Stitch Fix for Mapper

Bala Krishnamoorthy[1]
Department of Mathematics and Statistics, Washington State University

Nathaniel Saul[1]
Department of Mathematics and Statistics, Washington State University

Bei Wang[1]
School of Computing, Scientific Computing and Imaging Institute, University of Utah

1 Introduction

Mapper is one of the main tools in topological data analysis (TDA) and visualization used for the study of multivariate data [3]. It takes as input a multivariate function and produces a summary of high-dimensional data using a cover of the range space of the function. For a given cover, such a summary converts the mapping into a simplicial complex suitable for data exploration.

In this abstract, we take a *constructive* viewpoint of a multivariate function $f : \mathbb{X} \to \mathbb{R}^d$ and consider it as a vector of continuous, real-valued functions defined on a shared domain, $f = (f_1, f_2, \cdots, f_d)$, $f_i : \mathbb{X} \to \mathbb{R}$, where each f_i (referred to as a *filter function*) gives rise to a 1-dimensional mapper construction. We investigate a method for stitching a pair of mappers together and study a topological notion of information gain from such a process. In particular, we aim to assign a measure that captures information about how each filter function contributes to the topological complexity of the stitched result, and how the two filter functions are topologically correlated.

We are inspired by the ideas of stepwise regression for model selection and of scatterplot matrices for visualization. For a set of variables $x_1, x_2, \dots, x_d$, *stepwise regression* iteratively incorporates variables into a regression model based on some criterion. A measure of topological information gain can be used as a criterion for choosing filter functions, and to construct a single *best* mapper. The *scatterplot matrix* shows all pairwise scatterplots of the set of variables on a single $d \times d$ matrix. We introduce a topological analogue of the scatterplot matrix for a set of filter functions $f_1, f_2, \dots, f_d$, and study the degree of *topological correlation* between filter functions.

We define a composition (or stitching) operation for mappers (Definition 1) and show its equivalence to the standard construction (Theorem 2). We end by describing our ongoing effort in studying a topological notion of information gain and correlation between filter functions.

2 Preliminary Results

Given a space $\mathbb{X}$, a function $f : \mathbb{X} \to \mathbb{R}^d$, and a cover $\mathcal{U} = \{U_i\}$ of $f(\mathbb{X})$, we define the pullback cover $f^*(\mathcal{U})$ as the cover obtained by decomposing each $f^{-1}(U_i)$ into its path-connected components $U_i = \cup_{j=1}^{k_i} V_{ij}$. *Mapper* is then a simplicial complex defined as the nerve of this pullback cover $M(f, \mathcal{U}) := \mathrm{Nrv}(f^*(\mathcal{U}))$.

▶ **Definition 1** (Composition). Given $f, g : \mathbb{X} \to \mathbb{R}$ and covers of their images, $\mathcal{U} = \{U_i\}, \mathcal{V} = \{V_j\}$, we construct a composed cover $\mathcal{W}$ of $\mathbb{X}$ from $f^*(\mathcal{U})$ and $g^*(\mathcal{V})$ by taking the connected

[1] Funding: NSF DBI-1661348 and NSF DBI-1661375

components of the following set, where $u \in f^*(\mathcal{U})$, $v \in g^*(\mathcal{V})$ are path-connected cover elements of $\mathbb{X}$,

$$\{u \cap v \mid \forall u \in f^*(\mathcal{U}), \forall v \in g^*(\mathcal{V}), u \cap v \neq \emptyset\}.$$

We define the composed mapper as the nerve of this cover $\mathcal{W}$,

$$S(M(f,\mathcal{U}), M(g,\mathcal{V})) := \mathrm{Nrv}(\mathcal{W}).$$

Under certain assumptions, this composition S is equivalent to the classical method of constructing mappers from a pair of filter functions, as described by Theorem 2.

▶ **Theorem 2.** *If f and g are continuous real-valued functions, U_i, V_j, and $U_i \times V_j$ are simply connected for all i, j, then $S(M(f,\mathcal{U}), M(g,\mathcal{V})) = M((f,g),\mathcal{U} \times \mathcal{V})$, the mapper constructed in the traditional manner.*

Proof sketch. The proof follows directly from properties of continuous functions and connected sets. We provide a sketch here. Starting with the two covers associated with the two 1-dimensional mappers, $\mathcal{U}$ for f and $\mathcal{V}$ for g, we can show that the defined set $\mathcal{W}$ and the cover obtained from the traditional mapper construction are equivalent. Taking the nerve of each, we can conclude that the resulting mapper are equivalent as well. ◀

Furthermore, we give an algorithm that illustrates how the composition can be considerably simplified by directly incorporating simplex information from each of the two input mappers. The algorithm that combines (or stitches) two mappers together works by tracking vertices (i.e. representatives of the pull back cover elements as a result of the Nrv operation) of the first mapper in a breadth first search fashion, and combining them with vertices of the second mapper. The simplices in both mappers provide hints about which possible simplices could be in the composition. Using this information to avoid many explicit intersection checks, we can considerably simplify and speedup the composition process. While some simplicies from each 1-dimensional mapper can be added directly to the composition (the `stitch` step), others require explicitly checking the nerve condition in the mapper construction (the `fix` step).

3 Discussion

As part of our ongoing research, we propose measures of information gain (i.e., the increase in topological complexity) from the composition process as well as topological correlations between pairs of filter functions. By tracking the `stitch` and `fix` steps of the construction process, it is possible to quantify the relationship between filter functions.

With such measures in hand, we return to our topological analogues of the stepwise regression [1] and scatter plot matrix [2], which help to navigate topological relationships among multiple filter functions. A method for *stepwise stitching* would produce a mapper with optimal topological information by iteratively building a multi-dimensional mapper from topologically independent filter functions. A *topological scatter plot matrix* can reveal information about the filter functions such as topological dependencies and outliers by providing a visualization of the most information rich filter functions.

References

1 M.A. Efroymson. Multiple regression analysis. In Anthony Ralston and Herbert S. Wilf, editors, *Mathematical Methods for Digital Computers*. Wiley, 1960.

2 Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6), 2008.

3 Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3D object recognition. *Eurographics Symposium on Point-Based Graphics*, 22, 2007.