Application of Materials Informatics for Structure - Property Correlations

A THESIS

submitted by

ANKITA RAJARAM NAIK

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

DEPARTMENT OF METALLURGICAL AND MATERIALS ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

May 2017

THESIS CERTIFICATE

This is to certify that the thesis titled **Application of materials informatics for structure - property correlations**, submitted by **Ankita Rajaram Naik**, to the Indian Institute of Technology, Madras, for the award of the degree of **Master of Technology**, is a bonafide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Anand Kanjarla
Project Guide
Assistant Professor
Dept. of Metallurgical and Materials Engineering
IIT - Madras, 600 036

Place: Chennai

Date: 5th May 2017

ACKNOWLEDGEMENTS

'*The true sign of intelligence is not knowledge, but imagination*' - Albert Einstein

I would first like to thank my thesis advisor Prof. Anand Kanjarla at Indian Institute of Technology Madras who introduced me to one of the most imagination driven fields of Materials Science, Hierarchical Materials Informatics. The door to Prof. Kanjarla's office was always open whenever I ran into a trouble spot or had a question about my project. He not only guided me through the process, but more importantly gave me the freedom to explore ideas he knew wouldn't work from his experience.

I am very grateful to ICME lab, which gave me access to its computational facilities without which completing even a quarter of this project would have been a dream. I would like to thank my parents and my brother for their continuous motivation in the form of snacks, scoldings and love which motivated me to complete my project on time.

Additionally, I would like to thank my friend Arjun Iyer for the number of mathematical discussions we had from Discrete Fourier Transforms to Support Vector Machines he was present to listen to my rant and provide suggestions based on his knowledge derived from Physics. A special mention to Asmita Jana (Jana), without those long Chicken Biryani clad discussions it would not have been possible to see the fun aspect associated with my failures, Srihari Sundar who helped me be virtually present in ICME by helping me to access all computational facilities needed and my dual degree classmates Vaishnavi, Hari, Srikar ,Gopakumar ,Abhishek who gave me the motivation when I felt I was the only one whose results were light years away from success.

ABSTRACT

KEYWORDS: Microstructure; Principal Component Analysis; Two-point correlations; Reduced-order Models; Structure-property relationships.

Description of the material's internal structure lies at the core of materials science. Recent advances has made it possible to capture the 3-D material structure spanning a multitude of length scales (e.g. X-ray micro-tomography, automated serial sectioning, 3-D atom probe). Advances in material science have led to the development of hundred thousands of distinct engineered and natural materials of interest. A computationally efficient approach for archival, retrieval and real-time exploration of the microstructure datasets is required. Efforts in this direction are hindered by the lack of a rigorous mathematical definition of the internal structure or microstructure of a material. This thesis outlines a rigorous mathematical framework for definition of microstructure as a stochastic process. In this framework the microstructure can be thought of as a set of statistical rules that govern the spatial placement of microstructure features, and observed micrographs are different realizations of the overriding process. This interpretation of microstructure allows for the quantitative comparison of different materials based on structure and more importantly allows for the quantification of the observed variance in samples with the same nominal processing history.

Structure-property correlations are defining concepts in the field of material science, but we still lack to have a well-defined systematic framework for quantitative comparison of microstructures from different material classes. Thus, a novel microstructure quantification framework devised on the lines of framework, published by Niezgoda et. al. (Niezgoda *et al.*, 2013), that facilitates visualization of complex microstructure relationships within and across material classes has been mentioned here. In addition the framework will be used to link microstructure visualizations with properties to develop reduced-order microstructure-property linkages and performance models.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

DFT	Discrete Fourier Transform
IID	Independent and identically distributed random variable
MSD	Microstructure Sensitive Design
PCA	Principal Component Analysis
PSP	Process-Structure-Property
RVE	Representative Volume Element
SVD	Singular Value Decomposition
SVM	Support Vector Machines
WLS	Weighted Least Squares

NOTATION

$\mathbf{h}$	Local State
$\mathbf{H}$	Local State Space
$\mathbf{m_s^n}$	Probability of finding the material local state n in spatial bin s
$\mathbf{FZ}_\rho$	Fundamental Zone (i.e. all physically distinct crystal lattice orientations in the thermodynamic phase ρ that can theoretically occur in nature)
$\mathbf{g}$	Overall transformation matrix from the sample reference frame to the local crystal frame
ρ	Phase identifiers
$\mathbf{c_i}$	Chemical Composition of species identified by i
$\mathbf{C_i^\rho}$	Thermodynamically allowed solubility of the chemical species i in the phase ρ
ω	Experimental outcome, micrographs
$\mathbf{m}(\mathbf{x}, \mathbf{h})$	Microstructure function with local state h and x spatial point
$\mathbf{f^n}$	1-point statistics for state local state n
$\mathbf{S}$	Total number of Spatial bins
$\mathbf{f_r^{np}}$	2-point correlation of local state n and p along vector r
$\lvert\mathbf{M_k^n}\rvert$	Amplitude of DFT transform of point correlations
$\theta_\mathbf{k}^\mathbf{n}$	Phase of DFT transform of point correlations

CHAPTER 1

INTRODUCTION

1.1 Motivation

Advances in development of materials with enhanced performance characteristics has led to rapid development of technologies critical for their development. Such advancements in materials characteristics form the back-bone of economic prosperity in present day. Modern materials cannot be completely understood by their chemistry and bulk processing routes alone. A key to modern materials development lies in the in-depth understanding of the hierarchical three-dimensional (3-D) internal structure, or microstructure, which spans several disparate length scales and the optimization of its myriad details. Though, structure, properties and processing form the central theme of the field of material science it is just recently that tools necessary for exploring and harnessing multi-scale microstructure sensitive design have been devised. Due to the complexity of microstructure in virtually all engineering and natural materials, the traditional approaches to material development, relying on combinatoric experimentation guided by engineering principles and physical intuitions, have categorized and exploited a small handful of the readily accessible material microstructures.

Through the advent of sophisticated physics-based multi-scale modelling and simulation tools, material science is undergoing a transition from a data-limited field to a data-driven but analysis limited field. Large experiments and simulations are contributing to formation of data repositories loaded with information. But radically new approaches are required to convert these extensively large datasets into *useful materials knowledge*. Formation of a well-structured databases effectively capturing all the three legs of the triangle i.e. structure, properties and processing is critical for the successful utilisation of the database for materials design. While helping us reduce our dependency on slow combinatoric experimental approaches.

1.2 Scope of present research

While a significant amount of development, across the entire materials community, is required to even outline the structure and specific requirements of databases mentioned in section1.1, this thesis focuses on the definition and description of a computational framework and tools for the visualization, analysis, and quantitative comparison of microstructure datasets consisting of ensembles of 3-D micrographs collected for each material.The work addresses several aspects necessary for development and deployment of materials databases.

Following key ideas and features behind the proposed microstructure visualization space will be explored in detail throughout the thesis:

1. The inherent randomness possessed by microstructure suggests that the mathematics of stochastic processes can be used for description of the multi-scale variability of materials.
2. Such an approach leads to significant abstraction of the data and the resulting data lies on the surface of a complex high-dimensional manifold. In order to simplify the representation of data, dimensionality reduction techniques like principal component analysis (PCA) were used for reduced-order representation.
3. The reduced-order representation forms the basis of the visualization space. Each sample is represented by a unique point in this space. Points that are nearer to each other in this space correspond to samples with similar structure, whereas points further apart in the space represent appreciable structural differences. Also, it is envisioned that this space will serve as the interface between the large scale database of microstructure data and user. This space can provide tools which would help the user to interact with microsructure data through visual inspection of micrographs or 3-D datasets.
4. The space spanned by all characterized samples from a material or material class gives an indication of the microstructure variability with the material. This scatter in the visualisation space can be directly tied to the scatter in properties or response of the material.
5. Tasks such cluster analysis and classification are formulated based on the visualization space.
6. Explicit mapping of properties or performance characteristics into the visualization space can be performed to formulate invertible structure-property linkages.

1.3 Framework

The overall framework utilized in this thesis for setting up structure-property correlations can be denoted as follows:

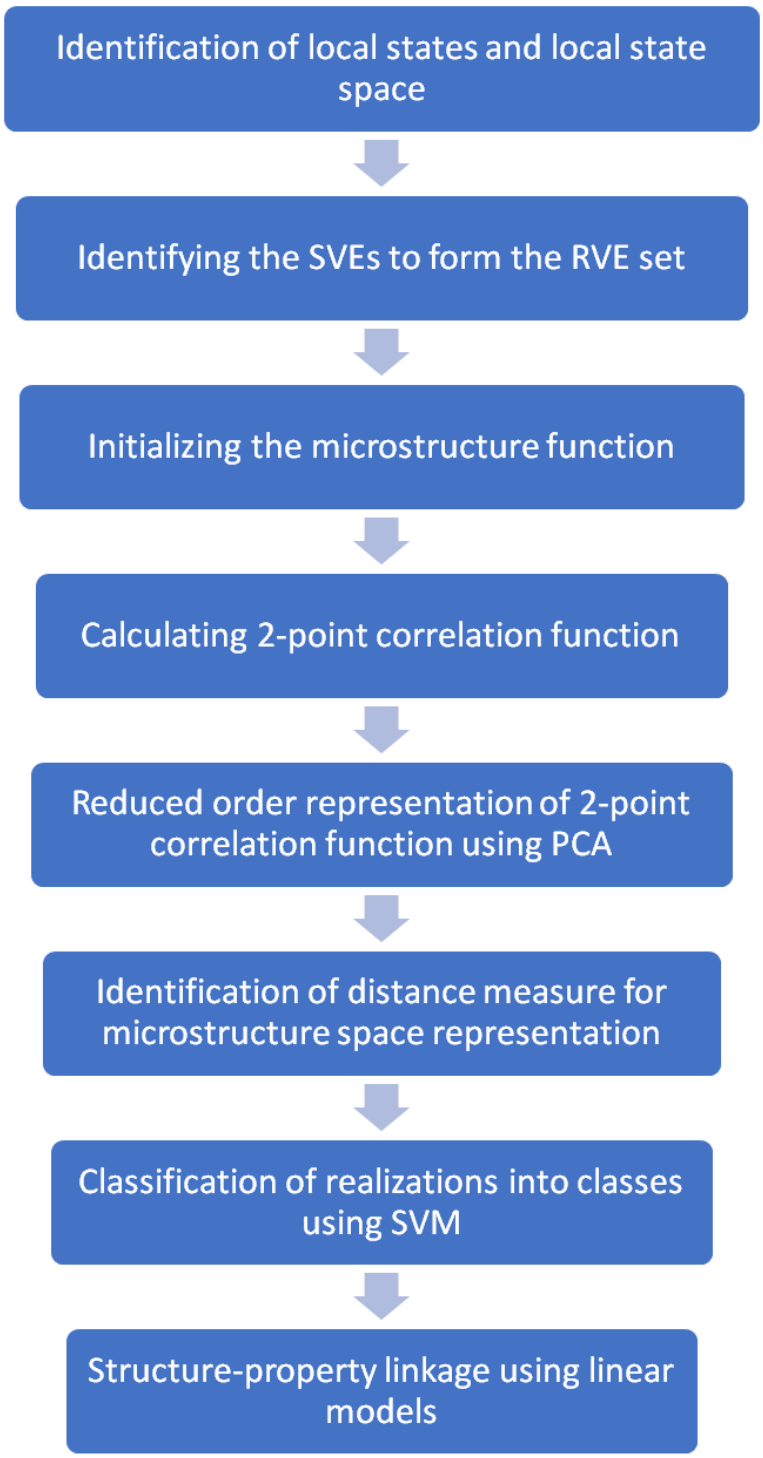

Figure 1.1: Framework for setting up structure-property linkages from micrograph realizations.

1.4 Organization of thesis

The key ideas described above are organized in this thesis in the following fashion:

Chapter 2 covers the background and prior work on this topic conducted by Prof.Kalidindi (Fullwood *et al.*, 2010) and Zabaras (Ganapathysubramanian and Zabaras, 2008; Sundararaghavan and Zabaras, 2005, 2004). It explains the steps of the Microstructure Design Framework and mentions the correlations between the framework and work in this thesis.

Chapter 3 introduces the common terminology used throughout the thesis. Following which an in-depth elaboration of the framework used to mathematically describe a microstructure is demonstrated using examples. Further the reduced order microstructure framework outlining the formation of the structure-property space and mathematical tools like principal component analysis necessary for the framework are covered.

Chapter 4 implements the reduced order microstructure framework defined in the previous chapter on porous composites. It explores different distance measures, classification algorithms and prediction methods to help establish structure-property correlations while validating the accuracy of the methods using simulated data.

Chapter 5 revisits the ideas mentioned in the section1.2 (scope of the thesis) thus summarizing the final outputs from the presented work. It is followed by a detailed description of limitations of the work outlined here and closes with a brief discussion of the future prospects of the work. Future prospects are listed majorly to address the limitations mentioned and possible solutions to be explored for the same.

CHAPTER 2

LITERATURE SURVEY

As a field material science is defined by the relationships between microstructure, properties and processing. Since the Bronze age, empirical experimentation guided by luck has largely contributed to the development of this field. The focus of materials science shifted from craft technologies to scientific investigation in the 1800s with William Gibb's ground breaking work 'On the Equilibrium of Heterogeneous Substances' (Gibbs, 1878) and William Roberts-Austen's delineation of the Iron-Carbon phase diagram (Tyrkel, 1968). Since then we have been slowly progressing towards the field of quantitative materials science.

Availability of micro-structure information and property measurements has driven the progress of this field via accumulation of databases linking structure and property (Ashby and Cebon, 1993; Selector, 2008; rep, 2006). Modern data mining techniques such as neural networks have been used to invert the relationship between microstructure and its properties to predict optimum structures (Lewis *et al.*, 2008; Sundararaghavan and Zabaras, 2009; Warde and Knowles, 1999; Mandal *et al.*, 2008). Development of such predictive models has been accelerated due to the availability of empirical micro-structure data. By combining such models, bridging various length scales and using a system approach, the first "material-by-design" concept was introduced in 1997 (Olson, 1997). Despite all our advances we are far away from developing truly predictive and invertible P-S-P (Process-Structure-Property) linkages required for success at tasks such as materials-by-design.

Following the lead Adams, Gamestani and Kalidindi developed the Micro-structure Sensitive Design (MSD) framework to place the material-by-design in a rigorous mathematical analysis and reduce the size of the design space when dealing with the structures at various length scales (Fullwood *et al.*, 2010). A broad overview of this approach is as follows:

1. Principal properties of interest in the design must be identified. For eg. Designer might be interested in the ultimate tensile strength of the material depending on the application it is being used for. Thus, UTS becomes the principal property.

2. Once the set of properties of interest are defined, a set of initial probable materials must be selected. Design handbooks or material selection tools such as Ashby charts (Ashby and Cebon, 1993) can be used to select a set of these materials.

3. A suitable mathematical definition of micro-structure must be applied to the materials. This definition must be consistent with the homogenization relationships and other structure/property models.

4. The design space of the material is determined by defining all possible micro-structures that may occur. This is called as micro-structure hull.

5. Relevant homogenization relationships must be identified for the properties of interest in terms of the micro-structure description. These constitute the length scales, local states, local space and representative volume element for the property of interest.

6. Homogenization relationships must be exercised over the complete micro-structure hull to specify all sets of possible combinations of principal properties.

7. Once the designer has searched the property closure for the optimal combinations relating to the engineering application of interest, the desired point or points are mapped back to the micro-structure hull to discover relevant optimal micro-structure for the application. This mapping may or may not be unique.

8. Finally, its not necessary that we may be aware about a processing route helping us arrive at the micro-structure evolved from theory. Hence, a framework must be developed for describing the effect of processing on micro-structure and subsequently on properties. This results into a set of path lines between the micro-structure hull and property closures that represent the evolution of the materials under the processes being modelled.

The success of the entire framework mentioned above depends on our ability to find a comprehensive and compact mathematical description of the micro-structure that is amenable to optimization. Early efforts at designing such a framework focused on volume fraction descriptions of the micro-structure (Fullwood *et al.*, 2010; Adams *et al.*, 2001; Kalidindi *et al.*, 2004), which through 1st-order homogenization techniques are sufficient to place only fairly wide bounds on most properties of interest to designers. Thus, the following work included in this report comprises sections motivated from Kalidindi's work (Kalidindi, 2015) to mathematically define the micro-structure and represent micro-structures in a format on which optimization can be carried out using correlation statistics.

As more characterized microstructure information and property measurements have become available, progress in this field has progressed via the accumulation of databases linking structure and properties (Ashby and Cebon, 1993; Selector, 2008) and modern data-mining techniques such as neural networks have been used to invert the relationships and predict optimum structure (Lewis *et al.*, 2008; Mandal *et al.*, 2008; Warde and Knowles, 1999; Sundararaghavan and Zabaras, 2009).

CHAPTER 3

MATHEMATICAL FRAMEWORK FOR MICROSTRUCTURE REPRESENTATION

The previous chapters along with exploring historically used methods for materials development also emphasized on the importance of the microstructure description and quantification, not only as a digital representation but also as the main foundational block in establishing high value PSP linkages needed to accelerate the materials development efforts. They comprise of a through account of how material science data has been previously used for prediction of properties using data analysis techniques. In this chapter, we would be understanding the basic terminology and establishing the mathematical rigorous framework for the representation of material microstructure.

3.1 Terminology

3.1.1 Microstructure

Microstructure refers to myriad features of internal structure in heterogeneous materials across many disparate length scales. At course length scales (μm - mm), materials are partitioned into regions of nearly continuous composition, phase or properties such as grains, fibers or precipitates. Deeper insights into smaller length scales reveals the details about the atomic structure of these partitions, such as lattice constants or lattice orientations. Still deeper inquiry (Å- μm), leads to the study of defects in the atomic structure of constituent phases including the state of vacancy or interstitial concentrations, the presence of fine scale precipitates, voids, grain boundaries, and the state of dislocation in the material. Beyond the atomic structure we encounter electronic and quantum states of the atoms (Fullwood *et al.*, 2010). Example of such a manifestation of hierarchical microstructure observed in nacre is shown in Figure 3.1. Depending on the application, some or all these features of the internal structure of materials impact the

properties and ultimate performance of a material. The collection of important structural features to properties of interest will collectively be termed the microstructure of a material.

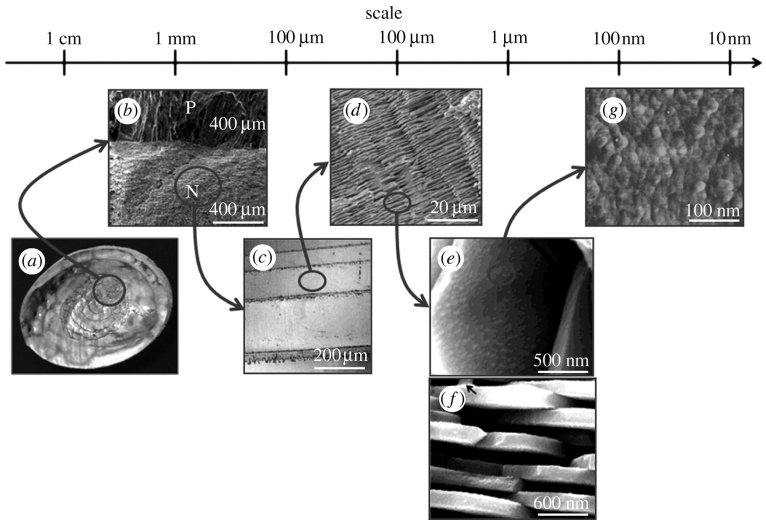

Figure 3.1: Hierarchical organization in nacre showing at least six structural levels (Luz and Mano, 2009).

Fortunately, a complete description of internal structure is not always required for modelling the microstructure-properties relationships of materials. But having an unambiguous understanding of the impact of the spatial distributions of microsructural features on properties and performance could serve the purpose of setting up such relationships.

3.1.2 Length Scales

The concept of homogenization required for the MSD framework mentioned in Chapter 2 requires the definition of well separated length scales. For eg. consider the hypothetical composite material system in Figure 3.2. The length scales of interest depicted in the figure are macroscale and mesoscale. Let L and l define the length scales of the material point at the two scales being investigated (i.e. macroscale and mesocale respectively).

The L and l are selected such that they represent the smallest acceptable homogenization length scales for macro and meso scales respectively. Define L such that it facilitates a sufficiently accurate description of the material's constitutive response for

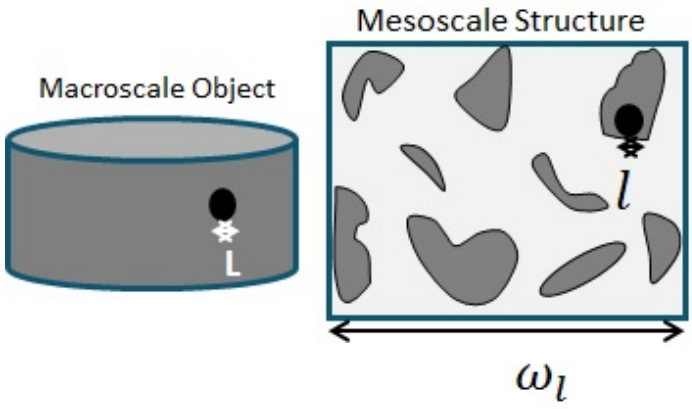

Figure 3.2: Schematic description of length scales involved in hierarchical description of microstructure (Kalidindi, 2015)

all design consideration at macroscale and l describes the same at mesosclae. Assume that the macroscale and mesoscale quantities do not vary much over length scale L and l respectively. For e.g. in an experiment the length scale is determined by the apparatus or for numerical simulations each material point used in simulation is used as length scale. Properties assigned to material point are implicitly assumed to reflect average values over volume defined by length scales. Thus, at any length scale there exists significant heterogeneity at lower length scales.

3.1.3 Representative Volume Element

Separated length scales is major concept invoking the concept of representative volume element . The commonly adopted definition of RVE in current literature implicitly identifies a finite region that is a 'statistically representative' of the whole, while effectively smoothing out the local spatial heterogeneities so that the macroscale properties of the sample are captured to a **desired accuracy** (Drugan and Willis, 1996; Kanit *et al.*, 2003; Gusev, 1997; Chen and Liu, 2004). The above definition focuses on the convergence in the prediction of selected marcoscale properties as opposed to explicitly considering whether or not the RVE has captured the desired microstructural details to sufficient accuracy as discussed by (Hill, 1963). Incidentally the classical definition of RVE provided by (Hill, 1963) requires the RVEs to be large enough to capture both the representative microstructure as well as its homogenised effective properties. By using the definition provided by Hill the RVE will be a function of both effec-

tive property under consideration and the contrast in local property values between the constitutive microstructure components. RVE for highly non-linear properties or high contrast composites (e.g. porous materials) will be substantially larger than for low contrast composites (e.g. phase separated metallic alloys) or linear properties (Niezgoda *et al.*, 2010). Thus, utilising such a description for a practical purpose would mean an extremely large RVE. So, an ensemble of smaller regions, often termed as statistical volume elements(SVEs), is used to approximate effective properties. As shown for the microstructure in figure3.3 the regions marked in yellow are the three SVEs which are being used for representing the microstructure.

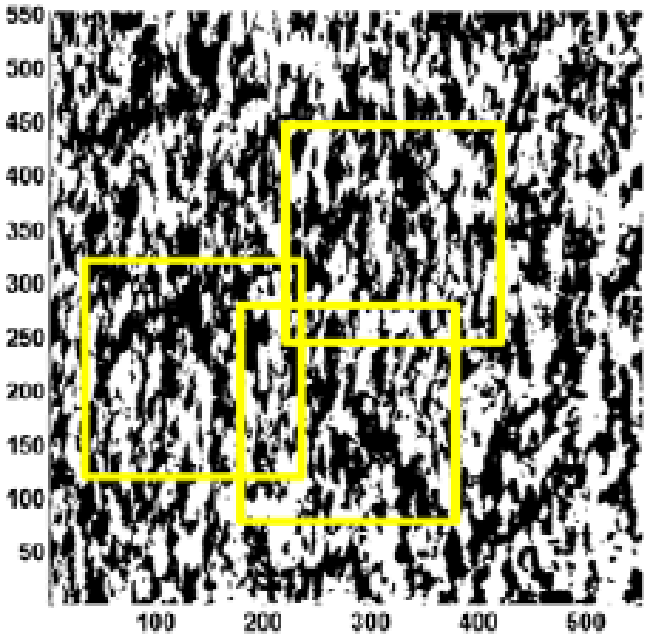

Figure 3.3: Yellow squares marking the smaller regions denoting three SVEs of the given microstructure are shown (Niezgoda *et al.*, 2010)

In this thesis we would be deviating from the classical definition of RVE proposed by Hill (Hill, 1963). We aim to define RVEs such that they replicate the measured spatial statistics of structure (as opposed to resulting effective properties). The properties of a material are complex function of its structure and previous work in constructing property closure has demonstrated that the mapping from microstructure to properties is many-to-one (Adams *et al.*, 2005; Kalidindi *et al.*, 2006). Thus, there is no guarantee that just because RVE captures effective property the underlying structure is also suitably captured. But it is expected that if the microstructure is represented to a high degree of accuracy, then a broad range of effective properties will also be captured

automatically.

Utilising an extremely large RVE for single simulation is a computationally expensive affair. At the same time using large number of SVEs for the complete representation is practically and computationally impossible. In this case, it would be beneficial to construct an optimized ensemble of SVEs whose members were not randomly sampled but selected to best represent the underlying material structure with as few members as possible. In this way, researchers could construct the best possible estimates of material behaviour while minimizing resources required. In this work, we will refer to such optimized ensembles as RVE sets.

3.1.4 Microstructure Hull

The space outlined by the SVEs from the RVE sets, for a particular material class is defined as the microstructure hull for that particular class. For e.g. if we are representing a material class in 3-D using a microstructure hull the points forming the microstructure hull will be utilised from the RVE set and these points are the SVEs of the microstructure. Each SVE should correspond to a point on the hull. Throughout the thesis we would be assuming that the SVEs selected by us for defining the RVE sets are sufficient to completely define the property under consideration. And each SVE is an iid (Independent and identically distributed) realisation of the material class.

3.1.5 Local States, Local State Spaces

As discussed in the sections above microstructure exhibits rich details that span several hierarchical length scales. Hypothesis of existence of homogenized properties at each length scale implicitly leads to definition of local states. A local state can be defined as a microstructure constituent that can be assigned a distinct local structure. A complete set of such local states that could theoretically be encountered across the length scales of interest is defined as the local state space.

At meso-scale local state (h) and local state space (H) which consider combination of phase identifier, chemical composition and crystal lattice orientation for polycrystalline

material can be defined as follows:

$$h = (\rho, c_i, g) \tag{3.1}$$

$$H = (\rho, c_i, g) \mid \rho \in (\alpha, \beta, \gamma, ...),\ c_i \in C_i^{\rho},\ g \in FZ_{\rho} \tag{3.2}$$

Where,

h : Local State

H : Local State Space

FZ_{ρ}: Fundamental Zone (i.e. all physically distinct crystal lattice orientations in the thermodynamic phase ρ that can theoretically occur in nature)

g : Overall transformation matrix from the sample reference frame to the local crystal frame

ρ : Phase identifiers

c_i : Chemical Composition of species identified by i

C_i^{ρ} : Thermodynamically allowed solubility of the chemical species i in the phase ρ

Points associated with distinct local states have different effective properties.

3.2 Microstructure as Stochastic Process

Success of the formation of an optimized RVE set depends on our ability to represent a microstructure mathematically. Thus, a rigorous framework for quantitative description of microstructure is required.

Microstructure is defined based on its internal structure and property dependence in section3.1.1. We also know that when a material undergoes a process it never produces the exact same micrograph during multiple trials (here micrograph is referred to the

structure obtained at the end of processing carried out at a particular trial). Only when we consider a large number of such events we tend to get the same result.

The theory used by (Niezgoda, 2010) to define a microstructure as a stochastic process will be used in this thesis. A stochastic process is defined as a set of rules that assigns to every experimental outcome a function. In the case of microstructure, each micrograph is an experimental outcome (ω), and there exists some set of rules that assigns a local state to each infinitesimal point in the micrograph, or equivalently there exists a stochastic process that assigns a local state field to each micrograph. This stochastic process will be referred to as the microstructure. The microstructure is a process; it is the set of rules that governs the placement of local states. Through characterization we view realizations of this process as micrographs or three-dimensional structure datasets. In turn, we can estimate the distributions associated with the microstructure from the micrographs. These ideas will be further clarified and put in a more formal mathematical framework in the upcoming sections.

In order to avoid confusion, with the frequent and necessary use of the word microstructure, when referring to any aspect of the internal topology, morphology or chemistry of a material, the term microstructure will be used when specifically referring to the stochastic process that describes the material structure, or equivalently the overall structure of the material as a whole. Micrograph will refer to features observed in a particular realisation or experimental outcome of this process.

3.2.1 Microstructure Function

Theoretically, it is possible to describe the observed microstructure of a single micrograph in a deterministic manner by exactly defining the local state of each point x. If we know the local state of each point exactly then we can completely describe the structure-property model. However, such a model doesnot take into consideration the hierarchy of material structure at various length scales. Thus, in order to represent microstructure in a statistical form Niezgoda (2010) has defined it in terms of concepts in stochastic processes. We would be using a similar description of microstructure in this work.

Consider a probability space defined by the ordered triplet (Ω, F, P) , where the set of all possible events F is the set of all possible arrangements of local states $h \in H$. The sample space (Ω) consists of an ensemble of micrograph realizations (either characterized, digital or even hypothetical), where each micrograph realization is considered an experimental outcome Ω. The microstructure function ($m(x, h)$) is then interpreted as the rules that assigns a function (local state field), $m(x, h, \omega)$, to every micrograph realization.

For a particular micrograph realisation ω, the function $m(x, h, \omega)$ is formally defined as

$$m(x, h, \omega)dh = \frac{dV}{V} \tag{3.3}$$

where, $\frac{dV}{V}$ indicates volume fraction of material associated with local state within a neighbourhood element dh about h, in a neighbourhood of volume V surrounding point x. Equation 3.3 normalised by volume of realisation ω such that

$$\frac{1}{vol(\omega)} \iint_{x \in \omega \; h \in H} m(x, h, \omega)dhdx = 1 \tag{3.4}$$

But the microstructure function is mostly considered to be discrete as majority of the microstruture identification techniques have limits on values of x recorded thus making it discrete. Also, measurements only identify a local state h with an uncertainty of dh circling it. Thus, making the microstructure function decently discrete with respect to h and x.

Eigen Microstructure

1. Material has a finite number of discrete local states
2. Local state at each of the sampling point is assigned fully to only one of the allowed local states
3. Discretized representation of the microstructure is binary

Most advanced material characterization equipment's produce eigen microstructure.

3.2.2 Digital Representation of Micrographs

Discretized microstructure function can be denoted as m_s^n and represents total volume fraction of all local states from bin n in spatial bin s.
Equations 3.3 and 3.4 can be represented as follows:

$$\sum_{n=1}^{N} m_s^n = 1 \qquad\qquad 0 \leq m_s^n \leq 1 \tag{3.5}$$

Where,

Spatial domain bins, s = $1, 2, 3, ..., S$

Discrete Local States, n = $1, 2, 3, ..., N$

Let us define the microstructure function for a checkerboard microstructure as shown in figure3.4

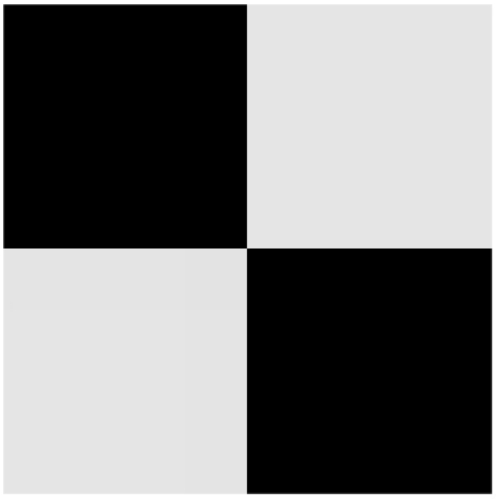

Figure 3.4: Schematic representation of an eigen microstructure with two local states and four spatial bins

Let,

n = 1 for region white

n = 2 for region black

s = 4 spatial bins

Thus, microstructure function:

$$m^1_{(1,1)} = 1$$
$$m^2_{(1,1)} = 0$$
$$m^1_{(1,2)} = 0$$
$$m^2_{(1,2)} = 1$$

Thus, it can be seen that bin (1,1) is filled with region white and bin(1,2) is filled with region black as per the above equations.

3.3 Statistical Quantification of Material Structure

The mathematical formalism of stochastic processes has been demonstrated to be a powerful tool for quantification of microstructure variability and variance (Niezgoda *et al.*, 2011). Having set up the mathematical framework for defining a microstructure this section will be used to introduce a rigorous statistical quantification of the material internal structure. This would help us to identify and quantify local patterns in the internal morphology of the material structure. Consequently, in this section m^n_s simply denotes the volume fraction of the material local state belonging to the local state bin indexed by n in the spatial bin indexed by s. Thus, making m^n_s making the probability of finding the material local state n in spatial bin s.

3.3.1 Spatial Correlations

1-point Statistics

1-point statistics is the overall volume fraction of local states identified as n in the material for an ensemble having J elements or the expected value of m^s_n.
Formula for 1-point statistics:

$$f^n_s = \frac{1}{J}\sum_{j=1}^{J} m^{n(j)}_s \tag{3.6}$$

Assumptions:

1. Spatial Homogeneity at all length scales i.e. $f_s^n = f^n$ thus f^n is expected value of m_s^n in any selected spatial bin.
2. Material hierarchical structure of interest exhibits well separated salient length scales.
3. Periodicity of microstructure
4. All the j microstructures have same number of spatial bins (S)

$$f^n = \frac{1}{SJ} \sum_s \sum_{j=1}^{J} m_s^{n(j)} \tag{3.7}$$

where,

S = Total number of spatial bins
s = vector index (indices of discretized points in the microstructure)
r = enumerates the corresponding discretized vectors that can be thrown into the same microstructure

1- point auto-correlation can be thought as the ratio of all the vectors r of length 1 which can be thrown into the microstructure such that their tail and head lies in white region to the total number of vectors r of length 1 which can be accommodated in the microstrucure. It gives us information about what we can expect to see at any spatial point in the microstructure. It does not carry any information about the relative spatial placement of local states in the materials internal structure.

2-point Statistics

Similar to 1-point statistics 2-point statistics can be thought to be the conditional probability of finding local states n at spatial bin s while simultaneously finding local state p at spatial bin $s + r$.
It can be mathematically represented as follows:

$$f_r^{np} = \frac{1}{JS_r} \sum_s \sum_{j=1}^{J} m_s^{n(j)} m_{s+r}^{p(j)} \tag{3.8}$$

All assumptions mentioned for 1-point statistics should hold true while calculating 2-point statistics.

Also, lower order correlations can be calculated from higher order correlations. E.g. 2-point statistics at r = 0 will give the 1-point statistics.

3.3.2 Redundancies in 2-point Correlations

Calculation of n-point statistics is a computationally expensive process. The convolution and correlation properties of Discrete Fourier Transform (DFTs) can be used to reduce the computational burden. As mentioned in appendixA, DFT co-efficients tend to have a one-to-one relationship with the function defined. Thus, reconstruction of the function from their DFT transforms can be envisaged.

Owing to such properties of DFT it becomes a very useful tool for calculating the n-point statistics of microstructure function. For utilisation of DFT we have to assume that the microstructure function in use is periodic.
The complex-values DFT for a microstructure function can be expressed as follows:

$$M_k^n = \Im(m_s^n) = \sum_{s=0}^{S-1} m_s^n \exp^{\frac{2sk\pi}{S}} = |M_k^n| \exp^{i\theta_k^n} \tag{3.9}$$

Where,

$|M_k^n|$= Amplitude of the DFT

$\theta_{\mathbf{k}}^{\mathbf{n}}$ = Phase of DFT

To understand the redundancies in 2-point correlations of microstructure function lets consider an example of a 3-D microstructure realisation with N local states. The complete set of these 2-point correlations contains N^2 correlations for each value of the 3-D vector index r. Thus, for a 3D microstructure with 2 local states (i.e. N=2 say black and white) four 2-point correlations will be obtained.

Simple reflection on the definition of 2-point correlations mentioned in section3.3.1 reveals the following redundancies (Gokhale *et al.*, 2005; Frisch and Stillinger, 1963):

1. The probability of finding ordered local states n and p in two spatial bins separated by r is the same as finding the same local states in the reverse order in row

spatial bins separated by $-r$.

$$f_r^{np} = f_{-r}^{np} \tag{3.10}$$

where, all the above symbols have standard meaning. For a 2-point correlation of 3-D microstructure based on the above redundancy cross-correlations when head of vector r is at local state black and its tail is at local state white is equivalent to having a vector $-r$ with its head at local state white and tail at local state black.

2. Sum of the row (an equivalent relation can be derived for summing up a column) of the 2-point correlation would recover the corresponding 1-point correlation.

$$\sum_{p=1}^{N} f_r^{np} = f^n \tag{3.11}$$

where, all the above symbols have standard meaning. The 2-point correlations for 2 local states would follow the relation $f_r^{11} + f_r^{12} = f^1$ i.e .probability I start from a white state (where n=1 represents a white local state) given my end can be either black/white is equivalent to probability of finding a white phase in the microstructure.

Based on the redundancies measured above it can be shown that for a microstructure with N local states only N-1 2-point correlations are independent (Niezgoda *et al.*, 2008*a*). Thus, for a structure with two local states only one 2-point correlation is independent.

3.3.3 Computation and Visualization of 2-point Correlations

Example of 1-point and 2-point Statistics

The pymks package of python (Wheeler *et al.*, 2016) was used to calculate the auto and cross correlation for a $8X8$ checkerboard eigen and periodic microstructure. The results of both have been depicted graphically in the following graphs.

Observations from the auto correlation graph:

1. Auto correlation at (0,0) is always 0.5. i.e. at r = 0 the conditional probability of finding a state black at tail and black at head for a periodic eigen microstructure is 0.5. Or conditional probability of finding a state white at tail and white at head for a periodic eigen microstructure is also 0.5.
2. Similarly the auto-correlations for all vectors r for state n can be obtained by placing vectors of length r on the graph of state n.

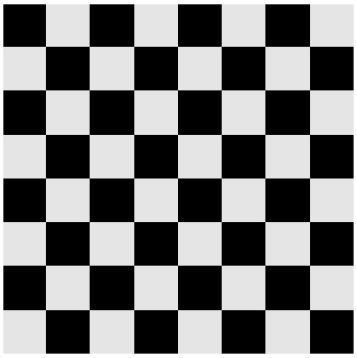

Figure 3.5: Schematic representation of a checkerboard eigen microstructure with two local states and 64 spatial bins

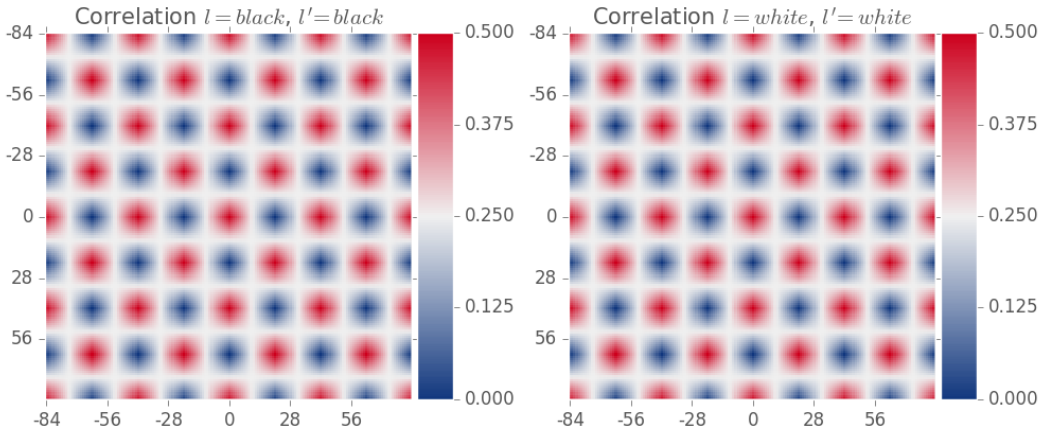

Figure 3.6: Autocorrelation visualisation of microstructure in figure3.5

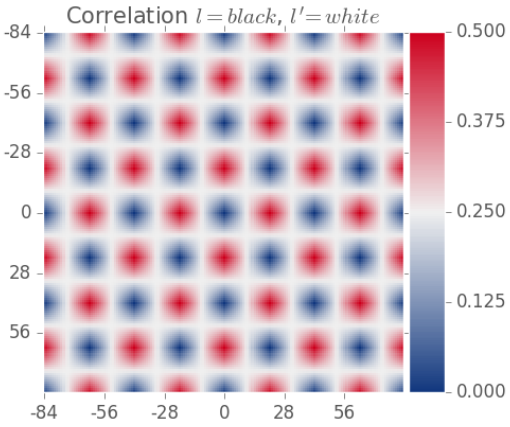

Figure 3.7: Crosscorrelation visualisation of microstructure in figure3.5

Observations from the cross correlation graph:

1. Cross correlation at (0,0) is always zero. i.e. at r = 0 the conditional probability of finding a state black at tail and white at head is 0.
2. Similarly the cross-correlations for all vectors r can be obtained by placing vectors of length r on the graph.

3.4 Reduced Order Representation of Microstructure Statistics using Principal Component Analysis

There are numerous approaches to dimensionality reduction including spectral methods, manifold learning approachs, metric multidimensional scaling, etc (Lee and Verleysen, 2007). The requirements of proposed microstructure space places strong constraints on the type and complexity of the dimensionality reduction scheme. In order to determine an appropriate low dimensional representation the following factors were considered:

1. The approach must be computationally robust and insensitive to the type or structure of the underlying data.
2. The representation must be able to be updated in real time as new information or datasets are added to the system.
3. The representation must be invertible in real time, so that the microstructure space can be used to explore new microstructures in both an interpolative or extrapolative manner.
4. In order to facilitate the computation of descriptive and relational statistics the representation should be an orthogonal decomposition of the data so that each dimension in the reduced frame can be considered as independent variables.

Based on this requirements and consideration of complexity and performance of PCA in previous work by Niezgoda *et al.* (2011), it was chosen for the task of dimensionality reduction.

3.4.1 Principal Component Analysis

The main purpose of PCA is to identify linear patterns and reduce the dimensions of the data with minimal loss of information. PCA reduces the dimensions of a d-dimensions data frame to a k-dimensions ($k < d$) data frame by projecting the d-dimensions data

frame onto a smaller subspace with k dimensions.

This procedure is carried out using the eigenvectors and eigenvalues of the initial data frame.

Step by Step implementation of PCA

1. Generating a sample
2. Data standardization
3. Calculate the Covariance Matrix, Eigenvectors and Eigenvalues
4. Choosing the principal components
   (a) Order the eigenvectors depending on their eigenvalues, highest to lowest
   (b) Form a feature vector by leaving the less important eigenvectors
   (c) Derive the new dataset by pre-multiply the feature vector with the row standardized data

Decomposition package from sklearn (Pedregosa *et al.*, 2011) was used for carrying out PCA in examples to follow in this thesis. The solver used for carrying out the eigenvector and eigenvalue identification was 'auto'. Thus, the SVD (Singular Value Decomposition) solver is selected by default policy based on the shape of input matrix and number of PCA components to be returned. If the input data is larger than 500x500 and the number of components to extract is lower than 80% of the smallest dimension of the data, then the more efficient 'randomized' method (Halko *et al.*, 2011) is enabled. Otherwise the exact full SVD is computed and optionally truncated afterwards.

CHAPTER 4

CASE STUDY: APPLICATION OF REDUCED ORDER MICROSTRUCTURE REPRESENTATION FOR STRUCTURE - PROPERTY CORRELATIONS

Previous chapter introduced the framework and methods required for the reduced order representation of a microstructure. The hypothesis that this reduced order representation can be used as features for establishing the structure-property linkage, mentioned in section1.2, will be tested in this chapter. The framework explained in Section1.3 will be applied on a three-dimensional two-phase microstructure. The major aim of this example is to demonstrate the use of the framework for classification of materials based on their inherent similarities and differences while further establishing the property-structure relationship. In the course of which we would be statistically quantifying the microstructure, reducing dimensions of its salient features using principal component analysis, visualising the reduced order form in the microstructure space and setting up the structure property relationship. Finally, the output from this chapter is expected to verify the key ideas and features listed in Section1.2.

4.1 Description of Microstructure Dataset

The dataset used for the example consists of 8 different material classes which were generated from thresholding random fields. The 8 material classes are all porous composites with constant pore density and varying degree of anisotropy in pore shape, connectivity and spatial distribution. The volume fraction of pores is all the material classes is maintained constant at 25%. Each material classes contains an ensemble of 50 realisations (each realisation can be thought of as a SVE in the RVE set to be constructed through this example 3.3.3) which would be used for training and testing of the classification and prediction algorithms. The mechanical response (i.e. C_{11} and σ_{yield}) of each volume element has been evaluated using an image based fast Fourier transform (FFT)

elastic-viscoplastic model, that directly uses three dimensional voxel material volumes as input. The three dimensional voxel material dataset and the mechanical response has been obtained from the work of (Niezgoda *et al.*, 2013). Thus, a complete statistical description of the formation of the dataset and the evaluation of mechanical response can be found in the paper (Niezgoda *et al.*, 2013). Figure4.1 gives a three dimensional visualisation of all the 8 material classes with the statistics used for there construction. These statistics highlight the relative anisotropy of pore shape and placement in the microstructure class. As shown in Figure4.1, Class 1 and 2 are isotropic, Class 3 and 5 are anisotropic along X-axis (i.e. e1 axis as shown in Figure4.1), Class 4 and 7 are anisotropic along Z-axis (i.e. e3 axis) and Class 6 and 8 are anisotropic along all the three dimensions.

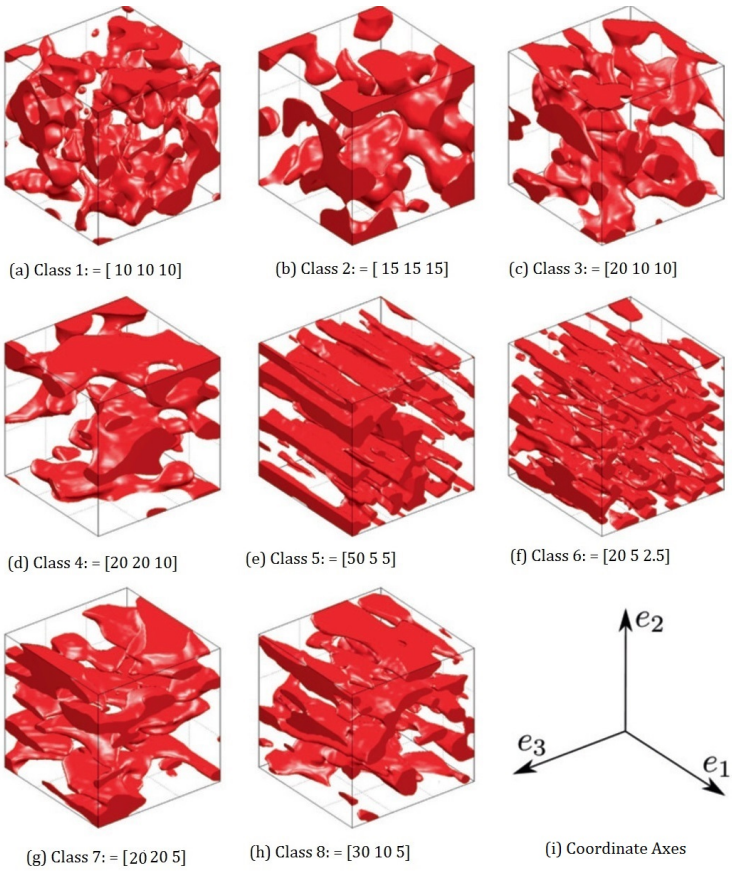

Figure 4.1: Representative examples of the 8 material classes of porous solids used in the study. The pores are shown in red while the isotropic matrix is transparent. The statistics used for formation of each classes has been mentioned from (a) to (h) as a subcaption below it. (i) shows the coordinate axes used to describe the material directions (Niezgoda *et al.*, 2013)

4.2 Statistics of Microstructure

1-point Statistics

As discussed in section3.3.1 the 1-point statistics of a microstructure manifests into being the volume fraction of the phases present. The volume fraction of pores for all samples across all classes is maintained constant. Thus, the 1-point statistic for all the samples is 0.25 for the red phase (i.e. pores) and 0.75 for the transparent phase (i.e. the solid composite).

2-point Statistics

The 2-point statistic of a microstructure is expected to capture the inherent structural variation like the variation induced in our example due to the spatial distribution of pores. It is expected that materials having similar statistics (i.e. statistics used to construct the microstructure) or orientation should have similar 2-point correlation function given the constituent phases and 1-point statistic remains constant. Thus, for verifying the observations mentioned in the previous section a comparative analysis of 2-point correlations of samples present in each of the 8 classes was done. Also, as mentioned in section3.3.2 due to redundancies present in 2-point correlations all characteristics of the correlation can be explained with a minimum of N-1 correlations (where, N is the number of phases). Thus, for our example (N = 2) only one correlation is sufficient for fully explaining the correlation characteristics of our microstructure samples. So, only the cross-correlations were utilised. PYMKS (Wheeler *et al.*, 2016) package utilised for 2-point correlation in the section3.3.3 showed erroneous result for 2-point correlation of 3D microstructure wherein the (0,0,0) value of the autocorrealtion was not equal to the volume fraction of the phases. Thus, algorithm suggested by Cecen *et al.* (2016) has been used for construction of 2-point correlation function for the 3-D microstructure realisations. Observations from the 2-point correlation functions:

1. Class 1 and Class 2 are isotropic materials thus the 2-point correlation function as seen in figure4.2 (a) to (c) and (d) to (f) respectively have similar structural patterns. The pattern is symmetric about (0,0,0) in all the directions.

2. Class 3 and Class 5 have anisotropy along X-axis. Thus, the 2-point correlations are only symmetric in the YZ plane as seen in figure4.3 (a) and (d). Thus, the image (a) and (d) is a mirror image about Y=Z.

3. Class 4 and Class 7 are anisotropic along Z-axis. A nature similar to the previous point is observed for these classes as well.

4. Class 6 and Class 8 have anisotropy in all the three dimensions and figure4.5 rightly demonstrates the same.

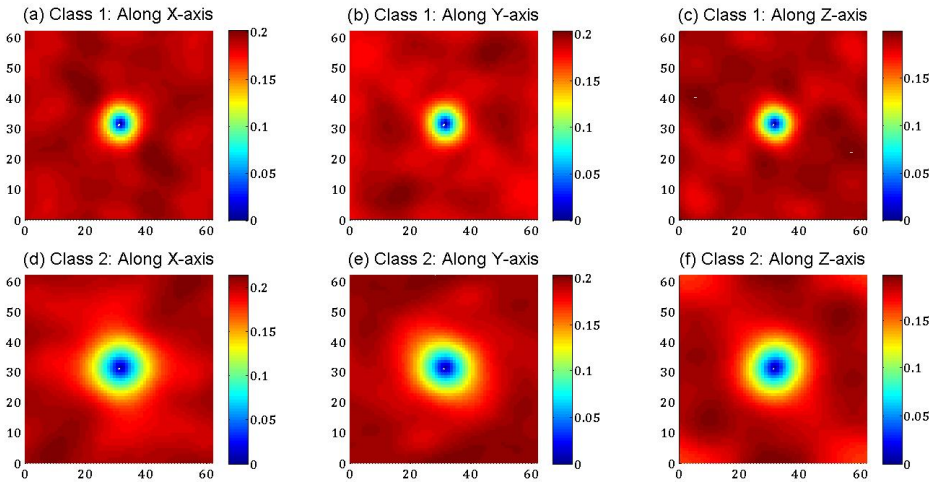

Figure 4.2: 2-point cross - correlation of Class 1 and 2 along all the three dimensions highlighting the isotropy in both the material classes. (a) to (c) captures variation along X-axis, Y-axis and Z-axis respectively for Class 1 whereas (d) to (f) covers the same for Class 2.

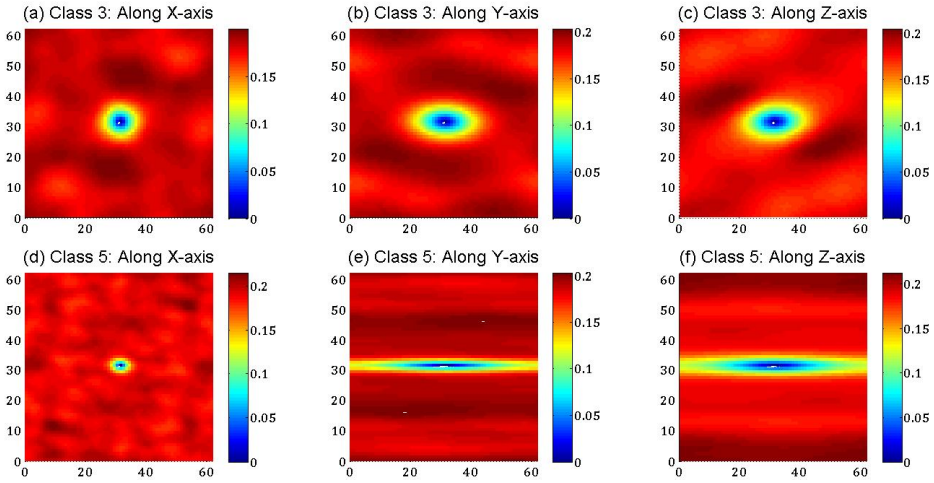

Figure 4.3: 2-point cross - correlation of Class 3 and 5 along all three dimensions highlighting the inherent anisotropy along X-axis. (a) to (c) captures variation along X-axis, Y-axis and Z-axis respectively for Class 3 whereas (d) to (f) covers the same for Class 5.

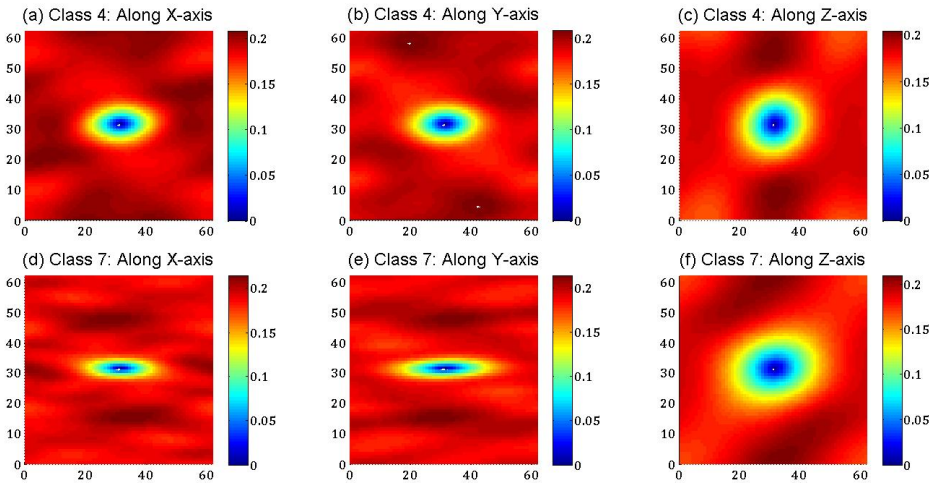

Figure 4.4: 2-point cross - correlation of Class 4 and 7 along all three dimensions highlighting the inherent anisotropy along Z-axis. (a) to (c) captures variation along X-axis, Y-axis and Z-axis respectively for Class 4 whereas (d) to (f) covers the same for Class 7.

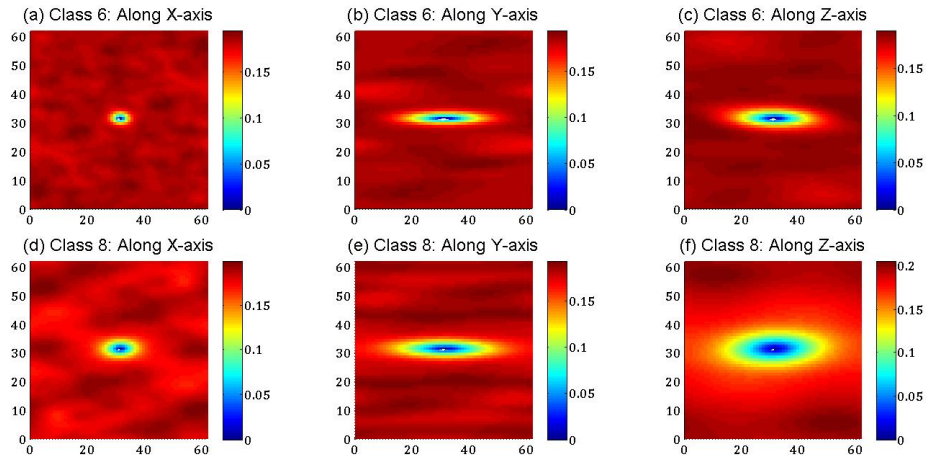

Figure 4.5: 2-point cross - correlation of Class 6 and 8 along all the three anisotropic dimensions. (a) to (c) captures variation along X-axis, Y-axis and Z-axis respectively for Class 6 whereas (d) to (f) covers the same for Class 8.

4.3 Microstructure Space

The n-point correlation functions lies on a very high dimensional space and thus it cannot be used for efficient computation, visualization or analysis (Niezgoda *et al.*, 2008*b*). For even simple material classes like the two phase example considered in this thesis and restricting to only the 2-point correlation we still require reduced order representation of the microstructure statistics. Based on the requirements of the proposed microstructure space explained in section3.4 principal component analysis was chosen for this study.

As discussed before the 2-point correlation data not only lies on a very high dimensional space but also has inherent non-linearity present in it (i.e. the data naturally lies on an embedded curved manifold in the high-dimensional space rather than on a hyperplane). However, as discussed by (Niezgoda *et al.*, 2013) it can be shown that when different materials are of same "family", such as the the various classes of porous composites explored in our case study, a suitable compact representation can be developed and useful visualisations could be obtained through a low dimensional representation.

For this study, two PCA decompositions were preformed to generate two-tired representation. An intra-class and inter-class representation was utilised to understand the relationship between the classes. The intra-class representation will be used to explore microstructure variance within the class and inter-class variance will be used to understand the relationships across multiple classes. A different distance measure (Mahalanobis distance explained in AppendixB) was used for the inter-class representation. Convex hulls bounding realizations of each microstructure class are plotted to help visualize the differences between volume of space occupied by the members of each class.

4.3.1 Visualizations in the microstructure space

An inter-class PCA is carried out on the microstructure dataset and the dataset is projected onto the first three principal directions. The advantage of projecting it on the first three principal directions is that these directions denote the highest variability directions. Thus, a representation from these three directions tends to capture the

maximum variance of the microstructure space. This does not, however, guarantee that the view is optimal for separating the different realizations by microstructure class. As seen in figure4.6 the different microstructure classes are indeed heavily overlapped in three dimensions, indicating that more terms are needed to delineate the different microstructure classes. At a glance figure4.6 shows the relationships between the different microstructure groups.

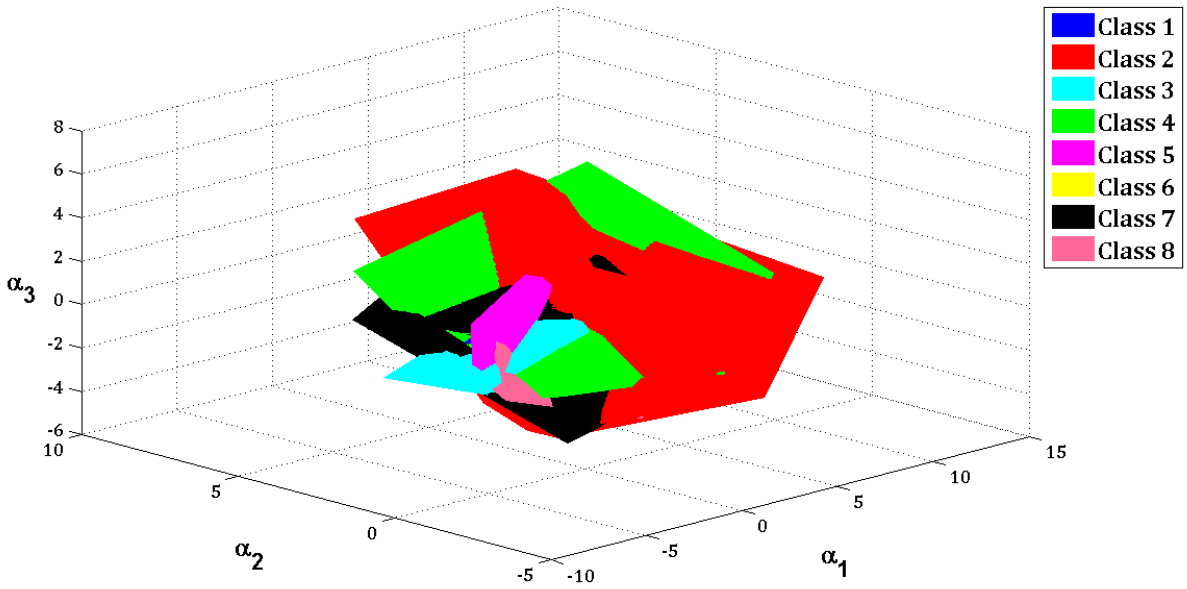

Figure 4.6: Three-dimensional microstructure hulls of all 8 microstructure classes.

Distances in this space serve as an indication of how similar or different structures are. Material classes whose bounding hulls are close (or overlapping) share more common features than those classes whose bounding hulls are well separated.

4.3.2 PCA on 2-point correlations versus microstructures

PCA reduction has been used by Sundaraghavan and Zabras for reduced order representation of microstructure to construct microstructure libraries(Sundararaghavan and Zabaras, 2004, 2005). In their work reduction was carried out on microstructure function directly rather than the statistical descriptors.

PCA is a linear transformation and thus effective dimensionality reduction can only be accomplished if the data can be approximately fitted to an embedded linear manifold (hyperplane) in the high dimensional space. Unfortunately, the underlying data in microstructure datasets rather lies on a highly non-linear surface and requires large number of principle components for reasonable representation (Ganapathysubramanian and Zabaras, 2008). Some basic properties of the n-point correlations, including having a natural origin at r = 0 and translation invariance, greatly reduce this non-linearity and adequate representation can be produced with only a few principal components. The eigenvalues of the decomposition give an indication of how significant each principal component is to the representation of the data-set, and the rate of decay of the eigenvalues gives an indication of how many terms must be kept for accurate representation. Thus, the rate of decomposition of eigenvalues or saturation of variance explained due to each principal component was plotted. Figure4.7 shows the rate of saturation for first 100 principal components for the inter-class PCA performed on microstructure images and PCA performed on spatial correlations.

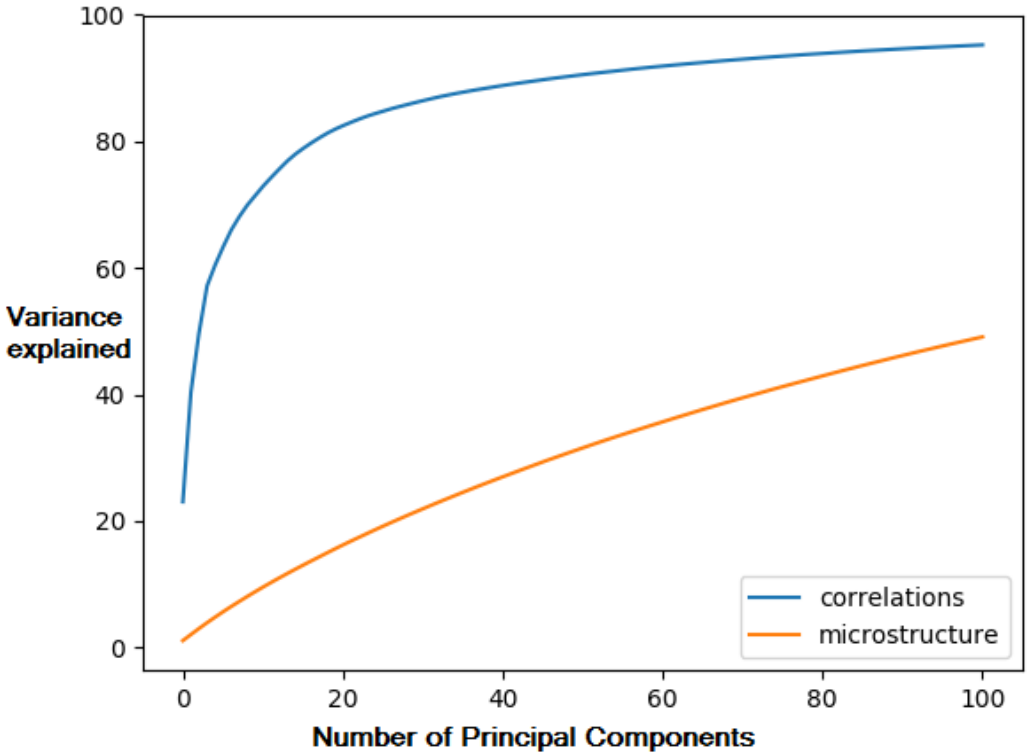

Figure 4.7: Rate of saturation of variance for decomposition on the microstructure images directly (red) and on the two point correlations (blue).

For the inter-class PCA on 2-point correlations there is a sharp increase in the variance when the number of principal components increase from 10 to 20 but the variance explained remains almost constant for PCA on the microstructure dataset directly. This verifies the results of mentioned by Zabras and lays the ground for utilisation of 2-point correlation rather than the microstructure datasets.

4.3.3 Relational Statistics in Microstructure Space

In order to effectively use the PCA space for microstructure quantification an appropriate distance measure should be defined. For intra-group comparison between individual points (i.e. how similar two points are) the Euclidean measure is sufficient. But for inter-group comparison we need a distance measure which should be able to distinguish between classes efficiently i.e. trying to increase the variance between the classes. Thus, Mahalanobis distance measure was used (AppendixB).

Consider a dataset of K material class ensembles where $\Omega_k = \omega_1^k, \omega_2^k, ..., \omega_p^k$ denotes the P_k members of the k^{th} ensemble. Each ω_p^k is represented by the J weights ${}^k\alpha_j^p$ by projecting them onto the PCA space. For compactness, let A_{pk} denote the J dimensional column vector of PCA weghts for the p^{th} member of the k^{th} ensemble. Further let $\overline{A_k}$ indicate the mean vector of PCA weights over all P_k members of ensemble k. The within group scatter is calculated by taking the intra-class sum of squares and cross product (SSCP) matrix, W, defined as follows:

$$W = \sum_{k=1}^{K} \sum_{p=1}^{P} (A_{pk} - \overline{A_k})(A_{pk} - \overline{A_k})^T \tag{4.1}$$

The between group scatter is characterized by the inter-class SSCP matrix defined as

$$B = \sum_{k=1}^{K} P^k (\overline{A_k} - \overline{A})(\overline{A_k} - \overline{A})^T \tag{4.2}$$

where $\overline{A}$ is the mean vector of PCA weights across all K ensembles which is by definition the zero vector. The representation of data in terms of canonical variables can be found by performing eigenvalue decomposition on $W^{-1}B$ hen projecting the PCA rep-

resentation onto the eigenvectors. Let $L = eigenvectors(W^{-1}B)$, then the canonical representation can then be found as $c_{pk} = LA_{pk}^{T}$. The canonical representation is useful for visualizing clustering. Figure4.8 and Figure4.9 show the projection of microstructure data onto the first two and first three canonical variables.

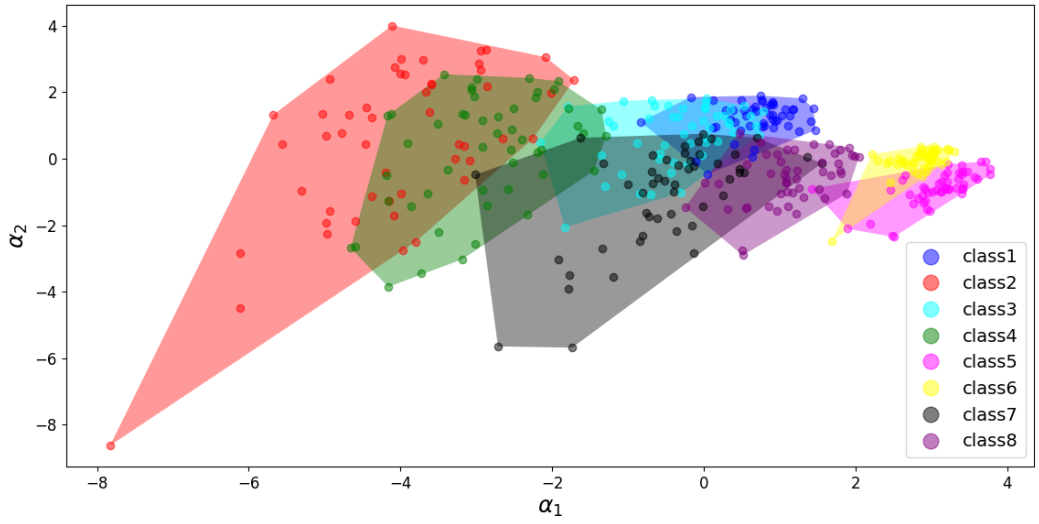

Figure 4.8: Projection of the microstructure data onto the first two canonical variables α_1 and α_2

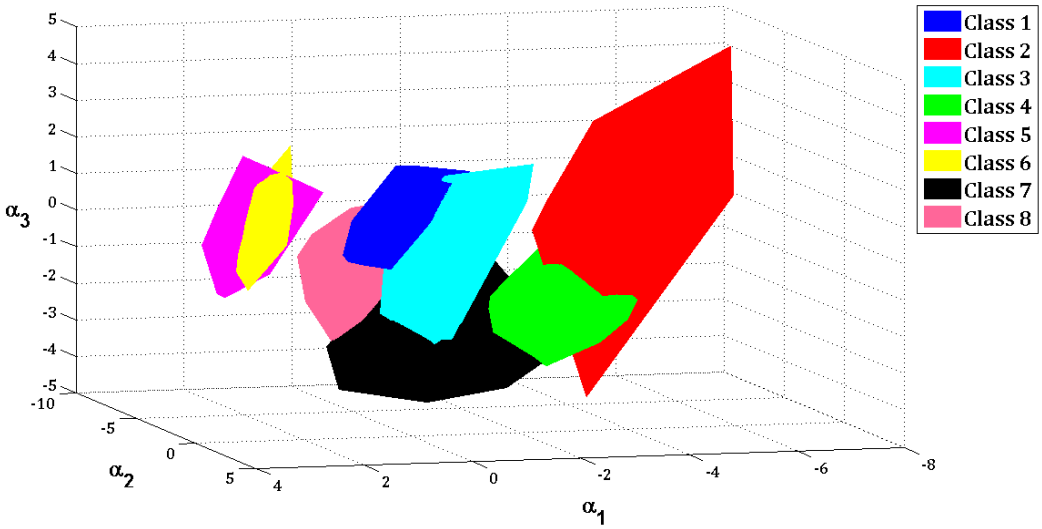

Figure 4.9: Projection of the microstructure data onto the first three canonical variables α_1 and α_2

It can be seen that the microstructure representation has lesser overlaps in 3-D than that in 2-D. Mahalanobis distance measure can be used to further explore the relationship between classes. A hierarchical cluster analysis was run on the class centroids

of 3-D mahalanobis distance measures (i.e. the class centroids derived from canonical variables). The dendogram for the hierarchical clustering showing the expected distance between groups is captured in figure4.10.

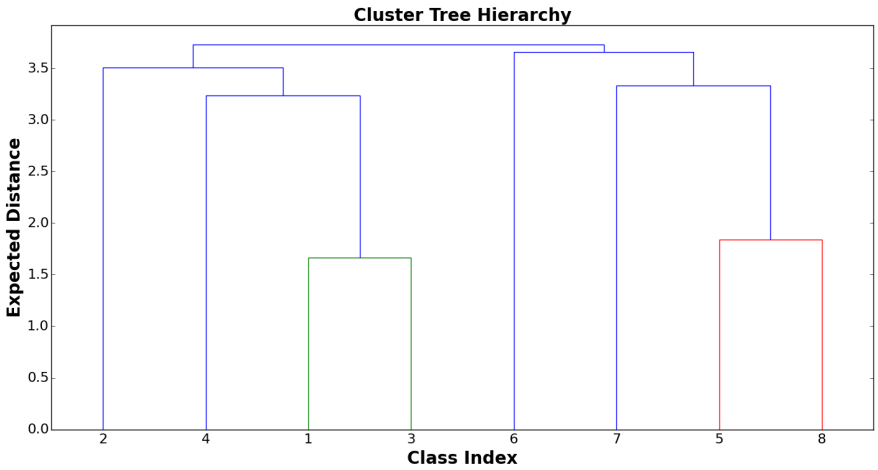

Figure 4.10: Cluster tree graph showing the relationship between the different microstructure classes.

In figure4.10 the height of the inverted U gives the measure of similarity between the classes. The cluster tree is interpreted by moving up on the y-axis, which shows that the expected Mahalanobis distance, height of the inverted U connecting two groups is the expected distance between those groups. Groups linked lower down are more similar than groups linked higher up. Observations from cluster tree:

1. Grouping follows the degree of anisotropy in the pore shape shown in figure4.1.
2. Classes 1-4 all have a low degree of anisotropy with the covariance ratio of 2 or less between the highest and lowest direction.
3. Classes 5-8 are more highly anisotropic.

4.4 Support Vector Machines for Classification of Microstructure

Figures4.9 and Figure4.9 showed that it is possible to obtained linear hyperplanes to separate the microstructure classes. If we consider the 3-D projection of the microstructure class on the canonical variables and say we wish to assign a new point to either class 5 or class 1 then multiple linear decision boundaries can be constructed between those two classes but to find the ideal linear decision boundary SVM classification algorihm will be used.

4.4.1 Support Vector Machines

Support Vector Machines (SVM) is a supervised machine learning algorithm which tries to search for an optimal hyperplane for binary linear classification. It attempts to find the separating hyperplane through the data such that the distance from it to the nearest data points on each side is maximised. Such a plane is called maximum-margin hyperplane. But in many cases such hyperplanes donot exist for e.g. when data is non-linearly separable or when there is noise present in data and the outliers from classes overlap. In such cases of outliers where data is nearly linearly separable, a penalty is introduced on points on the wrong side of the margin boundary that increases with the distance from it. This is known as soft margin SVM (Cortes and Vapnik, 1995). The soft margin method allows for an optimal classification without overfitting i.e. the margin is maximized even when an ideal linear separating hyperplane doesnot exists.

In cases where the data is not linearly separable non-linear kernel techniques can be used. Such, non-linear techniques tend to expand the variable space into a high-dimensional space and tries to find a linear separation is this vast high dimensional non-linear space. But due to the enhancement in the dimensions due to kernels these methods tend to be computationally much more expensive than using a linear decision boundary. A compromise between computational efficiency and accuracy should be struck to decide on the appropriate method to use. For multi-class problems (like our example) either a 1-against-1 classification or 1-against-rest classification can be carried out.

4.4.2 SVM on our dataset

As seen in figure4.9 and figure4.8 our dataset tends to be linearly separable in higher dimensions in canonical representation thus SVM with a linear kernel was used. Performing classification on canonical variables would have been more efficient as data tends to show a 100% separability in lower dimensions itself but as this setup is carried out to identify the least number of principal components required to capture range of microstructures in data PCA representation was used. SVM package from sklearn in python was used for the classification task (Pedregosa *et al.*, 2011). A stratified 80-20 split (i.e. 80% training data and 20% validation data) was used for carrying out the analysis. The 1-against-1 approach was found to perform better than 1-against-rest. Classification models were constructuted for increasing number of principal components. The accuracy of the classification model for both training and validation sets is shown in figure4.11.

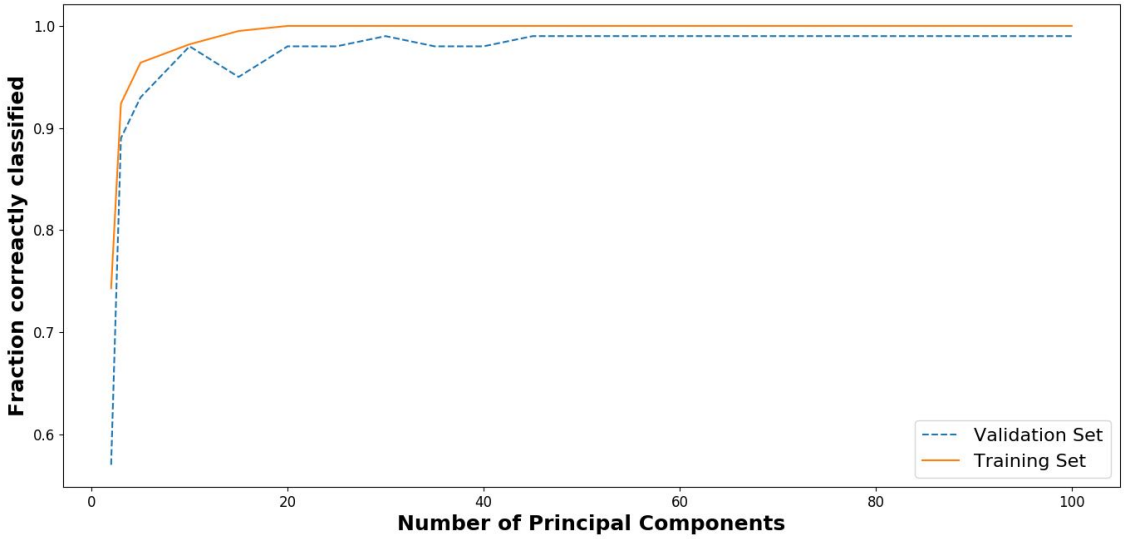

Figure 4.11: Accuracy of SVM classification model for both the raining and validation datasets as a function of the number of principal components used to generate the model.

As can be seen in 4.11 the classification accuracy rapidly increases with additional principal components. By approximately 15 components the training accuracy nears 100% and validation set accuracy reaches its highest of around 98%.

Following could be the reasons that the classification accuracy of the validation dataset does not approach 100% :

1. Outliers are present in the validation dataset
2. The additional material realizations contain some important information that is absent in the original dataset and is not captured in the original PCA decomposition.

4.5 Structure - Property Correlations

In this section we will be linking the material structure space with the property space. The Microstructure design framework covered in chapter2 (Fullwood *et al.*, 2010; Adams *et al.*, 2001), was predicated on delineating a microtructure design space termed the microstructure hull that bounded complete set of theoretically feasible microstructures of a given class and then projecting microstructures from the design space on the property space (Niezgoda *et al.*, 2008*b*). The major difficulty in such an approach is the construction of a microstructure hull that bounds all the theoretically feasible microstructure. As described in the process till now rather than constructing a such a microstructure hull a microstructure space is outlined using the 2-point correlations on microstructures obtained from digital simulations or characterizations. Instead of delineating property closures, or strict bounds on theoretical property values, we are exploring and mapping the property space that is currently achievable.

As mentioned in section4.1, the mechanical response of each realization for the 8 classes was obtained from the research work published by Niezgoda (Niezgoda *et al.*, 2013). Ashby charts or maps are well known design tools for visualising relationships between properties, performance and design constraints such as cost. Such maps tend to be invaluable and guiding material selection based on macroscale properties or performance criteria. An underlying goal of this thesis is to link microstructure and microstructre variance to property space for wider utilisation of microstructure sensitive design. In order to do so the properties and performance must be described as a function of microstructure function described in section3.2.1, or the PCA representation.

4.5.1 Prediction Algorithms

For setting up the structure-property correlation three prediction techniques were used. Each dataset was split by 80-20 rule (i.e. 80% training and 20% for testing). Prediction models were run on each class separately and an accuracy plot was constructed for each class.

4.5.2 Linear Regression

A basic linear homogenization model was fit between the effective modulus and yield strength and the first 50 PCA weights of each each realisation. The accuracy plots for effective modulus and yield strength is shown in figure4.12 and figure4.13.

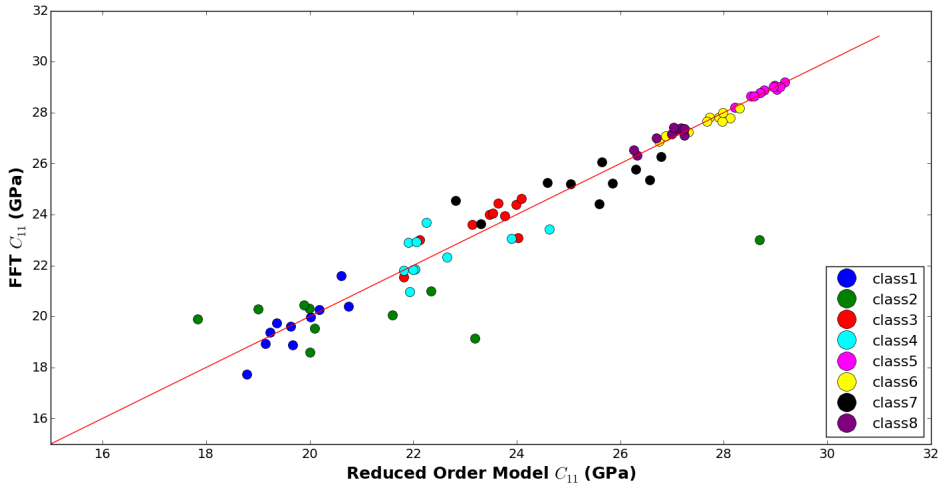

Figure 4.12: Linear homogenization weighted least square model validation for effective modulus with the $y = x$ line showing the ideal fit.

It can be seen that values for classes 5,6 and 8 are almost correctly predicted whereas there a substantial deviation for the rest classes with class 2 being the worst hit. A linear model is the most basic model which could be used. Whereas as the PCA representation captures the variability in the microstructure a way to assign weights to these directions can help us improve the prediction. With this view in mind weighted least squares technique was tested by Niezgoda.

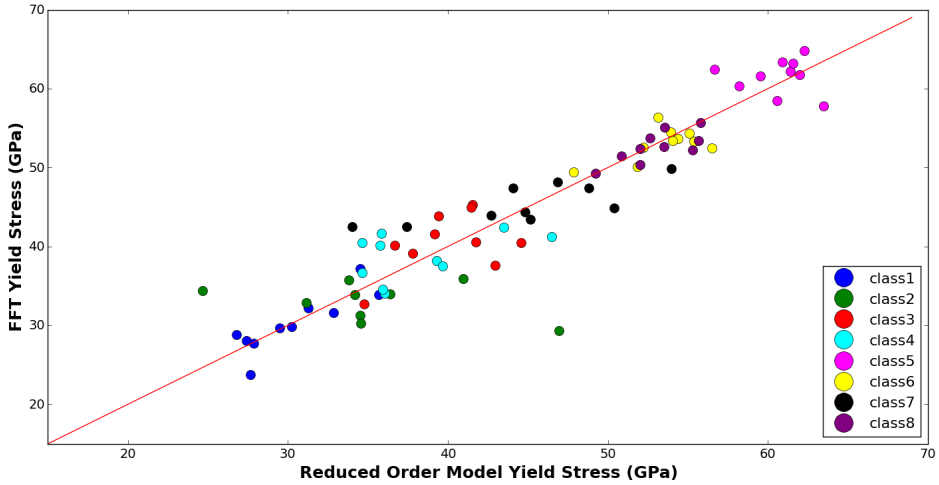

Figure 4.13: Linear homogenization weighted least square model validation for yield stress with the $y = x$ line showing the ideal fit.

Weighted Least Square Model

A linear weighted least squares model was fit between the effective modulus and yield strength and first 50 PCA weights of each realisation. The eigenvalues from inter-class PCA representation were chosen as the least square weights. The accuracy comparison for effective modulus and yield strength is shown in figure4.14 and figure4.15.

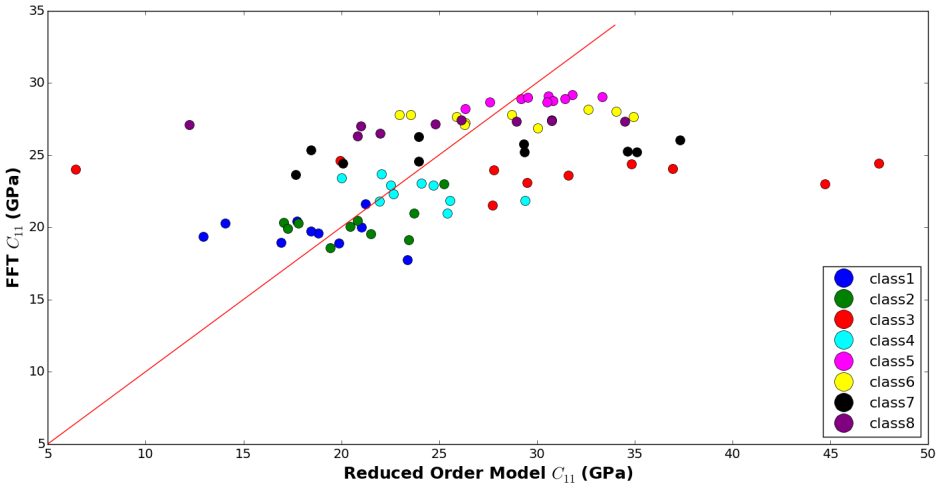

Figure 4.14: Linear homogenization weighted least square model validation for effective modulus with the $y = x$ line showing the ideal fit.

The results obtained by the above WLS technique perform poorly as compared to linear model. Ideally, the eigenvalues should be used as weights of the features (i.e.

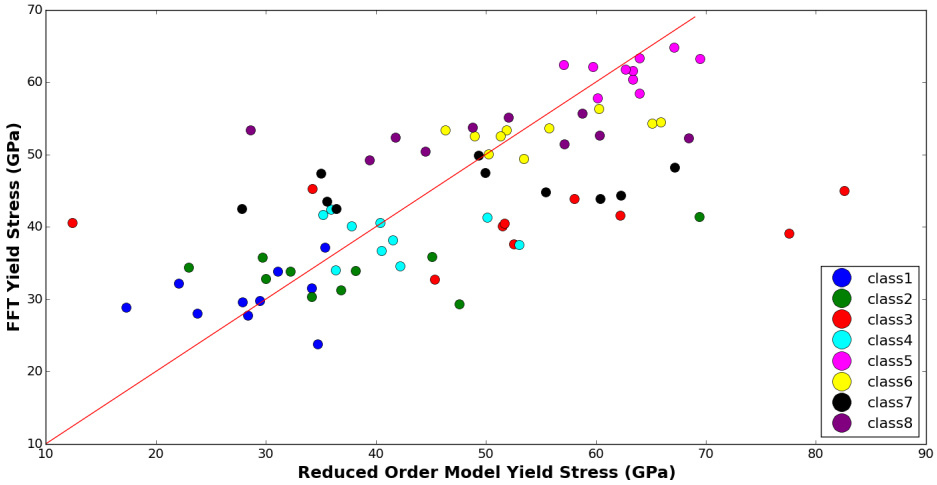

Figure 4.15: Linear homogenization weighted least square model validation for yield stress with the $y = x$ line showing the ideal fit.

the 50 principal directions used for prediction) rather than allocating these weights to the samples. Allocation of weights to the samples makes the algorithm depend greatly on the selection of the training and testing dataset. But the base WLS model used by Niezgoda (Niezgoda *et al.*, 2013) doesnot allow for the same. The model used by Niezgoda allows weight allocation to sample as per the basic WLS fit. Model similar to lasso or ridge regression wherein penalty can be assigned to model co-efficients can be used or algorithms sensitive to data scaling can be utilized.

SVM Regression Model

SVM has an effect of scaling as multiplication or division of all features by weights can either spread the points apart in microsrtructure space or bring them closer. As the decision boundary is decided based on the distance between the points SVM tends to be sensitive to data scaling. Thus, SVM with a linear kernel was used for prediction. Figure4.16 and figure4.17 shows the accuracy plot for effective modulus and yield stress using support vector machines regression.

It can be seen that the prediction for both c_{11} and ys by SVM is much better than linear model or weighted least square model used previously. But multiplying the principal components by their weights to enhance their importance or rather to give more weightage to the component with higher eigen-value doesn't seen to have a large im-

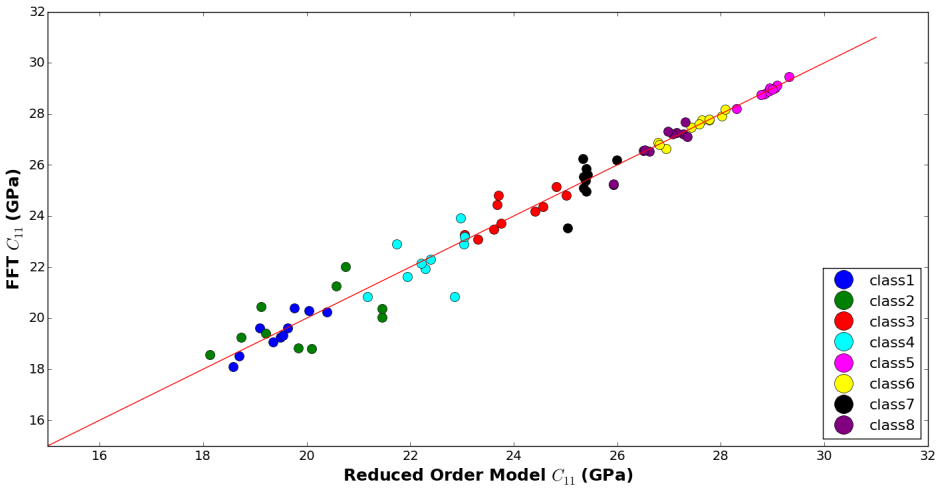

Figure 4.16: SVM regression model validation for effective modulus with the $y = x$ line showing the ideal fit.

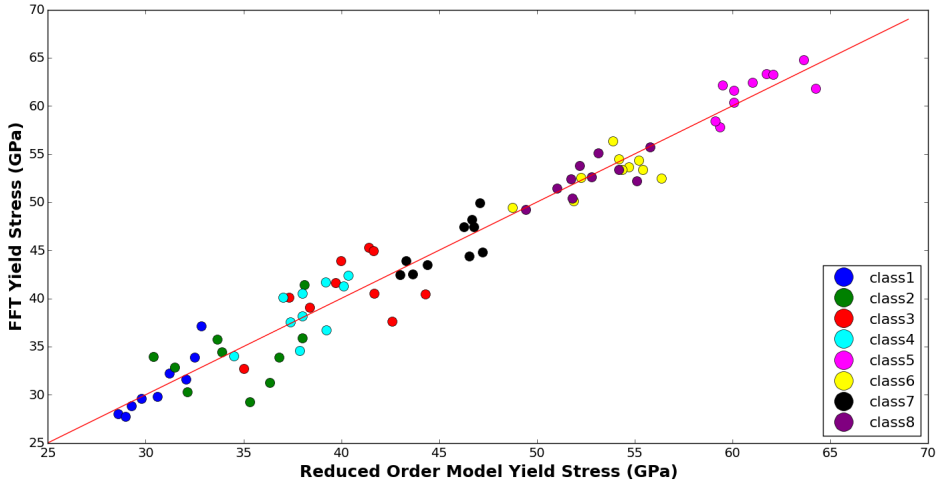

Figure 4.17: SVM regression model validation for yield stress with the $y = x$ line showing the ideal fit.

pact. The same can be verified from figure4.18 and figure4.19.

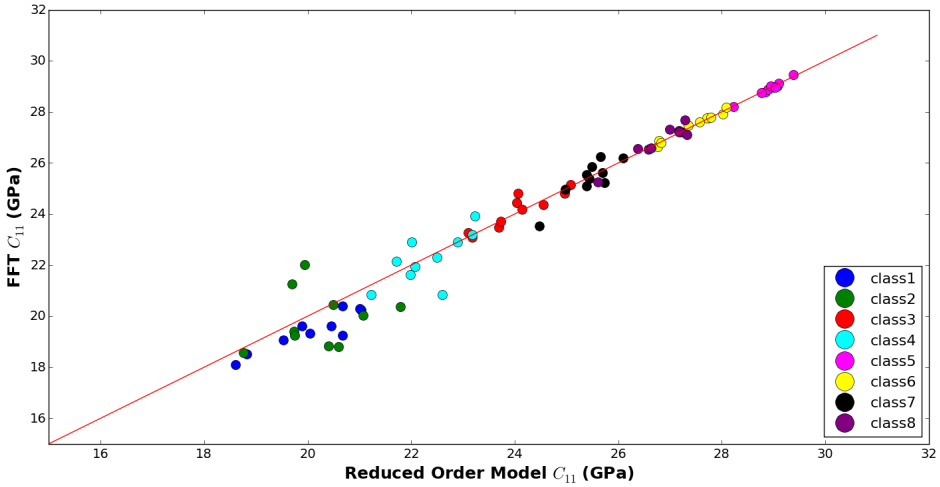

Figure 4.18: SVM regression model validation for effective modulus prediction through scaled PCA data with the $y = x$ line showing the ideal fit.

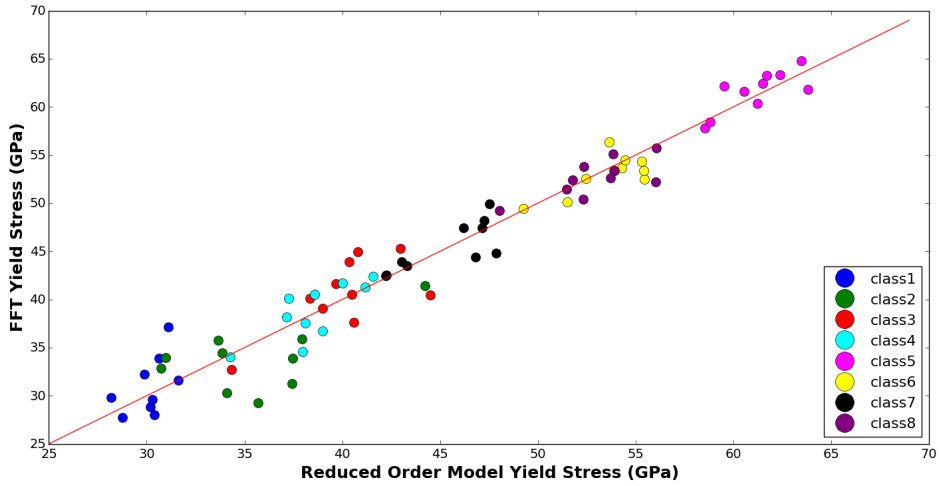

Figure 4.19: SVM regression model validation for yield stress prediction through scaled PCA data with the $y = x$ line showing the ideal fit.

Observations

1. SVM regression on scaled or unscaled data seems to give best accuracy as opposed to other models tested.

2. Classes like 5,6 and 8 which had high degree of anisotropy covariance (i.e. above 2) tend to have exact predictions for both mechanical properties as opposed to class 1,2,3,4 and 7.

CHAPTER 5

SUMMARY AND DIRECTION FOR FUTURE DEVELOPMENT

5.1 Summary

The work presented in this thesis is mainly motivated by the work of Niezgoda et. al. (Niezgoda *et al.*, 2013). The main concepts enumerated throughout the thesis are listed below:

1. The internal structure of engineering and natural materials can be represented as stochastic processes. This interpretation represents the microstructure as a set of probabilistic rules that dictate the placement of microstructural features in the material. The concept of microstructural function introduced by Adams et. al. (Adams *et al.*, 2001) was expanded to encompass the stochastic nature of the microstructure. The notation $m(x, h)$ for microstructure function was defined to represent a stochastic mapping of material constituents to a material sample.

2. The statistical properties of microstructure are not known apriori and must be quantified by observing many realizations of the microstructure. But a complete description of a microstructure using it s statistical properties would require a n^{th} order joint probability distribution (n-point correlation function) to be specified which is an impossible task given the limited information available about the internal structure of most materials. Thus, lower order spatial statistics, particularly 2-point correlation function, which can be more readily estimated using Fast Fourier Transform was explored.

3. The 2-point correlation function lies on a complex high-dimensional hyperplane which has many redundancies embedded in itself as opposed to the microstructure function representation. Thus, PCA was used on the 2-point correlation function of microstructure to reduce the dimensions to facilitate faster computation and visualization of the microstructure space embedded by the realizations under observation. Each set of correlations can be written as a linear combination of the principal components, and the associated weights provide an objective low dimensional representation of the data.

4. Each realization manifests as a single point in the microstructure visualization. An intra and inter-class analysis of the space using different distance measures showed that classes with similar structural properties or anisotropy relationships in the case of our example tend to form clusters nearer to eachother. Disparity in structural properties tends to clusters separated by large distance.

5. The canonical representation of the microstructure space using Mahalanobis distance measure for a well- separated inter-class segregation demonstrated an almost linear separability between all the classes under observation. Thus, a linear decision boundary classification using the SVM algorithm was utilised and the minimum number of principal components required for 98% classification accuracy was found to be 15.

6. The scatter in the visualization space was tied to the properties space by setting up linear homogenized models between effective modulus, yield strength and first 50 principal components of the reduced order 2-point correlation function. It was seen that off the algorithms used regression through SVM out performed linear and weighted least square regression. Classes with higher anisotropy measure could be predicted with greater accuracy as opposed to those having lower anisotropic characteristics.

5.2 Limitations of Framework and future direction

Though this work has been implemented on a 'toy example' it has the potential to transform the present methods of analysis and visualization of microstructure data. Presently, the framework doesnot help handle the variety of data types across multiple length scales. But a formalisation of such an approach could help scientist and engineers to model intrinsic randomness of microstructure which has a direct relation to property variance. But the a stochastic approach seems like a best approach to integrate microstructure databases collected at varied length scales.

Also, as mentioned at various locations in the thesis the 2-point correlation function utilised by us not only lies on a complex high-dimensional space but also has a non-linear structure. Thus, utilisation of PCA tends to capture lesser variance than what could be captured if non-linear transforms like kernel-PCA could be used.

Lastly, as discussed in the previous chapter a linear homogenized model which has a working like weighted least square but utilizes eigen vectors as weights for the principal components as opposed to the samples can be explored for a better accuracy of establishing the structure-property linkages.

APPENDIX A

DIGITAL REPRESENTATION OF A FUNCTION USING DISCRETE FOURIER TRANSFORM

Determining an unknown function $f(x)$, where x is a continuous variable will need that we know the value of $f(x)$ at infinite points or we have a prior information of the analytical distribution of the function. Occurrence of any of the aforementioned situation is improbable. Furthermore, since obtaining each value of the function incurs a certain cost, we need to develop an optimized framework for effectively recovering such function values.

Let $\{f_s = f(x_s); s = 1, 2, ..., S\}$ be discrete set of function values calculated at various points in the domain enumerated by variable s. Our major aim is to reconstruct function from this set f given.

For carrying out this analysis we need to express the original function using a suitable basis. We will be using Fourier Transforms for function reconstruction, where the basis functions are selected to provide global support. Fourier transforms are useful when function to be reconstructed is periodic. (As we will be utilizing this definition for micro-structure function which tend to demonstrate periodicity to a great extent Fourier Transforms form a good selection).
Assumptions about the function to be reconstructed:

1. Function to be known or reconstructed is periodic.
2. Function can be represented by series sum of harmonic functions (containing both sine and cosine functions) covering a broad range of frequencies.

An orthonormal basis is used. It produces a one-to-one mapping of the function from the original space to the frequency space. For example consider the following Fourier Transform of a function $f(x)$:

$$f(x) = \sum_{k=-\infty}^{\infty} F_k e^{ikx} \tag{A.1}$$

where,

Function $f(x)$ is assumed to be periodic with period 2π. Replacing x with $\frac{2\pi x}{p}$ will convert it to a function with periodicity p i.e. $f(x) = f(x+p)$ and $i = \sqrt{-1}$.

k = number of cycles over a length of 2π in x.

Here, F_k provides an equivalent representation of the original function in the frequency space. Their is one-to-one mapping between $f(x)$ and F_k. Given any one of the two representations, the other can be accurately reconstructed.

Fourier Co-efficients:

$$F_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx \tag{A.2}$$

The main disadvantage of using Fourier series is that the series expansion includes infinite terms. Thus, a special variant of the concept of Fourier Series called Discrete Fourier transforms (DFTs) will be used.

Discrete Fourier Transforms

Let $\{f_s | s = 0, 1, 2, ..., S-1\}$ represent the sampling of a function on a uniformly tessellated domain of interest (say, $x \epsilon (a, b)$). The function is implicitly assumed to be periodic.

The values of function in real and Fourier space is given by:

$$F_k = \Im(f_s) = \sum_{s=0}^{S-1} f_s e^{\frac{i2\pi ks}{S}} \tag{A.3}$$

$$f_s = \Im^{-1}(F_k) = \frac{1}{S} \sum_{s=0}^{S-1} F_k e^{\frac{-i2\pi ks}{S}} \tag{A.4}$$

where,

S = Periodicity of function

F_k = Amplitudes of the discrete frequencies defined on S

k = No. of cycles over the domain size associated with S

Nyquist Sampling Rate

The signal must be sampled atleast twice the highest frequency present in the signal. Thus, for spatial discretization of micro-structure the spatial bin should be smaller than half the size of the smallest feature to be captured in the micro-structure.

In practice, calculation of DFTs is computationally expensive. Thus, Fast Fourier transforms which capture the inherent relationship between the frequency amplitudes is utilized to calculate the DFTs.

Properties of DFT

$$\sum_{s=0}^{S-1} |f|^2 = \frac{1}{S} \sum_{s=0}^{S-1} |F_k|^2 \tag{A.5}$$

$$F_{s-k} = F_k^* \text{if all } f_s \text{ are real valued} \tag{A.6}$$

$$f_s \text{ and} g_s \text{are two discrete valued functions} \tag{A.7}$$

$$\sum_{s=0}^{S1} f_s g_s = \frac{1}{S} \sum_{k=0}^{S1} F_k G_k \tag{A.8}$$

$$\text{Convolutions : } conv(f,g)h_r = \sum_{s=0}^{S1} f_s g_{rs} = \Im^1(F_k G_k) \tag{A.9}$$

$$\text{Correlations : } corr(f,g)_r = \sum_{s=0}^{S1} f_{r+s} g_s = \Im^1(F_k G_k) \tag{A.10}$$

where, all symbols have the meaning mentioned above.

APPENDIX B

MAHALANOBIS DISTANCE MEASURE

Calculation of distance in multiple dimensions turns out to be very tricky. Euclidean distance measure is prone to scaling thus if variables with varied units are used then Euclidean distance measure tends to change based on the scaling parameters used. Euclidean distance is also prone to deviation based on the correlation between variables. Thus, Mahalanobis distance measure works well in such cases. It makes variables uncorrelated and scales the data before using Euclidean measure to calculate the distance.

Mahalanobis distance gauges the similarity between an unknown sample (or sample set) to a known or classified sample set (Stevens, 2012). Consider a multivariate vector $x = (x_1, x_2, x_3, ...)^T$, the mahalanobis distance between x and a distribution of vectors $f(y)$ with mean μ_y and covariance matrix C_y is defined as

$$D_M(x, f(y)) = \sqrt{(x - \mu_y)^T C_y^{-1} (x - \mu_y)} \tag{B.1}$$

The feature of the Mahalanobis distance is that it takes the shape of the known probability distribution into consideration, in that all the points which lie on the same iso-probability surface of the known distribution will have same Mahalanobis distance to that distribution regardless of the shape of the distribution (This is only strictly true for Gaussian or near-Gaussian 'well behaved' uni-modal distributions). Thus, it can be generalized to a similarity measure between two multivariate vectors x and y , where y comes from a distribution $f(y)$ with covariance C_y as

$$d(x, y) = \sqrt{(x - y)^T C_y^{-1} (x - y)} \tag{B.2}$$

When computing distance in PCA space the individual components of x and y, x_i and y_i are independent (orthogonal) variables and the covariance matrix is by definition diagonal. In this case the Mahalanobis distance becomes normalized Euclidean

distance,

$$d(x, y) = \sqrt{\sum_i \frac{(x_i - y_i)^2}{\sigma_i^2}} \tag{B.3}$$

where,

σ_i is the standard deviation of $f(y_i)$. In the limit that covariance matrix is an identity matrix, Euclidean distance measure is equivalent to Mahalanobis distance measure.

REFERENCES

1. *TPSX Material Properties Database*. Web Edition Version 4. NASA, 2006.

2. **Adams, B. L.**, **X. C. Gao**, and **S. R. Kalidindi** (2005). Finite approximations to the second-order properties closure in single phase polycrystals. *Acta Materialia*, **53**(13), 3563–3577.

3. **Adams, B. L.**, **A. Henrie**, **B. Henrie**, **M. Lyon**, **S. Kalidindi**, and **H. Garmestani** (2001). Microstructure-sensitive design of a compliant beam. *Journal of the Mechanics and Physics of Solids*, **49**(8), 1639–1663.

4. **Ashby, M. F.** and **D. Cebon** (1993). Materials selection in mechanical design. *Le Journal de Physique IV*, **3**(C7), C7–1.

5. **Cecen, A.**, **T. Fast**, and **S. R. Kalidindi** (2016). Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure. *Integrating Materials and Manufacturing Innovation*, **5**(1), 1.

6. **Chen, X.** and **Y. Liu** (2004). Square representative volume elements for evaluating the effective material properties of carbon nanotube-based composites. *Computational Materials Science*, **29**(1), 1–11.

7. **Cortes, C.** and **V. Vapnik** (1995). Support-vector networks. *Machine learning*, **20**(3), 273–297.

8. **Drugan, W.** and **J. Willis** (1996). A micromechanics-based nonlocal constitutive equation and estimates of representative volume element size for elastic composites. *Journal of the Mechanics and Physics of Solids*, **44**(4), 497–524.

9. **Frisch, H.** and **F. Stillinger** (1963). Contribution to the statistical geometric basis of radiation scattering. *The Journal of Chemical Physics*, **38**(9), 2200–2207.

10. **Fullwood, D. T.**, **S. R. Niezgoda**, **B. L. Adams**, and **S. R. Kalidindi** (2010). Microstructure sensitive design for performance optimization. *Progress in Materials Science*, **55**(6), 477–562.

11. **Ganapathysubramanian, B.** and **N. Zabaras** (2008). A non-linear dimension reduction methodology for generating data-driven stochastic input models. *Journal of Computational Physics*, **227**(13), 6612–6637.

12. **Gibbs, J. W.** (1878). On the equilibrium of heterogeneous substances. *American Journal of Science*, (96), 441–458.

13. **Gokhale, A.**, **A. Tewari**, and **H. Garmestani** (2005). Constraints on microstructural two-point correlation functions. *Scripta Materialia*, **53**(8), 989–993.

14. **Gusev, A. A.** (1997). Representative volume element size for elastic composites: a numerical study. *Journal of the Mechanics and Physics of Solids*, **45**(9), 1449–1459.

15. **Halko, N.**, **P.-G. Martinsson**, and **J. A. Tropp** (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, **53**(2), 217–288.

16. **Hill, R.** (1963). Elastic properties of reinforced solids: some theoretical principles. *Journal of the Mechanics and Physics of Solids*, **11**(5), 357–372.

17. **Kalidindi, S. R.**, *Hierarchical Materials Informatics: Novel Analytics for Materials Data*. Elsevier, 2015.

18. **Kalidindi, S. R.**, **M. Binci**, **D. Fullwood**, and **B. L. Adams** (2006). Elastic properties closures using second-order homogenization theories: case studies in composites of two isotropic constituents. *Acta Materialia*, **54**(11), 3117–3126.

19. **Kalidindi, S. R.**, **J. R. Houskamp**, **M. Lyons**, and **B. L. Adams** (2004). Microstructure sensitive design of an orthotropic plate subjected to tensile load. *International Journal of Plasticity*, **20**(8), 1561–1575.

20. **Kanit, T.**, **S. Forest**, **I. Galliet**, **V. Mounoury**, and **D. Jeulin** (2003). Determination of the size of the representative volume element for random composites: statistical and numerical approach. *International Journal of solids and structures*, **40**(13), 3647–3679.

21. **Lee, J. A.** and **M. Verleysen**, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

22. **Lewis, A.**, **C. Suh**, **M. Stukowski**, **A. Geltmacher**, **K. Rajan**, and **G. Spanos** (2008). Tracking correlations between mechanical response and microstructure in three-dimensional reconstructions of a commercial stainless steel. *Scripta Materialia*, **58**(7), 575–578.

23. **Luz, G. M.** and **J. F. Mano** (2009). Biomimetic design of materials and biomaterials inspired by the structure of nacre. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **367**(1893), 1587–1605.

24. **Mandal, S.**, **P. Sivaprasad**, **S. Venugopal**, **K. Murthy**, and **B. Raj** (2008). Artificial neural network modeling of composition–process–property correlations in austenitic stainless steels. *Materials Science and Engineering: A*, **485**(1), 571–580.

25. **Niezgoda, S.**, **D. Fullwood**, and **S. Kalidindi** (2008*a*). Delineation of the space of 2-point correlations in a composite material system. *Acta Materialia*, **56**(18), 5285–5292.

26. **Niezgoda, S.**, **D. Fullwood**, and **S. Kalidindi** (2008*b*). Delineation of the space of 2-point correlations in a composite material system. *Acta Materialia*, **56**(18), 5285–5292.

27. **Niezgoda, S. R.**, *Stochastic representation of microstructure via higher-order statistics: theory and application*. 2010.

28. **Niezgoda, S. R.**, **A. K. Kanjarla**, and **S. R. Kalidindi** (2013). Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data. *Integrating Materials and Manufacturing Innovation*, **2**(1), 1.

29. **Niezgoda, S. R.**, **D. M. Turner**, **D. T. Fullwood**, and **S. R. Kalidindi** (2010). Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics. *Acta Materialia*, **58**(13), 4432–4445.

30. **Niezgoda, S. R.**, **Y. C. Yabansu**, and **S. R. Kalidindi** (2011). Understanding and visualizing microstructure and microstructure variance as a stochastic process. *Acta Materialia*, **59**(16), 6387–6400.

31. **Olson, G. B.** (1997). Computational design of hierarchically structured materials. *Science*, **277**(5330), 1237–1242.

32. **Pedregosa, F.**, **G. Varoquaux**, **A. Gramfort**, **V. Michel**, **B. Thirion**, **O. Grisel**, **M. Blondel**, **P. Prettenhofer**, **R. Weiss**, **V. Dubourg**, *et al.* (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, **12**(Oct), 2825–2830.

33. **Selector, C.**, *Granta Material Intelligence*. 2008.

34. **Stevens, J. P.**, *Applied multivariate statistics for the social sciences*. Routledge, 2012.

35. **Sundararaghavan, V.** and **N. Zabaras** (2004). A dynamic material library for the representation of single-phase polyhedral microstructures. *Acta Materialia*, **52**(14), 4111–4119.

36. **Sundararaghavan, V.** and **N. Zabaras** (2005). Classification and reconstruction of three-dimensional microstructures using support vector machines. *Computational Materials Science*, **32**(2), 223–239.

37. **Sundararaghavan, V.** and **N. Zabaras** (2009). A statistical learning approach for the design of polycrystalline materials. *Statistical Analysis and Data Mining*, **1**(5), 306–321.

38. **Tyrkel, E.** (1968). âĂŹ, history of the development of the iron–carbon phase diagram. *Mashinostroenie, Moscow*.

39. **Warde, J.** and **D. Knowles** (1999). Use of neural networks for alloy design. *ISIJ international*, **39**(10), 1015–1019.

40. **Wheeler, D.**, **D. Brough**, **T. Fast**, **S. Kalidindi**, and **A. Reid** (2016). Pymks: Materials knowledge system in python (figshare, 2014).