the flow rates as the input to constructing the EPH, one could quickly and accurately identify the flow regimes.

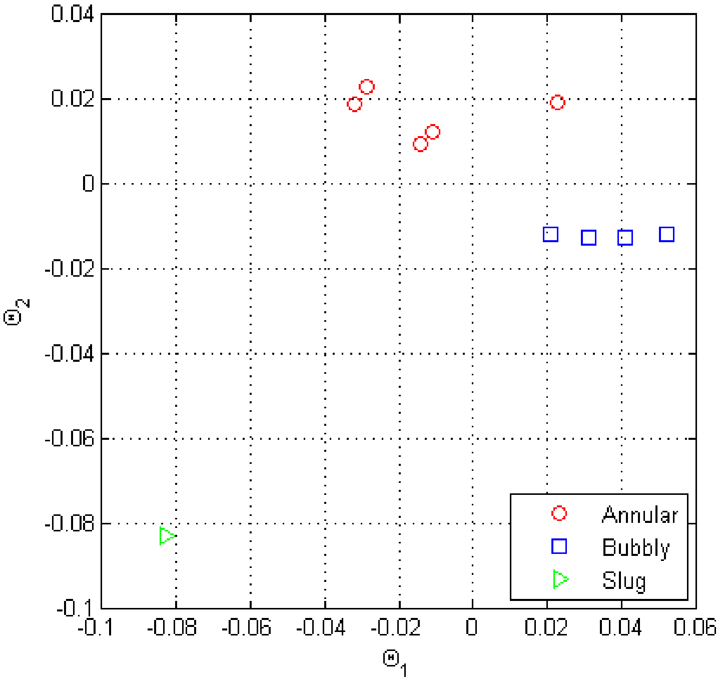


Figure 6.5 Plot of the First Two Diffusion Coordinates of Annular (red), Blue (bubbly), and Slug (green) Flow Regimes of Microgravity Two-Phase Flow.

6.2. 1g Two-Phase Flow

A database of high-speed flow regime videos was recorded using a test facility that had a 3/8 inch inner diameter acrylic tube connected to a piping system that was able to produce well characterized two-phase flow. One thousand frames of video were recorded for 1 second at a resolution of 116 x 256 pixels. Sixty-four videos were recorded and are presented in the montage shown in Figure 6.6. The montage consists of several frame captures from the 64 videos staring with bubbly flow at the top and ending with churn/annular flow at the bottom. Each video was played back at 30 frames per second and flow regime identification was carried out visually using five different regimes; annular, bubbly, bubbly-slug, slug, and churn. For each movie, the regime was identified on a superficial velocity map shown in Figure 6.7. The transition lines for the flow regime map shown in Figure 6.7 are misleading in that they appear to shown sharp transitions between each regime. Actual transitions are difficult to determine and this is described in the paper by Taitel and Dukler[147] and shown in the visual observations where a bubbly-slug regime is plotted due to the uncertainty of which regime exists.

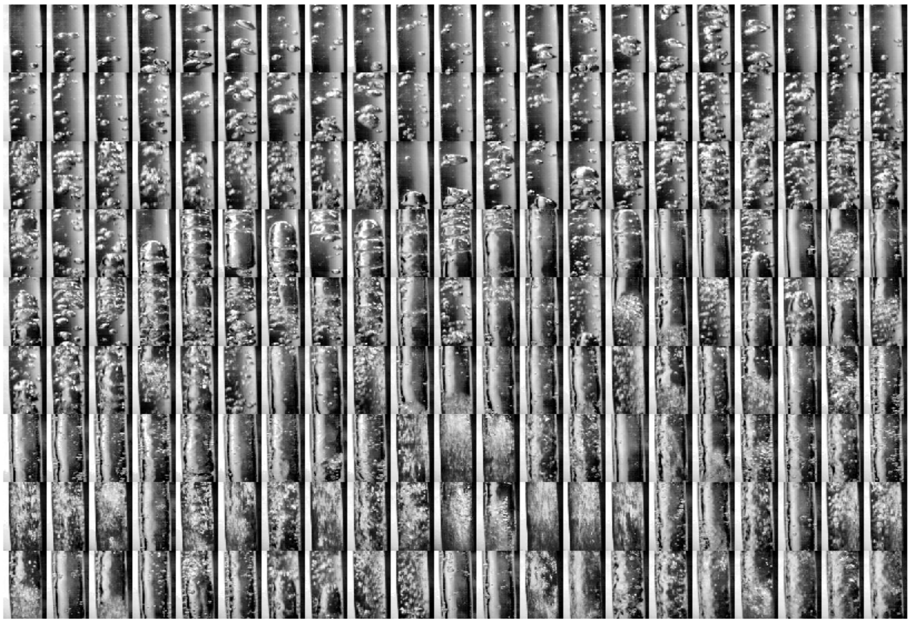


Figure 6.6 Montage of Frame Captures from Flow Regime Video Database.

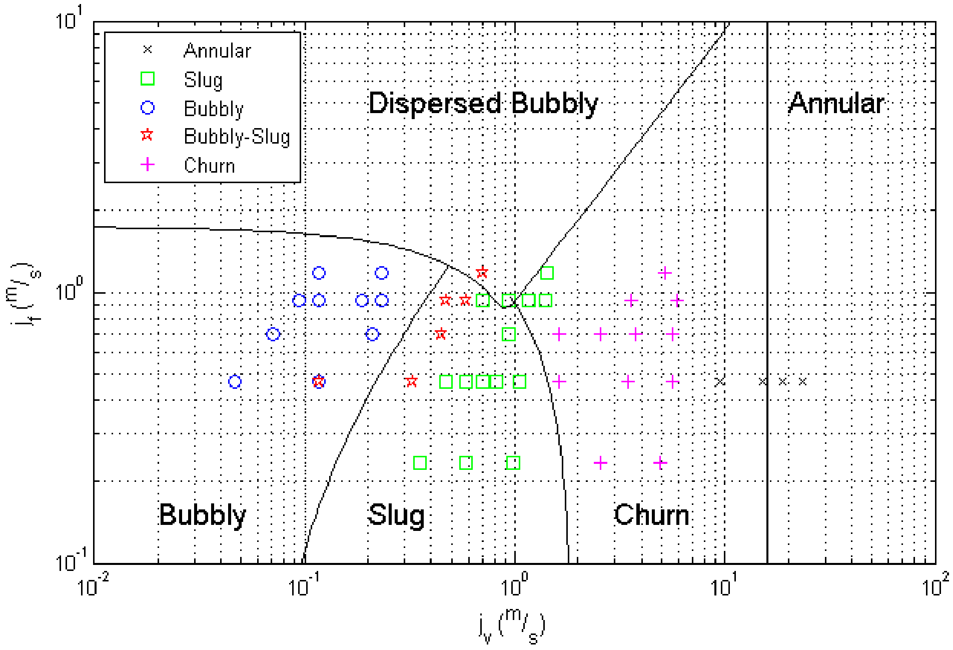


Figure 6.7 Superficial Velocity Map of 1g Vertical Air-Water Flow with Corresponding Regimes and Transition Lines Described by Taitel-Dukler[147].

Flow regimes are typically presented as a simple picture or cartoon that neglects both the subtle spatial and temporal features of the flow. Since these are constructs of the investigator, several regime definitions exists as shown in Table 6.1 and are not consistent. Further, the taxonomy is created prior to testing and is typically kept to a small number for simplicity. An approach that does not require a priori information would be extremely useful and would allow classification without subjective bias.

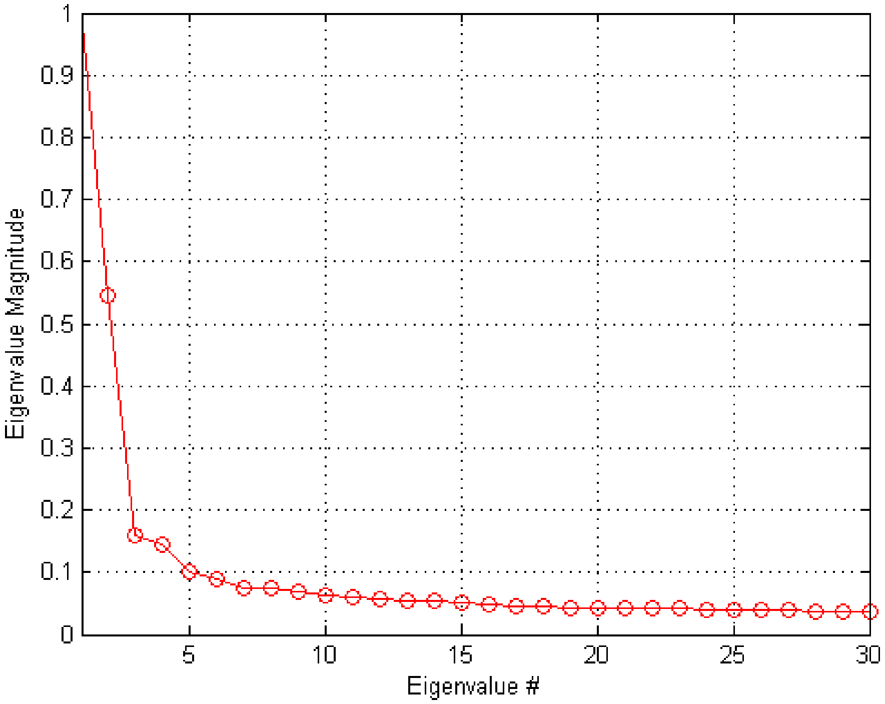


Figure 6.8 Plot of the First 30 Eigenvalues for Distance Matrix of Randomly Projected Images Shown in Figure 6.6.

The images shown in Figure 6.6 were projected using a random matrix to produce a vector of length 256. The vector for each image is embedded into a diffusion space and the first 30 eigenvalues are shown in Figure 6.8, which correspond to the underlying dimension of the data. The first eigenvalue is neglected[27] resulting in an approximate dimension of one or two. If one thinks of the eigenvalues as the principal components in PCA, the first two principal components provide most of the information.

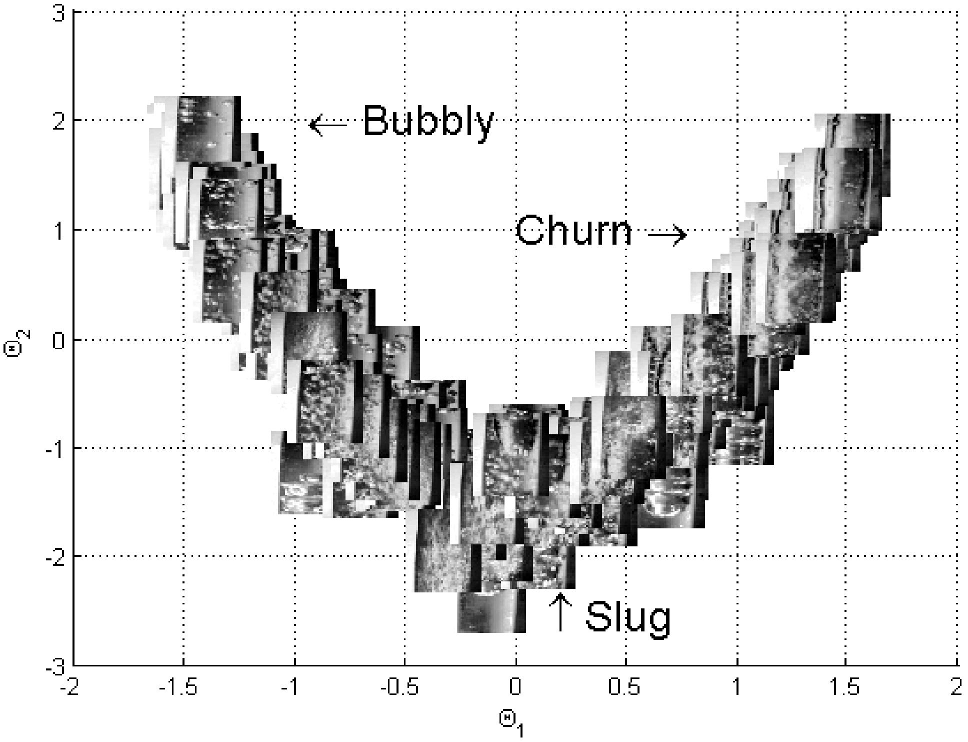
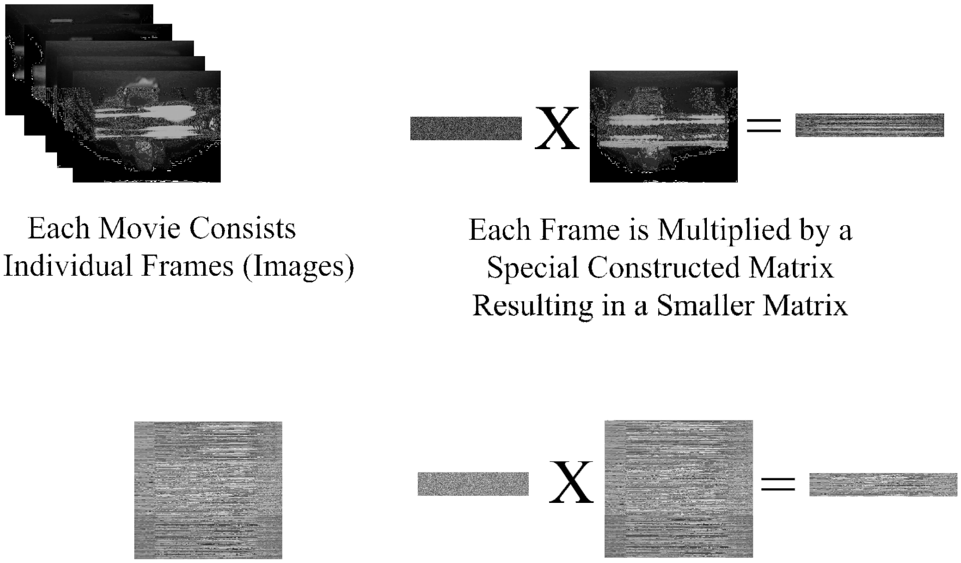


Figure 6.9 Embedding of Images in Figure 6.6 on the First 2 Diffusion Coordinates.

The images are arranged according to the first two diffusion coordinates and plotted. The resulting plot is shown in Figure 6.9 with the 189 pictures from 64 different flow conditions plotted using the corresponding diffusion coordinates. The embedding does quite well in organizing the flow regimes with a continuous curve starting with bubbly regimes and moving through slug and churn regimes. Unfortunately, the nature of two-phase flow requires observation through time to understand the intricacies of the particular flow regime.

Video analysis is quite difficult due to the extremely large amount of data. In order to reduce the dimension of the video data, both the spatial variables represented by each frame of the video and the spatial variables represented by each pixels value through time must be reduced. Performing the dimension reduction and embedding to organize video data was successfully shown in section 6.1 for microgravity two-phase flow. A graphical representation of the dimension reduction is shown in Figure 6.10.



Each Movie Consists
of Individual Frames (Images)

Each Frame is Multiplied by a
Special Constructed Matrix
Resulting in a Smaller Matrix

For Each Frame, The Resulting
Small Matrices are Concatenated
into Larger Matrix for Entire Movie

The Process is Repeated Resulting
in a Small Number of Meta-Variables
for Each Movie

Figure 6.10 Graphical Representation of the Dimension Reduction Process Used by the Methodology for Video Data.

Once the movie data has been reduced, a method of identifying the proper number of regimes must be determined. The meta-variables produced from the dimension reduction process or the embedding coordinates can be used to determine the appropriate number of clusters. Both approaches should produce the same number of clusters as well as produce similar flow regime maps. Several methods to determine clusters have been developed as well as validity measures to determine the optimal number of clusters.[30] To determine the number of flow regimes or clusters, several approaches are employed. The first is a graphical based approach that utilizes a dendrogram to illustrate the arrangement of data based on the Euclidean distance between the first two embedding coordinates. This hierarchical clustering approach constructs a tree with nodes corresponding to each video and the branches connecting the nodes reflecting the distance between each node and every other node. The arrangement of the nodes depends on a weighting function and a method of adding nodes or groups of nodes to the tree. The Euclidean distance between the meta-variables for each video is used as the weighting function and Ward's algorithm to link the nodes together. The clustering is performed by cutting the tree at a certain level with the resulting sub-trees corresponding to the flow regimes. Cutting the tree at a lower level will result in more clusters which introduce the problem of identifying the best cutting location and resulting number of clusters. K-means clustering also suffers from a similar problem in that one must specify a number of clusters prior to performing the clustering. Thus, the clustering process is an iterative approach.

To determine the number of needed embedding coordinates, one needs to evaluate the corresponding eigenvalues for each coordinate. A plot of the eignevalues is shown in Figure 6.11. Since the eignevalues are monotonic decreasing and the magnitude quickly decreases to near zero, only the first 12 are shown in Figure 6.11. It is quite apparent that only two embedding coordinates are needed. Thus, for the purposes of clustering, only two coordinates are utilized. Figure 6.12 is a plot of the first two embedding coordinates. The plot shows three clusters that have been color coded and labeled ‘Bubbly’, ’Slug’, and ‘Churn’. The labeling is determined both from hierarchical clustering and from a k-means clustering algorithm. Using the two embedding coordinates a dendrogram was produced which is shown in Figure 6.13. The tree nodes are labeled with the superficial velocities and observed flow regime for each test point and three clusters are color coded to match the coloring of Figure 6.12. The organization of the data shows that the embedding coordinates based on the reduced dimension meta-variables produce an accurate clustering of similar regimes.

To evaluate the choice for the number of clusters, a number of tests were performed as outlined in the reference by Balasko.[158] The first two are the Partition Coefficient and Classification Entropy which were calculated for several clusters. The Partition Coefficient and the Classification Entropy are shown in Figure 6.14. Since one is looking for a small number of clusters or flow regimes and that the Partition Coefficient should be monotonic decreasing and the Classification Entropy should be monotonic increasing, three clusters were chosen. As shown in Figure 6.14 by the local minima and maxima respectively. A stronger measure is the Dunn Index which is the

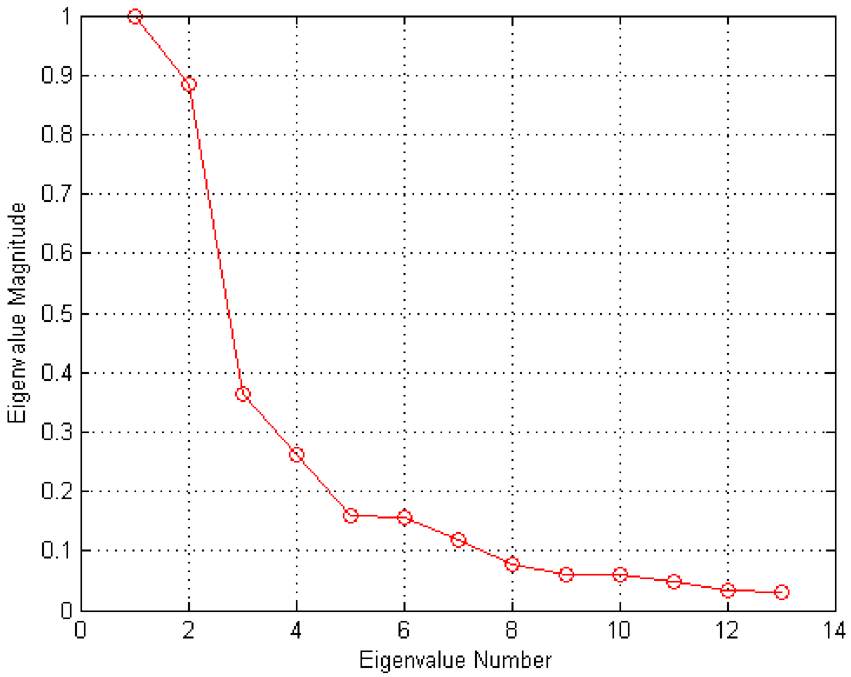


Figure 6.11 Magnitude of Eigenvalues for Each Embedding Coordinates.

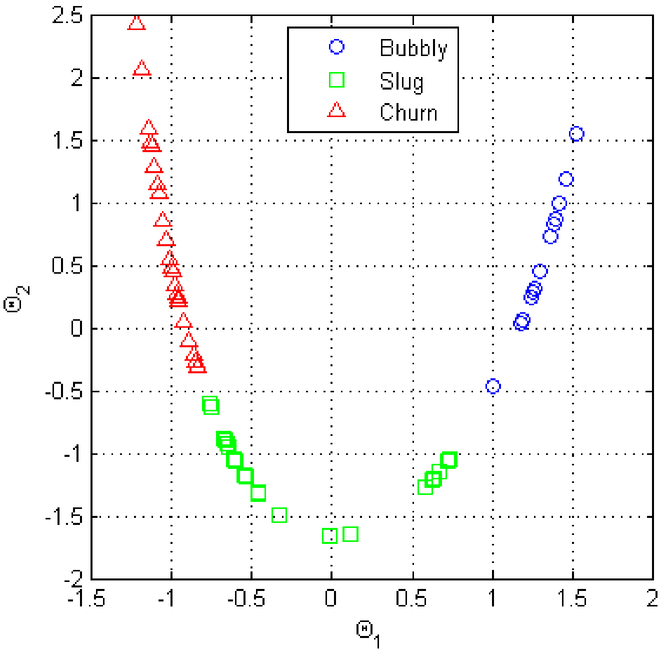


Figure 6.12 Plot of the First Two Embeding Coordinates for Flow Regime Video Data.

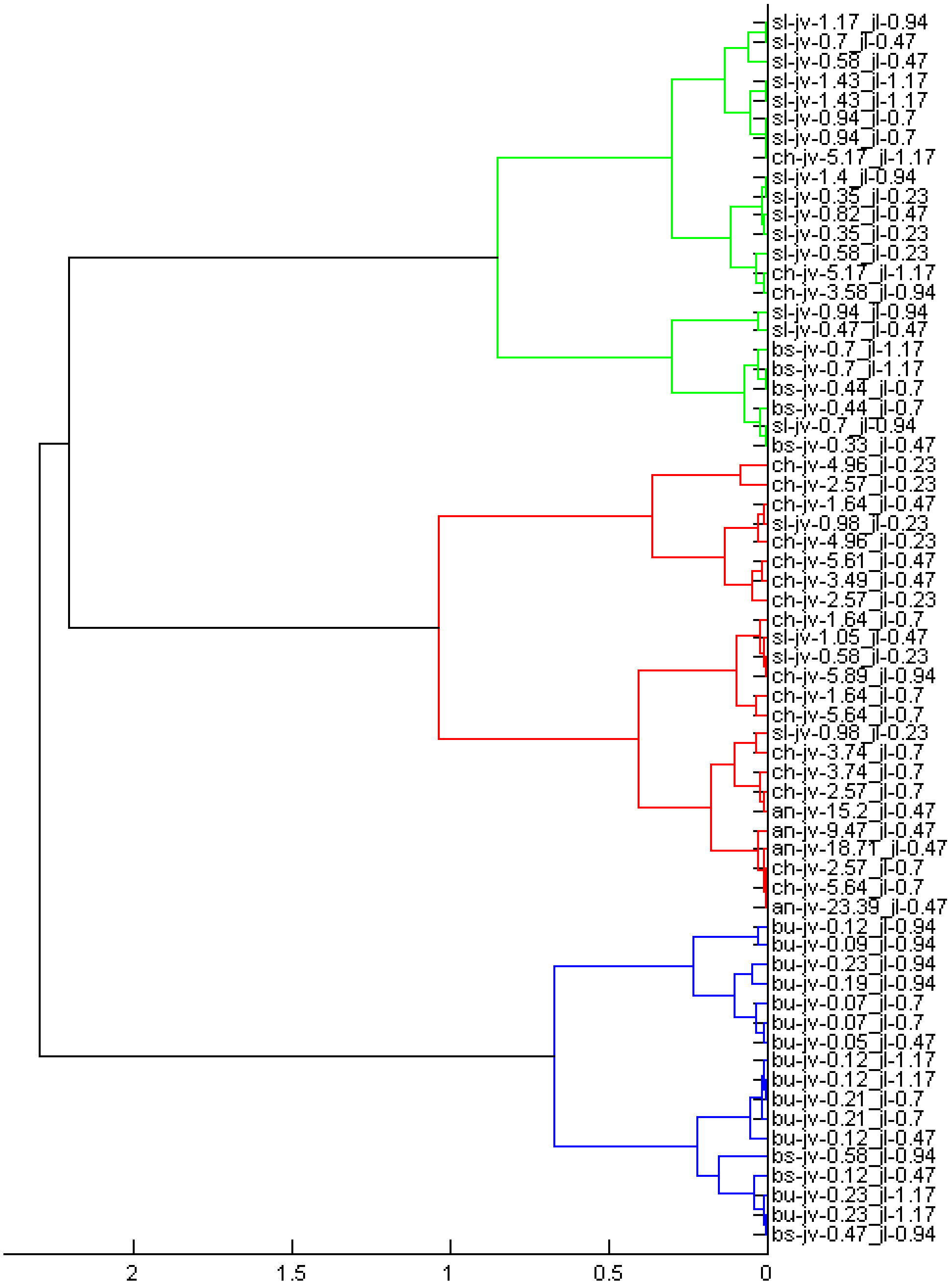


Figure 6.13 Dendrogram of Movie Data.

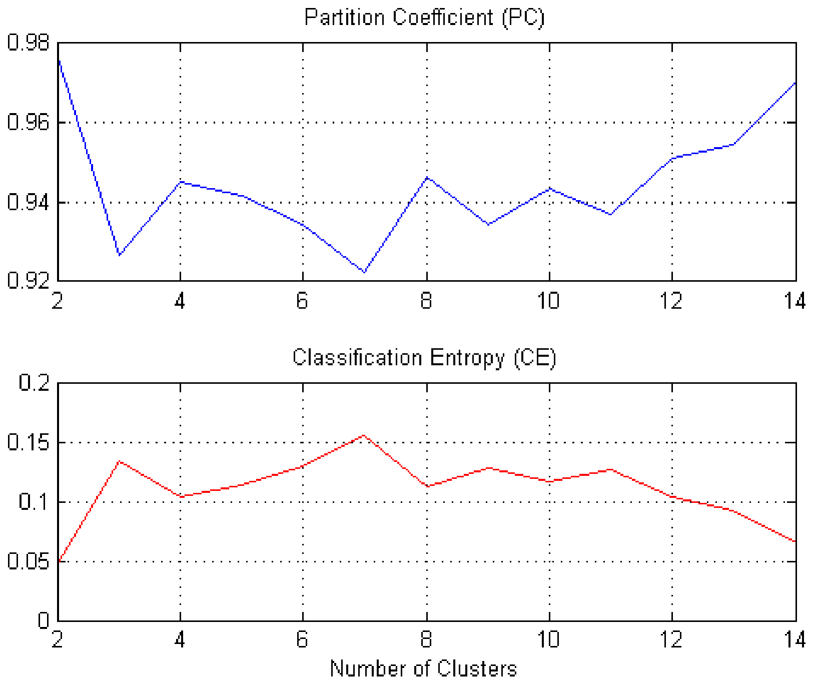


Figure 6.14 Plot of a) Partition Coefficient and b) Classification Entropy.

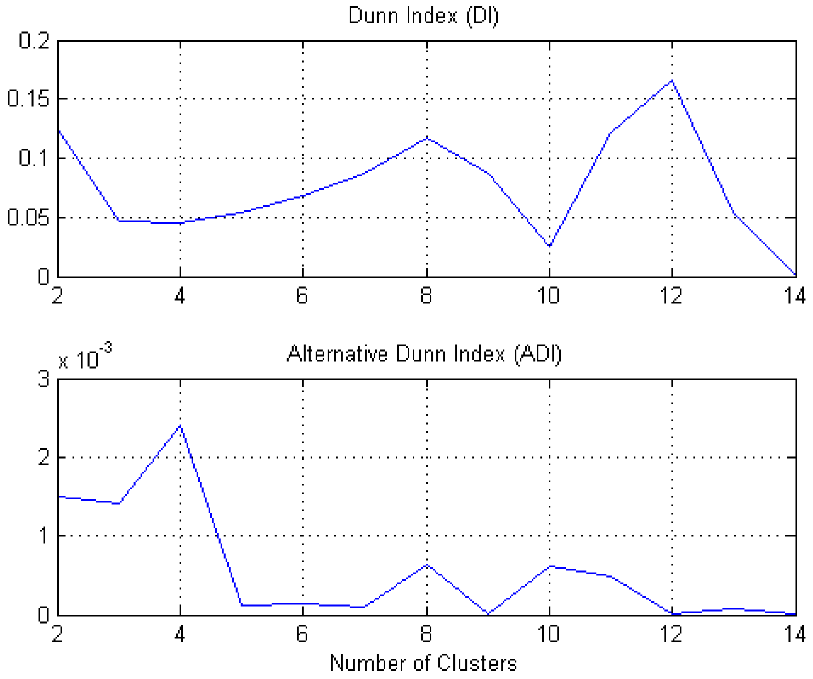


Figure 6.15 Plot of the a) Dunn Index and b) Alternative Dunn Index.

ratio between the minimal intracluster distance to maximal intercluster distance. The Dunn Index is shown along with the Alternative Dunn Index in Figure 6.15. Figure 6.15 shows a local minima at a value of three clusters. Based on the validity measures shown in Figures 6.14 and 6.15, three clusters were chosen based on the embedding of the meta-variables for each movie. The clusters closely match the observed data as well as the flow regimes predicted using the Taitel-Dukler map as shown in Figure 6.16. The flow regimes identified objectively from video data through the random projection-diffusion embedding process match the prediction map quite well. The major differences between the objectively clustered and the flow regime map can be found at the higher liquid superficial velocities where slug points can be found in the churn region and at lower superficial velocities where churn points are found in slug region. Further, the annular and dispersed bubbly flow regimes are not identified using the method. This is due to the fact that no dispersed bubbly data was recorded during testing and very few annular points were collected. Interestingly, another flow regime map[159] shown in Figure 6.17 has a slug region separating the dispersed bubbly and churn regions. The two plots illustrate the difficulty of evaluating two-phase flow regime experiments. As mentioned previously, the preponderance of different identification maps, number and type of regimes, and the subjective nature of human identification produce an inconsistent set of tools for analysis. The benefit of the methodology is the complete objective nature of the regime identification and the ability to utilize existing data to make projections of new test conditions.

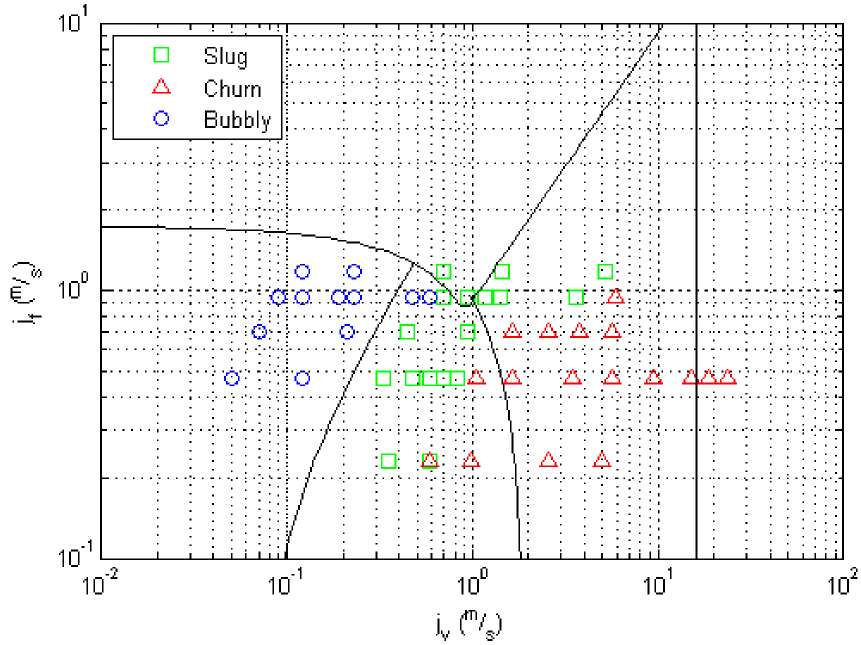


Figure 6.16 Flow Regime Map of Clustered Data.

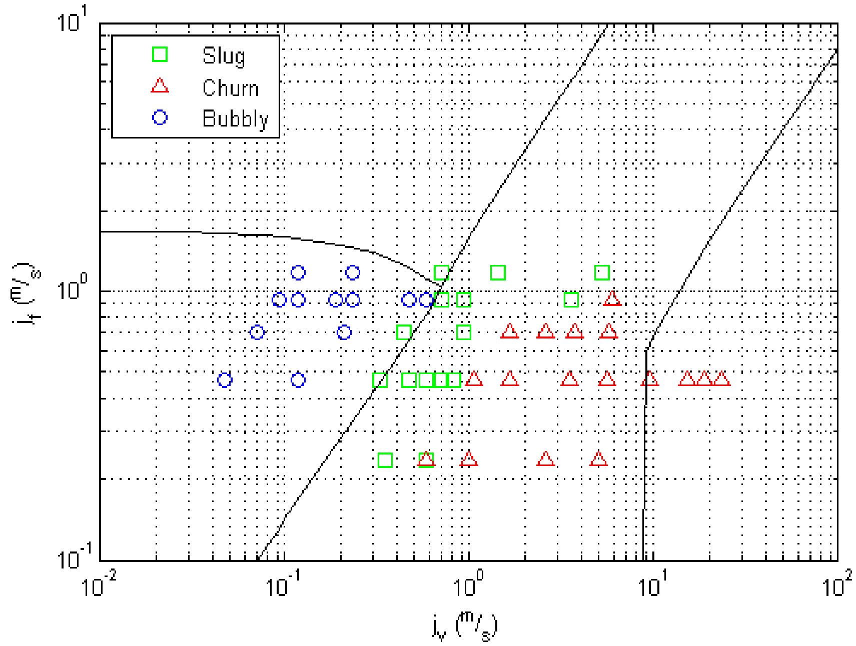


Figure 6.17 Alternative Flow Regime Map of Clustered Data.

To predict the flow regime for a new set of superficial velocities, one may utilize the meta-variables, embedded coordinates, the superficial velocities, or the volumetric flow rates for the new case. An example is shown in Figure 6.18 where 10,000 points were predicted given only the test data shown as black points and the corresponding flow regimes determined above. The color coding corresponds to the color coding in the previous plots; blue – bubbly, green – slug, and red – churn. The predicted flow regimes are based on the mean probability determined from the Experimental Probabilistic Hypersurface where the known input information was the superficial velocity pairs and their corresponding embedded diffusion coordinates. Queries are made for a number of gas and liquid superficial velocities. The predicted values match the flow regime map quite well as demonstrated by the similarity of the predicted bubbly-slug transition line with the flow regime map transition. Not only does prediction match the transition line close to test points, it fits well beyond the range tested flow conditions indicating an extrapolation capability. Further, using the superficial velocity data along with the clustering from the embedded meta-variables allow predictions to be made without the computational expense of the dimension reduction and embedding used for the raw video data. This may be useful when releasing the original data is not acceptable.

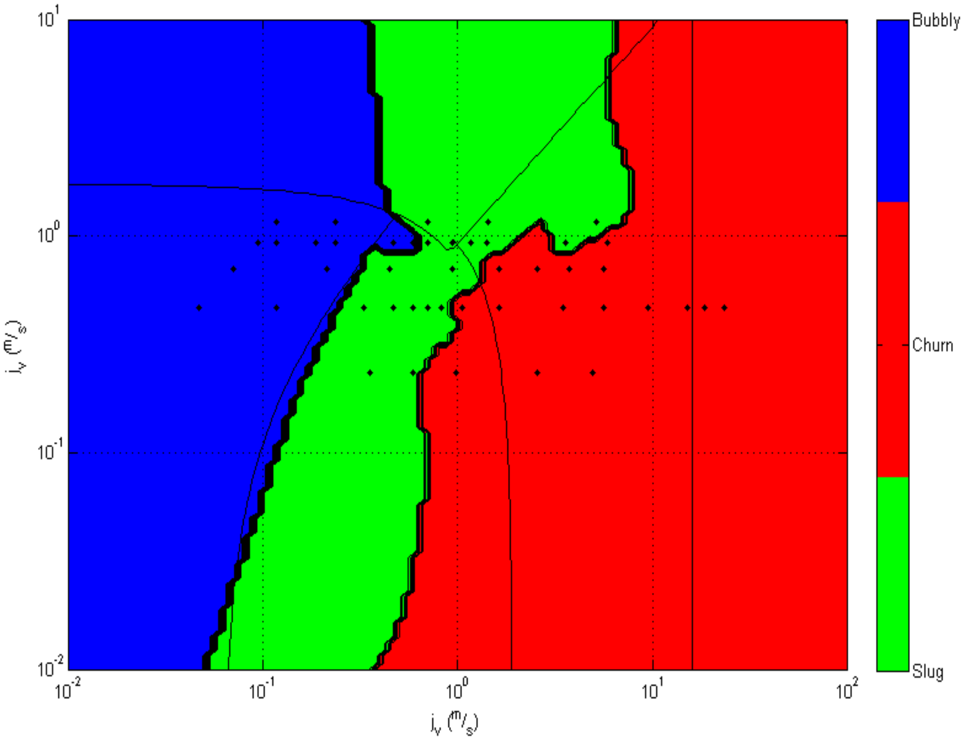


Figure 6.18 Predicted Flow Regime Map Using Information from Video Data.

The preceding analysis was carried out using video recorded at one thousand frames per second. Only 500 frames of the video were used yielding a actual frame rate of 500 frames per second. The resolution is 116 x 256 pixels and as mentioned previously, sixty-four videos were recorded as shown in the montage in Figure 6.6. Determining the size of the random projection matrices to reduce the video data was carried out iteratively

7. CONCLUSIONS AND RECOMMENDATIONS

A new approach to predicting engineering phenomena has been developed. The new approach is especially useful for problems that have a large number of variables and a low sample size that reduce the possibility of using classical statistical based modeling techniques. The application of this methodology is limitless:

- Prediction/Regression – The ability to utilize a small number of high dimensional data to predict a new set of conditions. The prediction not only provides the most likely result but an accompanying probability distribution that provides the information regarding how much one knows about the region of interest. Further, new data can be easily added to provide a more accurate probability distribution and resulting prediction.
- Classification - The labeling of sets of data that reside in a high dimensional space is an application with a number of areas of interest ranging from mechanics to physiology. The ability to take high dimensional diagnostic data and map the information to a few embedding coordinates that can be visualized graphically. These coordinates can then be labeled if known or groupings determined using classical statistical techniques. This low dimensional labeled data can then be used for new queries.
- Extrapolation – Due to the nature of high dimensional data, the methodology has shown a capability to extrapolate outside the range of known data. This capability is enhanced over typical approaches due to the probability

distribution, which provides information regarding how sure one is regarding the prediction.

These features which have been demonstrated through the application to various problems found in engineering provide a new, unique approach to solving high dimension problems in a quick, accurate manner.

Section 2 introduced the methodology and described the process by which high dimension data can be reduced to a lower dimension while maintaining critical information describing the phenomena of interest. The reduced dimension data can then be embedded into a new coordinate system that is used as input to construct a probability distribution for a new query. Section 3 demonstrated the utility and accuracy of this approach to simple low dimension data sets that are well understood. Both the friction factor and critical heat flux prediction using the methodology rival current techniques and the results provide confidence to apply the methodology to more complex problems. Section 4 dealt with the application of the methodology to predicting the most common result of a stochastic process. A criticality problem was predicted to reasonable success demonstrating the methodology's ability to deal with both continuous and categorical data as well as methods to make predictions over the entire domain of the manifold i.e. the probability of any configuration achieving a critical configuration. The last section, Section 5, applied the methodology to high dimension data; first, to perform a regression to predict the space shuttle pitch angle based on a small set of known images, second, the analysis of spectrum data for the prediction of cell irradiation and for the presence of Ovarian cancer, and third, the determination of heat flux in a flow boiling experiment where a limited number of videos are labeled. Section 6 used the methodology to objectively classify flow regimes using video data

7.1. Addition to the State-of-the-Art

A number of additions to the current state-of-the-art in machine learning have been presented herein. First, coupling modern dimension reduction techniques with the predictive power of the Experimental Probabilistic Hypersurface is new. Second, the embedding of random projections into a diffusion space allows the method to be applied to non-linear problems and helps deal with the problems of using the Euclidean distance with high dimensional data. In regard to engineering, this approach provides a new correlation driven approach that utilizes existing data to make prognostications for new conditions. Rather than providing a singular result, the method produces a probability density that can be interpreted accordingly. This approach allows for prediction outside the range of known data, extrapolation, with a built in diagnostic, the probability density, to indicate whether there is enough information to trust the result. As demonstrated in a number of examples throughout this work, the data driven method produces an accurate result with relatively few known data. This should be contrasted with classic correlation approaches that continue to be developed over years/decades of work with significantly larger datasets.

7.2. Predicting High Dimensional Data

Originally, the purpose of this work was to objectively identify flow regimes and provide a method for integrating data from multiple test conditions, fluids, etc. Several measurement techniques have been put forward but it was decided to utilize video imagery similar to what is done when evaluating flow regimes subjectively. The problems of dealing with an large number of variables required a method that had a dimension reduction component. The use of embedding the coordinates in a diffusion space arose from the non-linearity of the data and the requirement to develop a diagnostic for determining the number of coordinates to use for prediction. Finally, a probabilistic prediction tool was desired that would reflect the uncertainty found when discerning flow regime transitions. The resulting methodology was found to be quite useful for predicting flow regimes as well as a number of other problems ranging from predicting thermal-hydraulic phenomena to the analysis of protein serum data in the prediction of ovarian cancer.

7.2.1. Images

Image analysis was carried out through the arrangement of facial data and the prediction of facial pose using a corrupted image. The unique approach of this classic machine learning problem was the use of random projections to reduce the dimension of the image data while maintaining the Euclidean distance between images. The random projections of each image were embedded into a diffusion space where they were arranged according to rotation of the face. Thus, the random projections are able to capture a significant amount of important facial features in an unsupervised manner. Determining the rotation of the corrupted image was a simple operation of finding the facial pose with the closest embedding coordinates.

A more challenging approach was to utilize shuttle imagery to determine the pitch angle of the orbiter in relation to the International Space Station. By randomly projecting each image and embedding them into a diffusion space, a few embedding coordinates could be used for prediction of orbiter pitch angle. It was shown that only a few tens of randomly selected images could provide enough information to predict the pitch angle to within two degrees. Thus, one can use the embedding coordinates as inputs into the Experimental Probabilistic Hypersurface to accurately predict the result of a new query condition. The use of a few diffusion coordinates rapidly increases the speed of computation making this method feasible for a host of problems including video data.

7.2.2. Spectrum Data

Spectrum data is quite common in engineering and science. A number of applications where the methodology could be applied is the field nuclear forensics, thermal analysis, and health physics. The methodology was used to cluster the spectrum data which corresponded to phenomena of interest. The first was the identification of cell irradiation and the second was the classification of cancer or no cancer. A large amount of literature is devoted to identifying the exact channels which correspond to proteins or other unique identifiers. The interesting aspect of utilizing the methodology is that all channels are utilized and though dimension reduction and embedding, can be queried repeatedly to identify several phenomena of interest. Thus, one test can provide answers to multiple queries without performing individual tests for each. Also, previous clustering approaches relied on statistical methods and a priori information regarding the number of clusters. The methodology presented here is able to identify clusters objectively in an unsupervised fashion and rather than complicated clustering rules that minimize false positives, the methodology provides a probability density that allows the professional to interpret membership in the appropriate cluster (e.g. cancer or not). Rather than making a false positive, a follow up test can be carried out.

7.2.3. Video Data

The most challenging problems dealt with the video data that first generated the idea for this dissertation. The prediction of applied heat flux based on the observation of high speed flow boiling video was an interesting problem. Qualitatively, one can observe the presence and size of vapor bubbles and relate that observation to the known heat flux. However, the ability to accurately predict the heat flux for conditions not observed is challenging due to the non-linear nature of bubble size and frequency in relation to he applied heat flux. The simple approach of randomly projecting both the images in space and in time worked quite well. The first two embedding coordinates provided enough information to perform an accurate regression based on the known heat fluxes for each of the known movies. The two coordinates for the query video provide the input and the resulting mean of the probability density is the most likely heat flux.

Initial flow regime identification worked very well for the few microgravity videos used. These flow conditions were near ideal flow regimes and were accurately clustered regardless of the noise of the video recording. The vertical earth gravity tests provided a stronger challenge. First, the number of identified regimes was in dispute as well as the flow regime map to be used for comparison. The methodology was applied and the resulting diffusion coordinates were clustered based on two different approaches. The validity of the clusters was tested resulting in three clusters representing bubbly, slug, and churn. Using the superficial velocities as inputs and either the corresponding embedding coordinates or cluster number as the known data, queries over a large space of superficial velocities was performed that matched well with previous published flow regime maps. The unique thing about this result was that it was performed objectively using video data only. This approach can be extended to include other parameters of interest such as gravity level, tube inclination, fluid type, tube diameter, etc.

7.3. Recommendations for Further Work

The methodology outlined in this dissertation opens up a number of avenues for further research. First, an area of research can be directed toward finding and proving a basis for the selection of the minimum number of projections and number of embedding coordinates. The number of projections is loosely based on the Johnson-Lindenstrauss lemma and the work in the area of compressed sensing where the dimension of the reduced vector scales with the logarithm of the dimension of the original data. This needs to be further analyzed for the coupling of projections that are carried out in the reduction of video data. A further area of work is to merge the embedding coordinates with other data as a form of data fusion. For flow regime identification, the flow rate information could be added to the embedded coordinates for flow regime prediction and this data fusion could be potentially used for more accurate results.

The number of diffusion or embedding coordinates is typically evaluated based on the eigenvalues associated with each coordinate. However, these values are dependent on the radial basis function used. In the method put forward by Lafon,[6] the Gaussian window is set by a parameter that describes the width or the scaling of the distance measurement which is dependent on the number of known data used in performing the embedding. Another area of research should be directed toward the normalization and transformation of data used to produce the Experimental Probabilistic Hypersurface. Since the EPH requires the calculation of Euclidean distance, the scaling of the input data is critical to the accuracy of the method. Typically, common transforms were used such as taking the logarithm of the data and rescaling such that the data ranged between zero and one. This is common in statistics and has a strong foundation mathematically but how this effects the construction of the probability density is not clearly understood.

The work here indicates that this methodology can be utilized on a wide range of problems. Today's requirements that analysis techniques be fast and easily applied to new problems where the underlying physics is not well understood and that the number of variables exceed the number of measurements. Contrast this to some of the problems outlined in previous sections. The prediction of friction factor has been carried out over hundreds of years, the prediction of critical heat flux and flow regimes decades. This new approach was able to match the state-of-the-art methods quickly without complicated one-of-a-kind modeling approaches. Using the methodology with datasets similar to those used in the literature, prediction of thermal-hydraulic phenomena was carried out with results matching the best of the most current published correlations. What sets this apart from previous modeling approaches is the ability to produce a probability density that allows the investigator to evaluate the confidence in the prediction and potentially feed that information along with the most likely value. This is similar to the confidence intervals for other regressions but with the caveat that other regressions do not scale well with the number of variables. The resulting confidence intervals for a high dimension correlation would likely be unreasonable for a small number of data.

The use of visual observation has been an important part of experiments and modeling. The ability to cognitively evaluate results simplifies the development of predictive models. In fact, one's intuition is commonly used to project the results based on previous experience. The methodology presented here is the machine equivalent to this intuition. The method uses past data to make projections of future events. Since it is a machine driven process, the number of dimensions can be dramatically increased beyond the capability of human observers allowing a new type of modeling and of collecting data. The implication of this is tremendous and provides a fruitful area of research in applying this methodology to a wide array of problems across disciplines.

REFERENCES

[1] M. TURK and A. PENTLAND, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience,* vol. **3**, pp. 71, (1991).

[2] J. KRUSKAL and M. WISH, *Multidimensional Scaling*: Sage Publications, Beverly Hills, CA., (1978).

[3] E. FARIA and C. PEREIRA, "Nuclear Fuel Loading Pattern Optimization Using a Neural Network," *Annals of Nuclear Energy,* vol. **30**, pp. 603, (2003).

[4] J. B. TENENBAUM, V. D. SILVA, and J. C. LANGFORD, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science,* vol. **290**, pp. 2319, December 22, (2000).

[5] S. T. ROWEIS and L. K. SAUL, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. **290**, pp. 2323, Dec 22 (2000).

[6] S. LAFON and A. B. LEE, "Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. **28**, pp. 1393, Sep (2006).

[7] E. BINGHAM and H. MANNILA, "Random Projection in Dimensionality Reduction: Applications to Image and Text Data," *International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA., pp. 245-250, (2001).

[8] Y. TSAIG and D. DONOHO, "Extensions of Compressed Sensing," *Signal Processing,* vol. **86**, pp. 549, (2006).

[9] B. BEAUZAMY, "The Experimental Probabilistic Hypersurface," *Laboratoire de Mathématiques et Applications*: l'Université de Bretagne Sud, (2004).

[10] O. ZEYDINA, "L'Hypersurface Probabiliste Construction Générale et Applications - Rapport no 4 adressé à l'Institut de RadioProtection et de Sûreté Nucléaire," in *Laboratoire de Mathématiques et Applications*: l'Université de Bretagne Sud, (2007).

[11] K. M. HURLBERT, L. C. WITTE, F. R. BEST, and C. KURWITZ, "Scaling Two-Phase Flows to Mars and Moon Gravity Conditions," *International Journal of Multiphase Flow,* vol. **30**, pp. 351, Apr (2004).

[12] J. R. FORD, A. J. MASLOWSKI, R. A. REDD, and L. A. BRABY, "Radiation Responses of Perfused Tracheal Tissue," *Radiation Research,* vol. **164**, pp. 487, Oct (2005).

[13] C. ESTRADA-PEREZ, E. DOMINGUEZ-ONTIVEROS, H. AHN, N. AMINI, and Y. HASSAN, "PTV Experiments of Subcooled Boiling Flow Through a Rectangular Channel," in *16th International Conference on Nuclear Engineering*, Orlando, FL, 2008.

[14] J. YANG, M. WARD, E. RUNDENSTEINER, and S. HUANG, "Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets," *ACM International Conference Proceeding Series*, vol. **40**, pp. 19, (2003)

[15] R. KOHAVI and F. PROVOST, "Applications of Data Mining to Electronic Commerce," *Data Mining and Knowledge Discovery,* vol. **5**, pp. 5, Jan-Apr (2001).

[16] R. C. FAIR, *Predicting Presidential Elections and Other Things*, Stanford University Press, Stanford, CA., (2002).

[17] G. ZWEIGER, *Transducing the Genome: Information, Anarchy, and Revolution in the Biomedical Sciences,* McGraw-Hill, New York, (2001).

[18] K. FUKUNAGA, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic Press, Boston, MA., (1990).

[19] T. F. COX, "Multidimensional Scaling Used in Multivariate Statistical Process Control," *Journal of Applied Statistics,* vol. **28**, pp. 365, Mar-May (2001).

[20] W. S. TORGERSON, "Multidimensional Scaling 1. Theory and Method," *Psychometrika,* vol. **17**, pp. 401, (1952).

[21] B. SCHOLKOPF, A. SMOLA, and K. R. MULLER, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation,* vol. **10**, pp. 1299, Jul 1 (1998).

[22] J. SHAWE-TAYLOR and N. CRISTIANINI, *Kernel Methods for Pattern Analysis*, University Press, Cambridge, UK., (2004).

[23] M. BELKIN and P. NIYOGI, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computing,* vol. **15**, pp. 1373, (2003).

[24] D. L. DONOHO and C. GRIMES, "Image Manifolds Which are Isometric to Euclidean Space," *J. Math. Imaging Vis.,* vol. **23**, pp. 5, (2005).

[25] F. SHA and L. K. SAUL, "Analysis and Extension of Spectral Methods for Nonlinear Dimensionality Reduction," in *Proc. of the 22nd International Conference on Machine Learning* Bonn, Germany: ACM, (2005).

[26] N. J. NILSSON, *Learning Machines; Foundations of Trainable Pattern-Classifying Systems*, McGraw-Hill, New York, (1965).

[27] S. S. LAFON, "Diffusion Maps and Geometric Harmonics," Yale University, pp. 125, (2004).

[28] R. R. COIFMAN and S. LAFON, "Diffusion Maps," *Applied and Computational Harmonic Analysis,* vol. **21**, pp. 5, Jul (2006).

[29] K. Q. WEINBERGER, F. SHA, and L. K. SAUL, "Learning a Kernel Matrix for Nonlinear Dimensionality Reduction," in *Proc. of the Twenty-First International Conference on Machine Learning,* Banff, Alberta, Canada: ACM, (2004).

[30] J. A. HARTIGAN, *Clustering Algorithms*, Wiley, New York, (1975).

[31] O. ZEYDINA, "L'Hypersurface Probabiliste Construction Pratique - Rapport no 1 adressé à l'Institut de RadioProtection et de Sûreté Nucléaire," in *Laboratoire de Mathématiques et Applications*: l'Université de Bretagne Sud, (2006).

[32] O. ZEYDINA, "L'Hypersurface Probabiliste Nouvelle Construction - Rapport no 2 adressé à l'Institut de RadioProtection et de Sûreté Nucléaire," in *Laboratoire de Mathématiques et Applications*: l'Université de Bretagne Sud, (2007).

[33] O. ZEYDINA, "L'Hypersurface Probabiliste Construction explicite à partir du Code Cathare - Rapport no 3 adressé à l'Institut de RadioProtection et de Sûreté Nucléaire," in *Laboratoire de Mathématiques et Applications*: l'Université de Bretagne Sud, (2007).

[34] G. H. GOLUB and C. F. VAN LOAN, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD., (1996).

[35] S. DASGUPTA, "Experiments with Random Projection," in *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*: Morgan Kaufmann Publishers Inc., (2000).

[36] W. JOHNSON and J. LINDENSTRAUSS, "Extensions of Lipschitz Mappings into a Hilbert Space," *Contemporary Mathematics,* vol. **26**, pp. 1, (1984).

[37] D. ACHLIOPTAS, "Database-Friendly Random Projections," in *Proc. of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* Santa Barbara, CA., ACM, (2001).

[38] E. CANDES, J. ROMBERG, and T. TAO, "Stable Signal Recovery from Incomplete and Inaccurate Measurements," *Commun. Pure Appl. Math., 2005,* (2005).

[39] G. CORMODE, M. DATAR, P. INDYK, and S. MUTHUKRISHNAN, "Comparing Data Streams Using Hamming Norms (How to Zero in)," *IEEE Transactions on Knowledge and Data Engineering,* vol. **15**, pp. 529, May-Jun (2003).

[40] E. CANDES and T. TAO, "Near Optimal Signal Recovery From Random Projections and Universal Encoding Strategies," *Arxiv preprint math.CA/0410542,* (2004).

[41] E. CANDES, "Compressive Sampling," *International Congress of Mathematics*, vol. **3**, pp. 1433, (2006).

[42] J. AMADOR, "Random Projection and Orthonormality for Lossy Image Compression," *Image Vision Comput.,* vol. **25**, pp. 754, (2007).

[43] W. U. BAJWA, J. D. HAUPT, G. M. RAZ, S. J. WRIGHT, and R. D. NOWAK, "Toeplitz-Structured Compressed Sensing Matrices," pp. 294, (2007).

[44] P. INDYK, "Explicit Constructions for Compressed Sensing of Sparse Signals," in *Proc. of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms,* San Francisco, CA., (2008).

[45] S. KASKI, "Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering," *Neural Networks Proc. 1998. IEEE World Congress on Computational Intelligence*, Anchorage, AK, May (1998).

[46] S. DASGUPTA and A. GUPTA, "An Elementary Proof of the Johnson-Lindenstrauss Lemma," Tech. Rep. TR-99-06, Intl. Comput. Sci. Inst . (1999).

[47] D. GRAHAM and N. ALLINSON, "Characterizing Virtual Eigensignatures for General Purpose Face Recognition," NATO ASI Series F," *Computer and Systems Sciences,* vol. **163**, pp. 446, (1998).

[48] J. DATTORRO, *Convex Optimization & Euclidean Distance Geometry*, Lulu. Com,, http://meboo.convexoptimization.com/Meboo.html, (2006).

[49] P. SCHNITER, L. POTTER, and J. ZINIEL, "Fast Bayesian Matching Pursuit," *IEEE Transactions on Signal Processing,* vol. **56**, 1, pp. 326, (2008).

[50] M. BELKIN, "Problems of Learning on Manifolds," The University of Chicago, Chicago, IL., p. 75, (2003).

[51] U. VAIDYA, G. HAGEN, A. BANASZUK, S. LAFON, I. MEZIC, and R. R. COIFMAN, "Comparison of Systems Using Diffusion Maps," in *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05. 44th IEEE Conference*, pp. 7931, (2005).

[52] R. AGRAWAL, C. H. WU, W. I. GROSKY, and F. FOTOUHI, "Diffusion Maps-Based Image Clustering," in *Proc. of the 2006 International Workshop on Research Issues in Digital Libraries* Kolkata, India: ACM, (2007).

[53] D. WASSERMANN, M. DESCOTEAUX, and R. DERICHE, "Diffusion Maps Clustering for Magnetic Resonance Q-ball Imaging Segmentation," *Journal of Biomedical Imaging,* vol. **8**, pp. 1, (2008).

[54] R. XU, S. DAMELIN, B. NADLER, and D. C. W. II, "Clustering of High-Dimensional Gene Expression Data with Feature Filtering Methods and Diffusion Maps," in *Proc. of the 2008 International Conference on BioMedical Engineering and Informatics* - vol. **1**: IEEE Computer Society, pp. 245, (2008).

[55] E. T. JAYNES, "Information Theory and Statistical Mechanics," *Physical Review,* vol. **106**, pp. 620, (1957).

[56] V. BARTKUT and L. SAKALAUSKAS, "Experimental Probabalistic Hypersurface Construction by Gaussian Fields," http://icm.mcs.kent.edu/research/RMM/archives/, (2006).

[57] J. WEISBACH, *Lehrbuch der Ingenieur- und Maschinen-Mechanik, Vol. 1. Theoretische Mechanik* vol. **1**. Braunschweig, Vieweg und Sohn, (1845).

[58] G. O. BROWN, "The History of the Darcy-Weisbach Equation for Pipe Flow Resistance," *Proc. of the 150th Anniversary Conference of ASCE*, Washington DC, pp. 34, (2002).

[59] B. CASTELLI, *Della Misura Dell'acque Correnti*, Stamparia Camerale, Rome, (1628).

[60] O. REYNOLDS, "An Experimental Investigation of the Circumstances Which Determine Whether the Motion of Water Shall be Direct or Sinuous and of the Law of Resistance in Parallel Channel," *Phil. Trans. of the Royal Soc.,* vol. **174**, pp. 935, (1883).

[61] G. HAGEN, "Ueber die Bewegung des Wassers in Engen Cylindrischen Röhren," *Annalen der Physik,* vol. **122**, pp. 423, (1839).

[62] J. POISEUILLE, "Experimental Research on the Movement of Liquids in Tubes of Very Small Diameters," *CR Acad. Sci., Paris,* vol. **11**, pp. 961, (1840).

[63] H. BLASIUS, "Mitt," Forschungsarb, (1913).

[64] C. COLEBROOK and C. WHITE, "Experiments with Fluid Friction in Roughened Pipes," *Proc. of the Royal Society of London. Series A, Mathematical and Physical Sciences (1934-1990),* vol. **161**, pp. 367, (1937).

[65] L. MOODY, "Friction Factors for Pipe Flow," *Trans. ASME,* vol. **66**, pp. 671, (1944).

[66] H. ROUSE, "Evaluation of Boundary Roughness," *Proc., 2nd Hydraulics Conf.,* The University of Iowa Studies in Engineering, Bulletin **27**, Wiley, New York, pp. 105, (1943).

[67] F. WHITE and C. KRAMER, *Fluid Mechanics*, McGraw-Hill, New York, (1994).

[68] T. DREW, E. KOO, and W. MCADAMS, "The Friction Factor for Clean Round Pipes," *Transactions of the American Institute of Chemical Engineers,* vol. **28**, pp. 56, (1932).

[69] C. H. LEES, "On the Effect of the Form of the Transverse Section on the Frictional Resistance to the Motion of an Elongated Body Parallel to Its Length through a Fluid," *Proc. Roy. Soc.* , vol. **92**, 636, pp. 144, (1915).

[70] J. FREEMAN, "Author Experiments Upon the Flow of Water in Pipes and Pipe Fittings," *ASME,* (1941).

[71] H. OMBECK, " Pressure of Flowing Air in a Straight Cylindrical Pipes," *Forschungsarbeiten,* Berlin, VDI-Vlg., (1914).

[72] T. STANTON and J. PANNELL, "Similarity of Motion in Relation to the Surface Friction of Fluids," *Proc. of the Royal Society of London. Series A,* vol. **90**, pp. 394, (1914).

[73] M. H. CLAPP and O. FITZSIMONS, "The Effect of Heat Transfer on Friction Factors in Fanning's Equation," (1929).

[74] S. HERMANN-BURBACH, "Wärmeübergang in Rohren," *Akadem. Verlagsges., Leipzig,* (1930).

[75] A. K. JAIN, "Accurate Explicit Equation for Friction Factor," *Journal of the Hydraulics Division-Asce,* vol. **102**, pp. 674, (1976).

[76] T. SERGHIDES, "Estimate Friction Factor Accurately," *Chemical Engineering,* vol. **91**, pp. 63, (1984).

[77] J. O. BABATOLA, A. M. OGUNTUASE, I. A. OKE, and M. O. OGEDENGBE, "An Evaluation of Frictional Factors in Pipe Network Analysis Using Statistical Methods," *Environmental Engineering Science,* vol. **25**, pp. 539, (2008).

[78] J. SONNAD and C. GOUDAR, "Explicit Reformulation of the Colebrook-White Equation for Turbulent Flow Friction Factor Calculation," *Ind. Eng. Chem. Res,* vol. **46**, pp. 2593, (2007).

[79] E. ROMEO, C. ROYO, and A. MONZÓN, "Improved Explicit Equations for Estimation of the Friction Factor in Rough and Smooth Pipes," *Chemical Engineering Journal,* vol. **86**, pp. 369, (2002).

[80] F. CHAPEAU-BLONDEAU and A. MONIR, "Numerical Evaluation of the Lambert W Function and Application to Generation of Generalized Gaussian Noise with Exponent 1/2," *Signal Processing, IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. **50**, pp. 2160, (2002).

[81] J. O. WILKES and S. G. BIKE, *Fluid Mechanics for Chemical Engineers*, Prentice Hall PTR, Upper Saddle River, NJ., (1999).

[82] C. BRANAN, *Rules of Thumb for Chemical Engineers: A Manual of Quick, Accurate Solutions to Everyday Process Engineering Problems*, 3rd ed., Gulf Professional Pub., New York, (2002).

[83] C. R. COLEBROOK, "Turbulent Flow in Pipes with Particular Reference to the Transition Region Between the Smooth and Rough Pipe Laws," *J. Inst. Civ. Eng,* vol. **1**, pp. 5024, (1938).

[84] D. WOOD, "An Explicit Friction Factor Relationship," *Civil Engineering,* vol. **36**, pp. 60, (1966).

[85] S. W. CHURCHILL, "Empirical Expressions for the Shear Stress in Turbulent Flow in Commercial Pipe," *AIChE Journal,* vol. **19**, pp. 375, (1973).

[86] N. H. CHEN, "Explicit Equation for Friction Factor in Pipe," *Industrial & Engineering Chemistry Fundamentals,* vol. **18**, pp. 296, (1979).

[87] G. F. ROUND, "An Explicit Approximation for the Friction Factor-Reynolds Number Relation for Rough and Smooth Pipes," *Canadian Journal of Chemical Engineering,* vol. **58**, pp. 122, (1980).

[88] D. I. H. BARR, "Accurate Explicit Equation for Friction Factor," *Journal of the Hydraulics Division-ASCE,* vol. **103**, pp. 334, (1977).

[89] D. I. H. BARR, "Solutions of the Colebrook-White Function for Resistance to Uniform Turbulent Flow," *Proc. of the Institution of Civil Engineers Part 2: Research and Theory,* vol. **71**, (1981).

[90] D. J. ZIGRANG and N. D. SYLVESTER, "Explicit Approximations to the Solution of Colebrook Friction Factor Equation," *AIChE Journal,* vol. **28**, pp. 514, (1982).

[91] S. E. HAALAND, "Simple and Explicit Formulas for the Friction Factor in Turbulent Pipe-Flow," *Journal of Fluids Engineering-Transactions of the Asme,* vol. **105**, pp. 89, (1983).

[92] G. MANADILI, "Replace Implicit Equations with Signomial Functions," *Chemical Engineering,* vol. **104**, pp. 129, Aug (1997).

[93] S. G. KANDLIKAR, "A Theoretical Model to Predict Pool Boiling CHF Incorporating Effects of Contact Angle and Orientation," *Journal of Heat Transfer-Transactions of the ASME,* vol. **123**, pp. 1071, Dec (2001).

[94] B. THOMPSON and R. V. MACBETH, "Boiling Water Heat Transfer Burnout in Uniformily Heated Round Tubes: A Complimation of World Data with Accurate Correlations," Winfrith, United Kingdom AEEW-R-356, (1964).

[95] D. D. HALL and I. MUDAWAR, "Critical Heat Flux (CHF) for Water Flow in Tubes - I. Compilation and Assessment of World CHF Data," *International Journal of Heat and Mass Transfer,* vol. **43**, pp. 2573, July (2000).

[96] D. D. HALL and I. MUDAWAR, "Critical Heat Flux (CHF) for Water Flow in Tubes - II. Subcooled CHF Correlations," *International Journal of Heat and Mass Transfer,* vol. **43**, pp. 2605, July (2000).

[97] D. GROENEVELD, L. LEUNG, P. KIRILLOV, V. BOBKOV, I. SMOGALEV, V. VINOGRADOV, X. HUANG, and E. ROYER, "The 1995 Look-up Table for Critical Heat Flux in Tubes," *Nuclear Engineering and Design,* vol. **163**, pp. 1, (1996).

[98] A. BERGLES, "Subcooled Burnout in Tubes of Small Diameter," ASME Paper No. 63-WA-182, (1963).

[99] C. F. BONILLA and C. W. PERRY, "Heat Transmission to Boiling Binary Liquid Mixtures," *Trans. AIChE,* **37**, pp. 685-705, (1941).

[100] A. SAKURAI and M. SHIOTSU, "Temperature-Controlled Pool-Boiling Heat Transfer," *Proc. International Heat Transfer Conference,* Tokyo, Japan, vol. **4**, CONF-740925--P4, (1974).

[101] J. H. LIU, L. YE, and H. S. LIU, "An Experimental-Study on the Critical Flux (Chf) of Flow Boiling in a Highly Viscous-Fluid in Vertical Tubes," *Chemical Engineering and Processing,* vol. **34**, pp. 35, Feb (1995).

[102] F. INASAKA and H. NARIAI, "Critical Heat Flux and Flow Characteristics of Subcooled Flow Boiling in Narrow Tubes," *JSME International Journal,* vol. **30**, pp. 1595, (1987).

[103] H. NARIAI, F. INASAKA, and T. SHIMURA, "Critical Heat Flux of Subcooled Flow Boiling in Narrow Tube," *ASME-JSME Thermal Engineering Joint Conference*, vol. **5**, pp. 455, (1987).

[104] R. D. BOYD, "Subcooled Water-Flow Boiling Transition and the L/D Effect on Chf for a Horizontal Uniformly Heated Tube," *Fusion Technology,* vol. **18**, pp. 317, Sep (1990).

[105] ORNATSKI.AP, CHERNOBA.VA, N. A. LAZAREV, and V. S. FURAEV, "Investigation of Influence of Eccentricity on Burnout Heat Transfer in Annular Channels," *Thermal Engineering,* vol. **16**, pp. 111, (1969).

[106] X. CHENG, F. J. ERBACHER, E. STARON, and W. ZEGGEL, "Critical Heat Flux in Circular Tubes at High Pressures and High Mass Fluxes," *Proc. of NURETH-5,* Salt Lake City, UT., pp. 832, (1992).

[107] K. M. BECKER, G. HERNBORG, M. BODE, and O. ERIKSSON, "Burnout Data for Flow of Boiling Water in Vertical Round Ducts, Annuli and Rod Clusters," *AE-177, Aktiebolaget Atomenergi,* Stockholm, Sweden, (1965).

[108] V. DOROSHCHUK, L. LEVITAN, and F. LANTZMAN, "Investigation into Burnout in Uniformly Heated Tubes," ASME Publication No. 75-WA/HT-22, (1975).

[109] D. GROENEVELD, B. KIAMEH, and S. CHENG, "Prediction of Critical Heat Flux (CHF) for Non-Aqueous Fluids in Forced Convective Boiling," *Proc. of the Eighth International Heat Transfer Conference,* vol. **5**, (1986).

[110] G. CELATA, M. CUMO, and A. MARIANI, "Assessment of Correlations and Models for the Prediction of CHF in Water Subcooled Flow Boiling," *International Journal of Heat and Mass Transfer,* vol. **37**, pp. 237, (1994).

[111] W. LIU, H. NARIAI, and F. INASAKA, "Prediction of Critical Heat Flux for Subcooled Flow Boiling," *International Journal of Heat and Mass Transfer,* vol. **43**, pp. 3371, (2000).

[112] G. SU, K. FUKUDA, K. MORITA, M. PIDDUCK, D. JIA, T. MATSUMOTO, and R. AKASAKA, "Applications of Artificial Neural Network for the Prediction of Flow Boiling Curves," *Journal of Nuclear Science and Technology,* vol. **39**, pp. 1190, (2002).

[113] M. ARIK and A. BAR-COHEN, "Effusivity-Based Correlation of Surface Property Effects in Pool Boiling CHF of Dielectric Liquids," *International Journal of Heat and Mass Transfer,* vol. **46**, pp. 3755, (2003).

[114] Z. DENG, "Prediction of Critical Heat Flux for Flow Boiling in Subcooled and Saturated Regimes." Dissertation, Chemical Engineering and Applied Chemistry, Columbia University, (1998).

[115] H. Y. CHEH and C. F. FIGHETTI, "Flow Excursion Experimental Program Single Tube Uniformly Heated Tests," CU-HTRF-T4(1990).

[116] X. CUI, *Prediction of Critical Heat Flux in Bundles Using Tube Look-Up Table*: National Library of Canada, Ottowa, ON., (1999).

[117] K. BECKER, G. STRAND, and C. OSTERDAHL, "Round Tube Burnout Data for Flow of Boiling Water at Pressure Between 30 and 200 bar," *Royal Institute of Technology, Laboratory of Nuclear Engineering,* KTH-NEL-14, Stockholm, Sweden, (1971).

[118] D. LEE, "An Experimental Investigation of Forced Convection Burnout in High Pressure Water, Part III: Long Tubes with Uniform and Non-uniform Axial Heating*",* AEEW-R355, (1965).

[119] H. SWENSON, J. CARVER, and C. KAKARALA, "The Influence of Axial Heat-Flux Distribution on the Departure from Nucleate Boiling in a Water-Cooled Tube," ASME Paper N62-WA-297 (1962).

[120] J. GRIFFEL, "Forced-Convection Boiling Burnout for Water in Uniformly Heated Tubular Test Sections," Thesis, Mechanical Engineering, Columbia University, New York, NYO-187-7, US Atomic Energy Commission, (1965).

[121] H. CHEH, C. FIGHETTI, and E. MCASSEY, "Onset of Flow Instability and Critical Heat Flux Experiments," Report No. CU-HTRF-T8, (1992).

[122] A. ORNATSKII and A. KICHIGIN, "An Investigation of the Dependence of Critical Thermal Loading on Weight Velocity, Underheating and Pressure," *Teploenergetika,* vol. **8**, pp. 75, (1961).

[123] Y. ZEIGARNIK, N. PRIVALOV, and A. KLIMOV, "Critical Heat Flux with Boiling of Subcooled Water in Rectangular Channels With One-Sided Supply of Heat," *Teploenergetika,* vol. **28**, pp. 48, (1981).

[124] P. WEBER and K. JOHANNSEN, "Study of Critical Heat Flux Condition at Convective Boiling of Water: Temperature and Power Controlled Experiments," *Proc. 9th Int. Heat Transfer Conf.*, Jerusalem Vol. **2**, pp. 63 (1990).

[125] A. BERGLES and W. ROHSENOW, "Forced-Convection Surface-Boiling Heat Transfer and Burnout in Tubes of Small Diameter," Massachusetts Institute of Technology, Cambridge, MA., US Atomic Energy Commission DSR Report 8767-21, (1962).

[126] O. PESKOV, I. SUBBOTIN, B. ZENKEVICH, and N. SERGEYEV, "The Critical Heat Flux for the Flow of Steam-Water Mixtures Through Pipes," *Problems of Heat Transfer and Hydraulics of Two-phase Media,* Pergamon Press, Oxford, (1969).

[127] A. LEZZI, A. NIRO, and G. BERETTA, "Experimental Data of CHF for Forced Convection Water Boiling in Long Horizontal Capillary Tubes," *Proc. of the Tenth Internal Heat Transfer Conference, Institution of Chemical Engineers*, vol. **7**, pp. 491, (1994).

[128] G. P. CELATA, M. CUMO, and A. MARIANI, "Burnout in Highly Subcooled Water Flow Boiling in Small Diameter Tubes," *International Journal of Heat and Mass Transfer,* vol. **36**, pp. 1269, (1993).

[129] F. CAMPOLUNGHI, M. CUMO, G. FERRARI, R. LEO, and G. VACCARO, "An Experimental Study on Heat Transfer in Long, Sub-Critical Once-Through Steam Generators," in *Reactor Heat Transfer*, Karlsruhe, Germany, (1973).

[130] B. A. ZENKEVICH, O. L. PESKOV, and V. I. SUBBOTIN, "A Study of Critical Heat Flux Densities for Tubular Fuel Elements in Atomic Power Stations," *Thermal Engineering* vol. **11**, pp. 23, (1964).

[131] R. MOSTELLER, S. FRANKLE, and P. YOUNG, "Data Testing of ENDF/B-VI with MCNP: Critical Experiments, Thermal-Reactor Lattices and Time-of-Flight

Measurements," *Advances in Nuclear Science and Technology,* vol. **24**, pp. 131, (1997).

[132] T. GOORLEY, "Criticality Calculations with MCNP5: A Primer," *Los Alamos National Laboratory, X-5,* (2004).

[133] E. F. PETRICOIN, A. M. ARDEKANI, B. A. HITT, P. J. LEVINE, V. A. FUSARO, S. M. STEINBERG, G. B. MILLS, C. SIMONE, D. A. FISHMAN, E. C. KOHN, and L. A. LIOTTA, "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer," *Lancet,* vol. **359**, pp. 572, Feb 16 (2002).

[134] J. LIU, G. CUTLER, W. LI, Z. PAN, S. PENG, T. HOEY, L. CHEN, and X. LING, "Multiclass Cancer Classification and Biomarker Discovery Using GA-Based Algorithms," *Bioinformatics,* vol. **21**, pp. 2691, (2005).

[135] E. FRANK, M. HALL, L. TRIGG, G. HOLMES, and I. WITTEN, "Data Mining in Bioinformatics Using Weka," *Bioinformatics,* vol. **20**, pp. 2479, (2004).

[136] G. WRIGHT JR, "SELDI Proteinchip MS: a Platform for Biomarker Discovery and Cancer Diagnosis," *Expert Review of Molecular Diagnostics,* vol. **2**, pp. 549, (2002).

[137] E. J. CANDES and M. B. WAKIN, "People Hearing Without Listening: An Introduction to Compressive Sampling," *IEEE Signal Processing Magazine*, to appear, http://www.acm.caltech.edu/~emmanuel/papers/spm-robustcs-v05.pdf (2007).

[138] O. BAKER, "Simultaneous Flow of Oil and Gas," *Oil Gas J,* vol. **53**, pp. 185, (1954).

[139] D. L. XIE, Z. Y. HUANG, H. F. JI, and H. Q. LI, "An Online Flow Pattern Identification System for Gas-Oil Two-Phase Flow Using Electrical Capacitance Tomography," *IEEE Transactions on Instrumentation and Measurement,* vol. **55**, pp. 1833, Oct (2006).

[140] J. H. CHANG, "Statistical Comparison of Two-Phase Flow, Void Fraction Fluctuations in a Microgravity Environment," Thesis, Nuclear Engineering, Texas A&M University, (1997).

[141] L. VALOTA, "Microgravity Flow Pattern Identification Using Void Fraction Signals," Thesis, Nuclear Engineering, Texas A&M University, (2005).

[142] J. G. COLLIER and J. R. THOME, *Convective Boiling and Condensation*, 3rd ed. Clarendon Press, Oxford, (1994).

[143] P. SPEDDING and V. NGUYEN, "Regime Maps for Air Water Two-Phase Flow," *Chem. Eng. Sci,* vol. **35**, pp. 779, (1980).

[144] J. WEISMAN, D. DUNCAN, J. GIBSON, and T. CRAWFORD, "Effects of Fluid Properties and Pipe Diameter on Two-Phase Flow Patterns in Horizontal Lines," *International Journal of Multiphase Flow,* vol. **5**, pp. 437, (1979).

[145] H. T. BI and J. R. GRACE, "Regime Transitions: Analogy Between Gas-Liquid Co-Current Upward Flow and Gas-Solids Upward Transport," *International Journal of Multiphase Flow,* vol. **22**, pp. 1, (1996).

[146] J. MANDHANE, G. GREGORY, and K. AZIZ, "A Flow Pattern Map for Gas-Liquid Flow in Horizontal Pipes," *International Journal of Multiphase Flow,* vol. **1**, pp. 537, (1974).

[147] Y. TAITEL and A. E. DUKLER, "Model for Predicting Flow Regime Transitions in Horizontal and near Horizontal Gas-Liquid Flow," *AIChE Journal,* vol. **22**, pp. 47, (1976).

[148] J. W. COLEMAN and S. GARIMELLA, "Two-Phase Flow Regimes in Round, Square and Rectangular Tubes During Condensation of Refrigerant R134a," *International Journal of Refrigeration-Revue Internationale Du Froid,* vol. **26**, pp. 117, Jan (2003).

[149] J. EL HAJAL, J. THOME, and A. CAVALLINI, "Condensation in Horizontal Tubes, Part 1: Two-Phase Flow Pattern Map," *International Journal of Heat and Mass Transfer,* vol. **46**, pp. 3349, (2003).

[150] E. JASSIM, T. NEWELL, and J. CHATO, "Probabilistic Determination of Two-Phase Flow Regimes in Horizontal Tubes Utilizing an Automated Image Recognition Technique," *Experiments in Fluids,* vol. **42**, pp. 563, (2007).

[151] O. C. JONES and N. ZUBER, "The Interrelation Between Void Fraction Fluctuations and Flow Patterns in Two-Phase Flow," *International Journal of Multiphase Flow,* vol. **2**, pp. 273, (1975).

[152] L. VALOTA, C. KURWITZ, A. SHEPHARD, and F. BEST, "Microgravity Flow Regime Data and Analysis," *International Journal of Multiphase Flow,* vol. **33**, pp. 1172, Nov (2007).

[153] D. LOWE and K. REZKALLAH, "Flow Regime Identification in Microgravity Two-Phase Flows Using Void Fraction Signals," *International Journal of Multiphase Flow,* vol. **25**, pp. 433, (1999).

[154] M. HUBBARD and A. DUKLER, "The Characterization of Flow Regimes for Horizontal Two-Phase Flow," *Proc. Heat Transfer and Fluid Mech,* Ed. MA Saad, University Press, Moller, Stanford, CA., (1966).

[155] G. SAETHER, K. BENDIKSEN, J. MULLER, and E. FROLAND, "The Fractal Statistics of Liquid Slug Lengths," *International Journal of Multiphase Flow,* vol. **16**, pp. 1117, (1990).

[156] R. KOZMA, H. KOK, M. SAKUMA, D. DJAINAL, and M. KITAMURA, "Characterization of Two-Phase Flows Using Fractal Analysis of Local Temperature Fluctuations," *International Journal of Multiphase Flow,* vol. **22**, pp. 953, (1996).

[157] C. SUNDE, S. AVDIC, and I. PAZSIT, "Classification of Two-Phase Flow Regimes via Image Analysis and a Neuro-Wavelet Approach," *Progress in Nuclear Energy,* vol. **46**, pp. 348, (2005).

[158] B. BALASKO, J. ABONYI, and B. FEIL, "Fuzzy Clustering and Data Analysis Toolbox," *Freely available Matlab package. http://www.fmt.vein.hu/softcomp/fclusttoolbox*.

[159] C. CROWLEY and M. IZENSON, "Design Manual for Microgravity Two-Phase Flow and Heat Transfer," Creare Report no. AL-TR-89-027, (1989).

VITA

Name: Richard Cable Kurwitz

Address: Texas A&M University Dept. of Nuclear Engineering, MS3133,
College Station, TX. 77843-3133

Email Address: kurwitz@tamu.edu

Education: B.S., Nuclear Engineering, Texas A&M University, 1993
M.S., Nuclear Engineering, Texas A&M University, 1997
Ph.D., Nuclear Engineering, Texas A&M University, 2009