HARMONIC ANALYTIC GEOMETRY ON SUBSETS IN HIGH DIMENSIONS, "EMPIRICAL MODELS".

RONALD R. COIFMAN

Abstract. We describe a recent evolution of Harmonic Analysis to generate analytic tools for the joint organization of the geometry of subsets of $\mathbb{R}^n$ and the analysis of functions and operators on the subsets. In this analysis we establish a duality between the geometry of functions and the geometry of the space. The methods are used to automate various analytic organizations, as well as to enable informative data analysis. These tools extend to higher order tensors, to combine dynamic analysis of changing structures.

In particular we view these tools as necessary to enable automated empirical modeling, in which the goal is to model dynamics in nature, *ab initio*, through observations alone. We will illustrate recent developments in which physical models can be discovered and modelled directly from observations, in which the conventional Newtonian differential equations, are replaced by observed geometric data constraints. This work represents an extended global collaboration including, recently, A. Averbuch, A. Singer, Y. Kevrekidis, R. Talmon, M. Gavish, W.Leeb , J. Ankenman , G. Mishne and many more [36, 28, 35, 9, 10].

1. Introduction

We describe developments in Harmonic Analysis on subsets of $\mathbb{R}^n$, methodologies which integrate geometry, combinatorics, probability and Harmonic analysis, both linear and non-linear. We view the emerging structures,as providing natural settings to enable data driven Empirical models for observed dynamics.

Our initial focus is on methods applicable to discrete subsets viewed here as data samples on a continuous structure, a varifold, an infinite dimensional metric space etc. These samples could be generated through a discretization of a stochastic differential equation or through observations of natural or human driven processes .

The challenges of high dimensions, and the need to process massive amounts of seemingly unstructured clouds of points in $\mathbb{R}^n$ (sometimes data) forces us to introduce automated analytic methodologies to reveal the geometry of natural data, understand natural function spaces, or operators on such functions.

A basic insight is that the geometry of a subset is intimately connected to the geometry of functions on its points, (or sometimes operators on functions) not just the coordinate functions which are linear functions, or exponentials $e^{ix\cdot w}$, with random w or band limited functions and corresponding prolate functions or, more generally, the eigenvectors of natural operators such as graph Laplacians on the subset .

We exploit the fact that eigenvectors of the Laplace Beltrami operator on a manifold (or their discrete approximations), provide, both a high dimensional embedding of the manifold, and a coordinate system, opening the door to analysis.

2010 *Mathematics Subject Classification.* 42B35,42C20,42C40 (primary) and 62G08,62G86 (secondary).

Key words and phrases. Dual geometry, high dimensional approximation models, tensor Haar basis, matrix/tensor compression, Ab initio Empirical Models, Intrinsic variables.

Some of these ideas are well known classically, for Riemannian manifolds, where the Laplace operator, Dirac operators, pseudo differential operators, enable the passage from local properties, to global geometric invariants (as in Atiyah-Singer theories). In Harmonic Analysis, well known theorems, of G. David, S. Semmes and Peter Jones, show the equivalence of the existence of a bi-Lipschitz parameterization of a subset of $\mathbb{R}^n$ and the boundedness of the restriction of Calderon Zygmund operators, as well as some geometric multi-scale Carleson measure type deviation estimates.

The program described here , can be viewed as describing "unsupervised geometric machine learning" , and parallels some of the goals and methodologies of Deep Neural nets (such as variational auto encoders) , and Recurrent Neural nets, where a variety of algorithms strive to build generative models. for data clouds, see an overview by Yann LeCun, Yoshua Bengio , Geoffrey Hinton [**?**]. The duality (or triality) point of view described here can be seen as complementary, and necessary to provide better understanding of internal dependence structures.

One of our goals in this paper is to describe the interplay of such analytic tools with the geometry and combinatorics of data and information. We will provide a range of illustrations and application to the analysis of operators, as well as to the analysis of documents, questionnaires, and higher dimensional data bases viewed as tensors.

As will become apparent, the data geometry, or document organization point of view, can illuminate and inspire fundamental questions of geometry, such as duality and Heisenberg principles in Riemannian geometry, Carnot geometry etc, defining "dual metric" structures on the set of eigenvectors of the Laplace operator, (or sub-Laplace operator). Similarly the abstract organization of data bases, can inspire deep geometric organization of operators, their decompositions and analysis (following the Calderon Zygmund 'hard' Harmonic Analysis paradigm). In particular the tuning of the geometry to the nature of an operator , as well as the 3 tensor geometry that we discuss, could illuminate the variable geometric structures,which arise in solving nonlinear partial differential equations, defining "naturally evolving metric spaces ".

The following topics are interlaced in this presentation:

(a) Geometries of point clouds, and their graphs.
(b) From local to global, the role of eigenfunctions as integrators.
(c) Diffusion geometries in original coordinates, and organization in "intrinsic coordinates".
(d) Coupled dual geometries, Matrices of Data and Operators, duality between rows and columns, tensor product geometries.
(e) Harmonic Analysis, Haar systems, tensor Besov and bi-Hölder functions, Calderon Zygmund decompositions.
(f) Sparse grids and efficient processing of data.
(g) Applications; to Mathematics, organization of operators, the dual geometries of eigenvectors,
(h) Application to empirical modeling of natural dynamical systems through observations alone, defining **intrinsic latent variables**. Triality or, extensions of duality to 3 tensors .

2. Geometries of point clouds in $\mathbb{R}^n$

2.1. **Illustrative example.** Usually when considering a data set, each item or document is converted into a vector in high dimensional Euclidean space. For example a text document

could be converted to the vector, whose coordinates are, the list of occurrence frequencies of words in a lexicon. A particularly illuminating example carrying the complexity of issues we wish to address is a Corpus of text documents represented as a collection, or list of points in $\mathbb{R}^n$.

They have to be organized according to their mutual relevance. We can view this list either as a single cloud of documents or as a database matrix, in which each column is a document and each row, is the list of probabilities of occurrence of a given word in the various documents. We view the words as functions on the documents, and the documents as functions on the words. We will describe an ab initio geometric methodology to jointly assemble the language and the documents into a "smooth" coherent structure, in which documents are organized by context or topic, and vocabulary is organized conceptually by contextual occurrences. We will later describe an adapted tensor geometry and harmonic analysis of rows and columns that links concept and context by duality.

The naive approach to use the distance (or similarity) between two documents through their Euclidean distance or their inner product, is bound to fail, as already in moderate dimensions most points are far away, or essentially orthogonal. The distances in high dimensions are informative only when they are quite small, leading to the "connect the dots" diffusion geometry.

For this example if the distribution of the vocabulary in two documents are extremely close, we can infer that they deal with a similar topic. In this case we can link the two documents and weigh the link by a weight reflecting the probability that the documents are dealing with the same topic. This construction builds a graph of documents, as well as a corresponding random walk (or diffusion process) on the graph. The analogy with Riemannian geometry, in which we have a local metric, which defines a Laplace operator or a heat diffusion process is quite obvious, and will drive much of the initial discussion.

However; this approach to organize the documents as a cloud of points is by itself faulty as it does not account directly for the conceptual similarity , and dependencies between words , or between documents and their content. In order to untangle these relations we view the collection of documents as a matrix in row columns duality .

The columns are viewed as functions on the rows and the rows as functions on the columns. We organize the columns into a hierarchy of topics , (a partition tree of subsets.) These topical groups are then used to organize the vocabulary (rows) into a graph by their co-ocurrence in various document topics . This enables the organization of the vocabulary into a hierarchy of conceptual groups , which themselves can be reused to redefine the affinity between documents , (this process can be iterated as long as we gain in efficiency and precision of the representation) Coupling the construction of the two partition tree Hierarchies – on the columns and the rows – takes us away from the representation of the dataset as a point cloud in Euclidean space, towards representation of the dataset as a function on the product set $\{rows\} \times \{columns\}$. This natural document organization is quite abstract and will be quantified below, in particular it will become clear that the construction generalizes various methods of organization in Numerical Analysis ,and Harmonic Analysis, and extends naturally to higher order tensorial structures .

2.2. **Calculus.** The first fundamental point is that there is a natural reformulation of the basic concepts of Differential Calculus (or PDE) in terms of eigenvectors of appropriate linear transformations that will enable us to go from this local or infinitesimal description to an integrated global view of a data cloud. More generally it explains the ability to build data

driven empirical models, without the use of calculus. We start from a simple reformulation of the fundamental theorem of calculus, which is an observation of Amit Singer. A basic problem already posed by Cauchy is the following:

> **Sensor Localization Problem.** Assume we know some of the distances between a set of points in Euclidean space and assume these distances are known to determine the system, how does one map the points?

Think of the particular example where you know the distances of each city of a country to a few nearest neighbors: how would one manage to condense that information into a map of that country? There is a trivial answer: if enough local triangles with known lengths are given, then we can compute a local map which can be assembled bit by bit like a puzzle: this can be thought of as an analogue of integration. A more powerful method is obtained by writing each point p_i as the center of mass of its known neighbors, i.e.

$$p_i = \sum_{p_j \sim p_i} w_{ij} p_j \quad \text{where} \quad \sum_j w_{ij} = 1.$$

Observe that these equations are invariant under rigid motion and scaling. This tells us that the vector of $x-$coordinates of all points is an eigenvector corresponding to eigenvalue 1 of the matrix W. Similarly, the vector of y--coordinates and the vector all of whose coordinates are 1 are also in the same space. We thus see easily that the solution to the sensor localization problem is obtained by finding a basis of this eigenspace and expressing three points in this basis (using their mutual distances). Similarly, if we are given a set of points $(n, f(n)) \in \mathbb{R}^2$ and we know the differences $|f(n) - f(n-1)|$ and $|f(n) - f(n-2)|$, then we can determine f (which is a simple variant of the fundamental theorem of Calculus).

2.3. **Diffusion.** We now return to point clouds in $\mathbb{R}^n$. We can define a notion of local affinity, or similarity between elements of a set of points $\{p_1, \ldots, p_n\} \subset \mathbb{R}^n$ via the matrix

$$A_{ij} = \frac{\exp(-|p_i - p_j|^2/\varepsilon)}{\sum_{k=1}^n \exp(-|p_i - p_k|^2/\varepsilon)}.$$

This matrix can be interpreted as collecting the transition probabilities of a Markov process. $\varepsilon > 0$ is a parameter controlling the scale of influence (with small ε making a transition to very close neighbors likely while ε large allows for medium- and long-range transitions). Alternatively, it may be preferable to consider a notion of similarity given by

$$A_{ij} = \frac{\exp(-|p_i - p_j|^2/\varepsilon)}{\omega_i \omega_j}$$

where the weights ω_i, ω_j are chosen such that A is Markov matrix in both rows and columns (see N .Marshall [26]). Later we will correct it, or select a graph structure optimized for efficient analysis of functions on the data cloud, or to discover intrinsic Riemannian metrics. It is easy to verify that in the case that the points are uniformly distributed on a smooth submanifold of Euclidean space $\Delta = (I - A)/\varepsilon$ is an approximation (in a weak topology) of the Laplace-Beltrami operator on the manifold. Moreover, eigenvectors of A approximate the eigenvectors of the Laplace operator and powers of A correspond to diffusion on the manifold scaled by ε. See M.Belkin , P.Nyogi , S.Lafon [5, 25, 26].

Another more generic (non manifold) example consists of data generated through a stochastic Langevin equation, (a stochastic gradient descent differential equation) this kind of data can be also organized as above, with $\Delta = (I - A)/\epsilon$ approximating the corresponding Fokker Plank operator . [9, 10]

We can diagonalize A and use the eigenvectors of A to define powers of the diffusion

$$A^t(p_i, p_j) = \sum_{k=1}^{n} \lambda_k^t \phi_k(p_i)\phi_k(p_j).$$

This one-parameter family of diffusion defines an embedding Φ_t in $\mathbb{R}^n$ as follows:

$$\Phi_t(p_i) = \left\{\lambda_k^t \phi_k(p_i) : 1 \leq k \leq n\right\}.$$

We see that this embedding can be computed to any precision by restricting the eigenvector expansion to the first few eigenvectors (depending on the decay of the eigenvalues (λ_k)). This enables a lower-dimensional embedding of the data through what we call the diffusion map. The eigenvectors can also provide natural local coordinates on the manifold, see P.Jones ,M.Maggioni ,R Schul [**?**]. In the case of stochastic data the eigenfunctions approximate the eigenvectors of the Fokker Plank operator, they are supported on the main diffusion trails, and reveal latent variables. See B. Nadler , Y.Kevrekidis. [29]

The diffusion distance at time t is given, in the bi-stochastic symmetric case as

$$d_t^2(p, q) = A^t(p, p) + A^t(q, q) - 2A^i(p, q) = |\Phi_t(p) - \Phi_t(q)|^2$$

Where A^t represents the t power of A or the diffusion at time t .

3. Harmonic- Analysis of Databases-Matrices, and Tensors .

3.1. **Matrix organization in high dimensional Data analysis.** Our claim is that when dealing with a subset of $\mathbb{R}^n$ where n is large but the subset locally is of much lower dimension, exhibiting local correlations, for example if the subset is a subset of a Varifold, or the cloud is formed by stochastic orbits of dynamical systems, one wishes to understand and encapsulate the local constraints. Moreover linear functions such as the coordinates are not linear as functions on the set. In fact any collection of functions can provide us more coordinates. In particular, band limited functions such as $\exp(i\langle x, \xi\rangle)$ where $|\xi| < C$ are quite informative in revealing the geometry . More general plane waves as generated through deep neural nets can serve similar modeling functions.

As illustrated before on the example of a Corpus of text documents. It becomes productive to view the points as a matrix of data, or a discretized version of a kernel, both rows and columns could correspond to real-world variables or entities of enduring interest. The values of n (dimension) and p (number of points) are often of comparable magnitude, may both be large, and in an asymptotic analysis,may both be allowed to grow to infinity. The correlation or codependence structure of *both* rows and columns is of interest, this has been a main point of analysis, when viewing the data as a matrix of an operator, such as a Green operator ,or an eigendecomposition transform, discussed below.

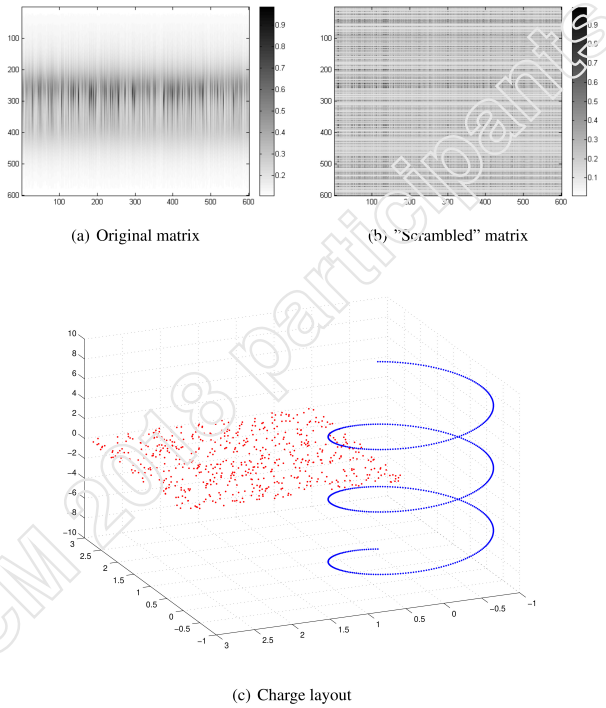
3.2. **Matrix organization in numerical analysis.** A bottleneck in many numerical analysis tasks involves the need to store very large matrices, apply them to vectors and compute functions of the operators they represent. For example the Fast Fourier Transform and the Fast Multipole Methods are explicitly based on exploiting the known geometrical organization of the row set and the column set of the transformation. A corresponding paradigm

in Harmonic Analysis is the organization of an operator as in Calderon-Zygmund theory, (Here we derive automatically the C-Z organization directly from the kernel of the operator or the data matrix).

Consider V. Rokhlin's Fast Multipole Method algorithm [20], which organizes a matrix

$$M_{i,j} = \|x_i - y_j\|^{-1}$$

of electrostatic or gravitational interactions between a known set of sources $\{x_i\} \subset \mathbb{R}^3$ and a known set of receivers $\{y_j\} \subset \mathbb{R}^3$, by exploiting the known geometry of the source set (the column set, say) and the receiver set (the row set). A similar approach yields fast wavelet transforms of linear operators [3]. There, too, the known organization of matrix rows and columns leads to efficient algorithms for storing, applying and computing functions of certain linear operators. Suppose however that we wish to apply an analog of the Fast Multipole



(a) Original matrix (b) "Scrambled" matrix

(c) Charge layout

FIGURE 1. Geometric unravelling of a scrambled matrix (random labels) of potential interactions (b). Charges are on the spiral, receivers in the plane. Our matrix organization reveals the two geometries and their internal structures.

method to a given matrix of electrostatic interactions, $M_{i,j} = \|x_i - y_j\|^{-1}$, where the sets $\{x_i\}$ and $\{y_j\}$ themselves are unknown. The order in which rows and columns are given is meaningless, yet the locations $\{x_i\}$ and $\{y_j\}$ remain encoded in M. (Figure 1) In this context, the theory developed below leads to data agnostic organizational methods which are able, even for some oscillatory potentials, such as $M(x_i, y_j) = M_{i,j} = cos(100\|x_i - y_j\|)\|x_i - y_j\|^{-1}$ to recover the underlying coupled source and receiver "geometric optics", (in the case of points sampled on a surface or a curve), and furthermore leads to orthonormal bases enabling the implementation of a corresponding fast transform, analogous to [3], the ℓ_1 norm of matrix coefficients in this basis measures the compression rate it is able to achieve: It can be easily proved that this norm controls the mixed smoothness of the matrix.

3.3. **Setup.** Let M be a matrix, we denote its column set by X and its row set by Y. M can be viewed as a function on the product space, namely

$$M : X \times Y \to \mathbb{R}$$

Our first step in processing M, regardless of the particular problem, is to simultaneously organize X and Y, or in other words, to construct a product geometry on $X \times Y$ in which proximity (in some appropriate metrics) implies predictability of matrix entries. Equivalently, we would like the function M to be "smooth" with respect to the tensor product geometry in its domain. As we will see, smoothness, compressibility, having low entropy, are all interlinked in this organization. We start by redefining the classical notions of smoothness in the context of tree metrics .

3.4. **Brief description of Haar Bases.** A hierarchical partition tree on a dataset X is an ordered collection of (finite) disjoint covers of the set where each cover is a refinement of the preceding cover, Such a structure allows harmonic analysis of real-valued functions on X, as it induces special orthonormal *Haar bases* [17]. The elements of the cover will be denoted as folders or nodes of the tree connecting a folder to the coarser folder containing it.

A Haar basis is obtained from a partition tree as follows. Suppose that a node (subset or folder) in the tree has n children, that is, that the set described by the node decomposes into n subsets in the next, more refined, level. Then this node contributes $n - 1$ functions to the basis. These functions are all supported on the set described by the node, are piecewise constant on its n subsets, all mutually orthogonal, and are orthogonal to the constant function on the set.

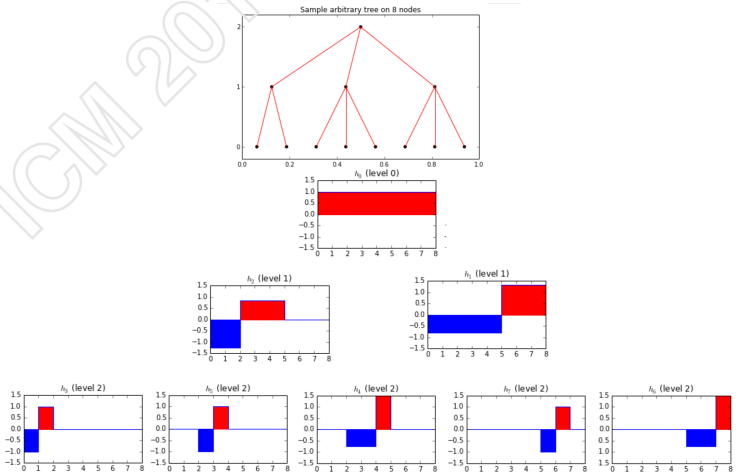


FIGURE 2. A partition tree on the unit interval starting with a partition into three subintervals,one of which is further devided in two and the other two into three subintervals . The corresponding Haar functions are orthogonal, measuring the variation of averages among neighbors, with the color corresponding to their sign .

Observe that just like the classical Haar functions, coefficients of an expansion in a Haar basis measure variability of the conditional expectations of the function in sub nodes of a given node.

Tasks such as compression of functions on the data set, as well as subsampling, denoising and "learning" such functions, can be performed in Haar coefficient space using methods familiar from Euclidean harmonic analysis and signal processing [17].

Some results for the classical Haar basis on $[0,1]$ extend to generalized Haar bases. Recall that the classical Haar functions based on the dyadic tree are given by

$$h_I(x) = \left(|I|^{-\frac{1}{2}}\right)(\chi_- - \chi_+)\ ,$$

where χ_- is the indicator of the left half of I and χ_+ is the indicator of the right half of I.

The classical Haar basis on $[0,1]$ is induced by the partition tree of dyadic subintervals of $[0,1]$. This tree defines a natural dyadic distance $d(x,y)$ on $[0,1]$, defined as the length of the smallest dyadic interval containing both points. Hölder classes in the metric d are characterized by the Haar coefficients $a_I = \int f(x)h_I(x)dx$:

$$|a_I| < c|I|^{\frac{1}{2}+\beta} \ \Leftrightarrow\ |f(x) - f(x')| < c \cdot d(x,x')^{\beta}\ .$$

A natural partition tree on a set of points in $\mathbb{R}^d$, is the vector quantization tree i,e a hierarchical organization into disjoint covers by subsets (folders) of approximate diameter $(1/2)^n$. We define a hierarchical tree distance between two points as being the diameter of the smallest folder containing both points .

The characterization of smoothness property holds for any Haar basis when d is the tree metric induced by the partition tree, and $|I| = \frac{\#I}{\#X}$ is the normalized size of the subset (folder) I. (We remark also that for $\beta < 1$ the usual Holder condition is equivalent to dyadic Holder for all shifted dyadic trees.)

We note that there are multiple ways to build partition trees (and corresponding smoothness spaces). The different construction methods can be divided into two classes: bottom-up construction and top-down construction. Broadly, a bottom-up construction begins with the definition of the lower levels, initially by grouping the leaves/samples, e.g., using k-means in the diffusion embedding . Then, these groups are further grouped in an iterative procedure to create the next levels, ending at the root, in which all the samples are placed under a single folder.

A top-down construction is typically implemented by an iterative clustering method, initially applied to the entire set of samples, then refined over the course of the iterations, starting with the root of the tree and ending at the leaves.

A simple blend is achieved by using the first few diffusion eigenvectors, to split the data into two groups using the first non-trivial eigenvector (approximate max-cut) then repeating on each subgroup using its own first non-trivial eigenvector, since the eigenvector computation is a bottom up iteration, this results in a binary tree, which is often well tuned to the internal data structures.

3.5. **Matrix organization through coupled partition trees.** To illustrate the basic concept underlying the simultaneous row-column organization, consider the case of a vector (namely, a matrix with one row). In this case, the only reasonable organization would be to bin the entries in decreasing order , (or in binary quantization tree). This decreasing function is obviously smooth outside a small exceptional set (being of bounded variation). Our approach extends this simple construction – which can be viewed as just a one-dimensional quantization tree – to a coupled quantization tree.

We now digress briefly to indicate a simple mathematical framework for joint row/column organization and analysis of a matrix.(quantization bi-trees) which renders an arbitrary matrix into a bi-Holder matrix, (extending the one row example). We start a hierarchical vector quantization tree on the set of columns, X,(as vectors in Euclidean space) with tree metric ρ_X .

The tree metric ρ_X is such that the rows are (tautologically) Lipschitz smooth in the tree metric, as functions of the columns. This implies that the Haar coefficients of the rows, relative to the tree on the columns, scale with the diameter. A similar hierarchical organization on the rescaled Haar coefficients of Y (the rows) as a function of the variable x, induces a similar tree metric ρ_Y on the rows with a similar smoothness property of the columns.

As we will see this implies that the full matrix is a bi-Lipschitz function i.e. it satisfies a *Mixed Lipschitz Hölder* condition

$$|M(x_0, y_0) - M(x_0, y_1) - M(x_1, y_0) + M(x_1, y_1)| \leq C \cdot \rho_X(x_0, x_1)^\alpha \cdot \rho_Y(y_0, y_1)^\alpha$$

This condition enables the estimation of one value in terms of three neighbors with a higher order error in the two metrics. (For the square in two dimensions, this would be a relaxation of the bounded mixed derivative condition $\left|\frac{\partial^2 M}{\partial x \partial y}\right| \leq C$, which has been studied in the context of approximation in high dimensions [33] [18] [7] [?]

This simple organization is not very effective in high dimension, as most points are far away from each other, leading us to explore various constructions of more intrinsic data driven metrics and trees, such as the diffusion metrics described above, or the corresponding "earth mover" metrics . One of our goals is to achieve higher efficiency in representing the matrix, and develop a Harmonic Analysis, or signal processing of functions on $X \times Y$. In particular we will see this as an automatic process to build a multiscale Harmonic Analysis of an operator, or Matrix.

We describe briefly elementary analysis of the Mixed Hölder function classes (as well as their Besov space duals) on an abstract product set equipped with a partition tree pair. A useful tool is an orthogonal transform for the space of matrices (functions on $X \times Y$), naturally induced by the pair of partition trees (or the tensor product of the corresponding martingale difference transforms). Specifically, we take the tensor product of the *Haar bases* induced on X and on Y by their respective partition trees,

The Mixed-Hölder arises naturally in several different ways. First, as seen above for vector quantization trees, any matrix can be given Mixed Hölder structure. Second,it can be shown that any bounded matrix decomposes into a sum of a Mixed Hölder part and a part with small support (as for the one row example). (of course the constants are pretty bad for random data in high dimensions)

3.6. **Coupled partition trees, optimized duality.** Our goal is to build coupled partition trees to optimize compression of the original function (Matrix) expanded in the tensor Haar basis, say by minimizing an l^1 norm of the tensor Haar coefficients. Such a task requires the discovery of both systems of Haar functions, it is clear that a unique minimizer does not exist in general . Moreover, the appropriate structure is a function of context and precision, as will become clear for various examples, in mathematics and beyond.

We now consider a matrix M and assume two partition trees – one on the column set of M and one on the row set of M – have already been constructed. Each tree induces a Haar

basis and a tree metric as above. The tensor product of the Haar bases is an orthonormal basis for the space of matrices of the same dimensions as M. We review some analysis of M in this basis.

Denote by $|R| = |I \times J|$ a "rectangle" of entries of M, where I is a folder in the column tree and J is a folder in the row tree. Denote by $|R| = |I||J|$ the volume of the "rectangle" R. Indexing Haar functions by their support folders, we write $h_I(x)$ for a Haar function on the rows. This allows us to index basis functions in the tensor product basis by rectangles and write $h_R(x,y) = h_I(x)h_J(y)$.

Analysis and synthesis of the matrix M is in the tensor orthonormal Haar basis is simply

$$\begin{aligned} a_R &= \int M(x,y)h_R(x,y)dxdy \\ M(x,y) &= \sum_R a_R h_R(x,y)\,. \end{aligned}$$

The characterization of Hölder functions mentioned above extends to mixed-Hölder matrices [12, 18]:

$$\Big|a_R\Big| < c\Big|R\Big|^{1/2+\beta} \Leftrightarrow \Big|M(x,y) - M(x',y) - M(x,y') + M(x',y')\Big| \leq c\rho_X(x,x')^\beta \rho_Y(y,y')^\beta$$

where ρ_X and ρ_Y are the tree metrics induced by the partition trees on the rows and columns, respectively. Observe that his condition implies the conventional two dimensional Holder condition

$$\Big|M(x,y) - M(x',y')\Big| \leq \rho_X(x,x')^\beta + \rho_Y(y,y')^\beta$$

Simplicity or sparsity of an expansion is quantified by an "entropy" such as

$$e_\alpha(M) = \Big(\sum \big|a_R\big|^\alpha\Big)^{1/\alpha}$$

for some $\alpha < 2$. We comment that this norm is just a tensor Besov norm that is easily seen to generalize Earth mover distances when scaled correctly, adding flexibility to our construction below. This norm can be generalized to the following family of Besov norms

$$e_{\alpha,\beta}(M) = \Big(\sum \big|R|^\beta |a_R\big|^\alpha\Big)^{1/\alpha}$$

for some α, β. Useful relations between this "entropy", efficiency of the representation in tensor Haar basis and the mixed-Hölder condition, is given by the following two propositions valid for "balanced trees" [12, 18].

Proposition. *Assume $e_\alpha(M) = \big(\sum \big|a_R\big|^\alpha\big) \leq 1$. Then the number of coefficients needed to approximate the expansion to precision $\varepsilon^{1-\alpha/2}$ does not exceed $\varepsilon^{-\alpha}\log(\varepsilon^{-1})$ and we need only consider large coefficients corresponding to Haar functions whose support is large. Specifically, we have*

$$\int \Big|M - \sum_{|R|>\varepsilon,\, |a_R|>\varepsilon} a_R h_R(x)\Big|^\alpha dx < \varepsilon^{1-\alpha/2}$$

The next proposition shows that $e_\alpha(M)$ estimates the rate at which M can be approximated by Hölder functions outside sets of small measure.

Proposition. *Let f be such that $e_\alpha \leq 1$. Then there is a decreasing sequence of sets E_ℓ such that $|E_\ell| \leq 2^{-\ell}$ and decompositions of Calderon Zygmund type $f = g_\ell + b_\ell$. Here, b_ℓ is supported on E_ℓ and g_ℓ is bi-Hölder $\beta = 1/\alpha - 1/2$ with constant $2^{(\ell+1)/\alpha}$. Equivalently, g_ℓ has Haar coefficients satisfying $|a_R| \leq 2^{(\ell+1)/\alpha}|R|^{1/\alpha}$.*

Thus we can decompose any matrix into a "good", or mixed-Hölder part, and a "bad" part with small support.

Mixed-Hölder matrices indeed deserve to be called "good" matrices, as they can be substantially sub-sampled. To see this, note that the number of samples needed to recover the functions to a given precision is of the order of the number of tensor Haar coefficients needed for that precision. For balanced partition trees, this is approximately the number of bi-folders R, whose area exceeds the precision ε. This number is of the order of $(1/\varepsilon)^\alpha \log(1/\varepsilon)$.

These remarks imply that the entropy condition quantifies the compatibility between the pair of partition trees (on the rows and on the columns) and the matrix on which they are constructed. In other words, to construct useful trees we should seek to minimize the entropy in the induced tensor Haar basis.

For a given matrix M, finding a partition tree pair, which is a global minimum of the entropy, is computationally intractable and not sensible, as the matrix could be the superposition of different structures, corresponding to conflicting organizations. At best we should attempt to peel off organized structured layers.

The iterative procedures for building tree pairs described previously for the text documents example, perform well in practice. These procedures alternate between construction of partition trees on rows and on columns . Each tree defines a Besov norms its dual (i,e functions on its nodes) which is used to reorganize the dual into a tree leading to a new tree on the original nodes .

A nice example in mathematics, is to view the matrix of eigenvectors of the Laplace operator on a compact Riemannian manifold as a data base, in which the columns are the points on the manifold and the rows are the values at the point of different eigenvectors .

We can organize the Riemannian geometry in a multiscale geometry, The construction described before builds Besov norms on functions on the manifold, which can be used to measure a distance between eigenvectors (the L^2 distance is useless being $= \sqrt{2}$),thereby inducing a distance on the "Fourier dual" of Laplace eigenvectors. Of course different geometries on the space, will give rise to different dual geometries.

To conclude, we see emerging an analysis or "Signal processing toolbobox" for digital data as a first step to analyse the geometry of large data sets in high-dimensional space and analyse functions defined on such data sets. The ideas described above are strongly related to nonlinear principal component analysis, kernel methods, spectral graph embedding, and many more, at the intersection of several branches of mathematics, computer science and engineering. They are documented in literally hundreds of papers in various communities. For a basic introduction to many of these ideas and more,as they relate to diffusion geometries. We refer the interested reader to the July 2006 special issue of Applied and Computational Harmonic Analysis, and references therein [13].

4. Empirical dynamics, or higher dimensional tensors.

The purpose of this section is to show that a corresponding generalization of the analysis to 3-tensors **by triality** enables the organization of dynamical systems as well as purely empirical modeling of natural dynamics . A particular implementation of these algorithms will allow a systematic realization of all these steps – inferring "natural geometries" from data, using just the data organization counterpart of the above discussion: similarity between nearby observations/measurements.

Recovering the underlying structure of nonlinear dynamical systems from data ("system identification") has attracted significant research efforts over many years, and several ingenious techniques have been proposed to address different aspects of this problem. These include methods to find nonlinear differential equations to discover governing equations from time-series or video sequences equation-free modeling approaches, and methods for empirical dynamic modeling . We present methods extending our prior discussion building upon the work of Y. Kevrekides and I.Mezic, [22, 27]. Our goal is the organization of observations originating from many different types of dynamical systems into a joint coherent structure, which should parametrize the various dynamical regimes and build empirical models of the whole observation space. Since we are comparing dynamical observations, which are distorted versions of each other, we are forced to discover variants of the *EMD*[1], which go beyond classical transforms in enabling data-driven comparisons between trajectories and their dynamics, see J.Ankenman, W. Leeb [4, ?]

4.1. **Problem Formulation and Toy Examples.** In our data agnostic setting, we think of time-dependent measurements which are the result of a number of experiments that we will call *trials*; during each trial, the (unknown "state") parameter values remain constant.

In this *black box* setting, the dynamical system is unknown, nonlinear and *autonomous*, and is given by

$$\frac{d\boldsymbol{x}}{dt} = f(\boldsymbol{x};\boldsymbol{p}) \tag{1}$$

$$\boldsymbol{y} = h(\boldsymbol{x}) \tag{2}$$

We do not have access to its state $\boldsymbol{x}$ nor to its parameter values $\boldsymbol{p}$; we also do not know the evolution law f, nor the measurement function h. We only have measurements (observations) $\boldsymbol{y}$ labelled by time t.

The black box is endowed with "knobs" that, in an unknown way, change the values of the parameters $\boldsymbol{p}$; so in every trial, for a new, but unknown, set of parameter values $\boldsymbol{p}$, we can observe $\boldsymbol{y}$ coming out of the box without knowing $\boldsymbol{x}$ or f or even h. We want to characterize the system dynamics by systematically organizing our observations (collected over several trials) of its outputs.

More specifically, we want to (a) organize the observations by finding a set of *state variables* and a set of *system parameters* that jointly preserve the essential features of the dynamics; and then (b) find the corresponding *intrinsic* geometry of this combined variable-parameter space, thus building a sort of normal form for the problem. Small changes in this *jointly intrinsic space* will correspond to small changes in dynamic behavior (i.e. to robustness). Having discovered a useful "joint geometry" we can then inspect its individual constituents. Inspecting, for example, the geometry of the discovered parameter space, will help identify regimes of different qualitative behavior. This might be different dynamic behavior, like hysteresis, or oscillations, separated by bifurcations; alternatively, we might observe transitions

between different sizes of the minimal realizations: regimes where the number of minimal variables/parameters necessary in the realization changes.

We can also inspect the identified state variable geometry, which will help us organize the temporal measurements in coherent phase portraits. In addition, if there exist regimes where the system becomes *singularly perturbed*, we expect we will be able to realize that the requisite minimal phase portrait dimension changes (reduces), and that the reduction in the number of state variables is linked with the reduction in the number of intrinsic parameters.

As an illustrative example, consider the following dynamical system, arising in the unfolding of the Bogdanov-Takens bifurcation[21]:

$$\begin{aligned} \frac{dx_1}{dt} &= x_2 \\ \frac{dx_2}{dt} &= \beta_1 + \beta_2 x_1 + x_1^2 - x_1 x_2. \end{aligned} \tag{3}$$

This set of differential equations defines a dynamical system with two parameters $\boldsymbol{p} = (\beta_1, \beta_2)$, two state variables $\boldsymbol{x} = (x_1, x_2)$, and two observables $\boldsymbol{y} = (y_1, y_2)$; at first we choose the observable to be the state variables themselves, i.e., $(y_1, y_2) = (x_1, x_2)$ with $h(\boldsymbol{x})$ being the identity function. It is known that the parameter space of this system (β_1, β_2) can be divided into 4 different regimes separated by one-parameter bifurcation curves [21]. Figure 1. shows this "ground truth" bifurcation diagram for our simulated 2D grid of parameter values. Each point $\boldsymbol{p} = (\beta_1, \beta_2)$ on the grid is colored according to its respective dynamical regime.

Our goal in this case would be to discover an accurate bifurcation map of the system in a data-driven manner purely from observations. These observations consist of several samples, where each sample is a single trajectory $\boldsymbol{y}(t)$ of the system initialized with unknown (possibly different) parameter values and initial values. In addition, we would like to deduce from these large number of realizations of trajectories $\boldsymbol{y}(t)$ arbitrarily and differently initialized that the system depends on only two parameters and can be realized with only two state variables; and to reconstruct the bifurcation diagram with its phase portraits.

4.2. **Learning dynamic structures and latent variables from observations.** Consider data arising from an autonomous dynamical system; we view the observations as entries in a three-dimensional tensor. One axis of the tensor corresponds to variations in the problem parameters, one to variations in the problem variables, and the third axis corresponds to time evolution along trajectories.

Formally, let $\mathcal{P}$ denote an ensemble of N_p sets of the d_p system parameters. Let $\mathcal{V}$ be a ensemble of N_v sets of initial condition values of the d_v state variables. For each $\boldsymbol{p} \in \mathcal{P}$ and $\boldsymbol{v} \in \mathcal{V}$, we observe a trajectory $Y(\boldsymbol{v}, \boldsymbol{p}, t)$ of length N_t in $\mathbb{R}^{d_v}$ of the system variables, where $t = 1, \ldots, N_t$ denotes the time sample. In summary, $\boldsymbol{p}$ is a label of the particular differential equations of the dynamical system, $\boldsymbol{v}$ is a label of the observations trajectory, and t is the time label.

Let $\mathbf{Y}$ denote the entire 3D tensor of observations of dimension $N_p \times N_v \times N_t$ consisting of all the data at hand. With respect to the black box setting described in the introduction, we emphasize that the identity of the parameters and variables is hidden; we only have trajectories of observations corresponding to various trials with possibly different hidden parameter values and with different hidden initial input coordinates.

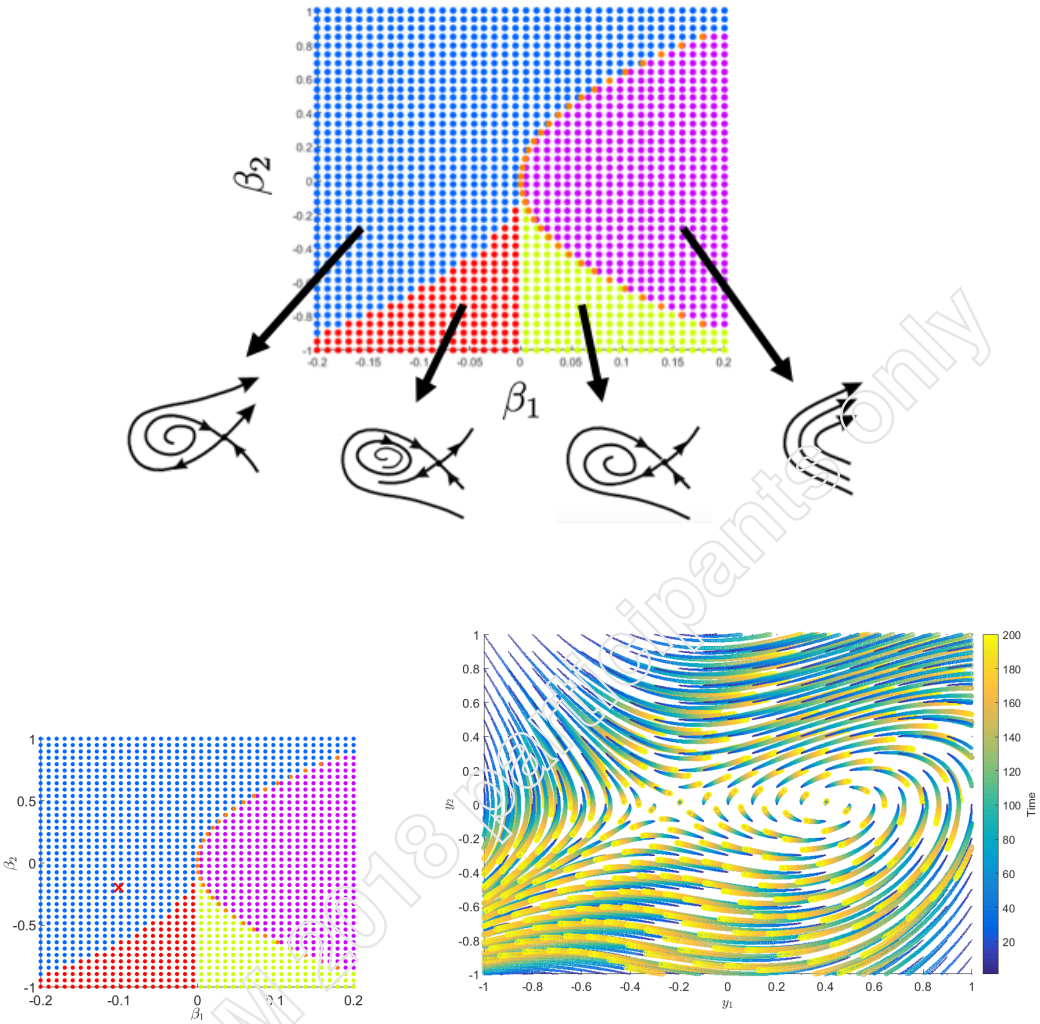


FIGURE 3. (up) The Bogdanov-Takens bifurcation maps with insets illustrating the typical phase-portraits in each dynamical regime. (left) The Bogdanov-Takens bifurcation map. (right) An example of the phase-portrait of the simulated trajectories of the Bogdanov-Takens system corresponding to the parameter set $(\beta_1, \beta_2) = (-0.1, -0.2)$, marked by red 'x' on the left.

To make the problem definition concrete we describe the setting of a specific example. Recall the Bogdanov-Takens dynamical system of two variables and two parameters, introduced in (3). We generate a set $\mathcal{P}$ of $N_p = 400$ different parameter values $\boldsymbol{p} = (\beta_1, \beta_2)$ from a regular fixed 2D grid, where $\beta_1 \in [-0.2, 0.2]$ and $\beta_2 \in [-1, 1]$, and additional 10 parameter values located exactly on the bifurcation. Similarly, we generate a set $\mathcal{V}$ of $N_v = 441$ different initial conditions $\boldsymbol{v} = (y_1(0), y_2(0))$ from a fixed 2D grid in $[-1, 1]^2$. For each $\boldsymbol{p} \in \mathcal{P}$ and $\boldsymbol{v} \in \mathcal{V}$, we observe a trajectory of the system for $N_t = 200$ time steps, where the interval between two adjacent time samples is $\Delta t = 0.004\,[\text{sec}]$ and collect all the trajectories into a single 3D tensor $\mathbf{Y}$. In this example, $N_p = 410, N_v = 441$ and $N_t = 200$ so overall we have $\mathbf{Y} \in \mathbb{R}^{410\times441\times200}$. For illustration purposes, Figure 3. (right) depicts $\beta_1 = -0.1$ and $\beta_2 = -0.2$ (marked by a red '$\times$' in Figure 3 (left)).

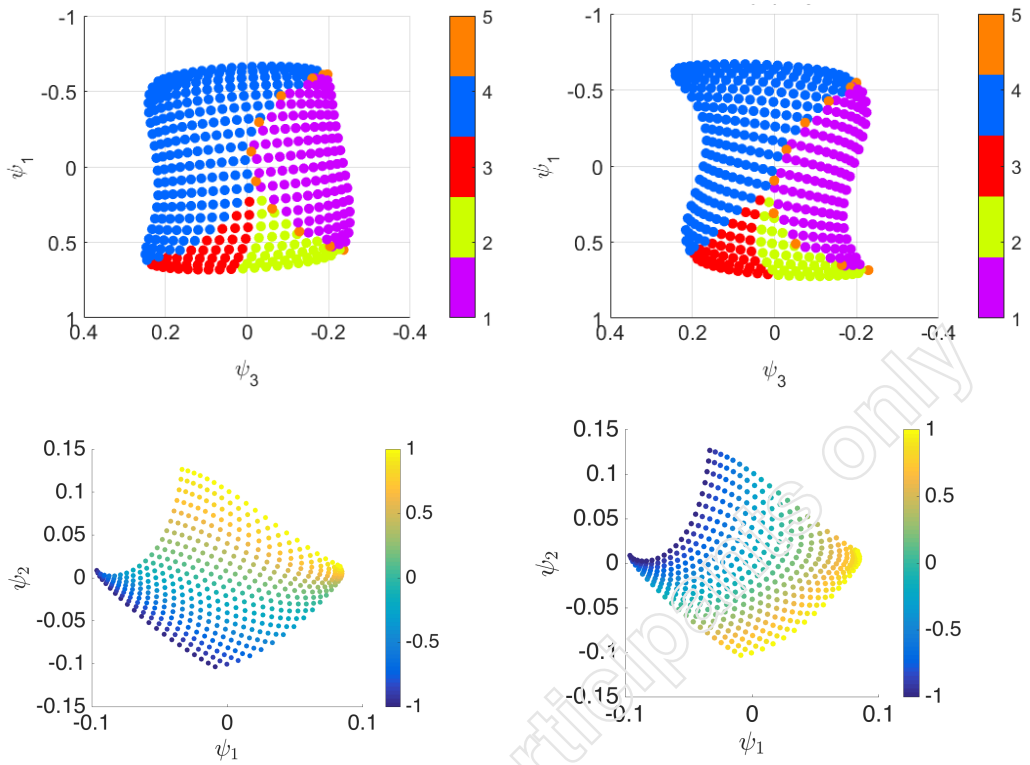


FIGURE 4. (left) Data-driven embedding of the parameters axis of the observations collected from the Bogdanov-Takens system (colored according to the true bifurcation map). Embeddings built from (a) state variable observations; and (b) observations through a nonlinear invertible function. (right) Data-driven embedding of the state variables axis (c) colored by the initial conditions of x_1, and (d) by the initial conditions of x_2.

We note that the trajectories (as illustrated in Figure 3) are long enough to partially overlap in phase space. Such an overlap induces the coupling between the time and variables axes, which is captured and exploited by our analysis. We wish to find a reliable representation of the hidden parameters, of the hidden variables, and of the time axis.

Define $\boldsymbol{y}_p = \{Y(\boldsymbol{v}, \boldsymbol{p}, t) | \forall \boldsymbol{v}, \forall t\}$ for each of the N_p vectors of hidden parameter values $\boldsymbol{p}$ in $\mathcal{P}$, namely, a data sample consisting of all the trajectories from a single trial. For simplicity of notation, we will use subscripts to denote both the appropriate axis and a specific set of entries values on the axis. We refer to $\{\boldsymbol{y}_p\}, \boldsymbol{p} \in \mathcal{P}$ as the data samples from the parameters axis viewpoint. In the Bogdanov-Takens example, Figure 3 depicts $\boldsymbol{y}_p$ for $\boldsymbol{p} = (\beta_1, \beta_2) = (-0.1, -0.2)$.

Similarly, let $\boldsymbol{y}_v$ and $\boldsymbol{y}_t$ be the samples from the viewpoints of the variables axis and the time axis, respectively, which are defined by

$$\begin{aligned} \boldsymbol{y}_v &= \{Y(\boldsymbol{v}, \boldsymbol{p}, t) | \forall \boldsymbol{p}, \forall t\}, \quad \boldsymbol{v} \in \mathcal{V} \\ \boldsymbol{y}_t &= \{Y(\boldsymbol{v}, \boldsymbol{p}, t) | \forall \boldsymbol{v}, \forall \boldsymbol{p}\}, \quad t = 1, \ldots, N_t. \end{aligned}$$

One way to accomplish our goal is to process the data *three successive times*, each time from a different viewpoint.

Here, we use a data-driven parametrization approach based on a kernel. From the trials (effectively, parameters) axis point of view, a typical kernel is defined by

$$k(\boldsymbol{y}_{p_1}, \boldsymbol{y}_{p_2}) = e^{-\|\boldsymbol{y}_{p_1} - \boldsymbol{y}_{p_2}\|^2/\epsilon}, \forall \boldsymbol{p}_1, \boldsymbol{p}_2 \in \mathcal{P} \tag{4}$$

based on distances between any pair of samples, where the Gaussian function induces a sense of locality relative to the kernel scale ϵ. To aggregate the pairwise affinities comprising the kernel into a global parametrization, traditionally, the eigenvalue decomposition (EVD) is applied to the kernel, and the eigenvalues and eigenvectors are used to construct the desired parametrization. The specific initial parametrization method that is used here is diffusion maps [**?**],

From three *separate* diffusion maps applications to the sets $\{\boldsymbol{y}_p\}$, $\{\boldsymbol{y}_v\}$, and $\{\boldsymbol{y}_t\}$, we can obtain three mappings as in (**??**), denoting the associated eigenvectors by $\{\boldsymbol{\psi}_\ell^{\mathcal{P}}\}$, $\{\boldsymbol{\psi}_\ell^{\mathcal{V}}\}$, and $\{\boldsymbol{\psi}_\ell^{\mathcal{T}}\}$, respectively.

However, such mappings do not take into account the strong correlations and co-dependencies between the parameter values and the dynamics of the variables which arise in typical dynamical systems. For example, in the Bogdanov-Takens system, the dynamical regime changes significantly depending on the values of the parameters.

To incorporate such co-dependencies, we extend the mutual metric learning algorithm described for matrices in order to build flexible EMD like distances on each axes of this 3 tensor . In the introduction of the affinity matrix in (4), we deliberately did not specify the norm used to compare between two samples. Common practice is to use the Euclidean norm. However, as pointed out by Lafon [25] anisotropic diffusion maps can be computed by using different norms. This issue has been extensively studied recently, and several norms and metrics have been developed for this purpose by Y. Kevrekidis , and Gal Mishne[28] [16].

Here, following Gal Mishne [28], we describe the 3- tensor extension of the preceding metric learning construction for matrix organization where the different axis geometries evolve together.

4.3. **Tensor Metric Construction, or "informed metrics".**

Partition Trees, and 3 tensor Besov spaces. The construction described previously for matrices, is easily extended to higher dimensional tensors, the only constraint is to define appropriate metrics on each coordinate axis, in the three tensor case a coordinate label defines a sub matrix, we match two labels 1 and 2 through the tensor Besov distance between them,

$$d_{\alpha,\beta}(M_1 - M_2) = \left(\sum |R|^\beta |a_R^1 - a_R^2|^\alpha \right)^{1/\alpha}$$

for some α, β.) Observe that the Besov distance for $\beta > 0$, can be computed without using the Haar functions, simply by replacing the Haar coefficient on the submatrix R by the average on R . see W. Leeb and J. Ankenman[**?**, 4] We build a partition tree for each axis based on this tensor product metric. Observe that this is a flexible metric generalizing earth mover to the context of matrices, where rows and columns have different smoothness geometries , it is not a conventional transportation metric.

4.4. **Iterative Metric Construction.**

The construction of the partition tree described above relies on a learning a metric between the samples on the different axis coordinates (sample labels), in which the construction of the tree relies on an iteratively evolving " metric"

induced by partition trees on the coordinates of the samples. Namely, the construction of $\mathcal{T}_v$ relies on a metric between the samples $\boldsymbol{y}_v$ which are matrices in the p,t labels, i.e., it uses the 2 tensor Besov distance or EMD .and the construction of $\mathcal{T}_t$ relies on a metric between the samples v,p Given $\mathcal{T}_v$ and $\mathcal{T}_t$, the informed metric between the samples $\boldsymbol{y}_p$ is constructed, and then, used to build a new partition tree $\mathcal{T}_p$ of the samples $\boldsymbol{y}_p$. In the second substep within the iteration, $\mathcal{T}_p$ can be used to construct refined metrics between $\boldsymbol{y}_v$ and between $\boldsymbol{y}_t$.

Once the metric is constructed, it can be used to build a partition tree $\mathcal{T}_p$ on the samples $\boldsymbol{y}_p$.

The construction of the informed metric between the samples $\boldsymbol{y}_p$ described above is repeated in an analogous manner to build informed metrics between the samples $\boldsymbol{y}_v$ and between the samples $\boldsymbol{y}_t$. Proving convergence for this "iterative, self-consistent re-normalization" of the coordinates, is the subject of current research.

We note that the particular choice of the specific Besov norm is explained in detail in W. Leeb [**?**] yet other L^2 type norms can be used depending on the application at hand.

First, the recursive procedure described above repeats in iterative manner, where in each iteration, three informed metrics are constructed one by one, based on the metrics *from the preceding iteration*. As the iterations progress, the metrics are gradually refined, and the dependency on the initialization is reduced.

Our method is applied to the 3D tensor of trajectories $\mathbf{Y}$ collected from the Bogdanov-Takens system. As described above, $\mathbf{Y}$ consists of (short) trajectories of observations arising from the system initialized with various initial conditions and with various parameters. We emphasize that the knowledge of the different regimes and the bifurcation map *were not taken into account* in the analysis; only the time-dependent data $\mathbf{Y}$ were considered.

Figure **??** (a) depicts the scatter plot of the two dominant eigenvectors representing the parameters axis. It consists of N_p points (the length of the eigenvectors), where each point corresponds to a single sample $\boldsymbol{y}_p \in \mathbb{R}^{N_v \times N_t}$, which is associated with parameters values $\boldsymbol{p} = (\beta_1, \beta_2)$ on the 2D grid depicted in Figure 3. Moreover, each point in Figure 1 is colored by the same color-coding used in Figure 2. We observe that our method discovers an *empirical bifurcation mapping of the system*. Indeed, the obtained representation of the parameters through the eigenvectors establishes a new coordinate system with a geometry, built solely from observations, which reflects the organization of the parameters space according to the true underlying bifurcation map – the "visual homeomorphism" (stopping short of claiming visual isometry) is clear.

To illustrate the generality of our method, we now apply a *nonlinear (yet invertible)* observation function

$$\boldsymbol{z}(t) = h(\boldsymbol{x}(t))$$

with $h_k(\boldsymbol{x}(t)) = \sqrt{\boldsymbol{a}_k^T \boldsymbol{x}(t) + \alpha_k}$, $k = 1, 2$, where $\boldsymbol{a}_k$ is a random observation vector and α_k is a constant set to guarantee positivity. Figure 4 (b) depicts the scatter plot of the two dominant eigenvectors representing the parameters axis obtained from the new set of nonlinear observations. An equivalent organization is clearly achieved.

Figure 4 (c) depicts the scatter plot of the two dominant eigenvectors representing the state variable axis. The plot consists of N_v points (the length of the eigenvectors), where each point corresponds to a single sample $\boldsymbol{y}_v \in \mathbb{R}^{N_p \times N_t}$, which is associated with a particular set of initial condition values $\boldsymbol{v} = (y_1(0), y_2(0))$. The embedded points are colored in Figure 4 (c) by the initial conditions of the variable y_1, and in Figure 4 (d) by the initial conditions of the variable y_2. The color-coding implies that the recovered 2D space corresponds to the 2D

space of the true variables of the system. In other words, the high dimensional samples $\boldsymbol{y}_v$ are embedded in a 2D space, which recovers a 2D structure accurately representing the true directions of the hidden, minimal, two state variables of the system.

4.5. **Two Coupled Pendula.** To demonstrate the ability to extract true physical parameters we simulate a system of two simple coupled pendula with equal lengths L and equal masses m, connected by a spring with variable constant $k(t)$, (corresponding to variable dynamics that we need recover)

To highlight the broad scope of our approach from a data analysis perspective, we assume that we do not have direct access to the horizontal displacement. Instead, we generate *movies* of the motion of the coupled pendula) . We now apply a *fixed, invertible, random projection* to each frame of the movie. In other words, each frame of the movie was multiplied by a fixed matrix, whose columns were independently sampled from a multivariate normal distribution and normalized to have a unit norm. The resulting movie with the projected frames can be found in the following link: youtube.com/watch?v=xz0hzQTyPCo.

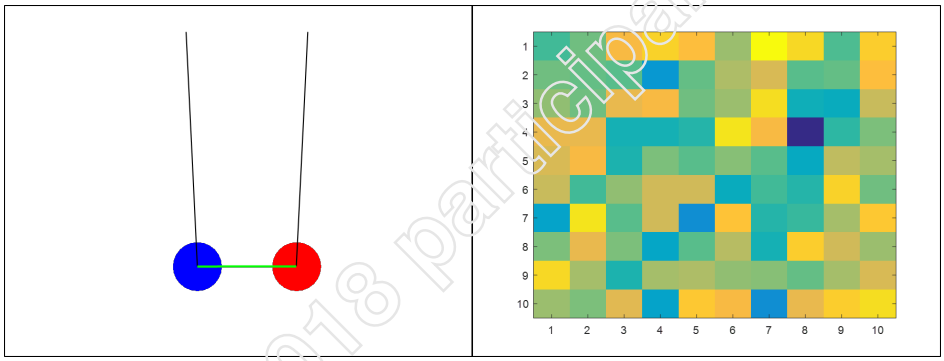


FIGURE 5. An example of a snapshot of the coupled pendulum system paired with its random projection counterpart.

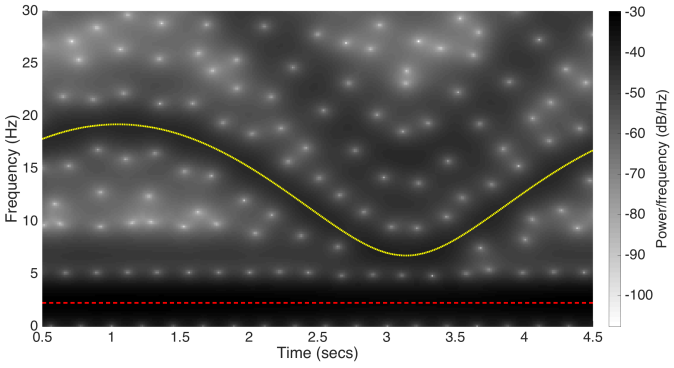


FIGURE 6. The Fourier spectrogram of the principal eigenvector representing the time axis. These results are based on the random projections of the movies frames with the same time-varying spring constant. The two frequencies ω_1 and $\omega_2(t)$ are overlayed on the spectrogram. The dashed red line corresponds to the fixed oscillation frequency ω_1 and the dotted yellow line corresponds to the time-varying oscillation frequency ω_2.

The pendulum model above was designed as an analogy to calcium fluorescence measuring neuronal activity in the motor cortex of a mouse repeating a task on multipls trials , the fluctuating pixels of the pendulum codes led us to latent variable which are the time variable normal modes. Identical processing on neuronal fluctuations should reveal internal latent controls

The setting in the figure below is identical to the one described above,see Figure 7, but we don't assume any equations and follow G.Mishne [28] in which the project is described

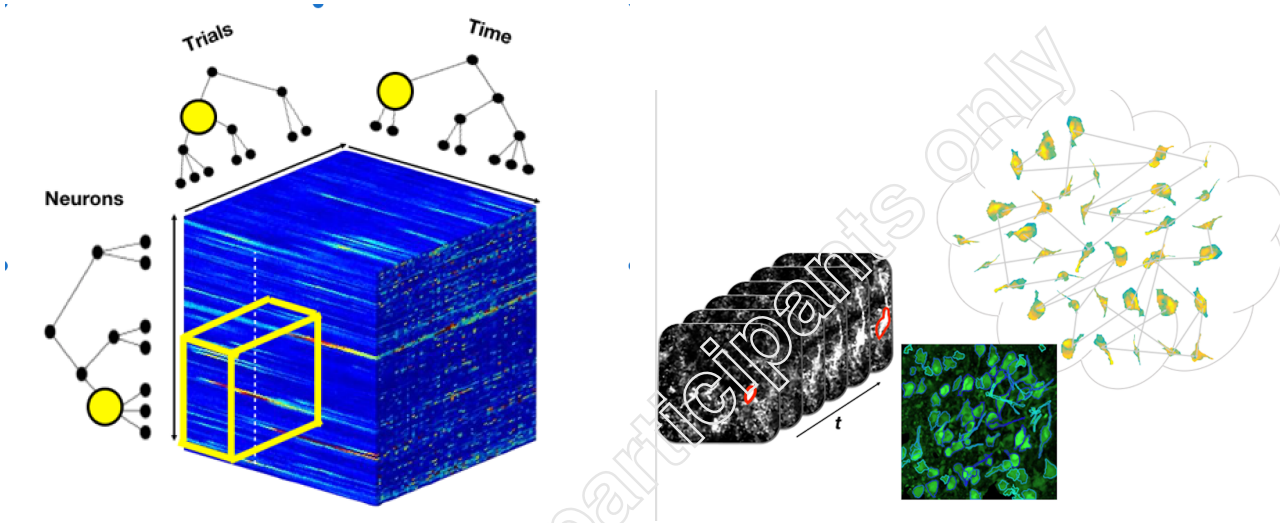


FIGURE 7. This figure illustrates the setting of our tri-geometry analysis of the collected trial-based neuronal activity from the motor cortical region. These measurements were taken from a behaving mouse in a single day of experiments. The data is composed of 60 trials. A single trial consists of 12 seconds acquired at 10Hz. The recordings are taken from 121 neurons located in M1 cortex. The entire data set of neuronal activity is therefore viewed as a 3-dimensional tensor(left), measuring a (121-dimensional) vector of neuronal activity at each time frame within each trial, and the neuronal activity is represented by the intensity level of the image (blue – no activity, red – high activity) The three trees in triality are plotted and a sub box in yellow corresponds to a group of neurons coactivity on a group of trials , at a fixed period . The data is visualized on the right as 2D slices of temporal neural images, with a clean neural map extracted in green , and the neuron graph above .

5. Learning Empirical Intrinsic Geometry, EIG.

In the preceding sections we have glossed over the basic problem of initializing the geometric affinity, we ignored the dependence of the eigenvectors on the coordinate system. We now describe with more detail, a simpler setup where empirical analysis reveals the underlying intrinsic latent geometric coordinates on which data is measured.

Our basic assumption is, that we are observing a stochastic time series governed by a Langevin equation on a Riemannian manifold, these observations are transformed through a nonlinear transformation into high dimensions in an ambient unknown independent noisy environment. Our goal is to show that we can recover the original Riemannian manifold as well as the potential, driving the dynamics of the observations. Moreover by building a

geometry capturing the normalized variabilities of local statistical histograms, we eliminate the effect of external noise interferences.

As an example consider a molecule (Alanine Dipeptide) consisting of 10 atoms and oscillating stochastically in water. It is known that the configuration at any given time is essentially described by two angle parameters. We assume that we observe five atoms of the molecule for a certain period of time, and five other atoms in the remainder time. The task is to describe the position of all atoms at all times, or more precisely, discover the angle parameters and their relation to the position of all atoms. See Figure 8.The main point is that the observations are quite different, perhaps using completely different sensors in different environments (but same dynamic phenomenon) and that we derive an identical intrinsic "natural" manifold parameterizing the observations.

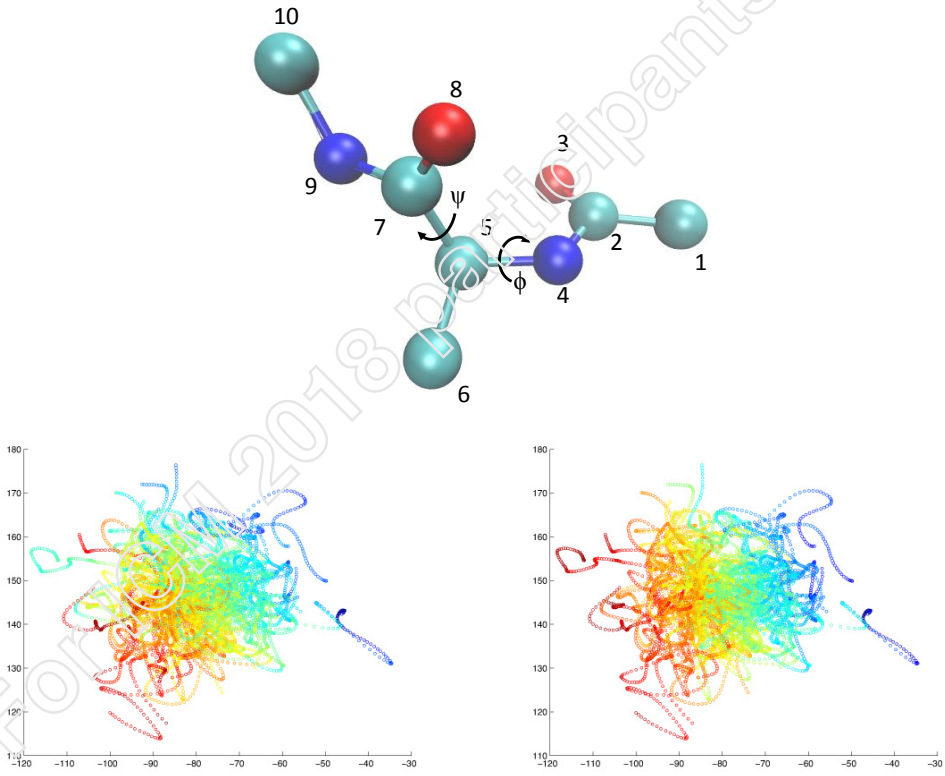


FIGURE 8. (a) A representative molecular structure of Alanine Dipeptide, excluding the hydrogens. The atoms are numbered and the two dihedral angles ϕ and ψ are indicated. (b)-(c): A 2-dimensional scatter plot of random trajectories of the dihedral angles ϕ and ψ. Based on observations of the corresponding random trajectories of merely five out of ten atoms of the molecule, we infer a model describing one of the angles. The points are colored according to the values of the inferred model from the five even atoms (b) and the five odd atoms (c). We observe that the gradient of the color is parallel to the x-axis, indicating an adequate representation of one of the angles. In addition, the color patterns are similar, indicating that the models are independent of the particular atoms observed, and describe the common intrinsic parameterization of the molecule dynamics.

An important remark is that we observe stochastic data constrained to lie on an unknown Riemannian manifold, that we need somehow to reconstruct explicitly, not having any coordinate system on the manifold . This is achieved through the explicit construction of the eigenvectors of an intrisic Laplace operator on the manifold (observations), these can be used to obtain a low dimensional canonical embeddings independent of observation modality, and obtain local charts on the manifold. This invariant description of the dynamics, is similar to the reformulation of Newton's law through invariant Hamiltonian equations see Talmon [**?**].

Broad outline.

To achieve this task and learn an intrisic Riemannian manifold structure, we assume that we observe stochastic clouds of points corresponding to some unknown standard brownian ensemble (as in the example of the Alanine molecule for short time intervals). More specifically this process has three scales:

The first identifies "local micro clouds" and converts them to statistical histograms .

The second relates clouds of histograms to each other using the affine invariant Mahalanobis metric between histograms,this metric is immune to the distortion due to independent noise.

The third builds the whole Riemannian manifold by integrating the local metrics

.

This provides an intrinsic Riemannian manifold that is both insensitive to noise and, invariant to changes of variables.

(This construction is a data driven version of information geometry see [**?**])

5.1. **detailed description.** Specifically and for simplicity of exposition, we consider a flat manifold for which we adopt the state-space formalism to provide a generic problem formulation that may be adapted to a wide variety of applications.

Let $\boldsymbol{\theta}_t$ be a d-dimensional underlying coordinates of a process in time index t. The dynamics of the process are described by normalized stochastic differential equations as follows[1]

$$d\theta_t^i = a^i(\theta_t^i)dt + dw_t^i,\ i = 1, \ldots, d, \tag{5}$$

where a^i are unknown drift functions and w_t^i are independent white noises. For simplicity, we consider here normalized processes with unit variance noises. Since a^i are any drift functions, we may first apply normalization without effecting the following derivation. See A. Singer ,R.Talmon[32, **?**] for details. We note that the underlying process is equivalent to the system state in the classical terminology of the state-space approach.

Let $\mathbf{y}_t$ denote an n-dimensional observation process in time index t, drawn from a probability density function (pdf) $f(\mathbf{y};\boldsymbol{\theta})$. The statistics of the observation process are time-varying and depend on the underlying process $\boldsymbol{\theta}_t$. We consider a model in which the clean observation process is accessible only via a noisy n-dimensional measurement process $\mathbf{z}_t$, given by

$$\mathbf{z}_t = g(\mathbf{y}_t, \mathbf{v}_t) \tag{6}$$

[1] x^i denotes access to the ith coordinate of a point $\mathbf{x}$.

where g is an unknown (possibly nonlinear) measurement function and $\mathbf{v}_t$ is a corrupting n-dimensional measurement noise, drawn from an unknown stationary pdf $q(\mathbf{v})$ and independent of $\mathbf{y}_t$.

The description of $\boldsymbol{\theta}_t$ constitutes a parametric manifold that controls the accessible measurements at-hand. Our goal is to reveal the underlying process $\boldsymbol{\theta}_t$ and its dynamics based on a sequence of measurements $\{\mathbf{z}_t\}$.

Let $p(\mathbf{z};\boldsymbol{\theta})$ denote the pdf of the measured process $\mathbf{z}_t$ controlled by $\boldsymbol{\theta}_t$, it satisfies the following property.

Lemma 1. *The pdf of the measured process $\mathbf{z}_t$ is a linear transformation of the pdf of the clean observation component $\mathbf{y}_t$.*

The proof is obvious, relying on the independence of $\mathbf{y}_t$ and $\mathbf{v}_t$, the pdf of the measured process is given by

$$p(\mathbf{z};\boldsymbol{\theta}) = \int\limits_{g(\mathbf{y},\mathbf{v})=\mathbf{z}} f(\mathbf{y};\boldsymbol{\theta})q(\mathbf{v})d\mathbf{y}d\mathbf{v}. \tag{7}$$

We note that in the common case of additive measurement noise, i.e., $g(\mathbf{y},\mathbf{v}) = \mathbf{y}+\mathbf{v}$, only a single solution $\mathbf{v}(\mathbf{z}) = \mathbf{z}-\mathbf{y}$ exists. Thus, $p(\mathbf{z};\boldsymbol{\theta})$ in (7) becomes a linear convolution

$$p(\mathbf{z};\boldsymbol{\theta}) = \int\limits_{\mathbf{y}} f(\mathbf{y};\boldsymbol{\theta})q(\mathbf{z}-\mathbf{y})d\mathbf{y} = f(\mathbf{z};\boldsymbol{\theta}) * q(\mathbf{z}).$$

The dynamics of the underlying process are conveyed by the time-varying pdf of the measured process. Thus, this pdf may be very useful in revealing the desired underlying process and its dynamics. Unfortunately, the pdf is unknown since the underlying process and the dynamical and measurement models are unknown. Assume we have access to a class of estimators of the pdf over discrete bins which can be viewed as linear transformations. Let $\mathbf{h}_t$ be such an estimator with m bins which is viewed as an m-dimensional process and is given by

$$p(\mathbf{z};\boldsymbol{\theta}_t) \overset{\mathcal{T}}{\mapsto} \mathbf{h}_t, \tag{8}$$

where $\mathcal{T}$ is a linear transformation of the density $p(\mathbf{z};\boldsymbol{\theta})$ from the infinite sample space of $\mathbf{z}$ into a finite interval space of dimension m. By Lemma 1 and by definition (8) we get the following results.

The process $\mathbf{h}_t$ is a linear transformation of the pdf of the clean observation component $\mathbf{y}_t$.

The process $\mathbf{h}_t$ can be described as a deterministic nonlinear map of the underlying process $\boldsymbol{\theta}_t$.

We can use histograms as estimates of the pdf, and we assume that a *sequence* of measurements is available. Accordingly, let $\mathbf{h}_t$ be the empirical local histogram of the measured process $\mathbf{z}_t$ in a short-time window of length L_1 at time t. Let $\mathcal{Z}$ be the sample space of $\mathbf{z}_t$ and let $\mathcal{Z} = \bigcup_{j=1}^{m} \mathcal{H}_j$ be a finite partition of $\mathcal{Z}$ into m disjoint histogram bins. Thus, the value of each histogram bin is given by

$$h_t^j = \frac{1}{|\mathcal{H}_j|}\frac{1}{L_1}\sum_{s=t-L_1+1}^{t} \mathbf{1}_{\mathcal{H}_j}(\mathbf{z}_s), \tag{9}$$

where $\mathbf{1}_{\mathcal{H}_j}(\mathbf{z}_t)$ is the indicator function of the bin $\mathcal{H}_j$ and $|\mathcal{H}_j|$ is its cardinality. By assuming (unrealistically) that infinite number of samples are available and that their density in each histogram bin is uniform, (9) can be expressed as

$$h_t^j = \frac{1}{|\mathcal{H}_j|} \int\limits_{\mathbf{z}\in\mathcal{H}_j} p(\mathbf{z};\boldsymbol{\theta}) d\mathbf{z}. \tag{10}$$

Thus, ideally the histograms are *linear transformations* of the pdf. In addition, if we shrink the bins of the histograms as we get more and more data, the histograms converge to the pdf

$$\mathbf{h}_t \xrightarrow[|\mathcal{H}_j|\to 0]{L_1\to\infty} p(\mathbf{z};\boldsymbol{\theta}). \tag{11}$$

In practice, since the computation of high-dimensional histograms is challenging, we preprocess high-dimensional data by applying random filters in order to reduce the dimensionality without corrupting the information.

5.2. **Mahalanobis Distance.** We view $\mathbf{h}_t$ (the linear transformation of the local densities, e.g. the local histograms) as feature vectors for each measurement $\mathbf{z}_t$. The process $\mathbf{h}_t$ satisfies the dynamics given by Itô's lemma

$$\begin{aligned} h_t^j &= \sum_{i=1}^{d} \left(\frac{1}{2} \frac{\partial^2 h^j}{\partial\theta^i \partial\theta^i} + a^i \frac{\partial h^j}{\partial\theta^i} \right) dt \\ &+ \sum_{i=1}^{d} \frac{\partial h^j}{\partial\theta^i} dw_t^i, \ j = 1,\ldots,m. \end{aligned} \tag{12}$$

For simplicity of notation, we omit the time index t from the partial derivatives. According to (12), the (j,k)th element of the $m \times m$ covariance matrix $\mathbf{C}_t$ of $\mathbf{h}_t$ is given by

$$C_t^{jk} = \mathrm{Cov}(h_t^j, h_t^k) = \sum_{i=1}^{d} \frac{\partial h^j}{\partial\theta^i} \frac{\partial h^k}{\partial\theta^i}, \ j,k = 1,\ldots,m. \tag{13}$$

In matrix form, (13) can be rewritten as

$$\mathbf{C}_t = \mathbf{J}_t \mathbf{J}_t^T \tag{14}$$

where $\mathbf{J}_t$ is the $m \times d$ Jacobian matrix, whose (j,i)th element is defined by

$$J_t^{ji} = \frac{\partial h^j}{\partial\theta^i}, \ j = 1,\ldots,m, \ i = 1,\ldots,d.$$

Thus, the covariance matrix $\mathbf{C}_t$ is a semi-definite positive matrix of rank d.

We define a nonsymmetric $\mathbf{C}$-dependent squared distance between pairs of measurements as

$$a_{\mathbf{C}}^2(\mathbf{z}_t, \mathbf{z}_s) = (\mathbf{h}_t - \mathbf{h}_s)^T \mathbf{C}_s^{-1} (\mathbf{h}_t - \mathbf{h}_s) \tag{15}$$

and a corresponding symmetric distance as

$$d_{\mathbf{C}}^2(\mathbf{z}_t, \mathbf{z}_s) = 2(\mathbf{h}_t - \mathbf{h}_s)^T \left(\mathbf{C}_t + \mathbf{C}_s\right)^{-1} (\mathbf{h}_t - \mathbf{h}_s). \tag{16}$$

Since usually the dimension d of the underlying process is smaller than the number of histogram bins m, the covariance matrix is singular and non-invertible. Thus, in practice we use the pseudo-inverse to compute the inverse matrices in (15) and (16).

The distance in (16) is known as the *Mahalanobis distance* with the property that it is invariant under linear transformations. Thus, by Lemma 1 , it is invariant to the measurement noise and functional distortion (e.g., additive noise or multiplicative noise). We note however that the linear transformation employed by the measurement noise on the observable pdf (7) may degrade the available information.

In addition, by Lemma 3.1 in [23], the Mahalanobis distance in (16) approximates the Euclidean distance between samples of the underlying process. Let $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_s$ be two samples of the underlying process. Then, the Euclidean distance between the samples is approximated to a second order by a local linearization of the nonlinear map of $\boldsymbol{\theta}_t$ to $\mathbf{h}_t$, and is given by

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_s\|^2 = d^2_{\mathbf{C}}(\mathbf{z}_t, \mathbf{z}_s) + O(\|\mathbf{h}_t - \mathbf{h}_s\|^4). \tag{17}$$

For more details see [32] and [23]. Assuming there is an intrinsic map $i(\mathbf{h}_t) = \boldsymbol{\theta}_t$ from the feature vector to the underlying process, the approximation in (17) is equivalent to the inverse problem defined by the following nonlinear differential equation

$$\sum_{i=1}^{m} \frac{\partial \theta^j}{\partial h^i} \frac{\partial \theta^k}{\partial h^i} = \left[C_t^{-1}\right]^{jk}, \ j,k = 1,\ldots,d. \tag{18}$$

This equation which is nothing more than a discrete formulation of the definition of a Riemannian metric on the manifold is empirically solved through the eigenvectors of the corresponding discrete Laplace operator. The approximation in (17) recovers the intrinsic distances on the parametric manifold and is obtained empirically from the noisy measurements by "infinitesimally" inverting the measurement function.

For further illustration, see Fig. 9.

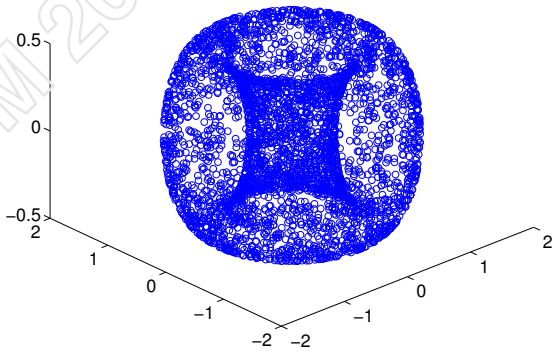


FIGURE 9. Consider a set of points on a 2-dimensional torus in $\mathbb{R}^3$ ("the manifold") which are samples of a Brownian motion on the torus. The geometric interpretation of the intrinsic notion is the search for a canonical description of the set, which is independent of the coordinate system. For example, the points can be written in 3 cartesian coordinates, or in the common parameterization of a torus using two angles, however, the intrinsic model (constructed based on the points) describing the torus should be the same. The mahalanobis distance attaches to each point a Riemannian metric that corresponds to a probability measure that is driven by the underlying dynamics (the Brownian motion in this particular case), and therefore, it is invariant to the coordinate system.

5.3. **Local Covariance Matrix Estimation.** Let t_0 be the time index of a "pivot" sample $\mathbf{h}_{t_0}$ of a "cloud" of samples $\{\mathbf{h}_{t_0,s}\}_{s=1}^{L_2}$ of size L_2 taken from a local neighborhood in time. Here we assume that a sequence of measurements is available, the temporal neighborhoods can be simply short windows in time centered at time index t_0.

The pdf estimates and the local clouds implicitly define two time scales on the sequence of measurements. The fine time scale is defined by short-time windows of L_1 measurements to estimate the temporal pdf. The coarse time scale is defined by the local neighborhood of L_2 neighboring feature vectors in time. Accordingly, we note that the approximation in (17) is valid as long as the statistics of the noise are locally fixed in the short-time windows of length L_1 (i.e., slowly changing compared to the fast variations of the underlying process) and the fast variations of the underlying process can be detected in the difference between the feature vectors in windows of length L_2.

According to the dynamical model in (5) and (12), the samples in the local cloud can be seen as small perturbations of the pivot sample created by the noise $\mathbf{w}_t$. Thus, we assume that the samples share similar local probability densities[2]and may be used to estimate the local covariance matrix, which is required for the construction of the Mahalanobis metric (16). The empirical covariance matrix of the cloud is estimated by

$$\begin{aligned} \hat{\mathbf{C}}_{t_0} &= \frac{1}{L_2}\sum_{s=1}^{L_2}\left(\mathbf{h}_{t_0,s}-\hat{\boldsymbol{\mu}}_{t_0}\right)\left(\mathbf{h}_{t_0,s}-\hat{\boldsymbol{\mu}}_{t_0}\right)^T \\ &\simeq \mathbb{E}\left[\left(\mathbf{h}_{t_0}-\mathbb{E}[\mathbf{h}_{t_0}]\right)\left(\mathbf{h}_{t_0}-\mathbb{E}[\mathbf{h}_{t_0}]\right)^T\right]=\mathbf{C}_{t_0} \end{aligned} \tag{19}$$

where $\hat{\boldsymbol{\mu}}_{t_0}$ is the empirical mean of the set.

As the rank of the matrix d is usually smaller than the covariance matrix dimension m, in order to compute the inverse matrix we use only the d principal components of the matrix. This operation "cleans" the matrix and filters out noise. In addition, when the empirical rank of the local covariance matrices of the feature vectors is lower than d, it indicates that the available feature vectors are insufficient and a larger cloud should be used.

6. Concluding remarks and bibliography.

This overview of various methodologies to learn and extract natural geometries, and latent variables from point clouds generated by , observations , computations, or mathematical processes , is by necessity superficial , and neglects to cover the massive amount of litterature and algorithms around machine learning. We refer to Yann Ollivier [**?**] who has pursued invariant geometric ideas, like the EIG approach, in the context of information geometry and natural deep learning. The use of tensors to extract features analogous to principal components ("tensor PCA") in the context of machine learning, or data processing is also quite extensive see A. Anandkumar. [**?**] .

Our goal here was to emphasize on the one hand the co-dependent geometries in duality or triality (duality in Besov spaces enables generalizing flexible earth mover distances) , and on the other hand to illustrate the essential interplay between geometry of point clouds with various analytic measures of smoothness. This construction enables both effective Harmonic analysis and dynamic metric constructions . We point out that in the case of matrices, or

[2]We emphasize that we consider the statistics of the feature vectors and not the feature vectors themselves, which are estimates of the varying statistics of the raw measurements.

even convolution operators on functions , it is not generally effective to use their eigenvectors or Fourier transform, in order to unravel its effect on functions.

The Calderon-Zygmund decompositions were introduced to gain an intimate understanding of the Hilbert transform, they have their wavelet analogs . We try to convey here , that this basic geometric organizationn philosophy is natural in the context of mathematical geometric learning. More generally given a class of geometric structures, such as curves or embedded surfaces it is natural to relate them through the properties of various operators intrinsic operators , such as a Diffusion , or other functions of the Laplace operator , and then use a distance between these operators , as a way of measuring similarity between the structures see Berard et al, who show that the distance between Riemanian manifolds can be measured [**?**]. This is our approach in the 3 tensor case , where we can view one axis as the label for the structures, and the other two as representing the corresponding operators, the metrics so defined are quite remarkable. A similar vision is developed to achieve "shape" matching by G. Peyre, M. Cuturi, J. Solomon. They measure the distance between affinity matrices ,or diffusions through appropriate Earth mover distances, see [**?**] and their references.

We should also mention that alternative multiscale data models were developped by M. Magggioni [2], as well as R Lederman and R Talmon [**?**]who developed a methodology to extract common latent variables between disparate sets of observations , this method can have a profound impact on the scientific discovery process.

References

[1] Approximate earth movers distance in linear time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[2] William K. Allard, Guangliang Chen, and Mauro Maggioni. Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. *Appl. Comput. Harmon. Anal.*, 32(3):435–462, 2012.

[3] B. Alpert, G. Beylkin, R. Coifman, and V. Rokhlin. Wavelet-like bases for the fast solution of second-kind integral equations. *SIAM J. Sci. Comput.*, 14(1):159–184, 1993.

[4] Jerrod Isaac Ankenman. *Geometry and Analysis of Dual Networks on Questionnaires*. ProQuest LLC, Ann Arbor, MI, 2014. Thesis (Ph.D.)–Yale University.

[5] M. Belkin and P. P. Niyogi. *Laplacian eigenmaps and spectral techniques for embedding and clustering*, volume 14 of *Advances in Neural Information Processing Systems*. 2001.

[6] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35:1798–1828, 2013.

[7] Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta Numer.*, 13:147–269, 2004.

[8] Eric C. Chi, Genevera I. Allen, and Richard G. Baraniuk. Convex biclustering. *Biometrics*, 73(1):10–19, 2017.

[9] R. R. Coifman. Perspectives and challenges to harmonic analysis and geometry in high dimensions: geometric diffusions as a tool for harmonic analysis and structure definition of data. In *Perspectives in analysis*, volume 27 of *Math. Phys. Stud.*, pages 27–35. Springer, Berlin, 2005.

[10] R. R. Coifman. Perspectives and challenges to harmonic analysis and geometry in high dimensions: geometric diffusions as a tool for harmonic analysis and structure definition of data. In *Perspectives in analysis*, volume 27 of *Math. Phys. Stud.*, pages 27–35. Springer, Berlin, 2005.

[11] R. R. Coifman and R. Rochberg. Another characterization of BMO. *Proc. Amer. Math. Soc.*, 79(2):249–254, 1980.

[12] Ronald R. Coifman and Matan Gavish. Harmonic analysis of digital data bases. In *Wavelets and multiscale analysis*, Appl. Numer. Harmon. Anal., pages 161–197. Birkhäuser/Springer, New York, 2011.

[13] Ronald R. Coifman and Stéphane Lafon. Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. *Appl. Comput. Harmon. Anal.*, 21(1):31–52, 2006.

[14] Ronald R. Coifman and Mauro Maggioni. Diffusion wavelets. *Appl. Comput. Harmon. Anal.*, 21(1):53–94, 2006.
[15] David L. Donoho and Carrie Grimes. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA*, 100(10):5591–5596, 2003.
[16] Carmeline J. Dsilva, Ronen Talmon, C. William Gear, Ronald R. Coifman, and Ioannis G. Kevrekidis. Data-driven reduction for a class of multiscale fast-slow stochastic dynamical systems. *SIAM J. Appl. Dyn. Syst.*, 15(3):1327–1351, 2016.
[17] M. Gavish, B. Nadler, and R. R. Coifman. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *Proceedings of the 27th International Conference on Machine Learning, ICML (2010)*.
[18] Matan Gavish and Ronald R. Coifman. Sampling, denoising and compression of matrices by coherent matrix organization. *Appl. Comput. Harmon. Anal.*, 33(3):354–369, 2012.
[19] Dimitrios Giannakis. Dynamics-adapted cone kernels. *SIAM J. Appl. Dyn. Syst.*, 14(2):556–608, 2015.
[20] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.
[21] John Guckenheimer and Philip Holmes. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, volume 42 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1983.
[22] Ioannis G. Kevrekidis, C. William Gear, James M. Hyman, Panagiotis G. Kevrekidis, Olof Runborg, and Constantinos Theodoropoulos. Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level analysis. *Commun. Math. Sci.*, 1(4):715–762, 2003.
[23] Dan Kushnir, Ali Haddad, and Ronald R. Coifman. Anisotropic diffusion on sub-manifolds with application to Earth structure classification. *Appl. Comput. Harmon. Anal.*, 32(2):280–294, 2012.
[24] Dan Kushnir, Ali Haddad, and Ronald R. Coifman. Anisotropic diffusion on sub-manifolds with application to Earth structure classification. *Appl. Comput. Harmon. Anal.*, 32(2):280–294, 2012.
[25] Stephane S. Lafon. *Diffusion maps and geometric harmonics*. ProQuest LLC, Ann Arbor, MI, 2004. Thesis (Ph.D.)–Yale University.
[26] N. F. Marshall and R. R. Coifman. Manifold learning with bi-stochastic kernels. 2018.
[27] Igor Mezic. On comparison of dynamics of dissipative and finite-time systems using koopman operator methods. *IFAC – Papers OnLine*, 49:454–461, 2016.
[28] Gal Mishne, Ronen Talmon, Ron Meir, Jackie Schiller, Uri Dubin, and Ronald R. Coifman. Hierarchical coupled geometry analysis for neuronal structure and activity pattern discovery. 11 2015.
[29] Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Difusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.*, 21(1):113–127, 2006.
[30] Carey E. Priebe, David J. Marchette, Youngser Park, Edward J. Wegman, Jeffrey L. Solka, Diego A. Socolinsky, Damianos Karakos, Ken W. Church, Roland Guglielmi, Roland R. Coifman, Dekang Lin, Dennis M. Healy, Marc Q. Jacobs, and Anna Tsao. Iterative denoising for cross-corpus discovery. In *COMPSTAT 2004—Proceedings in Computational Statistics*, pages 381–392. Physica, Heidelberg, 2004.
[31] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20–36, 2011.
[32] Amit Singer and Ronald R. Coifman. Non-linear independent component analysis with diffusion maps. *Appl. Comput. Harmon. Anal.*, 25(2):226–239, 2008.
[33] S. A. Smoljak. Quadrature and interpolation formulae on tensor products of certain function classes. *Dokl. Akad. Nauk SSSR*, 148:1042–1045, 1963.
[34] Arthur D. Szlam, Mauro Maggioni, and Ronald R. Coifman. Regularization on graphs with function-adapted diffusion processes. *J. Mach. Learn. Res.*, 9:1711–1739, 2008.
[35] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman. Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs. *IEEE Signal Processing Magazine*, 30:75–86, 2013.
[36] Or Yair, Ronen Talmon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Reconstruction of normal forms by learning informed observation geometries from data. *Proc. Natl. Acad. Sci. USA*, 114(38):E7865–E7874, 2017.

E-mail address: coifman@math.yale.edu

Department of Mathematics, Program in Applied Mathematics, Yale University