Available online at www.sciencedirect.com

ScienceDirect

Journal of Computational Physics 227 (2008) 6612–6637

JOURNAL OF COMPUTATIONAL PHYSICS

www.elsevier.com/locate/jcp

A non-linear dimension reduction methodology for generating data-driven stochastic input models

Baskar Ganapathysubramanian, Nicholas Zabaras *

Materials Process Design and Control Laboratory, Sibley School of Mechanical and Aerospace Engineering, 101 Frank H.T. Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, USA

Received 17 June 2007; received in revised form 19 December 2007; accepted 17 March 2008
Available online 28 March 2008

Abstract

Stochastic analysis of random heterogeneous media (polycrystalline materials, porous media, functionally graded materials) provides information of significance only if realistic input models of the topology and property variations are used. This paper proposes a framework to construct such input stochastic models for the topology and thermal diffusivity variations in heterogeneous media using a data-driven strategy. Given a set of microstructure realizations (input samples) generated from given statistical information about the medium topology, the framework constructs a reduced-order stochastic representation of the thermal diffusivity. This problem of constructing a low-dimensional stochastic representation of property variations is analogous to the problem of manifold learning and parametric fitting of hyper-surfaces encountered in image processing and psychology.

Denote by $\mathcal{M}$ the set of microstructures that satisfy the given experimental statistics. A non-linear dimension reduction strategy is utilized to map $\mathcal{M}$ to a low-dimensional region, $\mathcal{A}$. We first show that $\mathcal{M}$ is a compact manifold embedded in a high-dimensional input space $\mathbb{R}^n$. An isometric mapping $\mathcal{F}$ from $\mathcal{M}$ to a low-dimensional, compact, connected set $\mathcal{A} \subset \mathbb{R}^d (d \ll n)$ is constructed. Given only a finite set of samples of the data, the methodology uses arguments from graph theory and differential geometry to construct the isometric transformation $\mathcal{F} : \mathcal{M} \rightarrow \mathcal{A}$. Asymptotic convergence of the representation of $\mathcal{M}$ by $\mathcal{A}$ is shown. This mapping $\mathcal{F}$ serves as an accurate, low-dimensional, data-driven representation of the property variations.

The reduced-order model of the material topology and thermal diffusivity variations is subsequently used as an input in the solution of stochastic partial differential equations that describe the evolution of dependant variables. A sparse grid collocation strategy (Smolyak algorithm) is utilized to solve these stochastic equations efficiently. We showcase the methodology by constructing low-dimensional input stochastic models to represent thermal diffusivity in two-phase microstructures. This model is used in analyzing the effect of topological variations of two-phase microstructures on the evolution of temperature in heat conduction processes.

Keywords: Stochastic partial differential equations; Data-driven models; Non-linear embedding; Manifold learning; Model reduction; Collocation methods; Sparse grids; Microstructure

* Corresponding author. Tel.: +1 607 255 9104; fax: +1 607 255 1222.
E-mail address: zabaras@cornell.edu (N. Zabaras).
URL: http://mpdc.mae.cornell.edu/ (N. Zabaras).

0021-9991/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2008.03.023

1. Introduction and outline

With the rapid advances in computational power and easier access to high-performance computing platforms, it has now become possible to computationally investigate realistic multiscale, multidomain, multiphysics materials problems to an unparalleled extent. As a direct consequence of this computational ability, there has been growing awareness that the tightly coupled and highly non-linear systems that such problems are composed of are affected to a significant extent by the inherent uncertainties in material properties and system characteristics. To accurately predict the performance of such systems, it then becomes essential for one to include the effects of these input uncertainties into the model system and understand how they propagate and alter the final solution.

In most complex systems involving multiple coupled physical phenomena, the material (thermo-physical) properties as well as the material distribution (topology) vary at a length scale much smaller than the system size. Familiar systems include analysis of thermal transport through devices (nozzle flaps, gears, etc.) that are composed of polycrystals and/or functionally graded materials, hydrodynamic transport through porous media and chemical flow through packed filtration beds. In such problems, the only information that is usually available experimentally to quantify these variations are statistical correlations. This leads to an analysis of the problem assuming that the property and topological variations are random fields satisfying the experimental correlations. To perform any such analysis, one must first construct models of these variations to be used as inputs in the subsequent uncertainty analysis. The analysis of the effect of such uncertainties on the system can basically be broken down into two major steps: (i) construction of a stochastic model (preferably a low-dimensional, continuous mapping) that encodes and quantifies the variation of material topology and properties in a mathematically rigorous way, and (ii) using this model as an input to the corresponding stochastic partial differential equations (SPDEs) that describe the relevant physical phenomena and solving for the evolution of the desired dependant variables.

There have been very few previous investigations into developing stochastic input representations. The recent work in [1] looks at developing probabilistic models of random coefficients in SPDEs using a maximum likelihood framework. The random domain decomposition (RDD) method was used in [2–4] to construct probabilistic models for heterogeneous permeability distributions. This methodology has been shown to work very well in describing permeability variations in geological strata [2]. Nevertheless, almost all of these techniques for constructing input models are based on the concept of transforming experimental data and statistics into probability distributions of the property. These techniques are usually highly application specific, require expert knowledge in assigning probabilities and invariably require some amount of heuristic parameter fitting.

The work introduced in this paper utilizes the available statistical information about the variability of random media to construct a set of plausible realizations of the material topology and property. The framework subsequently encodes *these property realizations* into a low-dimensional continuous space that represents all the possible property variations permitted by the experimental data. By sampling over this low-dimensional equivalent surrogate space, one is essentially sampling over the random space of property variations that satisfy the experimental data. This low-dimensional representation is subsequently used as a stochastic input model in the uncertainty analysis. The major advantages of this framework are the enormous reduction in complexity due to the analysis in a low-dimensional space and more importantly the absence of the requirement of any expert knowledge. In addition, the generality of the mathematical developments results in a framework that can construct input models for any property variability, seamlessly meshing with any reconstruction algorithm and software that produce plausible data sets.

In our earlier work in [5], we developed a linear embedding methodology to model the topological variations of composite microstructures satisfying some experimentally determined statistical correlations. We were able to construct a model reduction scheme (based on principle component analysis (PCA)) to convert the large-dimensional space describing the class of microstructures to a low-dimensional approximation of the space. The low-dimensional model represents the class of allowable microstructures that satisfy the experimental correlations. This model was utilized as the stochastic input in a stochastic variational multiscale (SVMS) framework. Though this methodology proved to be extremely effective in analyzing the effect of thermal diffusivity variations in two-phase microstructures, it has the following drawbacks: (1) it cannot be naturally extended to multi-component materials, (2) the construction of the model requires the solution of computationally exacting quadratic and higher-order non-linear equality constraints, and (3) the PCA based model reduction scheme constructs the *closest linear subspace* of the high-dimensional input space. The third drawback listed above is the major motivation towards developing better strategies of constructing reduced-order models. These model reduction strategies only provide an approximate linear representation of the input space. But most of the data sets contain essential non-linear structures that are invisible to PCA. Hence, one has to go beyond a linear representation of the high-dimensional input space to accurately access the effect of its variability on the output variables.

We extend the concepts developed in [5] into a non-linear dimension reduction strategy to embed data variations into a low-dimensional manifold that can serve as the input model for subsequent analysis. The major contributions of the present work are as follows: A completely general methodology to generate viable, realistic, reduced-order stochastic input models based on experimental data is proposed. To the best knowledge of the authors, this is the first time that such a broad framework for generating models for *any* property variation has been developed. We provide a rigorous mathematical basis for the proposed framework. Furthermore, error estimates and strict convergence estimates of the reduced-order model are provided. This methodology is applied to construct a reduced-order model of thermal property variation of a two-phase microstructure. The model is subsequently utilized as a stochastic input model to study the effect of material uncertainty on thermal diffusion. A highly efficient, stochastic collocation based solution strategy is used to solve the corresponding SPDE for the evolution of temperature.

The basic model reduction ideas envisioned in this work are not just limited to generation of viable stochastic input models of property variations. This framework has direct applicability to problems where working in high-dimensional spaces is computationally intractable, for instance, in visualization of property evolution, extracting process-property maps in low-dimensional spaces, among others. Furthermore, the generation of a low-dimensional surrogate space has major ramifications in the optimizing of properties, processes and structures, making complicated operations like searching, contouring and sorting computationally much more feasible.

The outline of the paper is as follows: In the next section the problem of interest is defined. Following this, the central idea of the non-linear dimension reduction strategy is described in Section 3. Section 4 lays down the mathematical foundation of the proposed methodology. This is followed by Section 5 describing the numerical details of the implementation. An illustrative numerical example is presented in Section 6. We conclude in Section 7 with a brief discussion on future avenues of research.

2. Problem definition

The focused application of the developed framework is to analyze transport phenomena in heterogeneous random media. In this work, we are particularly interested in investigating thermal diffusion in two-phase heterogeneous media. In such problems, the topology of the heterogeneous structure (the microstructure) is only known in terms of a few statistical correlations. Denote this set of statistical correlations by $S = \{S_1, \ldots, S_p\}$. Any material structure that satisfies these statistical correlations is a valid realization of the microstructure. Consequently, the microstructural topology should be considered as a random field (satisfying some statistical correlations) and the microstructure in any arbitrary specimen as a realization of this field. The thermal diffusivity of the material obviously depends on the topology of the microstructure. We assume that the thermal diffusivity of the material is uniquely defined by its microstructure (e.g. each point in a realization of a two-phase medium is assumed to be uniquely occupied by one of the two phases (0 or 1) and that each phase has a given diffusivity). That is, for a microstructure specified as a distribution of two phases in a domain, the thermal diffusivity distribution is given by simply replacing the phase description (0 or 1) at each point on the domain by its corresponding thermal diffusivity (α_0 or α_1). From this simple scaling, for each realization of the microstructure, we can compute the corresponding realization of the diffusivity, α ($\alpha = k/\rho C_p$).

Let $\mathcal{M}_S$ be the space of all microstructures that satisfy the statistical properties S. This is our event space. Every point in this space is equiprobable. Consequently, we can define a σ-algebra $\mathcal{G}$ and a corresponding probability measure $\mathcal{P} : \mathcal{G} \rightarrow [0, 1]$ to construct a complete probability space $(\mathcal{M}_S, \mathcal{G}, \mathcal{P})$ of allowable microstructures. Corresponding to a microstructure realization $\omega \in \mathcal{M}_S$, we can associate a thermal diffusivity distribution $\alpha(\boldsymbol{x}, \omega)$. That is, the thermal diffusivity of the random heterogeneous medium is represented as

$$\alpha(\boldsymbol{x}) = \alpha(\boldsymbol{x}, \omega), \quad \boldsymbol{x} \in D, \quad \omega \in \mathcal{M}_S, \tag{1}$$

where $D \subset \mathbb{R}^{n_{sd}}$ is the n_{sd}-dimensional bounded domain that is associated with this medium. The governing equation for thermal diffusion in this medium can be written as:

$$\frac{\partial u(\boldsymbol{x}, t, \omega)}{\partial t} = \nabla \cdot [\alpha(\boldsymbol{x}, \omega) \nabla u(\boldsymbol{x}, t, \omega)] + f(\boldsymbol{x}, t), \boldsymbol{x} \in D, t \in [0, T_f], \omega \in \mathcal{M}_S, \tag{2}$$

where u is the temperature. Here, f is the thermal source/sink and is assumed to be deterministic without loss of generality.

The solution methodology is to first reduce the complexity of the problem by reducing the probability space into a finite-dimensional space [6]. In the present case, the random topology satisfies certain statistical correlation functions $S = \{S_1, \ldots, S_p\}$ (from now on referred to as the 'experimental statistics'). We utilize non-linear model reduction techniques (see Section 3) to decompose the random topology field into a finite set of uncorrelated random variables. Upon decomposition and characterization of the random inputs into d random variables, $\xi_i(\omega)$, $i = 1, \ldots, d$, the solution to the SPDE Eq. (2) can be written as

$$u(\boldsymbol{x}, t, \omega) = u(\boldsymbol{x}, t, \boldsymbol{\xi}), \quad \boldsymbol{\xi} = (\xi_1, \ldots, \xi_d), \tag{3}$$

where $\boldsymbol{\xi}$ is the d-tuple of the random variables. The domain of definition of $\boldsymbol{\xi}$ is denoted by Γ. Eq. (2) can now be written as a $(d + n_{sd})$-dimensional problem as follows:

$$\frac{\partial u(\boldsymbol{x}, t, \boldsymbol{\xi})}{\partial t} = \nabla \cdot [\alpha(\boldsymbol{x}, \boldsymbol{\xi}) \nabla u(\boldsymbol{x}, t, \boldsymbol{\xi})] + f(\boldsymbol{x}, t), \boldsymbol{x} \in D, t \in [0, T_f], \quad \boldsymbol{\xi} \in \Gamma. \tag{4}$$

For the sake of brevity, we will denote the system of equations consisting of Eq. (4) and appropriate initial and boundary conditions (here for simplicity assumed to be deterministic) as $\mathcal{B}(u : \boldsymbol{x}, t, \boldsymbol{\xi}) = 0$.

We utilize a sparse grid collocation framework (based on the Smolyak algorithm) that results in a set of decoupled deterministic equations [7]. In this collocation approach, a finite element approximation is used for the spatial domain and the multi-dimensional stochastic space is approximated using interpolating functions. One computes the deterministic solution at various points in the stochastic space and then builds an interpolated function that best approximates the required solution [7]. The collocation method collapses the $(d + n_{sd})$-dimensional problem to solving M (where, M is the number of collocation points, $\boldsymbol{\xi}_k, k = 1, \ldots, M$) deterministic problems in n_{sd}-space dimensions. The q-th order statistics (for $q \geqslant 1$) of the random solution can be obtained through simple quadrature operations on the interpolated function $u(\boldsymbol{x}, \boldsymbol{\xi}) = \sum_{k=1}^{M} u(\boldsymbol{x}, \boldsymbol{\xi}_k) L_k(\boldsymbol{\xi})$ (where L_k are the sparse grid interpolation functions) as:

$$\langle u^q(x) \rangle = \sum_{k=1}^{M} u^q(x, \boldsymbol{\xi}_k) \int_{\Gamma} L_k(\boldsymbol{\xi}) \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}, \tag{5}$$

where $\rho(\boldsymbol{\xi}) : \Gamma \to \mathbb{R}$ is the joint probability distribution function for the set of random variables $\{\xi_1, \ldots, \xi_d\}$. In the following sections, we describe the non-linear model reduction framework for computing $\alpha(\boldsymbol{x}, \boldsymbol{\xi})$. Details of the implementation of the Smolyak algorithm for Eq. (4) can be found in [5].

3. Non-linear model reduction: its necessity and some basic ideas

In our recent work [5], a linear model reduction strategy was developed to convert experimental statistics into a plausible low-dimensional representation of two-phase microstructures. The first step in that formulation was the conversion of the statistical information into a set of plausible 3D realizations. This feature of first converting the given experimental statistics into a data set of plausible 3D microstructures is continued in the developments featured here. This is motivated by the fact that there exists a variety of mature mathematical and numerical techniques that convert experimental data and statistics into multiple plausible 3D reconstructions of the topology and thus property variations. For instance, the GeoStatistical Modelling Library [8] (GSLIB) converts experimental statistics of the permeability (semi-variograms, correlations, etc.) into plausible 3D models of permeability variations. Similarly, there have been various techniques that have been developed [5,9–12], to convert experimental statistics into a plausible 3D reconstructions of two-phase composite microstructures as well as polycrystalline materials.

3.1. Linear model reduction: where does it fail?

As stated in Section 1, the PCA based linear model reduction technique that we had previously formulated [5] has some drawbacks. The most critical of these is that any PCA based approach can only identify the closest linear subspace to the actual space (which is possibly non-linear) in which the data reside. This directly translates into the fact that PCA tends to consistently over-estimate the actual dimensionality of the space. This is illustrated in Fig. 1 which shows a plot of the number of eigenvectors (using PCA) required to represent 80% of the information content of a set of images, as the number of such sample images are increased. Each image is a 3D two-phase microstructure that satisfies a specific volume fraction and two-point correlation. Notice that as the number of samples increases, the number of eigenvectors required for a moderately accurate representation of the data monotonically increases.

The issue brought out in the discussion above can be understood in a more intuitive way by looking at the simple surface shown in Fig. 2a. The set of points shown in Fig. 2a all lie on a spiral in 3D space. The global coordinates of any point on the spiral is represented as a 3-tuple. Any PCA based model tries to fit a linear surface such that the reconstruction error is minimized. This is shown by the green plane which is a 2D representation of the data. This clearly results in a bad representation of the original data. On the other hand,

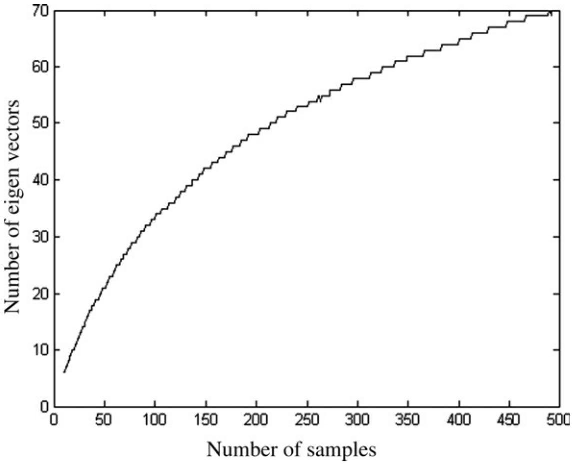

Fig. 1. Plot of the number of samples versus the number of eigenvectors required to represent 80% of the 'energy' spectrum contained in the samples.

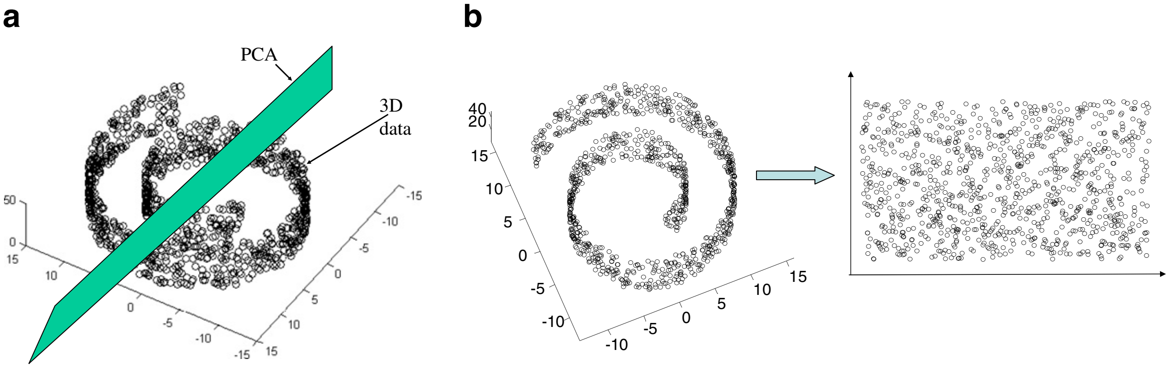

Fig. 2. A set of points lying on a spiral in 3D space. The global coordinates of any point on the spiral is represented as a 3-tuple. The figure on the left (a) shows the reduced-order model resulting from a linear PCA based reduction. The figure on the right (b) depicts a non-linear strategy that results in an accurate representation of the 3D spiral (data taken from [13]) that works by "unravelling and smoothing" the 3D spiral into a 2D sheet.

knowledge of the geometry of the 3D curve results in representing the global position of any point of the curve on a 2D plane, which is obtained by unravelling the spiral into a plane (Fig. 2b). This is essentially the intuitive concept of non-linear model reduction techniques: i.e. they try to unravel and smoothen curves lying in high-dimensional spaces, so that a low-dimensional representation naturally arises. Thus, while PCA based methods would require a 3D representation to accurately describe the spiral shown in Fig. 2, a non-linear model reduction method would identify the geometry of the curve, unravel it and provide a simple 2D representation of the data.

3.2. Non-linear model reduction: preliminary ideas

The essential idea of non-linear model reduction finds its roots in image compression and related digital signal processing principles. Fig. 3a shows multiple images of the same object taken at different left-right and up-down poses. Each image is a 64 × 64 gray-scale picture. Even though each image shown in the figure is defined using $64 \times 64 = 4096$ dimensional vector, each image is in fact uniquely parameterized by just two values, the right-left pose and the up-down pose. It follows that the curve (we will henceforth refer to this curve as the manifold) representing all possible pictures of this object is embedded in $\mathbb{R}^{4096}$ but is parameterized by a region in $\mathbb{R}^2$. The identification of the (small set of) parameters that uniquely define a manifold embedded in a high-dimensional space is called the 'manifold learning problem' [14–16]. This problem of estimating the low-dimensional representation of unordered high-dimensional data sets is a critical problem arising in studies in vision, speech, motor control and data compression.

Analogous to the problem defined above (using Fig. 3a, we define a problem based on the images in Fig. 3b. Fig. 3b shows multiple microstructures that satisfy experimentally determined two-point correlation and volume fraction. Each microstructure is a 65 × 65 × 65 binary image. Each microstructure that satisfies the given experimental statistics is a point that lies on a curve (manifold) embedded in $65 \times 65 \times 65 = 274{,}625$ dimensional space. The problem of 'manifold learning' or parameter estimation as applied to this situation is as follows:

Problem statement A: Given a set of N-unordered points belonging to a manifold $\mathcal{M}$ embedded in a high-dimensional space $\mathbb{R}^n$, find a low-dimensional region $\mathcal{A} \subset \mathbb{R}^d$ that parameterizes $\mathcal{M}$, where $d \ll n$. Classical methods in manifold learning have been methods like the principle component analysis (PCA), Karhunun-Loève expansion (KLE) and multi-dimensional scaling (MDS) [17]. These methods have been shown to extract optimal mappings when the manifold is embedded linearly or almost linearly in the input space. Recently two new approaches have been developed that combine the computational advantages of PCA with the ability to extract the geometric structure of non-linear manifolds. One set of methods preserve the local geometry of the data.

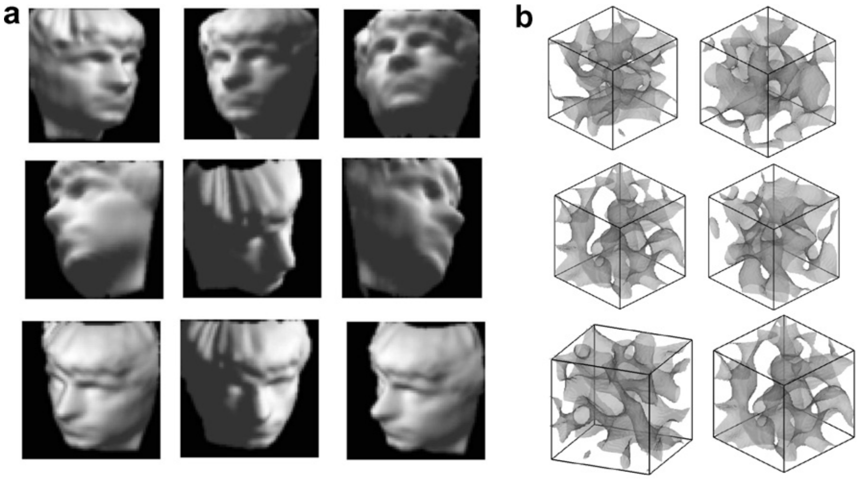

Fig. 3. Figure (a) on the left shows images of the same object (from [13]) taken at various poses (left-right and up-down) while Figure (b) on the right shows various two-phase microstructures that satisfy a specific volume fraction and two-point correlation.

They aim to map nearby points on the manifold to nearby points in the low-dimensional representation. Such methods, locally linear embedding (LLE) [14], Laplacian Eigen Maps, Hessian Eigen maps, essentially construct a homeomorphic mapping between local sets in the manifold to an open ball in a low-dimensional space. The complete mapping is a union of these local maps. An alternate set of approaches towards non-linear model reduction take a top-down approach. Such global approaches, like the Isomap and its numerous variants, attempt to preserve the geometry at *all scales* [15]. They ensure that nearby points on the manifold (with distance defined via a suitable metric) map to nearby points in the low-dimensional space and faraway points map to faraway points in the low-dimensional space. Though both approaches are viable, we focus our attention to global methods of non-linear dimension reduction. The global approach has been shown to provide a faithful representation of the global structure of the data. Furthermore, based on our developments, it is possible to prove tight error bounds and convergence estimates of the model reduction strategy using global methods. Finally, recent advances in such global approaches have made these strategies computationally very efficient.

The basic premise of the global methods (particularly the Isomap [16] algorithm) is that 'the geodesic distances reflect the true low-dimensional geometry of the manifold'. The geodesic distance (between two-points) on a manifold can be intuitively understood to be the shortest distance between the two-points *along the manifold*. Subsequent to the construction of the geodesic distance between the sample points ($\boldsymbol{x}$) in the high-dimensional space, the global methods construct the low-dimensional parametrization simply as a set of points ($\boldsymbol{y}$) lying in a low-dimensional space that most accurately preserve the geodesic distance. For example, the Isomap algorithm is an isometric transformation of the high-dimensional data into low-dimensional points. The global based methods solve the following version of the problem statement:

Problem statement B: Given a set of N-unordered points belonging to a manifold $\mathcal{M}$ embedded in a high-dimensional space $\mathbb{R}^n$, find a low-dimensional region $\mathcal{A} \in \mathbb{R}^d$ which is isometric to $\mathcal{M}$, with $d \ll n$.

The above discussion provides a basic, intuitive picture of the non-linear model reduction strategy. There are several features–the properties that the manifold satisfies, the notion of distances in the high-dimensional space, the optimality of the low-dimensional parametrization and the accuracy of this parametrization–that require a rigorous mathematical footing. We proceed to develop these in the next section.

4. Mathematical formulation

This section is divided into five parts. We first introduce some mathematical preliminaries and define an appropriate distance function $\mathcal{D}$ between two points in $\mathcal{M}_S$. For the straightforward construction of a transformation $\mathcal{F} : \mathcal{M}_S \rightarrow \mathcal{A}$ we ensure that $\mathcal{M}_S$ is topologically well-behaved, i.e. it is smooth and has no holes. This can be ensured by showing that $\mathcal{M}_S$ is compact (Section 4.1). The next step in the construction of the transformation is the estimation of the pair-wise geodesic distance between all the data points. The geodesic distance reflects the true geometry of the manifold embedded in the high-dimensional space. We utilize developments in the graph approximations to geodesics to do the same (Section 4.2). Following this, techniques for estimating the optimal dimensionality of the reduced-order model are developed (Section 4.3). Section 4.4 details the application of the Isomap algorithm (along with the estimate of the optimal dimensionality from Section 4.3) to construct the low-dimensional parametrization of the input data. Finally, Section 4.5 details a non-parametric mapping that serves as the reduced-order, data-driven model of the material topology and thus thermal diffusivity variation.

4.1. Some definitions and the compactness of the manifold $\mathcal{M}$

We provide a few definitions to make the subsequent presentation clear. Detailed proofs of the lemmas stated here can be found in Appendix.

Definition 4.1. By a microstructure $\boldsymbol{x}$, we mean a pixelized representation of a 3D topology. Without loss of generality, we assume that the number of pixels representing the microstructure is $n = q \times q \times q$.

Definition 4.2. Let $\mathcal{M}_{\mathrm{S}}$ denote the set of microstructures $\{\boldsymbol{x}_i\}$ satisfying a set of statistical correlations $\mathrm{S} = \{S_1, \ldots, S_p\}$, i.e.

$$\mathcal{M}_\mathrm{S} = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x} \text{ satisfies } \{S_1, \ldots, S_p\}\}. \tag{6}$$

Note that these correlations have an increasing hierarchy of information content. That is, the two-point correlation contains more information about the material topology and property distribution than, say, the volume fraction.

Definition 4.3. We denote as 'upper-correlation function', S_upper, the statistical correlation function which has a higher-information content than $\{S_1, \ldots, S_p\}$.For instance, if $\mathcal{M}_\mathrm{S}$ is the set of all microstructures satisfying a first-order constraint S_1 (volume fraction), the *upper-correlation function* for $\mathcal{M}_\mathrm{S}$ is the two-point correlation function.

Remark 4.1. In most realistic systems, it is possible to experimentally determine first-order (mean) and second-order (two-point correlation) statistics fairly easily [18]. Higher-order statistics, though feasible, are expensive to experimentally determine. Since we are particularly interested in converting these experimental statistics into viable stochastic models, we will henceforth limit our discussion to the set of two-phase microstructures that satisfy given first- and second-order statistical correlations. Nevertheless, the developments detailed below are in no way limited to second-order statistics and in fact can be extended to include higher-order statistics in a very straightforward manner.

For clarity of presentation, we restrict our analysis to the set of two-phase microstructures satisfying some volume fraction and two-point correlation. Each microstructure $\boldsymbol{x}$ is represented as a $n = q \times q \times q$ pixelized binary image. Each pixel can take values of 0 or 1, representing one of the two phases constituting the two-phase microstructure. We denote the set of microstructures satisfying the given first- and second-order statistics $S = \{S_1, S_2\}$ by $\mathcal{M}_{\mathrm{S}_2}$. The *upper-correlation function* for $\mathcal{M}_{\mathrm{S}_2}$ is the three-point correlation function S_3. That is, $S_3(a, b, c)$ is the probability of finding three points, forming a triangle with sides a, b, c that all belong to the same phase. Since the microstructure is discretized/pixelated, a, b, c take integer values. Also, since the microstructure is finite (defined on $q \times q \times q$ pixels), the number of such integer sets (a, b, c) is finite. Using various reconstruction techniques [19,20], it is possible to generate a set of samples $\boldsymbol{x}_i \in \mathcal{M}_{\mathrm{S}_2}$. The aim of the present work is to utilize these realizations to construct a low-dimensional model for the set $\mathcal{M}_{\mathrm{S}_2}$.

Definition 4.4. The function $\mathcal{D} : \mathcal{M}_{\mathrm{S}_2} \times \mathcal{M}_{\mathrm{S}_2} \to [0, \infty)$ is defined for every $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M}_{\mathrm{S}_2}$ as

$$\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = |S_\mathrm{upper}(\boldsymbol{x}_1) - S_\mathrm{upper}(\boldsymbol{x}_2)|. \tag{7}$$

Based on Remark 4.1, $\mathcal{D}$ is defined as

$$\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{(a,b,c)} |S_3(a, b, c)(\boldsymbol{x}_1) - S_3(a, b, c)(\boldsymbol{x}_2)|, \tag{8}$$

over all possible combinations of a,b,c.

Lemma 4.1. *$(\mathcal{M}_{\mathrm{S}_2}, \mathcal{D})$ is a metric space.*

Remark 4.2. Notion of distance and equivalence between two microstructures: The function $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ provides a notion of distance between two microstructures $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{M}_{\mathrm{S}_2}$. Since $\boldsymbol{x}_1$,$\boldsymbol{x}_2$ belong to $\mathcal{M}_{\mathrm{S}_2}$, both have identical volume fraction, S_1 and two-point correlation S_2. We naturally denote them as equivalent if they have the same *upper-correlation function*, S_3. Since we are dealing with statistically reconstructed microstructures, this definition of equivalence ensures that two microstructures are identical if their higher-order topological characterization is identical.

Remark 4.3. Any other mapping that satisfies the three properties of a metric [21] can be used as a measure of equivalence and distance. Since we are dealing with correlation statistics (that inherently result in limited information about the topology), using the upper-correlation function is a natural way of utilizing this limited information towards quantifying the difference between two points (microstructures) in $\mathcal{M}_{\mathrm{S}_2}$. An alternate idea of representing the distance between microstructures is to compute the pixel-wise difference between the two microstructures, i.e. $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{i,j,k=1}^{q} (\boldsymbol{x}_1(i, j, k) - \boldsymbol{x}_2(i, j, k))^2}$.

Lemma 4.2. *The metric space* $(\mathcal{M}_{\mathrm{S}_2}, \mathcal{D})$ *is totally bounded.*

Lemma 4.3. *The metric space* $(\mathcal{M}_{\mathrm{S}_2}, \mathcal{D})$ *is dense.*

Lemma 4.4. *The metric space* $(\mathcal{M}_{\mathrm{S}_2}, \mathcal{D})$ *is complete.*

Theorem 4.1. *The metric space* $(\mathcal{M}_{\mathrm{S}_2}, \mathcal{D})$ *is compact.*

Proof. Follows from Lemmas 4.2 and 4.4 (see Theorem 45.1 in [22]). □

4.2. Estimating the pair-wise geodesic distances: graph approximations

The Isomap algorithm attempts to find a low-dimensional representation, $\{\boldsymbol{y}_i\}$ of the points $\{\boldsymbol{x}_i\}$ such that $\{\boldsymbol{y}_i\}$ is isometric to $\{\boldsymbol{x}_i\}$ based on the geodesic distances between the points. It is therefore necessary to compute the pair-wise geodesic distance between all the input data points $\{\boldsymbol{x}_i\}$.

Definition 4.5. Denote the intrinsic geodesic distance between points in $\mathcal{M}_{\mathrm{S}_2}$ by $\mathcal{D}_M.\mathcal{D}_M$ is defined as

$$\mathcal{D}_M(\boldsymbol{x}_1, \boldsymbol{x}_2) = \inf_{\gamma}\{\text{length}(\gamma)\}, \tag{9}$$

where γ varies over the set of smooth arcs connecting $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We wish to remind the reader that length of the arcs in the equation above are defined using the distance metric $\mathcal{D}$.

It is important to appreciate the fact that we start off with no knowledge of the geometry of the manifold. We are only given N unordered samples $\{\boldsymbol{x}_i\}$ lying in $\mathcal{M}_{\mathrm{S}_2}$. Hence, the Definition 4.5 of the geodesic distance is not particularly useful in numerically computing the distance between two-points. An approximation of the geodesic distance is required to proceed further. Such an approximation is provided via the concept of graph distance. We subsequently show that this approximation asymptotically matches the actual geodesic distance (Eq. (9)) as the number of samples, N, increases (see Theorem 4.2 below).

The unknown geodesic distances in $\mathcal{M}_{\mathrm{S}_2}$ between the data points are approximated in terms of a *graph distance* with respect to a graph G constructed on the data points. This neighborhood graph G is very simple to construct [16]. Two points share an edge on the graph if they are neighbors. The neighborhood information is estimated in terms of the distance metric $\mathcal{D}$. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are neighbors if $\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \min_{j=1,\ldots,N}(\mathcal{D}(\boldsymbol{x}_1, \boldsymbol{x}_j))$. These neighborhood relations are subsequently represented as a weighted graph G over all the data points. The edges are given weights corresponding to the distance $\mathcal{D}(\boldsymbol{x}_i, \mathbf{x}_j)$ between points.

For points close to each other, the geodesic distance is well approximated by the distance metric $\mathcal{D}$. This is because the curve can be locally approximated to be a linear patch, and the distance between two points on this patch is the straight line distance between them. This straight line distance is given by the distance metric $\mathcal{D}$, which is just the edge length between the points on the graph, G. On the other hand, for points positioned faraway from each other, the geodesic distance is approximated by adding up a sequence of short hops between neighboring points [16]. These short hops can be computed easily from the neighborhood graph G. Denote the shortest path distance between two-points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on the graph G as $\mathcal{D}_G$. The key to constructing the low-dimensional representation is to approximate $\mathcal{D}_M$ as $\mathcal{D}_G$. As the number of input data points increases, the graph distance approximation approaches the intrinsic geodesic distance. The asymptotic convergence of the graph distance to the geodesic distance is rigorously stated in Theorem 4.2. This theorem utilizes some parameters for the quantification of the geometry of the manifold, particularly the minimal radius of curvature, r_o and the minimal branch separation s_o. For the sake of completeness, we state the theorem below. For the sake of brevity, we leave out the definitions of these abstract parameters (the interested reader is referred to [16,22] for discussion of these terms).

Theorem 4.2. *Let* $\mathcal{M}_{\mathrm{S}_2}$ *be a compact manifold of* $\mathbb{R}^n$ *isometrically equivalent to a convex domain* $\mathcal{A} \subset \mathbb{R}^d$. *Let* $0 < \lambda_1, \lambda_2 < 1$ *and* $0 < \mu < 1$ *be given, and let* $\epsilon > 0$ *be chosen such that* $\epsilon < s_o$ *and* $\epsilon \leqslant \frac{2}{\pi} r_o \sqrt{24\lambda_1}$. *A finite sample set* $\{\boldsymbol{x}_i\}$, $i = 1, \ldots, N$ *is chosen randomly from* $\mathcal{M}_{\mathrm{S}_2}$ *with a density* α, *with* $\alpha > \frac{\log\left(\frac{V}{\mu\eta_d(\lambda_2\epsilon/16)^d}\right)}{\eta_d(\lambda_2\epsilon/8)^d}$, *where* V *is the volume* *of $\mathcal{M}_{S_2}$ and η_d is the volume of the unit ball in $\mathbb{R}^d$. The neighborhood graph G is constructed on $\{\boldsymbol{x}_i\}$. Then, with probability at least $1-\mu$, the following inequalities hold for all $\boldsymbol{x}$, $\boldsymbol{y}$ in $\mathcal{M}_{S_2}$:*

$$(1-\lambda_1)\mathcal{D}_M(\boldsymbol{x},\boldsymbol{y}) \leqslant \mathcal{D}_G(\boldsymbol{x},\boldsymbol{y}) \leqslant (1+\lambda_2)\mathcal{D}_M(\boldsymbol{x},\boldsymbol{y}).$$

Proof. Follows from Theorem B in [23]. □

4.3. Estimating the optimal dimension, d, of the low-dimensional representation

$\mathcal{M}_{S_2}$ is intrinsically parameterized by a low-dimensional set $\mathcal{A} \subset \mathbb{R}^d$. As a first step towards constructing $\mathcal{A}$, the pair-wise geodesic distances, $\mathcal{D}_G$, are constructed from the neighborhood graph G. Before we can proceed further, the intrinsic dimensionality of this low-dimensional space, d has to be estimated.

We draw from the recent work in [24,25], where the intrinsic dimension of an embedded manifold is estimated using a novel geometrical probability approach. This work is based on a powerful result in geometric probability – the Breadwood–Halton–Hammersley [26] theorem where d is linked to the rate of convergence of the length functional of the minimal spanning tree of the geodesic distance matrix of the unordered data points in the high-dimensional space. Consistent estimates of the intrinsic dimension d of the sample set are obtained using a very simple procedure.

The sequel utilizes concepts from graph theory. We provide some of the essential definitions below. For a detailed discussion of graphs, trees and related constructs, the interested reader is referred to [27]. Consider a set of k points (vertices) and a graph defined on this set. A graph consists of two types of elements, namely vertices and edges. Every edge has two endpoints in the set of vertices, and is said to connect or join the two endpoints.

- A weighted graph associates a weight (here, this weight is the distance between the vertices) with every edge in the graph.
- A tree is a graph in which any two vertices are connected by exactly one path.
- A spanning tree of a graph (with k vertices) is a subset of $k-1$ edges that form a tree.
- The minimum spanning tree of a weighted graph is a set of edges of minimum total weight which form a spanning tree of the graph.

Definition 4.6. The geodesic minimal spanning tree (GMST) is the minimal spanning tree of the graph G. The length functional, $L(\{\boldsymbol{x}\})$ of the GMST is defined as $L(\{\boldsymbol{x}\}) = \min_{T\in T_G}\sum_{e\in T}|e|$, where T_G is the set of all spanning trees of the graph G and e are the edge-weights of the graph. This is in fact, simply the total weight of the tree.

The mean length of the GMST is linked to the intrinsic dimension d of a manifold $\mathcal{M}_{S_2}$ embedded in a high-dimensional space $\mathbb{R}^n$ through the following theorem:

Theorem 4.3. *Let $\mathcal{M}_{S_2}$ be a smooth d-dimensional manifold embedded in $\mathbb{R}^n$ through a map $\mathcal{F}^{-1}:\mathbb{R}^d \to \mathcal{M}_{S_2}$. Let $2 \leqslant d \leqslant n$. Suppose that $\{\boldsymbol{x}_i\}, i = 1,\ldots,N$ are random points in $\mathcal{M}_{S_2}$. Assume that each of the edge lengths $|e_{ij}|_M$ computed from the graph G converge to $|\mathcal{F}(\boldsymbol{x}_i)-\mathcal{F}(\boldsymbol{x}_j)|_2$ as $N\to\infty$ (i.e. the graph distance converges to the true manifold distance-This is guaranteed by Theorem 4.2). Then the length functional, $L(\{\boldsymbol{x}\})$ of the GMST satisfies:*

$$\lim_{N\to\infty} L(\{\boldsymbol{x}\})/N^{(d'-1)/d'} = \begin{cases} \infty & \text{if } d' < d, \\ \beta_m C & \text{if } d' = d, \\ 0 & \text{if } d' > d, \end{cases} \tag{10}$$

where β_m is a constant, and C is a non-zero function defined in $\mathcal{M}_{S_2}$.

Proof. Follows from Theorem 2 in [28]. □

Theorem 4.3, particularly the asymptotic limit given in Eq. (10) for the length functional of the GMST provides a means of estimating the intrinsic dimension of the manifold. It is clearly seen that the rate of convergence of $L(\{\boldsymbol{x}\})$ is strongly dependent on d. Following [28], we use this strong rate dependence to compute d. Defining $l_N = \log(L(\{\boldsymbol{x}\}))$, gives the following approximation for l_N (from Eq. (10))

$$l_N = a\log(N) + b + \epsilon_N, \tag{11}$$

where

$$a = \frac{d-1}{d}, \tag{12}$$

$$b = \log(\beta_m C), \tag{13}$$

and ϵ_N is an error residual that goes to zero as $N \to \infty$ [24,25,28]. The intrinsic dimensionality, d can be estimated by finding the length functional for different number of samples N and subsequently finding the best fit for (a, b) in Eq. (11).

4.4. Constructing the low-dimensional parametrization, $\mathcal{A}$

Section 4.2 presented the construction of the geodesic distance between all pairs of input points, while Section 4.3 provided an estimation of the dimensionality, d, of the low-dimensional representation. Denote as $\mathbf{M}$, the pair-wise distance matrix (based on the geodesic distance), with elements $d_{ij} = \mathcal{D}_G(\boldsymbol{x}_i, \boldsymbol{x}_j)$ where i, $j = 1, \ldots, N$. Multi-dimensional scaling [17,29] (MDS) arguments are used to compute the set of low-dimensional points that are isometric to these high-dimensional images. The objective of MDS is: *Given a matrix,* $\mathbf{M}$, *of pair-wise distances between N (high-dimensional) points, find a configuration of points in a low-dimensional space such that the coordinates of these N points yield a Euclidean distance matrix whose elements are identical to the elements of the given distance matrix* $\mathbf{M}$.

$$\texttt{Given}\ \mathbf{M},\ \texttt{find}\ \{\mathrm{y}_{\ i}\}, i = 1, \ldots, N\ \texttt{such that}\ \mathrm{y}_i \in \mathbb{R}^d\ \ \texttt{and}$$

$$\sum_{k=1}^{d} (y_{ik} - y_{jk})^2 = d_{ij}^2,\ \texttt{for all}\ i, j. \tag{14}$$

This is called 'coordinate recovery or parametrization': finding a set of points given only the pair-wise distance between the points.

Define the $N \times N$ symmetric matrix $\mathbf{A}$ with elements $A_{ij} = -\frac{1}{2} d_{ij}^2$. Define the $N \times N$ centering matrix [17,29], $\mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}'$ with elements $h_{ij} = \delta_{ij} - 1/N$. Define the $N \times d$ matrix of points $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^{\mathrm{T}}$. Assume without loss of generality that the centroid of the set of points is the origin $\left(\sum_{i=1}^{N} \mathrm{y}_i = \mathbf{0}\right)$. Define the $N \times N$ matrix $\mathbf{B}$ whose components are the scalar products $\mathrm{y}_i^{\mathrm{T}} \cdot \mathrm{y}_j$.

$$b_{ij} = \sum_{k=1}^{d} y_{ik} y_{jk} = \mathrm{y}_{\ i}^{\ \mathrm{T}} \cdot \mathrm{y}_j. \tag{15}$$

The problem at hand is to estimate the coordinates $\boldsymbol{y}_i$, given $\mathbf{M}$. We relate these points $\{\boldsymbol{y}_i\}$ to $\mathbf{B}$. $\mathbf{B}$ in turn can be represented in terms of the known distances $\mathbf{M} = \{d_{ij}\}$. It can easily be shown that [17,29]

$$\mathbf{B} = \mathbf{HAH}. \tag{16}$$

$\mathbf{B}$ is a positive definite matrix that can be decomposed as follows [29]:

$$\mathbf{B} = \mathbf{\Gamma\Lambda\Gamma}. \tag{17}$$

Here, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of the positive eigenvalues of $\mathbf{B}$ and $\mathbf{\Gamma} = (\gamma_1, \ldots, \gamma_N)^{\mathrm{T}}$ the matrix of the corresponding eigenvectors. Note that $\mathbf{B}$ will have a decaying eigen-spectrum. The required set of points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ is related to the d largest eigenvalues/eigenvectors of $\mathbf{B}$. $\mathbf{B}$ is by definition (see Eq. (15), the scalar product matrix)

$$\mathbf{B} = \mathbf{YY}^{\mathrm{T}}. \tag{18}$$

From Eq. (17) and (18), an estimate of $\mathbf{Y}$ in terms of the largest d eigenvectors of $\mathbf{B}$ follows: $\mathbf{Y} = \mathbf{\Gamma}_d \mathbf{\Lambda}_d^{\frac{1}{2}}$, where $\mathbf{\Lambda}_d$ is the diagonal $(d \times d)$ matrix of the largest d eigenvalues of $\mathbf{B}$ and $\mathbf{\Gamma}_d$ is the $N \times d$ matrix of the corresponding eigenvectors.

The knowledge of the optimal dimensionality, d, (from Section 4.3) is quite useful in truncating the eigen values of the $N \times N$ matrix **B**. In contrast, classical PCA techniques truncate the expansion based on representing some percentage of the eigen-spectrum i.e. choose the largest d eigenvalues that account for, say, 95% of the eigen-spectrum $\left(\sum_{i=1}^{d}\lambda_i/\sum_{i=1}^{N}\lambda_i > 0.95\right)$.

We have converted a finite set of data points $\{\boldsymbol{x}_i\}$, $i = 1,\ldots,N$ from $\mathcal{M}_{\mathrm{S}_2} \subset \mathbb{R}^n$ to points $\{\boldsymbol{y}_i\}$,$i = 1,\ldots,N$ in $\mathbb{R}^d$. These points belong to a connected convex subset $\mathcal{A} \subset \mathbb{R}^d$ (Theorem 4.2). This convex region can be numerically estimated as the convex hull of the set of points $\{\boldsymbol{y}_i\}$,$i = 1,\ldots,N$. Hence the low-dimensional parametrization of $\mathcal{M}_{\mathrm{S}_2}$ is given by

$$\mathcal{A} = \{\boldsymbol{\xi} \in \mathbb{R}^d | \boldsymbol{\xi} \in \text{ ConvexHull } (\{\mathrm{y}_{\ 1},\ldots,\mathrm{y}_N\})\}. \tag{19}$$

The stochastic collocation procedure for the solution of SPDEs involves computing the solution at various sample points, $\boldsymbol{\xi}$, from this space, $\mathcal{A}$.

4.5. Construction of a non-parametric mapping $\mathcal{F}^{-1} : \mathcal{A} \to \mathcal{M}_{\mathrm{S}_2}$

Given a set of samples $\{\boldsymbol{x}_i\}$, $i = 1,\ldots,N$ in $\mathcal{M}_{\mathrm{S}_2}$, the non-linear dimension reduction strategy (Section 4.2) coupled with the dimension estimation method (Section 4.3) convert these points into a set of points $\{\boldsymbol{y}_i\}$, $i = 1,\ldots,N$ belonging to a convex set $\mathcal{A}$. *This convex region $\mathcal{A} \subset \mathbb{R}^d$, defines the reduced representation of the space of microstructures, $\mathcal{M}_{\mathrm{S}_2}$, satisfying the given statistical correlations* S_2.$\mathcal{A}$ can be considered to be a surrogate space to $\mathcal{M}_{\mathrm{S}_2}$. One can access the complete variability in the topology and property distribution of microstructures in $\mathcal{M}_{\mathrm{S}_2}$ by simply sampling over the region $\mathcal{A}$. In the collocation based solution strategy for solving SPDEs, one constructs the statistics of the dependant variable by sampling over a discrete set of microstructures. Since we propose to utilize $\mathcal{A}$ as a reduced representation of $\mathcal{M}_{\mathrm{S}_2}$, we sample over a discrete set of points $\boldsymbol{\xi} \in \mathcal{A}$ instead. But we have no knowledge of the image of a random point $\boldsymbol{\xi} \in \mathcal{A}$ in the microstructural space $\mathcal{M}_{\mathrm{S}_2}$ (we only know that the points $\{\boldsymbol{y}_i\}$ for $i = 1,\ldots,N$ map to the microstructures $\{\boldsymbol{x}_i\}$ for $i = 1,\ldots,N$). *For a usable reduced-order model of the microstructure space, an explicit mapping $\mathcal{F}^{-1}$ from $\mathcal{A}$ to $\mathcal{M}_{\mathrm{S}_2}$ has to be constructed.*

There are numerous ways of constructing parametric as well as non-parametric mappings between two sets of objects. For instance, neural networks can be trained using the sample points ($\{\boldsymbol{y}_i\}$, $\{\boldsymbol{x}_i\}$), $i = 1,\ldots,N$ to construct a non-parametric mapping $\mathcal{F}^{-1}$. There have been recent reports of variants of the Isomap algorithm that along with constructing the reduced-order representation of the samples also construct an explicit mapping between the two sets [30]. But there are two significant issues that have to be considered when one utilizes such mapping strategies: (1) Most of these explicit mapping strategies are essentially some form of interpolation rule that utilize the sample set of values ($\{\boldsymbol{y}\}$,$\{\boldsymbol{x}\}$). One has to make sure that the interpolated result, $\mathrm{x}(\mathrm{x} = \mathcal{F}^{-1}(\boldsymbol{\xi}))$ for some arbitrary point $\boldsymbol{\xi} \in \mathcal{A}$ actually *belongs* to $\mathcal{M}_{\mathrm{S}_2}$. (2) Care must be taken to formulate the explicit mapping in a way that results in a computationally simple methodology of finding the images of points. This is very significant considering the fact that we will potentially deal with *very large* pixel sized images (pixelized microstructures or property maps) with pixel counts of the order of 128 × 128 × 128. Any strategy that involves performing non-trivial operations on large data sets of (high resolution) property maps would make the complete process very inefficient. We propose several strategies of constructing computationally simple mappings between the two spaces $\mathcal{M}_{\mathrm{S}_2}$ and $\mathcal{A}$ keeping in mind the issues raised above.

4.5.1. Method 1: nearest neighbor map

This is the simplest map (illustrated schematically in Fig. 4) that sets the image of an arbitrary point in the region $\mathcal{A}$ to the image of the nearest sampled point. The nearest point is the point that is the smallest Euclidean distance from the given arbitrary point. This method is particularly useful when the sampling density (α as defined in Theorem 4.5) is large, i.e. this method results in a reliable mapping when the number of microstructure samples $\{\boldsymbol{x}_i\}$, $i = 1,\ldots,N$ is large.

It is simple to construct error estimates for this mapping. Given an arbitrary point $\boldsymbol{\xi} \in \mathcal{A}$, approximating it by its closest neighbor results in an error, $e_{\mathcal{A}}$, given by $e_{\mathcal{A}} = \min_{i=1,\ldots,N}\sqrt{\sum_{k=1}^{d}(\xi_k - y_{ik})^2}$. From isometry, the

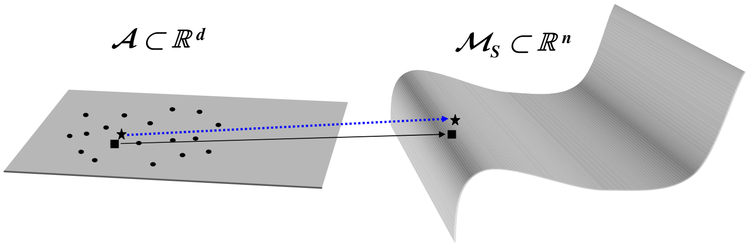
$\mathcal{A} \subset \mathbb{R}^d$
$\mathcal{M}_S \subset \mathbb{R}^n$

Fig. 4. The figure above illustrates the simplest possible mapping between the low-dimensional region $\mathcal{A}$ and the high-dimensional microstructural space $\mathcal{M}_{S_2}$. Given an arbitrary point $\boldsymbol{\xi} \in \mathcal{A}$, find the point $\boldsymbol{y}_k$ closest to $\boldsymbol{\xi}$ from the sampled points $\{\boldsymbol{y}_i\}, i = 1, \ldots, N$. Assign the image value of $\boldsymbol{y}_k$ i.e. $\boldsymbol{x}_k$ to the image of $\boldsymbol{\xi}$.

error between the actual image and the mapped image is given by $e = e_{\mathcal{A}}$. This error can be made arbitrarily small by increasing N.

4.5.2. Method 2: local linear interpolation

A simple non-parametric mapping based on the k-nearest neighbors of a point is defined (Fig. 5) as follows: Given any point $\boldsymbol{\xi} \in \mathcal{A}$, find the k-nearest neighbors, $\hat{\mathrm{y}}_i, i = 1, \ldots, k$ (k defined a priori) to $\boldsymbol{\xi}$ from the set $\{\boldsymbol{y}_i\}$. Compute the (Euclidean) distance $l_i = \sqrt{\sum_{p=1}^{d} (\xi_p - \hat{y}_{ip})^2}$ of $\boldsymbol{\xi}$ from $\hat{\mathrm{y}}_i$ $i = 1, \ldots, k$. The point $\boldsymbol{\xi}$ can be represented as a weighted sum of its k-nearest neighbors as

$$\boldsymbol{\xi} = \frac{\sum_{i=1}^{k} \frac{\hat{\mathrm{y}}_i}{l_i}}{\sum_{i=1}^{k} \frac{1}{l_i}}. \tag{20}$$

Utilizing the fact that the isometric embedding that generated the points $\boldsymbol{y}_i$ from $\boldsymbol{x}_i$ conserves distances, the image of $\boldsymbol{\xi}$ is then given by

$$\mathrm{x} = \frac{\sum_{i=1}^{k} \frac{\hat{\mathrm{x}}_i}{l_i}}{\sum_{i=1}^{k} \frac{1}{l_i}}. \tag{21}$$

That is, the image of $\boldsymbol{\xi}$ is the weighted sum of the images of the k-nearest neighbors of $\boldsymbol{\xi}$ (where the nearest neighbors are taken from the N-sampled points).

The linear interpolation procedure is based on the principle that a small region in a highly curved manifold can be well approximated as a linear patch. This is in fact one of the central concepts that result in local strategies of non-linear dimension reduction [14,15] (see Section 3 for a discussion of global versus local strategies of non-linear dimension reduction). This linear patch is constructed using the k-nearest neighbors of a point $\boldsymbol{\xi}$. As the sampling density (the number of sample points, N) used to perform the non-linear dimension reduction increases, the mean radius of the k-neighborhood of a point approaches zero ($\lim_{N\to\infty} \max_{i=1,\ldots,k} \|\boldsymbol{\xi} - \hat{\mathrm{y}}_i\|_2 \to 0$), ensuring that the linear patch represents the actual curved manifold arbitrarily well.

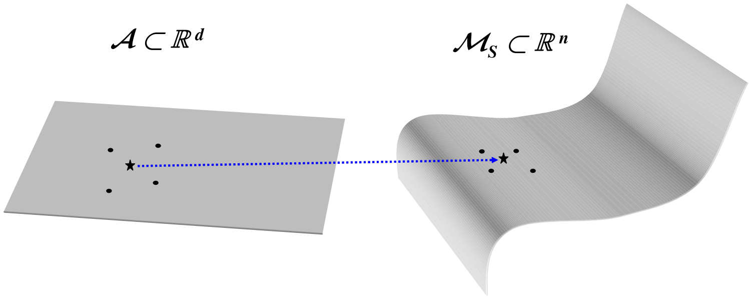
$\mathcal{A} \subset \mathbb{R}^d$
$\mathcal{M}_S \subset \mathbb{R}^n$

Fig. 5. The figure above illustrates a local linear (k-neighbor) interpolation mapping between the low-dimensional region $\mathcal{A}$ and the high-dimensional microstructural space $\mathcal{M}_{S_2}$.

It is possible to estimate approximate error bounds by computing the local curvature of the manifold. Denote by $\boldsymbol{x}$, the image of the point $\boldsymbol{\xi}$. As before, let the k-nearest neighbors of $\boldsymbol{\xi}$ be $\hat{\mathrm{y}}_i$ and their corresponding images in $\mathcal{M}_{\mathrm{S}_2}$ be $\hat{\mathrm{x}}_i$. The local curvature of the manifold in the neighborhood of $\boldsymbol{x}$ can be approximated from the geodesic distances between the points. Let r denote the radius of curvature of the manifold at $\boldsymbol{x}$. The error in the interpolation based representation is caused by considering the space to be locally linear, when it is curved (see Fig. 6). This error is approximated from simple geometry as $e = \max_{i=1,\ldots,k}(r - \sqrt{r^2 - l_i^2})$.

4.5.3. Method 3: local linear interpolation with projection

The local linear interpolation method (Method 2) can be made exact by simply projecting the image obtained after interpolation onto the manifold (shown schematically in Fig. 7). This ensures that the image lies on the manifold $\mathcal{M}_{\mathrm{S}_2}$.

The projection is numerically computed as follows: Denote the microstructure obtained after the interpolation step as $\boldsymbol{x}$. The locally linear interpolation provides a good approximation of the exact image, $\boldsymbol{x} \approx \boldsymbol{x}_{\mathrm{exact}}$, where $\boldsymbol{x}_{\mathrm{exact}}$ is that microstructure on $\mathcal{M}_{\mathrm{S}_2}$ whose geodesic distances from each of the k microstructures $\hat{\mathrm{x}}_i, i = 1,\ldots,k$ is l_i. The errors in this approximation are due to the fact that the approximation $\boldsymbol{x}$ does not *usually lie on the manifold* (as seen in Fig. 7). That is, $\boldsymbol{x}$ does not satisfy the statistical correlations $S = \{S_1,\ldots,S_p\}$ that all points on the manifold satisfy.

The projection operation essentially modifies the point $\boldsymbol{x}$ to satisfy these correlations. This can be achieved computationally by performing a stochastic optimization problem starting from $\boldsymbol{x}$ [11,18,31]. Since $\boldsymbol{x}$ is very close to $\boldsymbol{x}_{\mathrm{exact}}$, these algorithms are guaranteed to reach the local minima defined by $\boldsymbol{x}_{\mathrm{exact}}$. In the context of the numerical examples presented in this work, using two-phase microstructures (see Section 6), the stochastic optimization is done as follows: Starting from the approximate microstructure $\boldsymbol{x}$, compute the volume fraction and two-point correlation of this image. Change the pixel values of t sites in this microstructure such that the volume fraction matches the experimental volume fraction. Following this, randomly swap pixel values in the microstructure (accepting a move only if the error in the two-point correlation decreases), until the two-point correlation matches the experimental value. This can be considered as a version of simulated annealing, with the starting point, $\boldsymbol{x}$ being close to the optimal point, $\boldsymbol{x}_{\mathrm{exact}}$.

One question that arises from this mapping strategy is the following: The initial goal of the mapping was to construct a microstructure (lying on the manifold) that was a distance (geodesically) l_i from $\hat{\mathrm{x}}_i$. The interpolation step ensures this distance (but the microstructure does not lie on the manifold). How much deviation

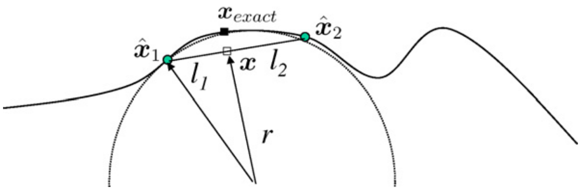

Fig. 6. Simple estimate for the interpolation error in Method 2: Let r denote the local radius of curvature of the function near the point $\boldsymbol{x}_{\mathrm{exact}}$ (the filled square). We approximate the curve by a linear patch, resulting in some error. This error is the distance between the approximate linear image, $\boldsymbol{x}$ (the unfilled square) and the actual point, $\boldsymbol{x}_{\mathrm{exact}}$ (the filled square). This distance can be approximated from simple geometry as a function of r and the local geodesic distance between the points.

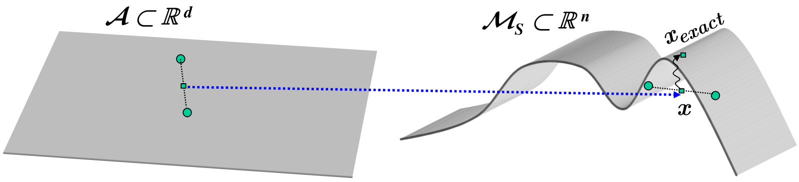
$\mathcal{A} \subset \mathbb{R}^d$
$\mathcal{M}_S \subset \mathbb{R}^n$
$\boldsymbol{x}_{exact}$
$\boldsymbol{x}$

Fig. 7. The local linear interpolation method can be made exact by simply projecting the image obtained after interpolation onto the manifold. This is illustrated in the figure above, where the botted blue line represents the linear interpolation and the curved line represents the projection operator that constructs the image lying on the manifold.

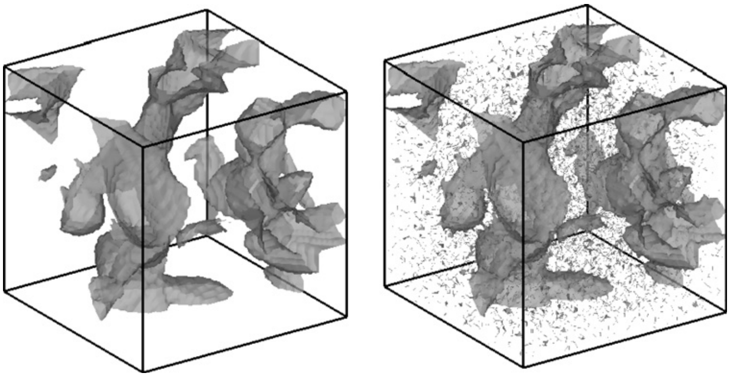

Fig. 8. Sample illustration of the projection step: The figure on the left is a microstructure after interpolation. Projecting it onto the manifold yields the microstructure on the right. There is negligible change in the three point correlation between the micorstructures.

from this distance does the projection operator cause? By swapping the pixels in the microstructure during the stochastic optimization step, the three-point correlation is being changed (which is a measure of distance). But in all of our numerical experiments, this optimization converged within 1000 such pixel flips, thus negligibly affecting the three-point correlation (since the three-point correlation is computed by sampling over $65 \times 65 \times 65 \sim 300{,}000$ points, changing 0.3% of the pixels is negligibly small). This is illustrated pictorially using a sample example shown in Fig. 8. The number of neighbors used for the local linear interpolation map is $k = 10$.

In all three methods detailed above, we only utilize the given input data $\{\boldsymbol{x}_i\}$ to construct the mapping for an arbitrary point $\boldsymbol{\xi} \in \mathcal{A}$. The mapped microstructure, $\mathcal{F}^{-1}(\boldsymbol{\xi})$ is constructed solely based on the available input microstructures and not based on direct reconstruction from moments. In the case of the projection operation used in method 3, the mapped microstructure is moved onto the manifold $\mathcal{M}_{S_2}$. This operation changes less that 0.3% of the microstructure having negligible effect on the thermal behavior of the microstructure.

In potential applications where such projection techniques are infeasible, there are two possible solution strategies that can be pursued: one can use an alternate definition of the distance between the images (see Remark 4.3) or one can utilize more computationally demanding reconstruction [19] or training frameworks [30] to construct the mapping $\mathcal{F}^{-1} : \mathcal{A} \rightarrow \mathcal{M}_{S_2}$.

4.6. The low-dimensional stochastic input model $\mathcal{F}^{-1} : \mathcal{A} \rightarrow \mathcal{M}_{S_2}$

$\mathcal{A}$ represents the space of d-tuples $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)$ that map to microstructures that satisfy the statistical properties $S = \{S_1, \ldots, S_p\}$.

Remark 4.4. In our theoretical derivations in this section, we have ensured that $\mathcal{A}$ is indeed a convex, connected and compact region of $\mathbb{R}^d$. Hence, starting from a large set of points $\{\boldsymbol{y}_i\}, i = 1, \ldots, N$ in $\mathcal{A}$, one can represent the complete region $\mathcal{A}$ as the convex hull of the points $\{\boldsymbol{y}_i\}$, $i = 1, \ldots, N$. Each of the microstructures in $\mathcal{M}_{S_2}$ (by definition) satisfies all required statistical properties, therefore they are equally probable to occur. That is, every point in the manifold $\mathcal{M}_{S_2}$ is equiprobable. This observation provides a way to construct the stochastic model for the allowable microstructures. Define the stochastic model for the topology variation as $\mathcal{F}^{-1}(\boldsymbol{\xi}) : \mathcal{A} \rightarrow M_{S_2}$ where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)$ is a uniform random variable chosen from $\mathcal{A}$. This low-dimensional stochastic model $\mathcal{F}^{-1}$ for the microstructure is the stochastic input in the SPDE (Eq. (4)) defining the diffusion problem.

5. Numerical implementation

This section contains recipes for the numerical implementation of the theoretical developments detailed in the previous sections. We divide this section into various subsections that sequentially discuss the data-driven strategy, starting with the generation of the samples given some limited microstructural information (Section 5.1), the algorithm for constructing the low-dimensional region $\mathcal{A}$ (Section 5.2) and the Smolyak algorithm for the solution of the SPDEs (Section 5.3).

5.1. Microstructure reconstruction: creating the samples $\boldsymbol{x}_i$

Given some experimentally determined statistical correlation functions of the microstructure, the goal is to reconstruct a large set of microstructures satisfying these correlation functions. This is the first step towards building a reduced-order model to the microstructural space. In this work, the microstructure is considered to be a level cut of a Gaussian Random Field (GRF). The statistical correlations are enforced during the reconstruction of the GRF using the given information [32–34]. With this method, a set of N 3D models of the property variations can be generated.

5.2. Constructing the low-dimensional region $\mathcal{A}$

The reconstruction procedure results in a large set of random samples $\boldsymbol{x}_i$ from the space $\mathcal{M}_{S_2}$. The following steps are followed to compute the corresponding points $\{\boldsymbol{y}_i\}$.

Step 1: Find the pair-wise distances, $\mathbf{P}$ between the N samples $\boldsymbol{x}_i, i = 1, \ldots, N$. This is done by first defining an appropriate distance metric, $\mathcal{D}$ between the microstructures. In the example (Section 6) using two-phase microstructure, we define the distance between two microstructures as the difference between their three-point correlations. This operation is obviously of $O(N^2)$ complexity.

Step 2: Construct the neighborhood graph, G, of this sample set. That is, determine which points are neighbors on the manifold based on the distance $\mathbf{P}(i,j)$. Find the nearest k-neighbors of each point. This is performed using a sorting algorithm ($O(NlogN)$ complexity). Connect these points on the graph G and set the edge lengths equal to $\mathbf{P}(i,j)$. The total complexity of this operation is $O(N^2 \log N)$.

```
for i=1:N
[z,I]=sort ($\mathbf{P}(i,:)$), z are the sorted distances
and I are the corresponding indices
$\mathbf{G}$ (1:k,i)=I(2:k+1)
```

Step 3: Estimate the geodesic distance $\mathbf{M}(i,j)$ between all pairs of points on the manifold. This can be done by computing the shortest path distances in the graph G. There are several algorithms to compute the shortest path on a graph. In our implementation, we utilize Floyd's algorithm to compute $\mathbf{M}(i,j)$. The complexity of this step is $O(N^3)$.

```
Initialize $\mathbf{M}(i,j)$ as
$\mathbf{M}(i,j) = \mathbf{P}(i,j)$ if i,j are neighbors or $\mathbf{M}(\mathrm{i},\mathrm{j}) = \infty$ otherwise
for k = 1:N
for each pair (i,j) in 1:N
$\mathbf{M}(i,j) = \min(\mathbf{M}(i,j), \mathbf{M}(i,k)+\mathbf{M}(k,j))$
```

Step 4: Decide on the optimal dimensionality, d, of the low-dimensional space $\mathcal{A}$. Using the graph G [28], estimate the average geodesic MST length. This is done as shown below:

```
Choose $Q$ integers $p = p_1,\ldots,p_Q$ between 1 and N
Randomly pick $p$ samples from the $N$ available samples
Compute the length of the MST of these samples, $L(p)$
Find the best least squares fit value of a for
$L(p) = a\ \log(p) + \epsilon_p$
The optimal dimension d is
$d = \mathrm{round}\left(\frac{1}{1-a}\right)$
```

We utilize the code in [35] to compute the length functional of the MST. The complexity of this step is $O(N \log N)$.

Step 5: Construct the d-dimensional embedding using MDS.

```
Using $\mathbf{M}$ compute $\mathbf{A}$: $A_{ij} = -\frac{1}{2}M_{ij}^2$
Compute $\mathbf{B} = \mathbf{HAH}$
Compute the eigenvalues of $\mathbf{B}$: $\mathbf{B} = \mathbf{\Gamma\Lambda\Gamma}$
Define $\mathbf{Y} = \mathbf{\Gamma\Lambda}^{\frac{1}{2}}$
```

The low-dimensional mapping of the input sample points is given by the first d components of $\mathbf{Y}$. That is, $y_{ik} = \mathbf{Y}(k,i)$ for $i = 1,\ldots,N$ and $k = 1,\ldots,d$. The complexity of this step is $O(nN^2)$, where n is the pixel count in each image. Here, $n = p \times p \times p \sim 128^3$. For number of samples $N \ll n$, MDS is computationally more feasible than PCA, which has a complexity of $O(n^2N)$.

Step 6: Following Remark 4.4, the low-dimensional region $\mathcal{A}$ that maps to the high-dimensional microstructural space $\mathcal{M}_S$ is given by the convex hull of the set of low-dimensional points $\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N\}$. We utilize the Qhull program [36] to compute the convex hull of multi-dimensional data sets.

```
$\mathcal{A}$ = convex hull($\{\mathrm{y}_1,\ldots,\mathrm{y}_N\}$)
```

5.3. Utilizing this reduced-order representation: Stochastic collocation

The above generated d-dimensional embedding is utilized as an input stochastic model for the solution of SPDEs. We utilize a sparse grid collocation strategy for constructing the stochastic solution [7]. The method essentially solves the problem at various points $\boldsymbol{\xi}$ on the stochastic space and constructs an interpolation based approximation to the stochastic solution. The sparse grid collocation strategy used utilizes piecewise multi-linear hierarchical basis functions as interpolation functions [7]. For a given set of stochastic collocation points $\boldsymbol{\xi}_i$, $i = 1,\ldots,M$, from $\mathcal{A}$, we find the corresponding images of these points (using the procedures detailed in Section 4.5) $\mathbf{x}_i = \mathcal{F}^{-1}(\boldsymbol{\xi}_i)$. These microstructural images are utilized as inputs (property maps) in the solution of the SPDEs.

Remark 5.1. In the stochastic collocation approach, the collocation points are usually given in the unit hypercube, i.e. $\boldsymbol{p} \in [0,1]^d$. As a first step, this point $\boldsymbol{p}$ must be mapped to a corresponding point $\boldsymbol{\xi} \in \mathcal{A}$.

```
for i = 1: M
Determine the collocation point $\boldsymbol{p}_\mathrm{i} \in [0,1]^\mathrm{d}$
Compute the point $\boldsymbol{\xi}_i \in \mathcal{A}$
Compute $\mathbf{x}_i = \mathcal{F}^{-1}(\boldsymbol{\xi}_i)$
Solve the PDE using $\mathbf{x}_\mathrm{i}$ as input
```

6. Illustrative example

In this section, we showcase the theoretical developments detailed in the previous sections with a realistic example.

6.1. Two-phase microstructures

The non-linear dimension reduction strategy is applied to construct a reduced-order model for the distribution of material in a two-phase metal–metal composite. The problem of interest is as follows:

Compute the PDFs of temperature evolution in a metal–metal composite microstructure when only limited statistics of the material distribution is known.

This topological uncertainty translates to uncertainties in the thermal diffusivity $\alpha(x)$ of the microstructure. For clarity of presentation, we divide the solution into multiple sections. (1) The first step is the extraction of topological statistics from the experimental image provided. These statistics are then utilized to reconstruct a large set of 3D microstructures $\{\boldsymbol{x}_i\}$, $i = 1,\ldots,N$. (2) The next step is to construct the low-dimensional representation of the class of microstructures utilizing the samples $\{\boldsymbol{x}_i\}$, $i = 1,\ldots,N$ in the input space. (3) The final step is to utilize the reduced-order representation of the microstructural topology (and hence, the thermal diffusivity coefficient) as an input stochastic model to solve for the evolution of the temperature statistics.

6.1.1. Data extraction and sample set construction

We start from a given experimental image of a microstructure. The image (204 µm × 236 µm), shown in Fig. 9, is of a Tungsten–Silver composite [37]. This is a well characterized system, which has been used to test various reconstruction procedures [20,38]. The first step is to extract the necessary statistical information from the experimental image. The image is cropped, deblurred and discretized. The volume fraction of silver is $p = 0.2$. The experimental two-point correlation is extracted from the image. The normalized two-point correlation ($g(r) = \frac{L_2(r)-p^2}{p-p^2}$), is shown in Fig. 10. The data extraction was performed in Matlab.

The next step is to utilize these extracted statistical relations (volume fraction and two-point correlation) to reconstruct a class of 3D microstructures. We utilize a statistics based reconstruction procedure based on Gaussian Random Fields (GRF). In this method, the 3D microstructure is obtained as the level cuts to a random field, $\varphi(\boldsymbol{x})$, $\boldsymbol{x} \in D$. The random field has a field–field correlation $\langle \varphi(0)\varphi(r) \rangle = \gamma(r)$. The statistics of the reconstructed 3D image can be matched to the experimental image by suitably modifying the field–field correlation function and the level cut values (see [20] for a detailed discussion). Following the work in [20], the GRF is assumed to satisfy a specified field–field correlation given by

$$\gamma(r) = \frac{\mathrm{e}^{-r/\beta} - (r_c/\beta)\mathrm{e}^{-r/r_c}}{1 - (r_c/\beta)} \frac{\sin(2\pi r/d)}{2\pi r/d}, \tag{22}$$

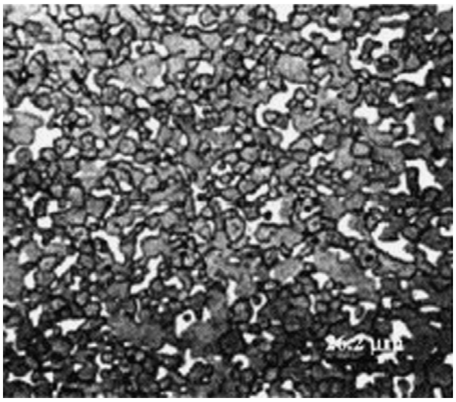

Fig. 9. Experimental image of a two-phase composite (from [37]).

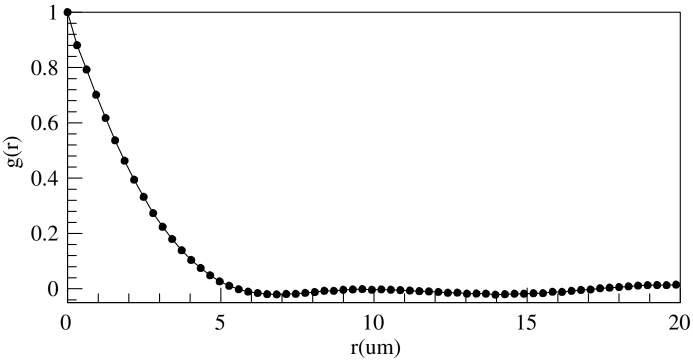

Fig. 10. The two-point correlation function.

where the field is characterized by the correlation length β, a domain scale d and a cutoff scale r_c. For a specific choice of (β, d, r_c), one can construct a microstructure from the resulting GRF. The (theoretical) two-point correlations corresponding to this reconstructed microstructures is computed. Optimal values of (β, d, r_c) are obtained by minimizing the error between the theoretical two-point correlation and the experimental two-point correlation. The theoretical two-point correlation corresponding to $(\beta, d, r_c) = (2.229, 12.457, 2.302)$ μm is plotted in Fig. 11.

Using the optimal parameters of the GRF (to match with the experimental data), realizations of 3D microstructure were computed. Each microstructure consisted of 65 × 65 × 65 pixels. This corresponds to a size of 20 μm × 20 μm × 20 μm. One realization of the 3D microstructure reconstructed using the GRF is shown in Fig. 12.

6.1.2. Non-linear dimension reduction and construction of the topological model

The GRF based reconstruction detailed above was used to generate a set of $N = 1000$ samples of two-phase microstructure. Each microstructure is represented as a 65 × 65 × 65 pixel image. The three-point correlations of all these samples are calculated. The three-point correlation is easily computed as follows: for a given value of (a, b, c), randomly place triangles of side lengths a, b, c on the microstructure. Count the number of times all three vertices of the triangle lie on the same phase. $S_3(a, b, c)$ is the ratio of the number of such successful placements over the total number of tries. In our computations, we randomly place 500,000 triangles to compute S_3 for each value of (a, b, c). The total computational time to reconstruct 1000 microstructures along with their S_3 was 30 min on 25 nodes of our in-house Linux cluster.

Based on the calculated S_3, the pair-wise distance matrix $\mathbf{P}$ is computed. This took 6 min to compute on a 3.8 GHz PC. From this, the geodesic distance matrix $\mathbf{M}$ and the graph G are computed. These are used to estimate the optimal dimensionality of the low-dimensional space by computing the length functional of

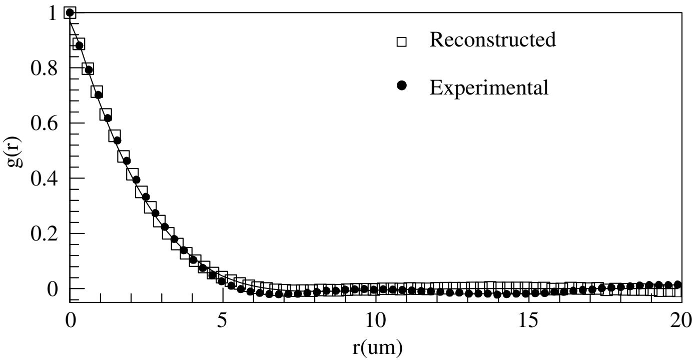

Fig. 11. Comparison of the two-point correlation function from experiments and from the GRF.

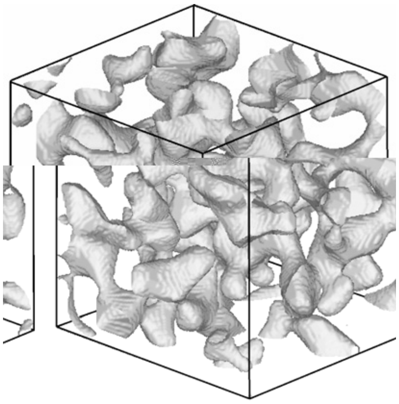

Fig. 12. One instance (realization) of the two-phase microstructure.

the MST of the graph G. The Matlab code available at [35] was utilized. Fig. 13 plots the length functional of the MST of the graph G for various sample sizes. The optimal dimensionality of the low-dimensional set is related to the slope of this line (see Eq. (11)). The slope of the curve is computed using a simple least squares fit. The optimal dimensionality was estimated to be $d = 9$ ($a \sim 0.885$). The total computational time to estimate the dimensionality was 8 min on a 3.8 GHz PC.

Multi-dimensional scaling is performed using the geodesic distance matrix $\mathbf{M}$. The nine largest eigenvalues and their corresponding eigenvectors are used to represent the input samples. The low-dimensional region $\mathcal{A}$ is constructed as the convex hull of these $N(=1000)$ nine-dimensional points ξ_i. This region coupled with the mappings developed in Section 4.5 define the reduced-order stochastic input model $\mathcal{F}^{-1} : \mathcal{A} \rightarrow \mathcal{M}_{S_2}$.

Fig. 14 illustrates the potential difficulty in choosing the dimensionality of the region $\mathcal{A}$ based on simple variance errors (i.e. choose the value of d that accounts for 90% of the variance in the data). Fig. 14a plots the eigenspectrum of the computed eigenvalues of $\mathbf{B}(\sum_{i=1}^{d}\lambda_i/\sum_{i=1}^{N}\lambda_i)$. All dimensionalities beyond $d = 4$ account for over 90% of the variance in the data. Hence, there arises some ambiguity in simply choosing the dimensionality of the reduced mode based on this plot. Fig. 14b plots the residual variance of the low-dimensional representation for various dimensionalities. The residual variance measures the difference between the intrinsic manifold distance matrix, $\mathbf{M}$ and the pair-wise Euclidian matrix, $\mathbf{D}_{\mathcal{A}}$, recovered from MDS (see Eq. (14)) for various dimensionalities d. It is defined in terms of the element-wise correlation

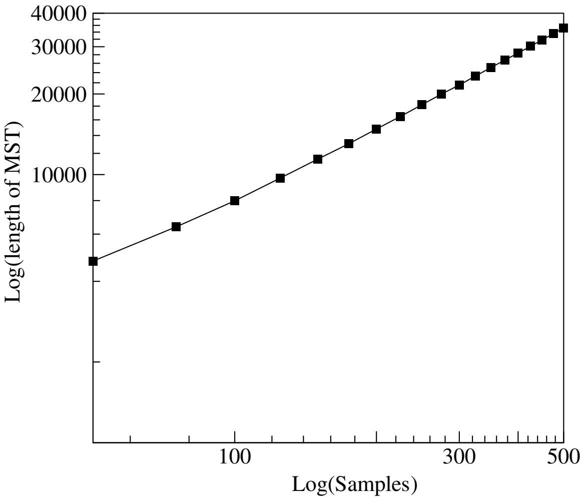

Fig. 13. Plot of the length functional of the MST of the graph G for various sample sizes.

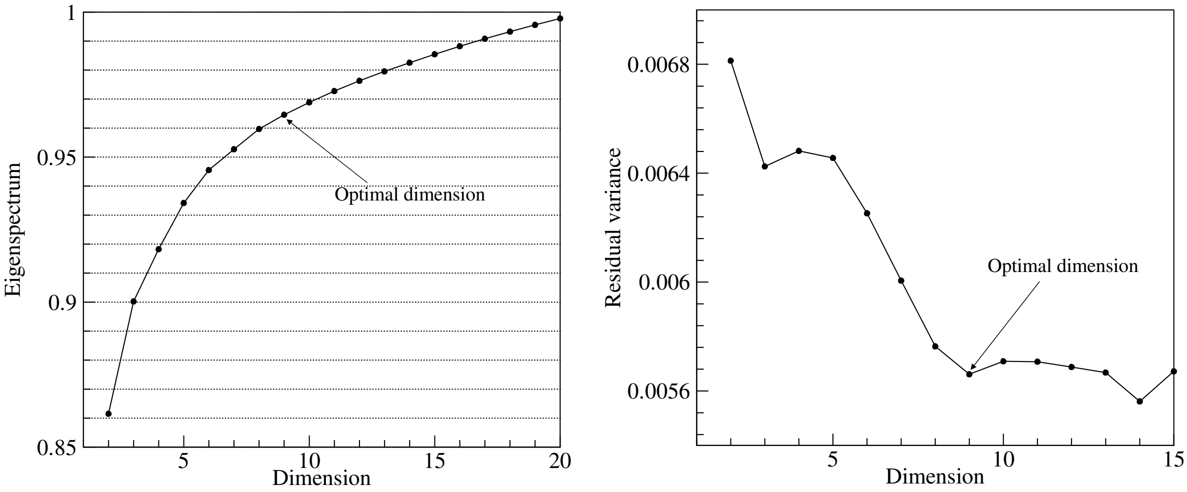

Fig. 14. (Left) The cumulative eigenspectrum of the data, $\sum_{i=1}^{d}\lambda_i/\sum_{i=1}^{N}\lambda_i$. (Right) The residual variance for different dimensionalities of the region $\mathcal{A}$, computed from MDS.

between the two matrices, $e = 1 - r^2(\mathbf{D}_{\mathcal{A}}, \mathbf{M})$, where r here is the standard linear correlation coefficient [39]. Notice that all the dimensions above $d = 8$ have fairly small variance, with $d = 14$ having the minimum. This ambiguity in choosing the dimensionality of the low-dimensional representation is overcome by using the ideas discussed in Section 4.3, resulting in an optimal dimensionality of $d = 9$.

6.1.3. Utilizing the model to solve a stochastic PDE: diffusion in random heterogeneous media

The procedure detailed above results in the function $\mathcal{F}^{-1}$. $\mathcal{F}^{-1}$ is a mapping from a nine-dimensional space $\mathcal{A}$ to the space of microstructures $\mathcal{M}_{S_2}$. $\mathcal{F}^{-1}$ along with $\xi \in \mathcal{A}$ serve as the stochastic input for the diffusion equation. A simple diffusion problem is considered (Eq. (4)). A computational domain of 65 × 65 × 65 is considered (this corresponds to a physical domain of 20 μm × 20 μm × 20 μm). The random heterogeneous microstructure is constructed as a 65 × 65 × 65 pixel image. The steady-state temperature profile, when a constant temperature of 0.5 is maintained on the left wall and a constant temperature of −0.5 is maintained on the right wall, is evaluated. All the other walls are thermally insulated. The axis along which the temperature boundary conditions are imposed is denoted as the x-axis (left-right) while the vertical axis is the z-axis.

The construction of the stochastic solution is through sparse grid collocation strategies (Smolyak algorithm). A level 5 interpolation scheme is used to compute the stochastic solution in nine dimensions. The stochastic problem was reduced to the solution of 26,017 deterministic decoupled equations. Fifty nodes (each with two 3.8G CPUs) of our 64-node Linux cluster were utilized to solve these deterministic equations. These are dual core processors with hyper-threading capabilities thus each node was used to perform the computation for four such problems. The total computational time was about 210 min . Each deterministic problem involved the solution of a diffusion problem on a given microstructure using an 64 × 64 × 64 element grid (uniform hexahedral elements).

The reduction in the interpolation error with increasing depth of interpolation is shown in Fig. 15. The interpolation error is defined as the variation of the interpolated value of the function from the actual function value $|u - I(u)|$. This is measured in terms of the hierarchical surpluses, w^i (here taken as the sum of the absolute value of the hierarchical surpluses in all stochastic dimensions), where i is the depth of interpolation [7]. Define the error as $e = \max_{j=1:n_{no}}(w^i)$, where n_{no} is the number of nodes in the finite element discretization of the spatial domain, D. As the level of interpolation increases, the number of sampling points used to construct the stochastic solution increases [7]. Notice that there is a slight jump in the error going from an interpolation of depth 2 to depth 3. This is probably due to the presence of some highly localized fluctuations of the stochastic solution that is captured only when the depth of interpolation reaches 3. Nevertheless, the error reduction shown above follows the theoretical convergence estimates for using Smolyak based interpolation [7].

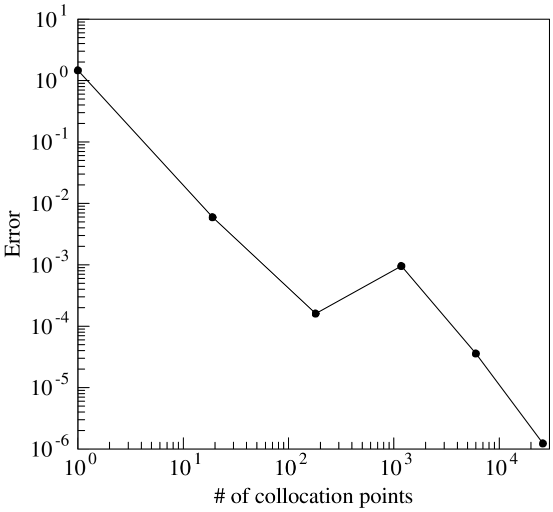

Fig. 15. Reduction in the interpolation error with increasing number of collocation points.

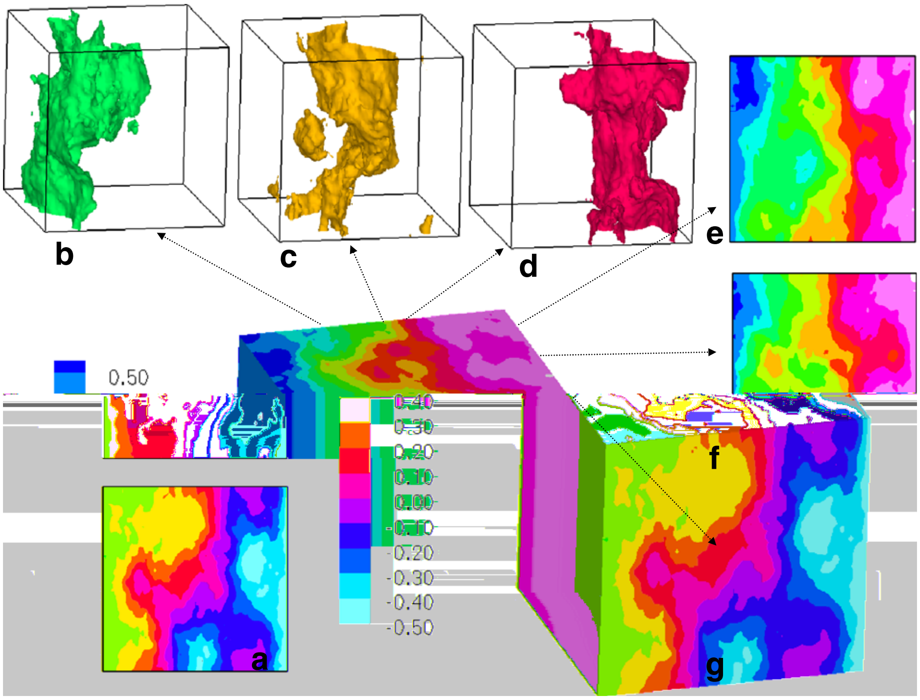

Fig. 16. Steady-state mean temperature: (a) temperature contour, (b–d) temperature iso-surfaces, and (e–g) temperature slices.

The mean temperature is shown in Fig. 16. The figure plots iso-surfaces of temperatures −0.25 (Fig. 16b), 0.0 (Fig. 16c) and 0.25 (Fig. 16d). The figure also shows temperature slices at three different locations of the xz plane: $y = 0$ (Fig. 16e), $y = 8$ μm (Fig. 16f) and $y = 16$ μm (Fig. 16(g)).

The standard deviation and other higher-order statistics of the temperature variation are shown in Fig. 17. Fig. 17a plots standard deviation iso-surfaces. Fig. 17d–f plot slices of the temperature deviation at three different planes $y = 0, y = 8$ μm, $y = 16$ μm, respectively. The standard deviation reaches 48% of the maximum temperature difference maintained. A point from a region of high-standard deviation ($A = (4, 4, 20)$ μm) is chosen and the PDF of temperature at this point is determined. Fig. 17c plots the PDF for the point.

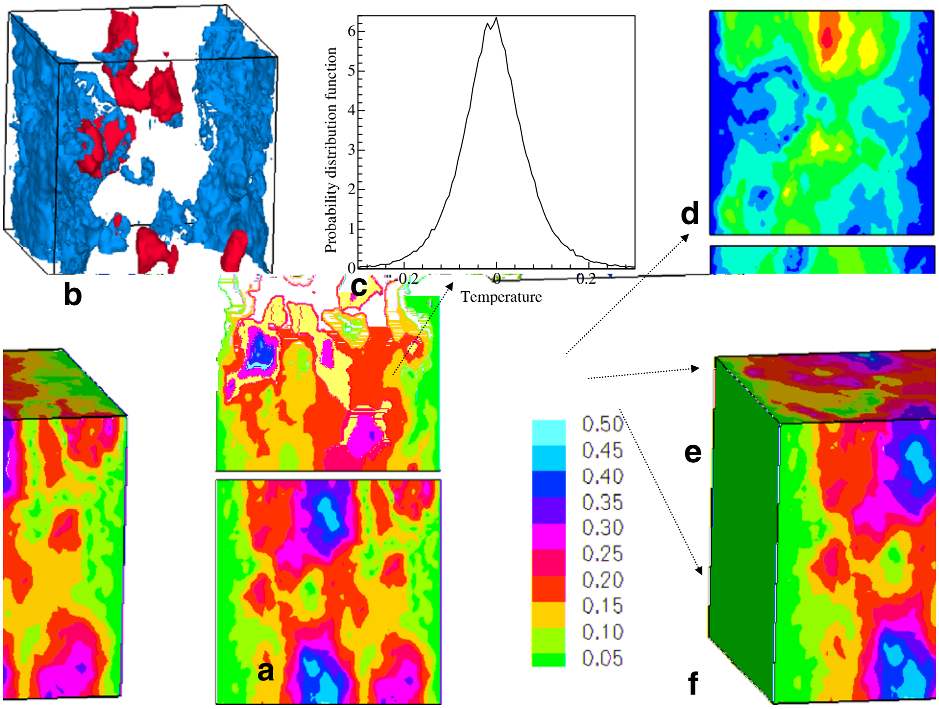

Fig. 17. Standard deviation of temperature: (a) standard deviation contours, (b) standard deviation iso-surfaces, (c) temperature PDF at a point, and (d–f) standard deviation slices.

We conclude this section by making a few observations. The physical features of the topology variation will have an effect on the low-dimensional model. If the correlation length (of the two-point correlation) decreases, the optimal dimensionality of the model will increase and vice versa. The model reduction strategy developed is data-driven: the transformation only converts the given finite input data set into a set of low-dimensional points. In case the input data, $\{\boldsymbol{x}_i\}$ all belong to a localized region of $\mathcal{M}_{S_2}$, the model reduction strategy framework will construct a low-dimensional parametrization of only this localized region. By increasing the amount of data utilized (i.e. N), one can make sure that the complete space, $\mathcal{M}_{S_2}$ is sampled, ensuring that the data-driven model reflects the variability in the complete space.

7. Conclusions

A non-linear model reduction technique for converting experimentally determined statistics into viable, realistic stochastic input models of property variability has been developed in the present work. The major advantages of the proposed developments are: it seamlessly meshes with any reconstruction method, directly converts samples into an equiprobable low-order model of the property, and is applicable to *any* property variation (for instance, property variation in polycrystalline materials, permeability variation in heterogeneous porous media, etc.).

The current developments borrow generously from ideas in image processing and psychology where the problem of manifold learning is frequently encountered. Ideas from differential geometry are employed to show the accuracy and asymptotic convergence of the reduced-order model. We showcase the framework developed to construct a realistic reduced-order stochastic model that describes the material and property variation in a two-phase microstructure (starting from an experimental image of the microstructure). We utilize this stochastic model as an input in the solution of a SPDE governing diffusion in random heterogeneous media. The solution provides an understanding of how uncertainty in the topology of the microstructure affects the evolution of a dependant variable (temperature).

The basic model reduction ideas envisioned in this work are not limited to generation of viable stochastic input models of property variations. This framework has direct applicability to problems where working in high-dimensional spaces is computationally intractable, for instance, in visualization of property evolution, extracting process-property maps in low-dimensional spaces, among others. Furthermore, the generation of a low-dimensional surrogate space has major ramifications in the optimizing of properties-processes and structures, making complicated operations like searching, contouring and sorting computationally much more feasible. These potentially exciting areas of application of the non-linear model reduction framework developed here offer fertile avenues of further research.

Different reduction techniques (for instance, locally linear embedding, kernel PCA, self organizing maps) can be incorporated into the general model reduction strategy formulated here. This is an area that is unexplored and could potentially result in very efficient, real time, data-driven, stochastic reduced-order model generation techniques. In addition to the importance of such models in process modeling of heterogeneous materials (polycrystals, composites, concrete, etc.), many other technological applications in modeling multiscale thermal/flow transport in geological media, soil contamination and reservoir engineering remain to be explored.

Acknowledgments

This research was supported by the Computational Mathematics program of AFOSR (grant F49620-00-1-0373). The computing was conducted in part using the resources of the Cornell University Center for Advanced Computing.

Appendix. Properties of the manifold

Lemma 4.1. *$(\mathcal{M}_{S_2}, \mathcal{D})$ is a metric space.*

Proof. For a function $\mathcal{D}$ to be a metric defined over the set $\mathcal{M}_{S_2}$, it must satisfy the properties of non-negativity, symmetry and the triangle inequality [21].

1. *Positive definiteness*: $S_3(a,b,c)(\boldsymbol{x})$ is a non-negative, continuous function. For any $\boldsymbol{x},\boldsymbol{y} \in \mathcal{M}_{S_2}$ $\mathcal{D}(\boldsymbol{x},\boldsymbol{y}) \geqslant 0$ follows from $|S_3(a,b,c)(\boldsymbol{x}) - S_3(a,b,c)(\boldsymbol{y})| \geqslant 0$. Also, $\mathcal{D}(\boldsymbol{x},\boldsymbol{x}) = 0$, by definition.
2. *Symmetry*: $\mathcal{D}(\boldsymbol{x},\boldsymbol{y}) = \mathcal{D}(\boldsymbol{y},\boldsymbol{x})$ from the definition of $\mathcal{D}$.
3. *Triangle inequality*: for any $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathcal{M}_{S_2}\mathcal{D}(\boldsymbol{x},\boldsymbol{y}) = \sum |S_3(a,b,c)(\boldsymbol{x}) - S_3(a,b,c)(\boldsymbol{y})| = \sum |S_3(a,b,c)(\boldsymbol{x}) - S_3(a,b,c)(\boldsymbol{z}) + S_3(a,b,c)(\boldsymbol{z}) - S_3(a,b,c)\boldsymbol{y})| \leqslant \sum |S_3(a,b,c)(\boldsymbol{x}) - S_3(a,b,c)(\boldsymbol{z})| + \sum |S_3(a,b,c)(\boldsymbol{z}) - S_3(a,b,c)(\boldsymbol{y})|dr = \mathcal{D}(\boldsymbol{x},\boldsymbol{z}) + \mathcal{D}(\boldsymbol{z},\boldsymbol{y})$. □

Lemma 4.2. *The metric space $(\mathcal{M}_{S_2}, \mathcal{D})$ is totally bounded*

Proof. A metric space is bounded iff

$$\exists r_d \in \mathbb{R}^+ \texttt{such that } \mathcal{D}(\boldsymbol{x},\boldsymbol{y}) \leqslant r_d, \quad \forall \boldsymbol{x},\boldsymbol{y} \in \mathcal{M}_{S_2}.$$

For each (a,b,c),$S_3(a,b,c)(\boldsymbol{x})$ denotes the probability of a randomly placed triangle with sides (a,b,c) having its three vertices belong to the same phase. It follows that $0 \leqslant S_3(a,b,c)(\boldsymbol{x}) \leqslant 1$. Now, for any $\boldsymbol{x},\boldsymbol{y} \in \mathcal{M}_{S_2}$

$$\mathcal{D}(\boldsymbol{x},\boldsymbol{y}) = \sum_{(a,b,c)} |S_3(a,b,c)(\boldsymbol{x}) - S_3(a,b,c)(\boldsymbol{y})| \leqslant \sum_{(a,b,c)} |S_3(a,b,c)(\boldsymbol{x})| + \sum_{(a,b,c)} |S_3(a,b,c)(\boldsymbol{y})|$$
$$\leqslant 2 \times \max\left\{\sum_{(a,b,c)} |S_3(a,b,c)(\boldsymbol{x})|, \sum_{(a,b,c)} |S_3(a,b,c)(\boldsymbol{y})|\right\} \leqslant 2\sum_{(a,b,c)} 1 = r_d.$$

Hence, we have shown that the metric space $(\mathcal{M}_{S_2}, \mathcal{D})$ is bounded. r_d is an upper bound of the 'diameter' of the metric space. □

We can define a volume corresponding to this diameter, i.e $(\mathcal{M}_{S_2}, \mathcal{D})$ can be inscribed into a n-ball (*n*-dimensional sphere) of diameter r_d. Denote this volume as $V(\mathcal{M}_{S_2}).V(\mathcal{M}_{S_2})$ is given by the volume of the n-ball with radius $r_d/2, V(\mathcal{M}_{S_2}) = (\pi^{n/2}(r_d/2)^n)/\Gamma(1+n/2)$, where $\Gamma(1+n/2)$ is the Gamma function [40]. Now, given any $\epsilon > 0$, one can cover $\mathcal{M}_{S_2}$ by a finite number of ϵ-balls (i.e. *n*-balls of radius ϵ). An estimate on the number of such balls required is given by $\frac{r_d^n}{\epsilon^n}$. Hence, the metric space $(\mathcal{M}_{S_2}, \mathcal{D})$ is totally bounded.

Lemma 4.3. *The metric space* $(\mathcal{M}_{S_2}, \mathcal{D})$ *is dense.*

Proof. A metric space is dense iff

$$\texttt{Given } \delta > 0, \texttt{ for any } \boldsymbol{x} \in \mathcal{M}_{S_2},$$
$$\exists \texttt{ at least one } \boldsymbol{y} \in \mathcal{M}_{S_2} \texttt{ such that } \mathcal{D}(\mathbf{x}, \mathbf{y}) < \delta.$$

For any given $\boldsymbol{x} \in \mathcal{M}_{S_2}$, compute $S_3(\boldsymbol{x})(a,b,c)$. Set $S_3(\boldsymbol{y})(a,b,c) = S_3(\boldsymbol{x})(a,b,c) + \frac{2\delta}{r_d}$. Now using any appropriate reconstructing methodology (see Section 5.1), construct a microstructure $\boldsymbol{y}$ satisfying S_1,S_2 as well as $S_3(\boldsymbol{y})(a,b,c)$. Since $\boldsymbol{y}$ satisfies S_1,S_2, it belongs to $\mathcal{M}_{S_2}$ and since it satisfies $S_3(\boldsymbol{y})(a,b,c)$,

$$\mathcal{D}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{(a,b,c)} |S_3(a,b,c)(\boldsymbol{x}) - S_3(a,b,c)(\boldsymbol{y})| = \sum_{(a,b,c)} \left| S_3(a,b,c)(\boldsymbol{x}) - S_3(a,b,c)(\boldsymbol{x}) + \frac{2\delta}{r_d} \right| = \sum_{(a,b,c)} \frac{2\delta}{r_d} \leqslant \delta.$$

Hence, $(\mathcal{M}_{S_2}, \mathcal{D})$ is dense. □

Lemma 4.4. *The metric space* $(\mathcal{M}_{S_2}, \mathcal{D})$ *is complete.*

Proof. Consider any Cauchy sequence, $\{\boldsymbol{x}_n\}$, in $\mathcal{M}_{S_2}$. A Cauchy sequence in $\mathcal{M}_{S_2}$ satisfies the following:

$$\texttt{Given } \epsilon > 0, \exists N \texttt{ such that } \mathcal{D}(\boldsymbol{x}_n, \boldsymbol{x}_m) < \epsilon, \texttt{ whenever } m,n > N.$$

That is, $\mathcal{D}(\boldsymbol{x}_{n+k}, \boldsymbol{x}_n) \to 0$ uniformly in k as $n \to \infty$. We have defined two microstructures to be equivalent if $\mathcal{D}(\boldsymbol{x}, \boldsymbol{y}) = 0$ (see Remark 4.2). Hence, by this definition of equivalence of microstructures, it follows that the sequence converges to a point $\boldsymbol{x} \in \mathcal{M}_{S_2}$ (the limit point of this sequence satisfies $S = \{S_1, S_2\}$, hence it belongs to $\mathcal{M}_{S_2}$). Since every Cauchy sequence in $\mathcal{M}_{S_2}$ converges in $\mathcal{M}_{S_2}$, the metric space $(\mathcal{M}_{S_2}, \mathcal{D})$ is complete (Lemma 43.1 in [22]). □

References

[1] C. Desceliers, R. Ghanem, C. Soize, Maximum likelihood estimation of stochastic chaos representations from experimental data, Int. J. Numer. Meth. Eng. 66 (2006) 978–1001.

[2] L. Guadagnini, A. Guadagnini, D.M. Tartakovsky, Probabilistic reconstruction of geologic facies, J. Hydrol. 294 (2004) 57–67.

[3] C.L. Winter, D.M. Tartakovsky, Mean flow in composite porous media, Geophys. Res. Lett. 27 (2000) 1759–1762.

[4] C.L. Winter, D.M. Tartakovsky, Groundwater flow in heterogeneous composite aquifers, Water Resour. Res. 38 (2002) 231.

[5] B. Ganapathysubramanian, N. Zabaras, Modelling diffusion in random heterogeneous media: data-driven models, stochastic collocation and the variational multi-scale method, J. Comput. Phys. 226 (2007) 326–353.

[6] I. Babuska, R. Tempone, G.E. Zouraris, Galerkin finite elements approximation of stochastic finite elements, SIAM J. Numer. Anal. 42 (2004) 800–825.

[7] B. Ganapathysubramanian, N. Zabaras, Sparse grid collocation schemes for stochastic natural convection problems, J. Comput. Phys. 225 (2007) 652–685.

[8] GSLIB: Geostatistical Software Library, available at: <http://www.pe.utexas.edu/Geosci/Software/GSLIB/gslib.html>.

[9] N. Zabaras, S. Sankaran, An information-theoretic approach to stochastic materials modeling, IEEE Comput. Sci. Eng. (CISE) (March/April) (2007) 50–59 (Special issue of 'Stochastic Modeling of Complex Systems').

[10] V. Sundararaghavan, N. Zabaras, Classification of three-dimensional microstructures using support vector machines, Comput. Mater. Sci. 32 (2005) 223–239.

[11] S. Sankaran, N. Zabaras, A maximum entropy approach for property prediction of random microstructures, Acta Mater. 54 (2006) 2265–2276.

[12] N. Zabaras, S. Sankaran, Computing property variability of polycrystals induced by grain size and orientation uncertainties, Acta Mater. 55 (2007) 2279–2290.

[13] A Global Geometric Framework for Nonlinear Dimensionality Reduction, freely downloadable software available at: <http://isomap.stanford.edu/>.

[14] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[15] V. deSilva, J.B. Tenenbaum, Global versus local methods in nonlinear dimensionality reduction, Adv. Neural Inform. Process. Syst. 15 (2003) 721–728.

[16] J. Tenenbaum, V. DeSilva, J. Langford, A global geometric framework for nonlinear dimension reduction, Science 290 (2000) 2319–2323.

[17] J.T. Kent, J.M. Bibby, K.V. Mardia, Multivariate Analysis (Probability and Mathematical Statistics), Elsevier, 2006.

[18] S. Torquato, Statistical description of microstructures, Ann. Rev. Mater. Sci. 32 (2002) 77–111.

[19] S. Torquato, Random Heterogeneous Materials: Microstructure and Macroscopic Properties, Springer, 2002.

[20] A.P. Roberts, E.J. Garboczi, Elastic properties of a tungsten–silver composite by reconstruction and computation, J. Mech. Phys. Solids 47 (1999) 2029–2055.

[21] C. Goffman, G. Pedrick, First Course in Functional Analysis, Prentice-Hall, 2002.

[22] J.R. Munkres, Topology, second ed., Prentice-Hall, 2000.

[23] M. Bernstein, V. deSilva, J.C. Langford, J.B. Tenenbaum, Graph approximations to geodesics on embedded manifolds, December 2000, Preprint may be downloaded at <http://isomap.stanford.edu/BdSLT.pdf>.

[24] J.A. Costa, A.O. Hero, Entropic graphs for manifold learning, in: Proceedings of the IEEE Asilomar Conference on Signals, Systems, and Computers, Pacific Groove, CA, November, 2003.

[25] J.A. Costa, A.O. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, IEEE Trans. Signal Process. 52 (2004) 2210–2221.

[26] J. Breadwood, J.H. Halton, J.M. Hamersley, The shortest path through many points, Proc. Cambridge Philos. Soc. 55 (1959) 299–327.

[27] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, The MIT Press, 2001.

[28] J.A. Costa, A.O. Hero, Manifold Learning with Geodesic Minimal Spanning Trees, 2003, arXiv:cs/0307038v1.

[29] W. Hardle, L. Simar, Applied Multivariate Statistical Analysis, 2004, available online at <http://www.quantlet.com/mdstat/scripts/mva/htmlbook/>.

[30] C.-G. Li, J. Guo, G. Chen, X.-F. Nie, Z. Yamg, A version of Isomap with explicit mapping, in: Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, August 2006.

[31] C.L.Y. Yeong, S. Torquato, Reconstructing random media II. Three-dimensional media from two-dimensional cuts, Phys. Rev. E 58 (1998) 224–233.

[32] A.P. Roberts, M.A. Knackstedt, Structure property correlations in model composite materials, Phys. Rev. E 54 (1996) 2313–2328.

[33] L. Arleth, S. Marc̆elja, T. Zemb, Gaussian random fields with two level-cuts– Model for asymmetric microemulsions with nonzero spontaneous curvature, J. Chem. Phys. 115 (8) (2001) 3923–3936.

[34] P.S. Koutsourelakis, G. Deodatis, Simulation of multi-dimensional binary random fields with application to modeling of two phase random media, J. Eng. Mech. 132 (2006) 619–631.

[35] Intrinsic Dimension and Entropy Estimation in Manifold Learning, Matlab codes available at: <http://www.eecs.umich.edu/hero/IntrinsicDim/>.

[36] The Qhull source code available at <http://www.qhull.org/>.

[37] S. Umekawa, R. Kotfila, O.D. Sherby, Elastic properties of a tungsten–silver composite above and below the melting point of silver, J. Mech. Phys. Solids 13 (1965) 229–230.

[38] J. Aldazabal, A. Martin-Meizoso, J.M. Martinez-Esnaola, Simulation of liquid phase sintering using the Monte-Carlo method, Mater. Sci. Eng. A 365 (2004) 151–155.

[39] P.J. Bickel, K.A. Doksum, Mathematical Statistics – Basic Ideas and Selected Topics, Prentice Hall, 2001.

[40] H. Hotelling, Tubes and Spheres in n-Spaces and a Class of Statistical Problems, Am. J. Math. 61 (1939) 440–460.