Manifold coordinates with physical meaning

Samson Koelle
Department of Statistics
University of Washington
Seattle, WA 98195
sjkoelle@uw.edu

Hanyu Zhang
Department of Statistics
University of Washington
Seattle, WA 98195
hanyuz6@uw.edu

Marina Meilă
Department of Statistics
University of Washington
Seattle, WA 98195
mmp2@uw.edu

Yu-Chia Chen
Department of Electrical & Computer Engineering
University of Washington
Seattle, WA 98195
yuchaz@uw.edu

Abstract

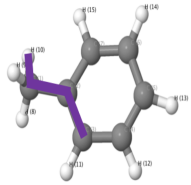
One of the aims of both linear and non-linear dimension reduction is to find a reduced set of *collective variables* that describe the data manifold. While algorithms return abstract coordinates such as spaces spanned by eigenvectors of data-dependent matrices, one can often associate these with features of the data, and hence with domain-related meaning. Usually, finding these domain-related or physical meanings is done via visual inspection by an expert. Our work formulates this problem as *sparse, non-parametric, non-linear* recovery of the manifold coordinates over a user-defined dictionary of domain-related functions. We show that the original problem can be transformed into a *linear Group Lasso* problem, and demonstrate the effectiveness of the method on molecular simulation data.

1 Motivation: manifold learning for collective variables

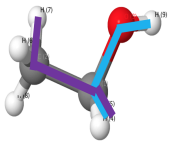
Our motivating application is the understanding of the *slow dynamic modes* of molecules and other atomic systems from molecular dynamics simulations. In such simulations, the positions of atoms within a molecule are sampled as they proceed through time from some initial conditions. Even though the vector of atomic coordinates can take any value, due to interatomic interactions, the relative positions of atoms within the molecule lie near a low-dimensional manifold.

Manifold Learning (ML) methods have become the framework of choice for finding these collective variables in molecular systems in a data-driven way. These variables correspond to macroscopically interesting transformations of the system, and can explain some of its properties [Clementi et al., 2000, Noé and Clementi, 2017]. Figure 1 illustrates several manifolds learned from molecular dynamics simulations. The learned collective variables are, in these cases, identified by visual inspection as corresponding to *bond torsions*, also known as dihedral angles.

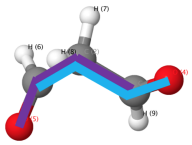
Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.



(a) Toluene

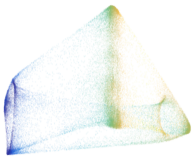


(b) Ethanol

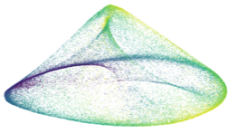


(c) Malonaldehyde

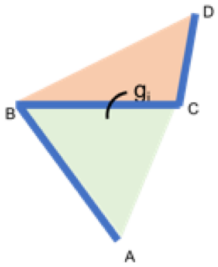
(d) g_1



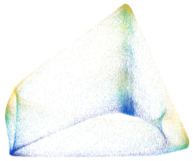
(e) g_1



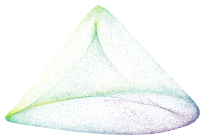
(f) g_1



(g) Torsion example



(h) g_2



(i) g_2

Figure 1: Collective coordinates with physical meaning in Molecular Dynamics (MD) simulations. 1a-1c Diagrams of the toluene (C_7H_8), ethanol (C_2H_5OH), and malonaldehyde ($C_3H_4O_2$) molecules, with the carbon (C) atoms in grey, the oxygen (O) atoms in red, and the hydrogen (H) atoms in white. Bonds defining important torsions g_j are marked in purple and blue. The bond torsion is the angle of the planes inscribing the first three and last three atoms on the line (1g). 1d Embedding of the configurations of toluene into $m = 2$ dimensions, showing a manifold of $d = 1$. The color corresponds to the values of the purple torsion g_1. 1e, 1h Embedding of the configurations of the ethanol in $m = 3$ dimensions, showing a manifold of dimension $d = 2$, respectively colored by the blue and purple torsions in Figure 1b. 1f, 1i. Embedding of the configurations of the malonaldehyde in $m = 3$ dimensions, showing a manifold of dimension $d = 2$, respectively colored by the blue and purple torsions in Figure 1c. Data is from Chmiela et al. [2017].

2 Problem formulation

We propose to replace such visual interpretation with a statistical procedure. We make the standard assumption that the observed data $\mathcal{D} = \{\xi_i \in \mathbb{R}^D : i \in 1 \dots n\}$ are sampled i.i.d. from a *smooth Riemannian manifold* [1] $(\mathcal{M}, \mathbf{id})$ of intrinsic dimension d embedded in a feature space $\mathbb{R}^D$ by the inclusion map, with $\mathbf{id}$ the identity metric with respect to $\mathbb{R}^D$. We assume that the intrinsic dimension d of $\mathcal{M}$ is known. Furthermore, we assume the existence of a smooth *embedding map* $\phi : \mathcal{M} \to \phi(\mathcal{M}) \subset \mathbb{R}^m$, where typically $m << D$. That is, ϕ restricted to $\mathcal{M}$ is a diffeomorphism onto its image, and $\phi(M)$ is a submanifold of $\mathbb{R}^m$. We call the coordinates $\phi(\xi_i)$ in this m dimensional ambient space the *embedding coordinates*; let $\Phi = [\phi(\xi_i)^T]_{i=1:n} \in \mathbb{R}^{n \times m}$. In practice, the mapping of the data $\mathcal{D}$ onto $\phi(\mathcal{D})$ represents the output of an embedding algorithm, and we only have access to $\mathcal{M}$ and ϕ via $\mathcal{D}$ and its image Φ.

[1] The reader is referred to Lee [2003] for the definitions of the differential geometric terms used in this paper.

In addition, we are given a *dictionary* of user-defined and domain-related smooth functions $\mathcal{G} = \{g_1, \dots g_p$, with $g_j : U \subseteq \mathbb{R}^D \to \mathbb{R}\}$. Our goal is to express the embedding coordinate functions $\phi_1 \dots \phi_m$ in terms of functions in $\mathcal{G}$. More precisely, we assume that $\phi(x) = h(g_{j_1}(x), \dots\ g_{j_s}(x))$, where $h : O \subseteq \mathbb{R}^s \to \mathbb{R}^m$ is a smooth function of s variables, defined on a open subset of $\mathbb{R}^s$ containing the ranges of $g_{j_1}, \dots\ g_{j_s}$. Let $S = \{j_1, \dots\ j_s\}$, and $g_S = [g_{j_1}(x), \dots\ g_{j_s}(x)]^T$. The problem is to discover the set $S \subset [p]$ such that $\phi = h \circ g_S$. We call S the *functional support* of h, or the *explanation* for the manifold $\mathcal{M}$ in terms of $\mathcal{G}$. For instance, in the toluene example, the functions in $\mathcal{G}$ are a set of torsions in the molecule, $s = 1$, and $g_S = g_1$ is the explanation for the 1-dimensional manifold traced by the configurations.

3 The MANIFOLDLASSO Algorithm

We start from the well-known mathematical fact that, for any differentiable functions f, g, h, when $f = h \circ g$, the differentials Df, Dh, and Dg at any point are in the *linear* relationship $Df = DhDg$. The key idea of our method is to compose differentials of functional covariates to reconstruct the differentials of the manifold embedding coordinates.

The MANIFOLDLASSO algorithm implements this idea. The Algorithm takes as input data $\mathcal{D}$ sampled from an unknown manifold $\mathcal{M}$, a dictionary $\mathcal{G}$ of functions defined on an open subset of the ambient space $\mathbb{R}^D$ that contains $\mathcal{M}$, and an embedding Φ in $\mathbb{R}^m$. The output of MANIFOLDLASSO is a set S of indices in $\mathcal{G}$, representing the functions in $\mathcal{G}$ that explain $\mathcal{M}$.

The first part of the algorithm calculates the necessary gradients, while the second finds the support S by solving the following Group Lasso [Yuan and Lin, 2006] problem:

$$\arg \min_{\beta \in \mathbb{R}^{mn'p}} \sum_{i \in I} \sum_{k=1}^{m} ||y_{ik} - x_i \beta_{ik}||_2^2 + \frac{\lambda}{\sqrt{mn'}} \sum_{j=1}^{p} ||\beta_j||_2. \tag{1}$$

In the above, $y_{ik} = \operatorname{grad}_{\mathcal{M}} \phi_k(\xi_i) \in \mathbb{R}^d$ and $x_i = \operatorname{grad}_{\mathcal{M}} \mathcal{G}(\xi_i) \in \mathbb{R}^{d \times p}$ are the gradients of the embedding coordinates and dictionary functions on $\mathcal{M}$, and $\beta_j = \frac{\partial h_{1:m}}{\partial g_j}(\mathcal{D}) \in \mathbb{R}^{mn'}$ are the learned coefficients corresponding to dictionary function g_j. The need for regularization arises in the general case when $p > d$. The algorithm can be run on subsets of points $I \subset 1 : n$ with $|I| = n'$, and therefore has runtime controllable independently of ϕ. The learned coefficients are the partial derivatives of the embedding coordinates with respect to the dictionary. The regularization term encourages entire β_j groups to be identically 0. The need for regularization is clear, since in general $p > d$. We base our choice of λ on matching the cardinality of the support to d; that is, we increase λ until only d functions are selected. In the extended paper, the use of Group Lasso for sparse functional regression was introduced for the first time, and recovery conditions for the set S were given [Meila et al., 2018].

3.1 Preprocessing dictionary gradients

Dictionary gradients x_i are assumed to be analytically available with respect to $\mathbb{R}^D$, but several preprocessing steps are required. First, since the intrapoint planar angle featurization we use to generate the embeddings in 1 is redundant with respect to the set of molecular shapes, gradients of dictionary functions are not well-defined. We thus utilize the method of Addicoat and Collins [2010] to project gradients obtained using automatic differentiation into the *shape space* embedded in $\mathbb{R}^D$. We then normalize all dictionary tangent bundles to have norm one. This is necessary to ensure that functions with larger gradients are not favored by our penalty. We then estimate tangent coordinates T_i to the data manifold $\mathcal{M}$ using weighted laplacian PCA [Chen et al., 2013], and project the normalized gradients onto the manifold to obtain the gradients on $\mathcal{M}$. This favors dictionary functions whose gradients are tangent to the manifold $\mathcal{M}$, and penalizes the g_j's which have large gradient components perpendicular to $\mathcal{M}$.

3.2 Estimating coordinate gradients

The coordinate gradients y_{ik} are not analytically available. Instead, we estimate y_{ik} using the Riemannian metric $G_i \in \mathbb{R}^{m \times m}$ of the embedding with respect to the original high-dimensional data [Perraul-Joncas and Meila, 2013]. The matrix G_i is the estimated value at point i of the *pushforward*

Riemannian metric $\mathbf{g}$, which is the unique Riemannian metric on $\phi(\mathcal{M})$ so that $(\phi(\mathcal{M}), \mathbf{g})$ is isometric to $(\mathcal{M}, \mathbf{id})$. This enables estimation of y_{ik} in the local basis T_i:

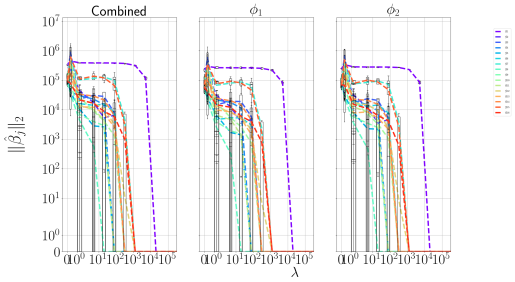
$$y_{i,1:m} = (A_i A_i^T)^{-1} A_i B_i^T G_i, \tag{2}$$

where

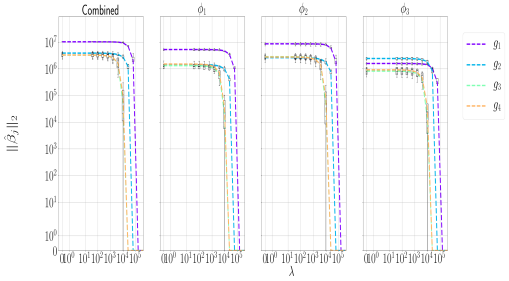
$$A_i = \left[T_i^T(\xi_{i'} - \xi_i)\right]_{i' \in \mathcal{N}_i} \in \mathbb{R}^{d \times k_i}, B_i = \left[\phi(\xi_{i'}) - \phi(\xi_i)\right]_{i' \in \mathcal{N}_i} \in \mathbb{R}^{m \times k_i}, \tag{3}$$

and $\mathcal{N}_i$ are the neighbors of datapoint ξ_i [Luo et al., 2009].

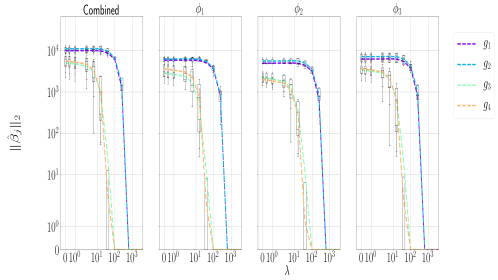
4 Experiments



(a) Toluene



(b) Ethanol



(c) Malonaldehyde

Figure 2: The left panels display $||\beta_j||_2$ $j \in 1:p$ as a function of the regularization parameter λ, for each of the three datasets shown in Figure 1. Colors represent individual dictionary functions (bond torsions). MANIFOLDLASSO selects the bond torsions visualized in Figure 1. The remaining panels display, each for one of the embedding coordinates ϕ_k, the norm of the vector $\vec{\beta}_{I,j,k}$ $j \in 1:p$. These confirm the visual association of g_1 with $\phi_{1,2}$ and g_2 with ϕ_3 in ethanol. Axes are linear between 0 and 1, and logarithmic above 1. Error bars summarize the outcomes of repetitions of sampling I.

5 Contributions

We have presented a novel paradigm for assigning meaning to the output of dimension reduction algorithms. In our paradigm, the scientist inputs a dictionary $\mathcal{G}$ of functions to be considered as possible collective coordinates. This relieves domain experts from visually examining every possible function in $\mathcal{G}$, and extends beyond mapping single coordinates to single functions to instead associating smooth maps of a subset of functions to a set of coordinates. This approach is very general: we do not rely on a particular embedding algorithm, and do not assume a parametric relationship between the embedding and the functions in the dictionary $\mathcal{G}$. Simplified versions of MANIFOLDLASSO can be used on domains that are not manifolds, such as the output of PCA, and non-linear sparse functional regression. Experiments show that the MANIFOLDLASSO Algorithm is robust to noise and successfully replaces visual inspection. The gradient group lasso approach and use of the Riemannian metric estimate G_i to pull back vectors between the tangent bundles of $\mathcal{M}$ and $\phi(\mathcal{M})$ are also original and of independent interest.

References

Matthew A. Addicoat and Michael A. Collins. Potential energy surfaces: the forces of chemistry. In Mark Brouard and Claire Vallance, editors, *Tutorials in Molecular Reaction Dynamics*, chapter 2, pages 28–49. Royal Society of Chemistry Publishing, London, 2010.

Guangliang Chen, Anna V. Little, and Mauro Maggioni. *Multi-Resolution Geometric Analysis for Data in High Dimensions*, pages 259–285. Birkhäuser Boston, Boston, 2013. ISBN 978-0-8176-8376-4. doi: 10.1007/978-0-8176-8376-4_13. URL `https://doi.org/10.1007/978-0-8176-8376-4_13`.

Stefan Chmiela, Alexandre Tkatchenko, Huziel Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, March 2017.

C. Clementi, H. Nymeyer, and J.N. Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *Journal of molecular biology*, 2000. says topology (of protein) more important than energy wells.

John M. Lee. *Introduction to Smooth Manifolds*. Springer-Verlag New York, 2003.

Chuanjiang Luo, Issam Safa, and Yusu Wang. Approximating gradients for meshes and point clouds via diffusion metric. *Comput. Graph. Forum*, 28(5):1497–1508, July 2009.

Marina Meila, Samson Koelle, and Hanyu Zhang. A regression approach for explaining manifold embedding coordinates. (1811.11891), 2018. URL `http://arxiv.org/abs/1811.11891`.

Frank Noé and Cecilia Clementi. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.*, 43:141–147, April 2017.

D. Perraul-Joncas and M. Meila. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery. *ArXiv e-prints*, May 2013.

M Yuan and Y Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.*, 2006.