Circular Coordinate Methods with Generalized Penalty Functions

Anonymous Author(s)
Affiliation
Address
email

Abstract

The circular coordinate representation performs dimension reduction and visualization for high-dimensional datasets on a torus using persistent cohomology. In this work, we propose a method to adapt the circular coordinate framework to take into account sparsity in high-dimensional applications. We use a generalized penalty function instead of an L_2 penalty in the traditional circular coordinate algorithm.

1 Introduction

Dimension reduction has been a major research subject in mathematics, statistics and computer science (Elad, 2010; Candès, 2014; Wilkinson and Luo, 2020). One of the major challenges in this field has been how to preserve the topological and geometrical structures of a high-dimensional, nonlinear dataset through the dimension reduction. The non-linear dimensionality reduction (NLDR) literature (Donoho and Grimes, 2005) consists of various attempts to address the problem of representing high-dimensional datasets, in terms of low-dimensional coordinate mappings. These methods have good use in both exploration and visualization of data.

To formally state dimension reduction problem, for a dataset $X \subset \mathbb{R}^d$ in form of $X = \{x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,d}) \in \mathbb{R}^d,\ i = 1, \cdots, n\}$ one assumes that X lives on a manifold M and attempts to find a collection of coordinate mappings $\Theta \coloneqq \{\theta_1, \cdots, \theta_k\}$, $\theta_j : \mathbb{R}^d \to \mathbb{R}$, $j = 1, \cdots, k$ with $k \leq d$. The reduced dataset can be written as $\Theta(X) = \{(\theta_1(x_i), \theta_2(x_i), \cdots, \theta_k(x_i)),\ i = 1, \cdots, n\} \subset \mathbb{R}^k$ through the coordinate mappings. A good choice of coordinate mappings would preserve the main distinctive geometric properties of the manifold, and hence better assist the user in data analysis tasks.

When the underlying manifold M has some nontrivial topological structures, these structures cannot be preserved by linear dimension reduction methods. Motivated by this, circular coordinates are proposed (de Silva et al., 2011) to take non-trivial topology of M into account when building the coordinate mappings. The paradigm of circular coordinate representation reveals the intrinsic structure of the high-dimensional data (Wang et al., 2011).

The *circular coordinates* are coordinate mappings with circular values in $S^1 \cong [0, 1]/\{0, 1\}$. The resulting coordinates maps the dataset $X \subset \mathbb{R}^d$ on a k-torus $\mathbb{T}^k = \left(S^1\right)^k$ through coordinates $\Theta = \{\theta_1, \cdots, \theta_k\}$, $\theta_j : \mathbb{R}^d \to S^1$, $j = 1, \cdots, k$. It has been shown that this representation retains significant topological features while reducing topological noise. While circular coordinates preserve the topological structure of the dataset, we also want it to accommodate the sparsity in high dimensional datasets.

Submitted to the Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

In this paper, we propose to impose a generalized penalty for circular coordinates representation, to accommodate the sparsity in the dataset from a theoretic perspective. Simulations and real data examples that show the effect of the choice of penalty function could be found in Luo et al. (2020b).

1.1 Circular coordinate representation

Like standard Topological Data Analysis (TDA) techniques, we approximate the underlying space M by constructing an *approximating complex* Σ, like *Vietoris-Rips complex* or *Čech complex* (Carlsson, 2009). From de Silva et al. (2011), we know that we can compute the associated *persistent cohomology*, and choose an S^1-valued function on Σ, known as the *circular coordinate function*, for each 1-cocycle in the computed persistent cohomology. Intuitively speaking, the circular coordinates are S^1-valued coordinate functions, which reflect the non-trivial topology of the approximating complex Σ. These S^1-valued functions serve as coordinate maps θ in the low-dimensional representation. We use the symbol α to denote a cocycle defined on the underlying complex Σ. The pipeline of the circular coordinate representation can be described as follows:

1. Construct a filtered Vietoris-Rips complex Σ to approximate the underlying space where the dataset X lives.
2. Use persistent cohomology and topological summary to identify those significant 1-cocycles and discard noise.
   e.g. We choose a cocycle $[\alpha] \in H^1(\Sigma, \mathbb{Z}_p)$ with *persistence* greater than a *significant threshold*.
3. For each 1-cocycle, we lift the 1-cocycle $[\alpha]$ into $H^1(\Sigma, \mathbb{Z})$ with integer coefficients.
4. For each 1-cocycle, we replace the integer valued cocycle α by a smoothed cohomologous cocycle $\bar{\alpha}$.
5. For each 1-cocycle, we integrate the function $\bar{\alpha}$ to obtain a corresponding S^1-valued function $\theta : \Sigma \to S^1$.

Remark. It is important to stress that when the $H^1(\Sigma, \mathbb{Z})$ is trivial, or equivalently there is no significant 1-cocycle in the complex with prescribed threshold, the circular coordinate methodology cannot be applied.

Using the terminologies in algebraic topology (Hatcher, 2001), we can describe the theoretical reasoning behind Step 3 in more detail. The chosen cocycle α can be smoothed to obtain a cohomologous cocycle $\bar{\alpha}$ that minimizes L_2 penalty by solving the following *cohomologous optimization problem*

$$\bar{f} = \arg\min_f \{\|\bar{\alpha}\|_{L_2} \mid f \in C^0(\Sigma, \mathbb{R}), \bar{\alpha} = \alpha + \delta_0 f\}. \tag{1}$$

In other words, we are trying to minimize the L_2 norm of a cocycle (function) α within the collection of cohomologous cocycles (functions) and the resulting $\bar{\alpha}$ can be proven to be harmonically smooth.

1.2 Sparsity and penalty functions

Circular coordinates are powerful in visualizing and discovering high-dimensional topological structures (Wang et al., 2011). As a non-linear dimensionality reduction approach, we want to explore its ability to handle challenges from high-dimensional data analysis.

In particular, we are interested in how and when circular coordinates correctly encode the sparsity present in the original high-dimensional data. Sparsity occurs naturally in high-dimensional datasets or due to a sampling scheme, which is often a difficult problem to handle. In presence of sparsity, a good low-dimensional representation of the dataset would have few non-zero coordinates accommodating the sparsity in the original dataset (Vershynin, 2018).

In the regression setting, sparsity is important in discovering the structure of data (Hastie et al., 2015). For a normal linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \sigma^2 > 0,$$

the least square estimates of regression coefficients $\hat{\boldsymbol{\beta}}$ is obtained by solving the *regression optimization problem* $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{L_2}^2$, which leads to the (ordinary) least square theory. When the coefficient $\boldsymbol{\beta}$ is defined in a high-dimensional parameter space, especially when the number of covariates is larger than the sample size, the linear model will encounter problems (Tibshirani et al., 2005).

To address the above problem in high dimensions, the LASSO model (Tibshirani, 1996) makes use of the L_1 norm in the regression optimization problem instead of L_2 norm. The above optimization problem is replaced by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{L_2}^2$ s.t. $\|\boldsymbol{\beta}\|_{L_1} \leq t$ with a predetermined radius $t > 0$. In Lagrangian multiplier forms, the LASSO regression optimization problem can be phrased as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta},\lambda} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{L_2}^2 + \lambda\|\boldsymbol{\beta}\|_{L_1}.$$

The data can be represented with the LASSO model by the covariates, with most regression coefficients being zeros due to the L_1 regularization. An important generalization of the LASSO method is the elastic net method (Zou and Hastie, 2005), which redefines the constraint to avoid including highly correlated covariates, a problem that might arise when using an L_1 regularization. In the elastic net method, the Lagrangian multiplier form of the optimization problem becomes

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta},\lambda} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{L_2}^2 + \lambda\|\boldsymbol{\beta}\|_{L_1} + (1-\lambda)\|\boldsymbol{\beta}\|_{L_2}^2.$$

With these penalized variants of the L_2 regression optimization problem, the regression model can effectively represent the sparse data with few non-zero regression coefficients. However, this representation is still of linear nature and cannot be applied to a non-linear dataset without loss of non-linearity.

We propose to modify the circular coordinate representation to accommodate sparsity by using a generalized penalty function. We will introduce new penalty functions below in Section 2.

2 Generalized penalty for circular coordinate representation

In this section, we explain how the cohomologous optimization problem that arises in the circular coordinates procedure can be solved with generalized penalty functions, which leads to what we call the *generalized penalty for circular coordinates*.

As we previously discussed, circular coordinates can be obtained by solving the *cohomologous optimization problem* (1). When using the L_2 penalty, de Silva et al. (2011) proved that the constructed coordinates possess harmonic smoothness and other well-behaved properties. Usually, on a low-dimensional dataset with significant topological features, this L_2 penalty works well and detect features by showing changes in coordinate values.

To address statistical sparsity in high dimensional datasets, we propose to use a generalized penalty function in the optimization problem (1) to accommodate sparsity. If the sparsity in a high-dimensional dataset is well utilized, the circular coordinates are expected to have mostly constant values and the rapidly changing non-constant is more localized compared to the L_2 penalty. The sparse circular coordinate for a 1-cocycle α will be the solution of the following optimization problem:

$$\bar{f} = \arg\min_{f}\{(1-\lambda)\|\bar{\alpha}\|_{L_1} + \lambda\|\bar{\alpha}\|_{L_2} \mid f \in C^0(\Sigma, \mathbb{R}), \bar{\alpha} = \alpha + \delta_0 f\}. \tag{2}$$

In particular, when $\lambda = 1$, the penalty reduces to L_2 penalty. When $\lambda = 0$, we have the following form using only an L_1 penalty function,

$$\bar{f} = \arg\min_{f}\{\|\bar{\alpha}\|_{L_1} \mid f \in C^0(\Sigma, \mathbb{R}), \bar{\alpha} = \alpha + \delta_0 f\}. \tag{3}$$

Note that these two problems (1) and (3) above, can be formalized as a restrained optimization problem, since coboundary maps δ are linear operators by definition. Although the harmonic smoothness in the resulting coordinates is lost when we use a generalized penalty other than L_2 (i.e., $\lambda \neq 1$), we will show by simulation studies that the topological features can still be preserved in the circular coordinates with few non-constant values.

Based on simulation studies and real data analysis (Luo et al., 2020b), circular coordinates with generalized penalty function may as well provide an informative reference when we are interested in the sampling scheme on high dimensional datasets. In short, under different sampling schemes, the circular coordinates under L_2 penalty functions:

1. would not create qualitative differences in the distribution of coordinate values, and hence the distribution of constant edges.
2. would usually display significant differences in the correlation plots associated with the circular coordinates.

The simulation studies could also provide additional evidence to the claim that the sampling scheme of the dataset is an important factor in TDA (Niyogi et al., 2008; Tausz and Carlsson, 2011). It also brings up a new question that how the topology in the approximating complex Σ could reflect the empirical distribution. In asymptotics, when the sample size $n \to \infty$, we expect that the approximating complex Σ would have the same topology as M; and we also expect that the empirical distribution would converge to the true density. Therefore, it remains an interesting question how these two aspects of the dataset interact (Luo et al., 2020a).

3 Discussion

Our contribution in this paper is that we propose a novel topological dimension reduction method that allows us to take explicitly into account the sparsity in high-dimensional datasets. And we explore the behavior of generalized penalty functions with experiments in Luo et al. (2020b) and show how they can be applied in a non-standard setting.

The circular coordinate (de Silva et al., 2011) is a non-linear dimension reduction method, which is capable of providing a topology-preserved low-dimensional representation of high-dimensional datasets using significant 1-cocycles selected from persistent cohomology based on the dataset.

With a generalized penalty function, the circular coordinate becomes a non-linear dimension reduction method with explicit sparsity control. The circular coordinate representation depends on the penalty function, and the sparsity control is achieved by choosing a generalized penalty function in the cohomologous optimization problem in form of:

$$\bar{f} = \arg\min_{f} \{(1-\lambda)\|\bar{\alpha}\|_{L_p} + \lambda\|\bar{\alpha}\|_{L_q} \mid f \in C^0(\Sigma, \mathbb{R}), \bar{\alpha} = \alpha + \delta_0 f\}. \quad (4)$$

analogous to the usage of generalized penalty function in a standard regression setting as explained in Section 2. In terms of extending our idea of using a different penalty in the smoothing procedure, it would be interesting to explore other kinds of penalty functions already established in a regression or topological setting, for example, fused LASSO (Tibshirani et al., 2005) and "minimal edits" $\|\delta_0\|$. On the other hand, it would also be important to explore the theory behind circular coordinates with a generalized penalty under different sampling schemes.

Beyond the S^1 coordinate functions, it is of interest to explore whether the idea of the penalized smoothing could be extended to coordinate functions with values in a general topological space other than S^1. In this direction, we want to explore the idea of generalized penalty functions with Eilenberg-MacLane coordinates, of which S^1 coordinates is a special case (Polanco and Perea, 2019). This line of research is motivated by TDA literature extending the circular coordinate framework.

As we observed, the computational cost for computing circular coordinates is high. One common way of reducing the computation cost is to use sub-samples instead of full samples in the construction of complexes (Otter et al., 2017). From the perspective of data analysis, such a sub-sampling will introduce more uncertainty and also lose some information. While we know that sub-sampling preserve most topological features in a dataset, it is unclear how other (non-linear) dimension reduction methods behave under a sub-sampling scheme. This line of research aims at exploring how sub-sampling can be utilized in topological dimension reduction tasks, and would be of interest for both statisticians and topologists alike.

Moreover, we know that the real coordinates in classical multi-dimensional scaling have an absolute scale that depends on the particular dataset. Circular coordinates have no absolute scale since their domain is specified to be S^1. The circular coordinates, along with penalty functions, provides algebraically topologically independent circle coordinates. It will be of great practical and theoretical interest to investigate the interaction between algebraic independence and probabilistic independence in multi-dimensional scaling (de Silva et al., 2011).

References

Candès EJ (2014) Mathematics of sparsity (and a few other things). In: Proceedings of the International Congress of Mathematicians, Seoul, South Korea, Citeseer, pp 1–27

Carlsson G (2009) Topology and data. Bulletin of the American Mathematical Society 46(2):255–308

de Silva V, Morozov D, Vejdemo-Johansson M (2011) Persistent cohomology and circular coordinates. Discrete & Computational Geometry 45(4):737–759

Donoho D, Grimes C (2005) New locally linear embedding techniques for high-dimensional data. In: Proceedings of the National Academy of Sciences, vol 100.10, pp 7426–7431

Elad M (2010) Sparse and redundant representations: from theory to applications in signal and image processing. Springer: New York

Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press: Boca Raton

Hatcher A (2001) Algebraic Topology. Cambridge University Press: Cambridge

Luo H, MacEachern S, Peruggia M (2020a) Asymptotics of lower dimensional zero-density regions. arXiv:200602568 pp 1–27

Luo H, Patania A, Kim J, Vejdemo-Johansson M (2020b) Generalized penalty for circular coordinate representation. arXiv:200602554 pp 1–39

Niyogi P, Smale S, Weinberger S (2008) Finding the homology of submanifolds with high confidence from random samples. Discrete & Computational Geometry 39(1-3):419–441

Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA (2017) A roadmap for the computation of persistent homology. EPJ Data Science 6(1):1–38

Polanco L, Perea JA (2019) Coordinatizing data with lens spaces and persistent cohomology. arXiv:190500350

Tausz A, Carlsson G (2011) Applications of zigzag persistence to topological data analysis. arXiv:11083545

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58(1):267–288

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B 67(1):91–108

Vershynin R (2018) High-dimensional probability: An introduction with applications in data science. Cambridge University Press: Cambridge

Wang B, Summa B, Pascucci V, Vejdemo-Johansson M (2011) Branching and circular features in high dimensional data. IEEE Transactions on Visualization and Computer Graphics 17(12):1902–1911

Wilkinson L, Luo H (2020) A distance-preserving matrix sketch. arXiv:200903979 pp 1–44

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B 67(2):301–320