Observing the Sun: From start to finish.

PhD dissertation by

Pablo Santamarina Guerrero

Instituto de Astrofísica de Andalucia (IAA-CSIC)

Programa de Doctorado en Física y Matemáticas (FisyMat)
Universidad de Granada

A thesis submitted in fulfillment
of the requirements of the degree of
Doctor of Philosophy

March 14, 2024

PhD thesis supervised by

Dr. David Orozco Suárez
Dr. Julián Blanco Rodríguez

Acknowledgements

dkshfgsidgfisgdfisdfs Agradecimientos

Resumen

Resumen de la tesis

Summary

Summary of the thesis

CONTENTS

CHAPTER 1

INTRODUCTION

1.1   Motivation of our work

DESIGNING AND CALLIBRATING AN INSTRUMENT.

2.1 Spectropolarimeters: TuMag's design

2.2 Calibration of TuMag

CHAPTER 3

OPERATION AND DATA REDUCTION.

3.1   TuMag's Observation modes and data reduction.

3.2   Challenges in data reduction: Etalon Cavity Map.

<center>

CHAPTER 4

</center>

<center>

SCIENTIFIC EXPLOITATION.

</center>

4.1 Persistent Homology in Solar Magnetograms.

4.1.1 Persistent Homology

Persistent homology stands out as a prominent technique within the topological data analysis toolkit, primarily for its capacity to capture the shape and distribution information of a dataset. The algorithm is rooted in the mathematical framework of homology groups. In topology, these groups measure the number of n-dimensional holes in a data set, or in other words, the number of connected components for a 0th dimensional analysis, holes or rings for a 1st dimensional analysis, spherical voids for the 2nd dimensional analysis, and so on.

The primary objective of persistent homology is not only to compute the homology groups of a given dataset but also to study how they vary at different scales. To achieve this, the input data undergoes a process of division into a series of sequential subspaces, with each subspace encompassing the previous one. This sequential process, known as filtration, begins with a starting subspace comprising a single point from the original dataset. Subsequent subspaces are then constructed by incrementally adding points to the previous subspace until the final subspace includes all points of the original dataset.

After the filtration process is performed, persistent homology algorithms shift their focus to analyzing the evolution of topological features across the different subspaces. Specifically, they record the filtration value at which a new feature appears, meaning that it is absent in the previous subspace, and when it disappears, meaning that it is no longer present in the following subspaces. These two events are known as the birth and death of a topological feature, respectively.

In a nutshell, the n-dimensional persistent homology of a dataset with a given filtration can be described as the aggregation of all n-dimensional features (homology groups) that were created (birth) and subsequently eliminated (death) during the filtration process (Hensel et al., 2021).

When applying persistent homology on a greyscale image, our focus lies in filtering the data according to the pixel values. Multiple filtering approaches exist, with the most extended ones being sublevel and superlevel filtrations, both based on the concept of thresholding. In these filtrations, the image is cropped to a specific value, forming a subspace that

<center>7</center>

includes all pixels with values higher than this value in a superlevel filtration, or lower in a sublevel filtration. This cropping value (i.e. the filtration value) is systematically varied from the lowest to the highest values of the image, or vice versa, thus generating a different subspace for each value. As a result, the persistence homology analysis captures and examines the evolution of topological features across different thresholds, enabling insights into the image's structural properties at various scales (Barnes et al., 2021).

A more formal way of defining these filtrations can be done by considering an image as a discrete representation of a function f, defined over a two-dimensional space $\mathbb{X}$, such that:

$$f : \mathbb{X} \longrightarrow \mathbb{R} \ . \tag{4.1}$$

Let $\mathbb{S}_\phi$ be the subspace of $\mathbb{X}$ for a filtration value of ϕ. In such a case, a filtration can be expressed as:

$$\mathbb{X} : \mathbb{S}_{\phi_0} \subset \mathbb{S}_{\phi_1} \subset \mathbb{S}_{\phi_2} \subset ... \subset \mathbb{X} \ . \tag{4.2}$$

With this formulation, a topological feature with birth-death coordinates:

$$(B, D) = (\phi_I, \phi_{II}) \ , \tag{4.3}$$

corresponds to a feature that appears for the first time during the filtration process at the subspace $\mathbb{S}_{\phi_I}$, and *persists* until the subspace $\mathbb{S}_{\phi_{II}}$, where it ceases to exist.

In a sublevel filtration, each subspace can be expressed as:

$$\mathbb{S}_\phi = f^{-1}\left((-\infty, \phi]\right) \ , \tag{4.4}$$

where ϕ_0 is selected as the lowest value for any given pixel and its value is increased until the subspace includes all pixels. On the contrary, in a superlevel filtration, the subspaces can be expressed as:

$$\mathbb{S}_\phi = f^{-1}\left([\phi, \infty)\right) \ , \tag{4.5}$$

where ϕ_0 is selected as the highest value for any given pixel and its value is decreased along the filtration.

Various methods exist for representing the information derived from a persistent homology analysis, including Betti numbers, persistence bars, and persistent diagrams (PDs) (Cohen-Steiner et al. 2005, Aktas et al. 2019), among many others. For this study, we will utilize the PDs as our chosen approach due to their straightforward interpretation and extended use. A n-dimensional PD is a multiset of Birth-Death pairs, (B_i, D_j), with multiplicity k, where each pair measures the number (k) of n-dimensional components that have been born at the filtration subspace $\mathbb{X}_i$ and died in $\mathbb{X}_j$, that is usually represented in a 2D scatter plot.

The process of generating a PD of a greyscale image is as follows. We start by selecting the filtration direction (sublevel or superlevel) and the dimension of the analysis (either 0 or 1). We initialize a threshold as the highest or lowest value from the image, depending on the choice of filtration. We then perform the filtration by systematically adjusting the threshold and creating a binary image for each threshold. This process divides the image into two sets: pixels with values above the threshold and pixels with values below it. The
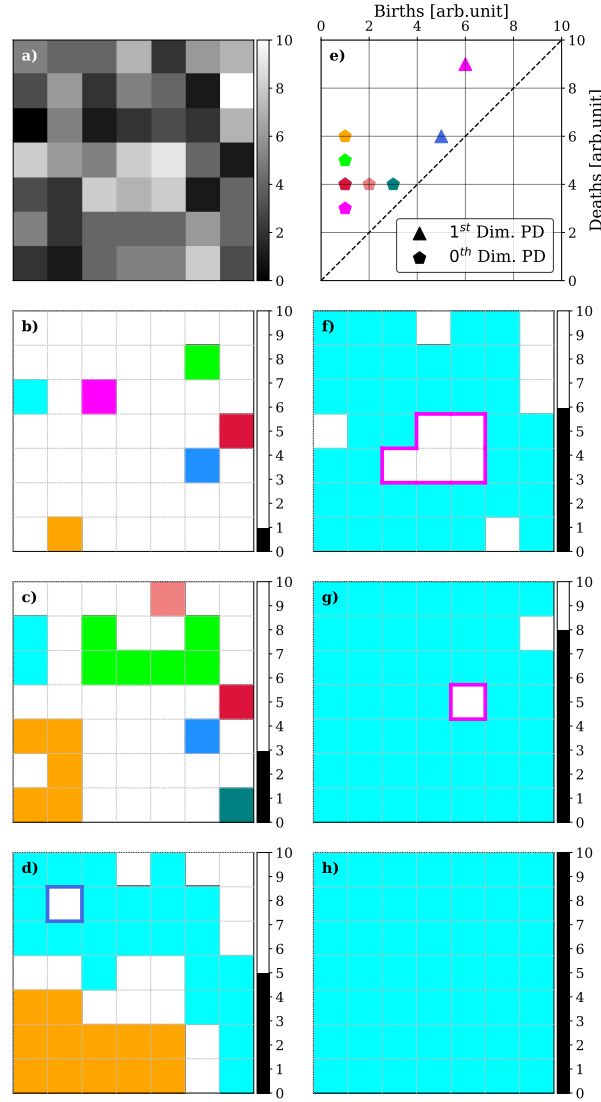
Figure 4.1 Sublevel filtration of a greyscale image and PDs of the 0th and 1st dimensions. Panel a) shows the input data. In panels b), c), d), f), g), and h) different snapshots of the filtration process are shown. The value of the filtration parameter, ϕ is shown in the color bar at the right of each image. Only pixels with a value lower than the filtration value (colored pixels) belong to the subspace shown in each snapshot. Different homology groups are represented with different colors at each snapshot. For connected components (0th dimensional homology groups) the whole pixel is shown with the corresponding color. For rings, (1st dimensional homology groups), only the border of each hole is colored. In panel e) the PDs of both dimensions are shown. The color of each point in the diagram is the same as the one used to plot the corresponding topological feature in other panels. An animation depicting the whole filtration process is also provided as part of the article.

choice of filtering determines which of the two sets makes up the subspace. We then look for the existing topological features within each of these subdivisions. The specific process by which these features are identified is detailed in the next paragraph, where the structures corresponding to both dimensions are illustrated using the example shown in Fig. 4.1. We repeat this process until the threshold reaches the opposite limit to that from which it started. Along this process, we follow the appearance, merging, and disappearance of connected components. When two components merge, the longer-lived one (*i.e.* the first to appear along the filtration process) absorbs the younger one, thus resulting in the death of the second (Edelsbrunner & Harer, 2022). We determine the birth and death for each component based on the thresholds at which these events occur. Finally, we construct a scatter plot where the horizontal axis represents the birth values and the vertical axis represents the death values. Each point on this plot corresponds to a persistence point, whose coordinates reveal the scales at which the corresponding topological feature is present.

An example of this process with a sublevel filtration is shown in Fig. 4.1. Panel a) displays the input data, panels b), c), d), f), g), and h) show some of the key steps of the filtration process, and, lastly, panel e) displays the PD for a 0th and 1st dimensional analyses. These plots illustrate how connected components and rings are born and then die as we increase the filtration level. As these components (shown in different colors) increase in size and come into contact with other components, one absorbs the other, thus resulting in the death of the second. This phenomenon is shown in panels b) to d), where we observe the progression of the components until only the blue and orange connected components remain. Additionally, the diagrams also reveal the appearance of two rings in the data (panels f) and g)). These rings are found when pixels that do not belong to the subspace are surrounded by a connected component, and die when those pixels are included in the component as the threshold increases (blue ring in panel f)). Finally, the PD (panel e)) displays the birth and death values (i.e. the filtration value) of all the features, of dimensions 0 and 1, that have been identified (birth) and subsequently eliminated (death) throughout the filtration process.

4.1.2 Persistent Images

The PD displayed in Fig. 4.1 contains only a limited number of points due to the simplicity of the input image. However, when analyzing real data, these diagrams can consist of hundreds or even thousands of birth-death pairs with high multiplicities, simply due to the size of the images. Additionally, features not only representing the genuine behavior of the data but also reflecting the distribution of noise appear on the diagrams. To address this complexity, several strategies have been developed to simplify the information from PDs, such as persistence curves (Chung & Lawson, 2019), persistence landscapes (Bubenik et al., 2015), or persistence images (PI) (Adams et al., 2017). In this study, we will focus on the latter, due to its noise filtering capabilities and because the representation of the results remains in a Birth-Death diagram, allowing for easy interpretation of the results, similar to a persistence diagram.

PIs are a condensed form of a persistence diagram, offering a concise and easy-to-
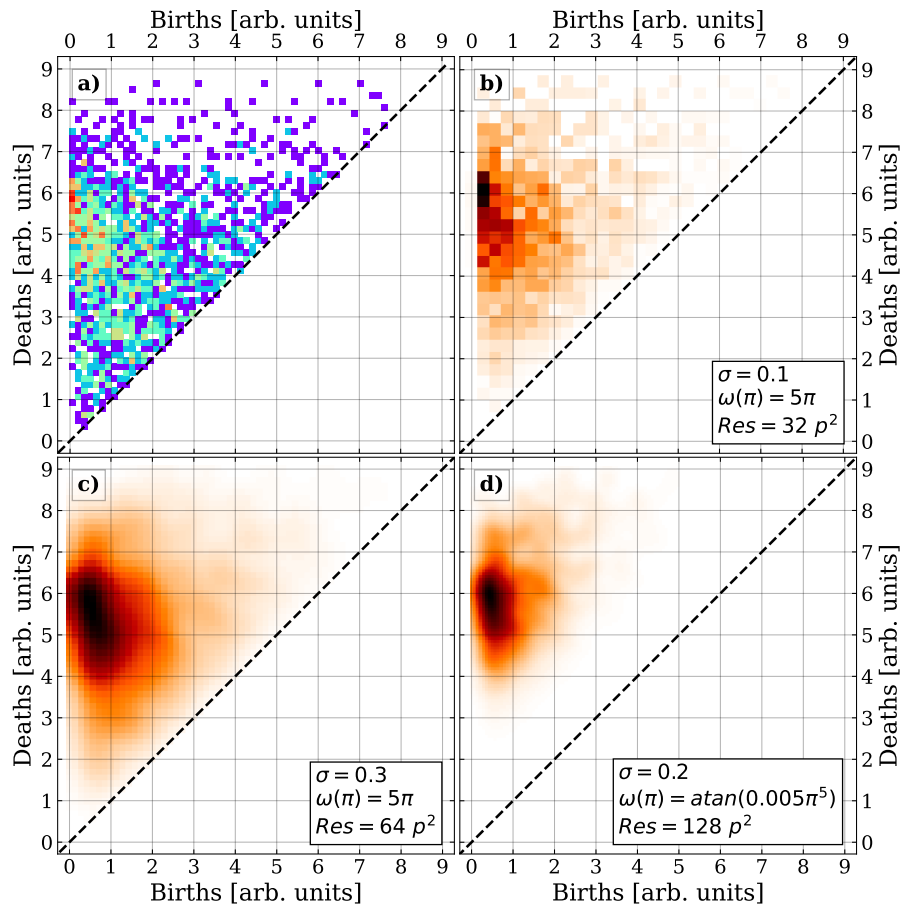
Figure 4.2 Panel a) shows an example of a persistence diagram as a 2D histogram. The color of each bin represents its multiplicity, with the red spots corresponding to higher values. Panels b), c), and d) show three different examples of PIs. The three parameters given in the legends of the PIs are: the standard deviation (σ) of the Gaussian kernel ($K(z)$), the weighting function, and the resolution of the image PI in pixels (p).

understand representation of its topological features. They capture the spatial distribution and persistence information of these features, allowing for the enhancement of the most relevant ones and filtering of the others. A PI is constructed using the concept of persistence. Each topological feature, represented by a point in a PD, has a persistence, π, of:

$$\pi = D - B \tag{4.6}$$

where (B, D) are the corresponding birth-death coordinates in the diagram. A feature with a large persistence is present at different scales in the data and therefore is more likely to represent the real behavior of the data. On the contrary, short-lived features are typically associated with the noise distribution and usually do not provide much information about the data.

When constructing a PI, a weighting function, $\omega(\pi)$, is employed to assign weights to

each point in the diagram, ensuring that longer-lived features have greater weights than shorter-lived ones. There are multiple choices for the shape of the weighting function, which are entirely dependent on the aims of the study and data type. The simplest example is often a linear or power-law relation ($\omega(\pi) = a\pi^{b}$), where a and b can be tuned to assign progressively higher weights to higher persistencies, thus focusing the study on the longer-lived components. On the other hand, if the objective is to filter out noise while assigning similar weights to all non-noise points so that all points have a similar relevance in the analysis, the chosen function is usually an arc-tangent.

The PI is then generated by dividing the persistence diagram plane into a grid with a desired resolution. Within each grid region (or pixel), the weighted features of the diagram within the region are added up using a kernel density estimation. The kernel function, $K(z)$, can be tuned to suit the nature and objectives of the analysis, with Gaussian functions being the most common approach.

The resultant PI is a 2D matrix, wherein each pixel corresponds to a specific area in the persistence diagram, and its value represents the cumulative weight of the topological features found within that area. In Figure 4.2, three examples of PI (panels b), c), and d)) are presented for the same persistence diagram (panel a)), where distinct choices of resolution, kernel function, and weighting function have been applied to each image.

All the PDs, PIs, and the rest of the analysis tools presented in this work, have been computed using the Homcloud python package (Obayashi et al., 2022).

Chapter 5

Summary and conclusions

The conclusions are …

Appendix A

Profile derivatvies

(Hale et al., 1919)

Bibliography

Adams, H., Emerson, T., Kirby, M., et al. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research* **18**, 1 (2017).

Aktas, M. E., Akbas, E., & Fatmaoui, A. E. Persistence homology of networks: methods and applications. *Applied Network Science* **4**, 1 (2019).

Barnes, D., Polanco, L., & Perea, J. A. A comparative study of machine learning methods for persistence diagrams. *Frontiers in Artificial Intelligence* **4**, 681174 (2021).

Bubenik, P., et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16**, 77 (2015).

Chung, Y., & Lawson, A. Persistence curves: a canonical framework for summarizing persistence diagrams. arxiv. *arXiv preprint arXiv:1904.07768* (2019).

Cohen-Steiner, D., Edelsbrunner, H., & Harer, J. Stability of persistence diagrams. *Proceedings of the twenty-first annual symposium on Computational geometry*, 263–271 (2005).

Edelsbrunner, H., & Harer, J. L. (2022), Computational topology: an introduction (American Mathematical Society).

Hale, G. E., Ellerman, F., Nicholson, S. B., & Joy, A. H. The magnetic polarity of sun-spots. *Astrophysical Journal, vol. 49, p. 153* **49**, 153 (1919).

Hensel, F., Moor, M., & Rieck, B. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence* **4**, 681108 (2021).

Obayashi, I., Nakamura, T., & Hiraoka, Y. Persistent homology analysis for materials research and persistent homology software: Homcloud. *journal of the physical society of japan* **91**, 091013 (2022).
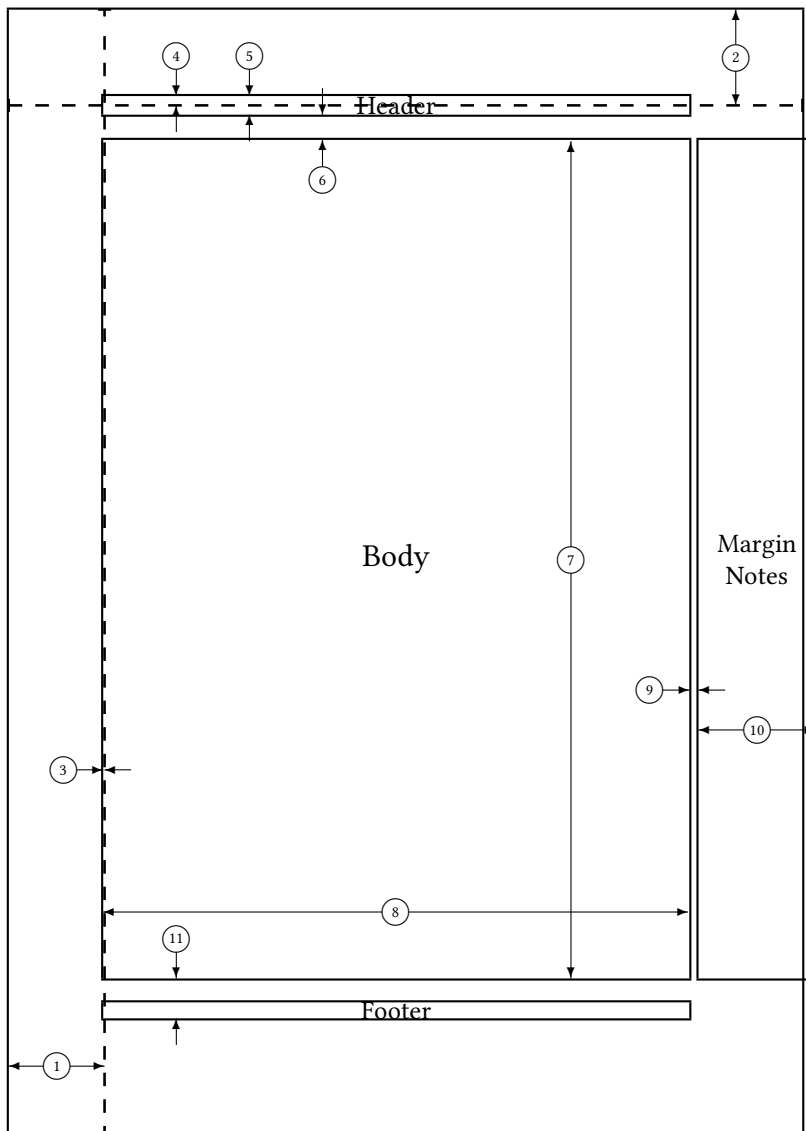
```
 1   one inch + \hoffset        2   one inch + \voffset
 3   \oddsidemargin = -1pt      4   \topmargin = -7pt
 5   \headheight = 14pt         6   \headsep = 19pt
 7   \textheight = 631pt        8   \textwidth = 441pt
 9   \marginparsep = 7pt       10   \marginparwidth = 88pt
11   \footskip = 30pt               \marginparpush = 7pt (not shown)
     \hoffset = 0pt                 \voffset = 0pt
     \paperwidth = 597pt            \paperheight = 845pt
```
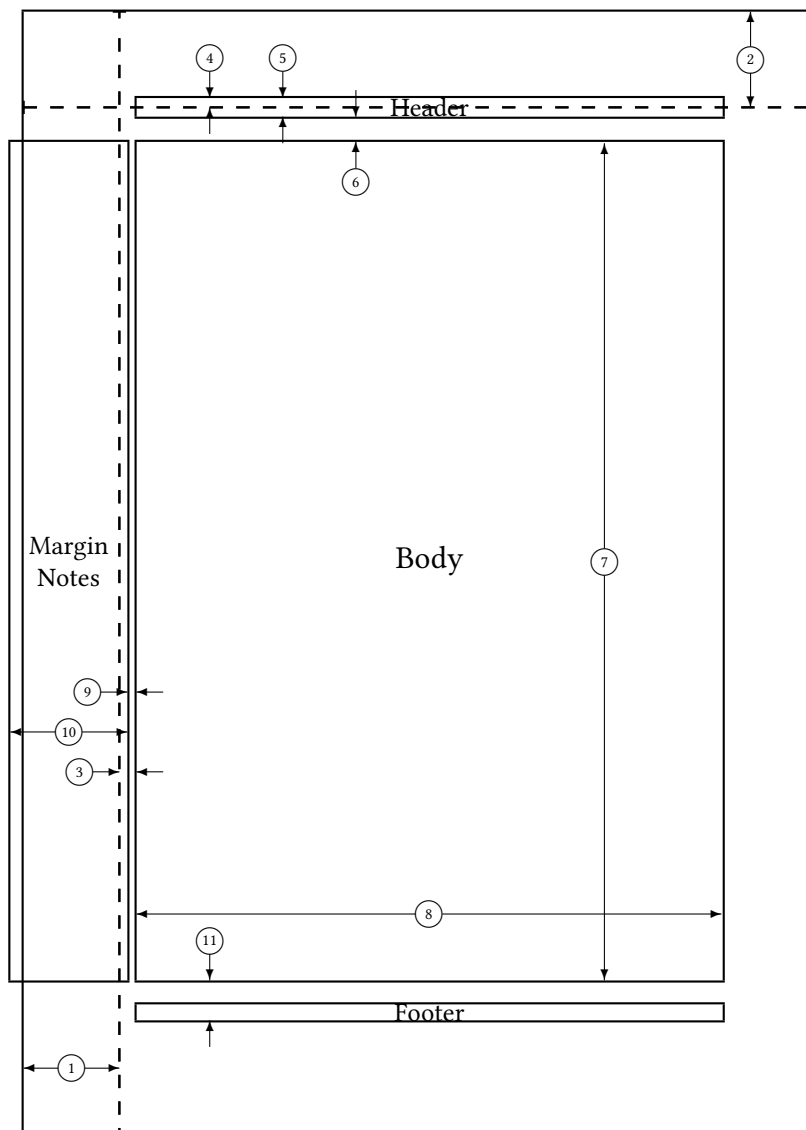
1	one inch + \hoffset	2	one inch + \voffset
3	\evensidemargin = 13pt	4	\topmargin = -7pt
5	\headheight = 14pt	6	\headsep = 19pt
7	\textheight = 631pt	8	\textwidth = 441pt
9	\marginparsep = 7pt	10	\marginparwidth = 88pt
11	\footskip = 30pt		\marginparpush = 7pt (not shown)
	\hoffset = 0pt		\voffset = 0pt
	\paperwidth = 597pt		\paperheight = 845pt