Estimation of the Mach Number Using Persistent Homology and Regression Analysis

Basil Ahmad
Rutgers University
New Brunswick, New Jersey
ba205@scarletmail.rutgers.edu

ABSTRACT

In [1], the phenomenon of the impact between photoelastic particles and granular material is studied. It is shown that pertinent properties of the system, namely the speed and the spatial structure of the forces just after impact, depend on a dimensionless constant known as the mach number. It is our goal to find a novel way to estimate this physical constant using techniques from algebraic topology and machine learning.

We are given data in the form of videos of live experiments where the impact is observed. Using a mathematical technique known as persistent homology, we can transform each video into a collection of four graphs known as persistence diagrams. These graphs give insight into the creation of new loops between the forces, as well as new connected components made by the granular material as time passes, which we hope will capture the essence of the nonlinear structure of these experiments.

From our persistence diagrams, we obtain 18 possible summary statistics based off of the mean, median, and mode of the following values (per frame of the video): number of loops, number of connected components, birth time of loops, birth time of connected components, death time of loops, death time of connected components. Using lasso regression, we find that the only features correlated with the mach number include the mean, median, and mode of the number of loops and connected components per frame of the video. Afterwards, we compare lasso, ridge, and kernel ridge on this truncated dataset, and analyze the average mean squared error of each of these respective methods over 100000 different splits of the dataset into testing and training data.

Keywords

Mach Number; Regression; Persistence Homology; Persistence Diagrams

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

1. INTRODUCTION

In [1], the authors ran experiments where an intruding particle made contact with granular material. The forces which propagated were nonlinear and complex, and it was hypothesized that the speed and the spatial structure of these forces is reflected in a dimensionless quantity called the mach number, denoted M'.

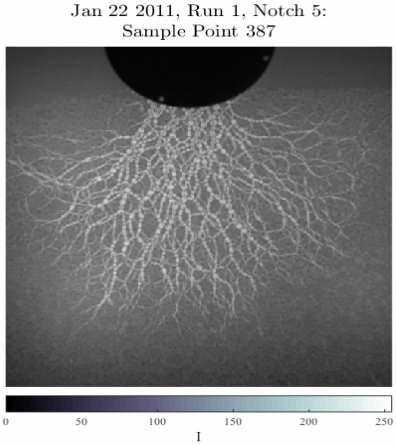
Whereas the physicists tried to estimate M' experimentally by using a proxy P: the fraction of grains beneath the intruder exhibiting a strong force. Here, we will try and estimate M' using persistent homology and regression techniques. Persistent homology will be used to extract meaningful information from videos of the experiments, with which we can engineer features suitable for a regression. Afterward, we will try various regression methods and compare their accuracy on different test sets with different training data.

Specifically for this project, we were only working with "soft particles," and the intruder was dropped from higher and higher heights, increasing the velocity. In all, I had 14 experiments over 9 different velocity profiles to study.

2. METHODOLOGY

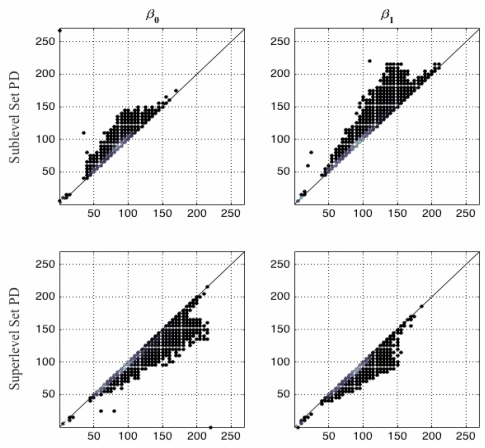
2.1 Persistent Homology

As noted previously, the force propagation is highly nonlinear, as you can see from the following image. Visually, we



A snapshot from an experiment video.

can see how the forces turn in on themselves forming "loops," as well as closing off parts of the granular medium into islands or "connected components." Persistent homology is a technique which takes each frame in a video and captures all of the "loop" and "connected component" information and represents them in a mathematical object called persistence diagrams, as shown below. Each point on the β_0 diagrams



The four Persistence Diagrams generated from the previous image.

denotes a connected component generated in a single frame of video, whereas each point on the β_1 diagrams denotes a loop. The $x-$axis denotes birth time, whereas the $y-$axis denotes death time. Now, we have a concrete representation of our visual interpretation of the experiment data.

2.2 Feature Engineering

Persistent homology gives us the number of loops and connected components per frame for each experimental video given, which is valuable insight, but still unsuited for Regression, because there's no inherent ordering to be made in the Birth-Death plane. As a result, we need to think of any possible useful summary statistics to estimate M' based off this information. To this end, I have selected the following 18 features as prime suspects:

- Mean, Median, and Mode of the number of connected components per frame in the video
- Mean, Median, and Mode of the number of loops per frame in the video
- Mean, Median, and Mode of the birth times of connected components per frame in the video
- Mean, Median, and Mode of the birth times of loops per frame in the video
- Mean, Median, and Mode of the death times of connected components per frame in the video
- Mean, Median, and Mode of the death times of loops per frame in the video

Thus, we have a well-posed supervised learning problem with input features from $\mathbb{R}^{18}$, and output features being $M' \in \mathbb{R}$.

2.3 Regression

As known in the literature, Lasso is a very good method for feature selection, so I first applied Lasso to our new well-posed supervised learning problem to do this. I found that the coefficients for anything that did not have to do with the number of connected components and the number of loops went straight to zero, that is, any summary statistic having to do with the birth and death rates were irrelevant.

Thus, our dataset essentially now has vectors in $\mathbb{R}^6$ as the features, with each entry corresponding to the mean, median, or mode of the the number of connected components/loops per frame. Then, I split this simplified dataset into test and training sets randomly, 100000 times for each regression I was about to try: Lasso, Ridge, and Kernel Ridge. For each split of test and training data, I performed each regression and calculated it's mean squared error using the following formula:

$$\frac{1}{2}\sum_{i=1}^{m}(f(x_i) - M_i')^2$$

Here, f denotes the hypothesis function gained from the particular regression. I took the averages of the mean squared error over the $100,000$ splits for each regression, and came up with the following results:

- Ridge: 8.1895946665003923
- Kernel Ridge: 0.95602784938954766
- Lasso: 0.13364179997462949

Thus, it seems Lasso does the best on average for predicting M', which is surprising, because that suggests that the relationship is still very close to linear.

3. CONCLUSIONS/FUTURE DIRECTIONS

Future directions for this project include handling more varied types of particles, and handling more data in general, as this project required 9 GB of data on my home machine for 14 experiments. Another interesting insight is whether we can cut down on features and find a better regression model for M'.

One thing is certain: the number of connected components and loops created in the system definitely correlates with M' in some way. While Lasso does pretty well from a prediction standpoint, I'm confident that there is more to this relationship that can be explored in future studies.

4. ACKNOWLEDGMENTS

I would like to acknowledge Professor Awasthi, for giving me a solid background in Machine Learning to attempt the regression part of this project. I would also like to thank Professor Konstantin Mischaikow for teaching me about Computational Topology and giving me the idea to use it in this project, as well as Rachel Levanger, for helping me with all of the small stuff that comes with these types of project, while keeping me focused on the big picture.

I would also like to acknowledge thank the Physicists at Duke for lending their experimental video data to my Computational Topology course so I could even begin tackling

anything, as well as the creators of Scikit-learn for creating a robust open-source Regression library.

APPENDIX

A. REFERENCES

[1] Abram H. Clark, Alec J. Petersen, Lou Kondic, and Robert P. Behringer. *Nonlinear Force Propagation During Granular Impact.* http://arxiv.org/pdf/1408.1971v2.pdf

[2] Tomasz Kaczynski, Konstantin Mischaikow, and Marian Mrozek. *Computational Homology.* Applied Mathematical Sciences 157, Springer-Verlag, 2004.

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning.* Data Mining, Inference, and Prediction, Second Edition.

[4] Christopher Bishop *Pattern Recognition and Machine Learning.* Information Science and Statistics, 2006.