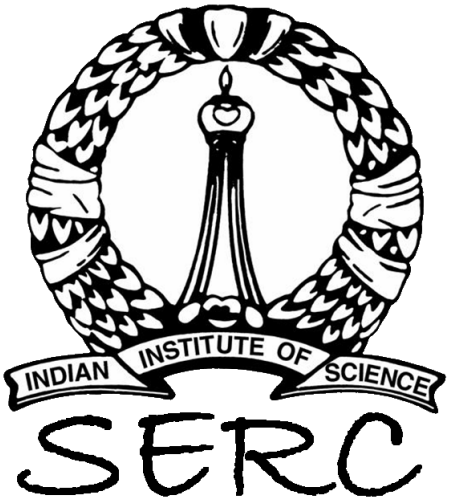
Term-Project

SE294: Data Analysis and Visualization

Morse Smale Regression

Students:
Satish Kumar
Vijay Kumar

Instructors:
Vijay Natarajan
R. Venkatesh Babu
Partha Talukdar



Supercomputer Education and Research Centre
Indian Institute of Science, Bengaluru

Jan-Apr, 2015

Abstract

Work in this project is built upon the results of Gerber et al.[Journal of Computational and Graphical Statistics,22(1),2013] to exploit Morse-Smale(MS) complex for partition based regression. **msr**, an R package, already contains routine to perform linear model approximations over Segments obtained from MS Complex. In this project, We have studied higher order model approximation over segments. We have also implemented quadratic model version of MS Regression. A comparative study of our implementation with the linear model have been done on few higher dimensional data sets from real world.

Acknowledgements

I want to thank Dr.Vijay Natrajan for his guidance to me throughout the project. His unconditional help, stimulating suggestions and encouragement helped me in all the time of my work.

Contents

1 Introduction **1**
- 1.1 Partition Based Regression . . . 1
- 1.2 Morse Smale Complex . . . 1
  - 1.2.1 Persistence . . . 2
- 1.3 Literature Reviews . . . 2

2 Problem Formulation **3**
- 2.1 Higher order regression . . . 3
- 2.2 Case Study . . . 3

3 Methodology **5**
- 3.1 Persistence Selection and Partition Prediction . . . 5
- 3.2 Regression model . . . 6
- 3.3 msr: a package in R . . . 6

4 Results and Analysis **7**
- 4.1 Results . . . 7
- 4.2 Analysis . . . 7

5 Conclusion **10**

Bibliography **11**

List of Figures

1.1 Illustration of Morse Smale Complex 2
1.2 Illustaration of persistance 2

4.1 Cells obtained from Morse Smale Complex 8
4.2 2D Scalar Function and Persistance Plot 8
4.3 Linear vs Quadratice 8
4.4 Regression Curve for UCI Concrete data set 9

Chapter 1

Introduction

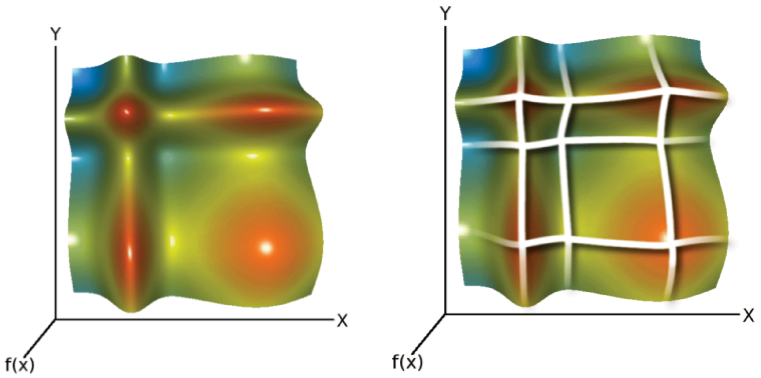
Recent advancement in computational topology have led to many approaches to perform analysis and visualization of data sets based on their topological properties. One such approach is Partition based regression for regression analysis onto multivariate data.

1.1 Partition Based Regression

In Partition based regression, the independent variable is partitioned into various segments, and a separate regression model is fitted on each segments. This approach of regression provides a flexible trade-off between the simplicity and interpret-ability of parametric models and the predictive capabilities of non-parametric methods of regression analysis. Most important aspect of this approach is the quality of the segmentation itself. Various partitioning technique have been used in past e.g. Regression tree and principal Hessian direction (PHD) tree. In this project, segmentation of the domain induced by the Morse-Smale complex have been employed for partitioning.

1.2 Morse Smale Complex

Morse Smale(MS) Complex provides topologically meaningful decomposition of domain. Idea is to decompose(partition) the domain (independent variables) into regions of uniform gradient flow such that the interior of each sub-domains (segments) does not contain any critical points. The decomposition leads to various monotonic sub-domain (i.e. Single maximum and single minimum withing sub-domain). Figure 1.1 illustrates the MS Complex over 2D scalar function.



(A) 2D Scalar Function (B) MS Complex Segments

FIGURE 1.1: Illustration of Morse Smale Complex

1.2.1 Persistence

Persistence is a measure of the amount of change in the function value f required to remove a critical point, and thus, to merge two or more partitions. Figure 1.2 illustrates the idea using 1D scalar function. Note that, Quality of partition depends on persistence value chosen. Since, with increase in persistence value no of partitions decreases.Thus, persistence introduces a notion of scale at which the Morse-Smale complex of f is considered.

1.3 Literature Reviews

Gerber et. al. [1] in 2013, introduced the Morse Smale regression that incorporates topological information. For partitioning, a discrete approximation of the Morse-Smale complex for data sets, as implemented in *msr* package on R, have been used. For fitting the model, they used linear model. Also, they did a comparative analysis of topological accuracy with other partitioning techniques.

Gerber et. al. [2] in 2012, presents the R package *msr* for exploratory data analysis and visualization of multivariate scalar functions based on the Morse-Smale complex.

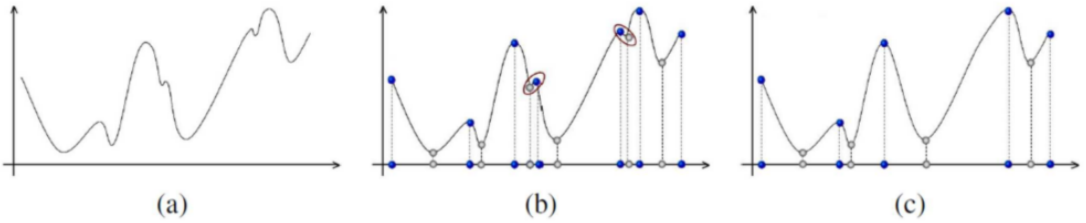


FIGURE 1.2: Illustaration of persistance

Chapter 2

Problem Formulation

Morse Smale complex has been proven as a state of the art technique of partitioning [1]. Gerber et. al. [1] already showed the advantages of using Morse-Smale Complex based regression approach, but their work involved fitting linear curve onto each segments.

2.1 Higher order regression

Need of higher order regression arises when linear model doesn't fit quite well withing sub-domains. One reason can be due to larger sub-domain with non-linear data nature. Although Going for higher dimensional model is not recommended while doing partition based regression, exploring the quadratic and cubic order of approximation can be a reasonable option, atleast for data sets with larger sub-domain.

While doing higher order regression, Important thing to take care is that fitted model should be monotonic within each partitions. Adding this another constraints to least square problem will suffer in RMSE on train data, but expectation is that it will should work well with the test data.

2.2 Case Study

To validate the performance of our model, We have worked with few higher dimensional scalar data sets. Data sets used for reporting the result is as follows:

- **fourpeaks:** This is a 2D data generated from analytical function given below:

$$f(x) = \frac{1}{2}(e^{x_1-\frac{1}{4})^2/0.3^2} + e^{x_2-\frac{1}{4})^2/0.3^2} + e^{x_1-\frac{3}{4})^2/0.1^2} + e^{x_2-\frac{3}{4})^2/0.1^2})$$

The function has 4 maxima and 9 minima. A program to generate fourpeaks data is already available in R. We are using 5000 data points for reporting the result.

- **camera_estimation:** It is a 9D data of size 10,000 generated from an Energy Function of a Camera Estimation Problem as described in [2].
- **gaussian:** It is a data of size 10,000 generated from a mixture of four 2D Gaussian, generated randomly. To generate the data, we are using a program called Gauss-Droper, made in the Visualization and Graphics lab (VGL) of IISc, Bengaluru.
- **UCI Concrete:** It is a real data taken from UCI Machine Learning data repository. The data set contains 1030 instances of 8D data. Here, Compression strength of concrete is the target.

For evaluation of performance, we have split the labeled data available equally into training and test data and evaluated RMSE for predicted test data.

Chapter 3

Methodology

Morse Smale regression consists of two parts: 1. computing Morse Smale partitions and 2. fitting a regression model for each partitions. For part 1., we are going to use algorithms as proposed in paper [1] by Gerber et. al. For testing the data, first for each data, we have to predict which partition it belongs to and then predict the output accordingly.

3.1 Persistence Selection and Partition Prediction

Correct choice of persistence is very essential for quality decomposition of domain. Though persistence value depends on scale of the data output, To choose correct persistence value without much knowledge of scale, Idea is to perform Hierarchical Partitioning based on different persistence level and choose a persistence level such that going beyond that level requires a larger change in persistence value.

For partition prediction on test data, Idea is to employ support vector machine (SVM) to build a multi-class classifier using the one-against-one [1]. The SVM is trained on the partition assignments of the Morse-Smale complex. To estimate partition probabilities from the classifier, a logistic regression is fitted to the decision values of the SVM.

For each test data, we get probability of the data belonging to each of the partition obtained previously. Then One can have two options for predicting output: 1. Using model from the partition having highest probability (Hard bound) or 2. Use weighted sum from each model, with weight as probability obtained (Soft bound). We have used Hard bound of data, since It is amenable to interpretation.

3.2 Regression model

Model chosen for performing regression on each partition is as follows:

$$\tilde{f}_l = a_0 + \sum_{i=1}^{p} b_i x_i + \sum_{i=1}^{p} \sum_{j=i}^{p} c_{ij} x_i x_j \quad \forall x \in C_l$$

Here, x_i denotes i^{th} coordinate of data x, $\tilde{f}_l$ denotes predicted output for partition l and p denotes dimension of independent variables. Note that, even for quadratic model, no. of parameters is $1 + p + \frac{p * (1 + p)}{2}$.

3.3 msr: a package in R

R-software is being used for implementation, computation and visualization. We have used *msr* package of R [2] for Computing Morse Smale complex. Description of various routines can be found in the documentation [2].

Chapter 4

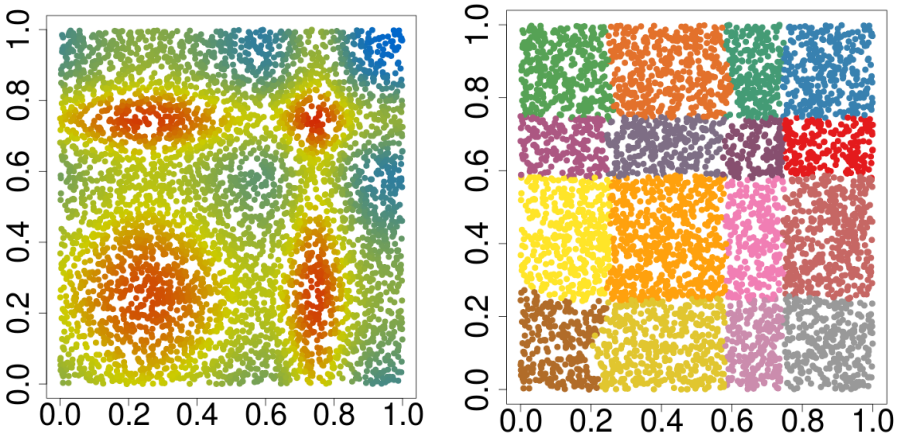
Results and Analysis

4.1 Results

RMSE values obtained for each of the data are as follows:

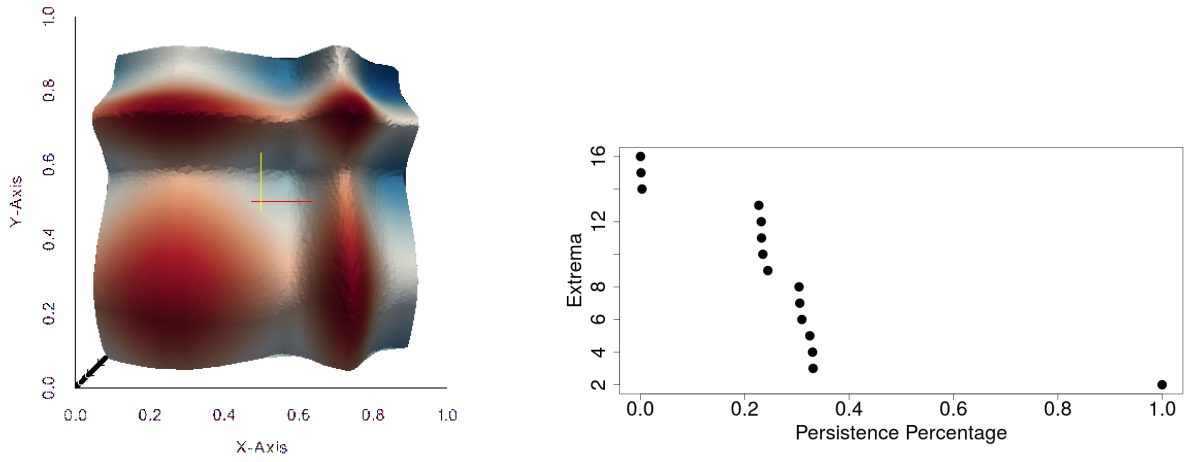
	fourpeaks	Gaussian	Camera Estimation	UCI Concrete
Linear	0.04310	0.34807	0.09013	0.06392
Quadratic	0.01957	0.10031	0.05033	0.05843

Figures shown below are for the fourpeaks data set. Here, Figure 4.1 shows visualization of 2D scalar data, and Morse smale cells generated. Figure 4.2 contains 2 plots. Figure 4.2a shows 3D visualization of fourpeaks data-sets. Figure 4.2b shows persistence plot obtained while doing Hierarchical Morse Smale computations. Also, Figure 4.3 shows a visual analysis of linear and quadratic model fitted for fourpeaks data.



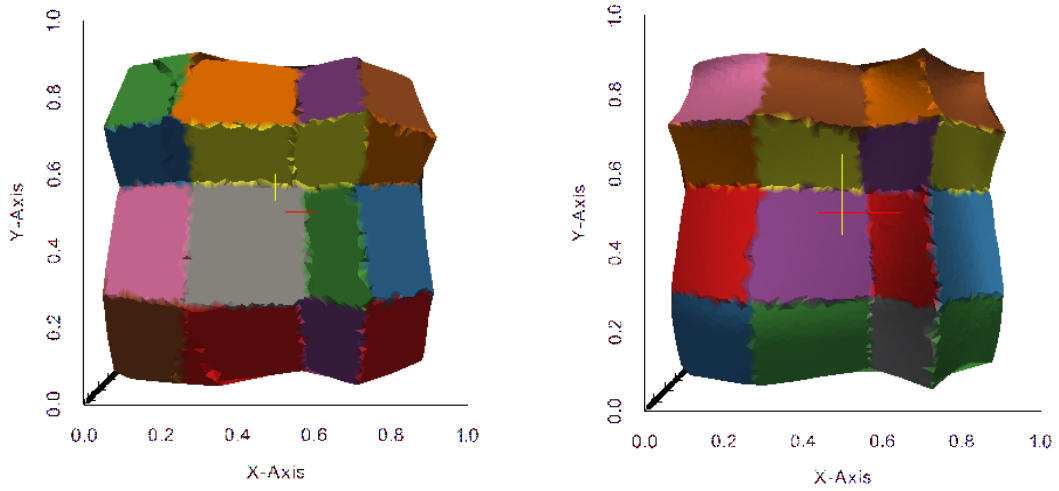
(A) 2D Scalar Function (B) Morse Smale Complex

FIGURE 4.1: Cells obtained from Morse Smale Complex



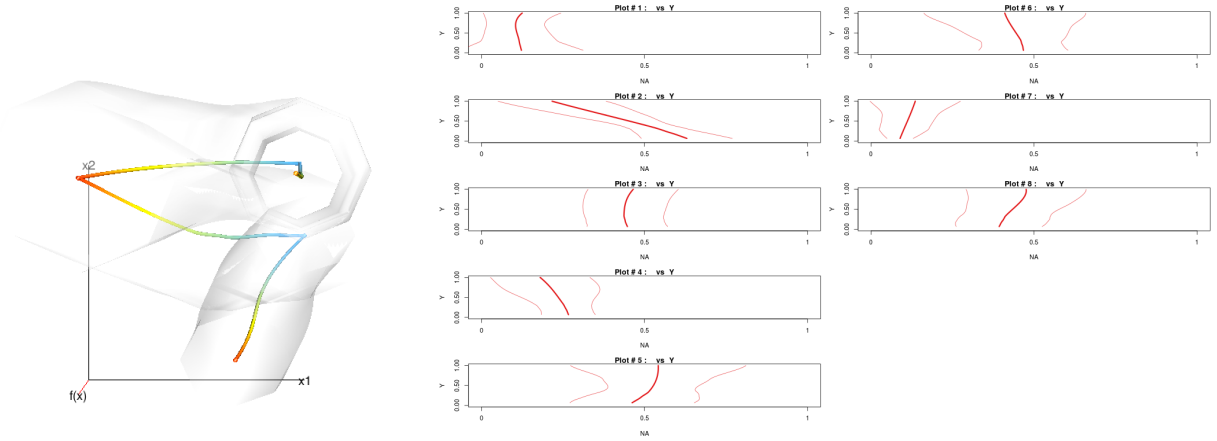
(A) 2D Scalar Function (B) Persistance Plot

FIGURE 4.2: 2D Scalar Function and Persistance Plot



(A) Linear Model Fit (B) Quadratic Model fit

FIGURE 4.3: Linear vs Quadratice



(A) Regression Curve (B) 2D curve for one partition

FIGURE 4.4: Regression Curve for UCI Concrete data set

For visualization of high dimensional data, Regression curve as described in [3] is being used. For UCI concrete data set, Regression curve is as shown in the Figure 4.4.

4.2 Analysis

RMSE values obtained shows that quadratic model works much better that linear model.

Also, From the Figure 4.3, We can see that, due to choosing hard bound for partition prediction, data on partitions boundary doesn't fits quite smoothly for both the cases.

Note that, Making smooth curve at boundary may affect accuracy of prediction.Also, We are doing partition based regression where we intend to exploit the distinct feature of each partition separately.

From plot of persistence level Figure 4.2b, We can see that there is a high jump in persistence value between level 3 and level 4. So, level 3 will be a good choice of persistence level for analysis. Obviously, with more higher level, persistence value increase significantly, but that leads to merging the cells.

Chapter 5

Conclusion

If data set is having non-linear nature, and Partitions obtained are larger in size, It is better to use quadratic model.

Choice of persistence value is very critical to the accuracy of analysis. Hence, Hierarchical MS Complex can be used for current persistence value.

Going beyond quadratic model for high dimensional data, increases computational complexity significantly.

Bibliography

[1] Samuel Gerber, Oliver Rübel, Peer-Timo Bremer, Valerio Pascucci, and Ross T Whitaker. **Morse–Smale Regression**. *Journal of Computational and Graphical Statistics*, 22(1):193–214, 2013. URL http://www.ncbi.nlm.nih.gov/pubmed/23687424.

[2] Samuel Gerber and Kristin Potter. **Data Analysis with the Morse-Smale Complex: The msr Package for R**. *Journal of Statistical Software*, 50(2):1–22, 7 2012. ISSN 1548-7660. URL http://www.jstatsoft.org/v50/i02.

[3] S. Gerber, P. Bremer, V. Pascucci, and R. Whitaker. **Visual Exploration of High Dimensional Scalar Functions**. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1271–1280, Nov 2010. ISSN 1077-2626. doi: 10.1109/TVCG.2010.213. URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5613467&tag=1.