Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.167

Bump hunting by topological data analysis

Max Sommerfeld[a], Giseon Heo[b], Peter Kim[c], Stephen T. Rush[d] and J. S. Marron[e]*

Received 13 September 2017; Accepted 24 September 2017

A topological data analysis approach is taken to the challenging problem of finding and validating the statistical significance of local modes in a data set. As with the SIgnificance of the ZERo (SiZer) approach to this problem, statistical inference is performed in a multi-scale way, that is, across bandwidths. The key contribution is a two-parameter approach to the persistent homology representation. For each kernel bandwidth, a sub-level set filtration of the resulting kernel density estimate is computed. Inference based on the resulting persistence diagram indicates statistical significance of modes. It is seen through a simulated example, and by analysis of the famous Hidalgo stamps data, that the new method has more statistical power for finding bumps than SiZer. Copyright © 2017 John Wiley & Sons, Ltd.

Keywords: bootstrap; kernel density estimation; mode hunting; persistent homology; SiZer

1 Introduction

A long-standing and intuitively appealing challenge in the field of exploratory data analysis was aptly termed *bump hunting* by Good & Gaskins (1980). That is the task of identifying statistically significant peaks in an estimated probability density function. This task is important because finding unexpected peaks can lead to the discovery of new scientific phenomena. Statistical significance of discovered bumps is an important component, to avoid wasted effort on investigating spurious peaks that eventually turn out to be mere artefacts of sampling variation.

While the simple histogram remains the most commonly used density estimator, it is seen for example in Section 2.2 of Silverman (1986) and Section 3.2.7 of Scott (2015) to be very slippery for bump hunting applications. As noted in those monographs, and also the monographs Devroye & Gyorfi (1985), Wand & Jones (1994) and Simonoff (2012), an intuitively more appealing approach to density estimation is the *kernel density estimate* (KDE). Given a *kernel function* K, which integrates to 1, and a *bandwidth* $h > 0$, the KDE for a set of data $X_1, \ldots, X_n$ is

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h (x - X_i) ,$$

where $K_h(\bullet) = \frac{1}{h} K\left(\frac{\bullet}{h}\right)$. As discussed in the aforementioned monographs, the main behaviour of the estimate is driven by the bandwidth h, sometimes called the window width, as it determines the critical amount of local averaging, that

[a]Felix Bernstein Institute for Mathematical Statistics in the Biosciences, University of Göttingen, Göttingen 37077, Germany
[b]School of Dentistry, University of Alberta, Edmonton, Alberta T6G 2R7, Canada
[c]Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario N1G 2W1, Canada
[d]School of Medical Sciences, Örebro Universitet, Örebro SE-701 82, Sweden
[e]Department of Statistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
*Email: marron@unc.edu

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.167

The ISI's Journal for the Rapid Dissemination of Statistics Research

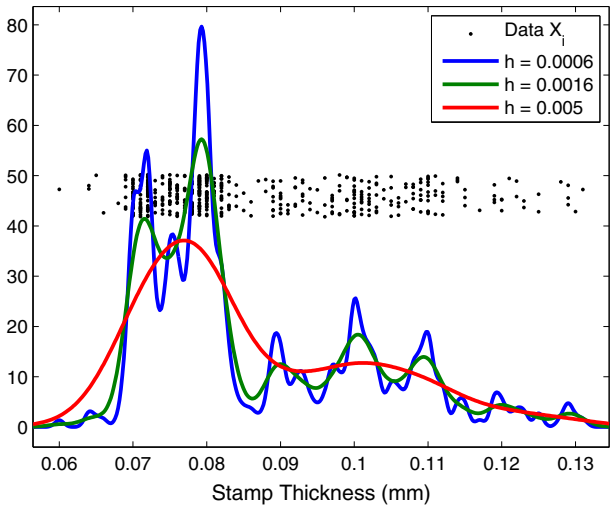


Figure 1. Three kernel density estimates of the Hidalgo stamp thickness data, using quite different bandwidths. Raw data are shown as black dots. Kernel density estimates reveal different types of structure at different scales.

is, *level of smoothing*. The kernel function K is relatively important, although it should integrate to 1 to make $\hat{f}_h(x)$ a reasonable probability density estimate (e.g. to integrate to 1). In this paper, we take K to be standard Gaussian, because of its *variation diminishing* property. As carefully discussed by Chaudhuri & Marron (2000), this means that the number of modes of the estimate is a decreasing function of the bandwidth h.

Figure 1 shows some KDEs of the intriguing and widely studied Hidalgo stamp data, brought to the statistical literature by Izenman & Sommer (1988). The data $X_1, \ldots, X_n$ are thicknesses (in mm) of $n = 485$ postage stamps produced in Mexico during the nineteenth century. The paper thicknesses of the stamps vary widely, motivating philatelists to question the number of paper sources. Each stamp thickness is shown as a black dot in Figure 1, with stamp thickness on the horizontal axis, and a random value (called jitter) to visually separate the dots on the vertical axis. The jitter plot already suggests clumps of high density, that is, clusters in the data, thought to correspond to separate factories producing the paper. This impression is sharpened by the three KDEs, using the bandwidths indicated in the legend. The red KDE with $h = 0.005$ shows just two modes, that is, suggests that there are two factories. A bootstrap conclusion of two modes in this data set can be found in Section 16.5 of Efron & Tibshirani (1994). The green KDE with $h = 0.0016$ represents substantially less smoothing (i.e. local averaging) and suggests that there may be seven modes, although the two on the right are very small and may include too few data points to conclude they are actual clusters. Analyses indicating seven modes include Izenman & Sommer (1988), Basford et al. (1997) and Fisher & Marron (2001). An even less smooth KDE is the blue curve using $h = 0.006$, which shows quite a few more modes, many of which seem likely to be spurious artefacts of sampling variation. However, some researchers have been interested in at least some of these, such as the mode that appears between the two largest modes. In particular, Minnotte & Scott (1993) suggested that there may be up to 10 modes. Other approaches have led to other answers for this data; for example, three modes indicated by Walther (2002), three to five by Chaudhuri & Marron (1999) and five by Minnotte (2010).

A time-honoured approach to the challenge of the divergent answers to the question of which modes represent important underlying structure is data based *bandwidth selection*. As noted by Jones et al. (1996b, 1996a), there is a large literature on this topic. For the Stamps data, the Sheather–Jones Plug-In Bandwidth recommended by Jones

(wileyonlinelibrary.com) DOI: 10.1002/sta4.167

et al. (1996a) is $h = 0.0012$, somewhere between the green and blue curves. As the green $h = 0.0016$ bandwidth is already turning up questionable modes, classical data-based bandwidth solution is not very effective in the context of bump hunting. This is perhaps not surprising, as this approach attempts to optimize an L^2 norm, which is a rather different goal from finding modes.

This type of consideration motivated the SiZer approach to bump hunting proposed by Chaudhuri & Marron (1999). The first major contribution of the SiZer idea is to skirt the bandwidth selection problem by taking a *scale space* approach. Scale space is a concept from computer vision based on extracting information from a digital image using a family of Gaussian kernel smooths. The idea is that features from more smoothed versions represent coarse-scale macroscopic aspects of the image, while smaller windows correspond to taking a more-fine-scale detailed view. Note that from this perspective, it does not make sense to choose a single bandwidth, but instead, all scales contain different types of useful information, which should not be discarded. This motivated SiZer, which is a multi-scale view of data combined with relevant statistical inference. A SiZer analysis of the Hidalgo stamp data is shown in Figure 2. The top panel includes the jitter plot black dots from Figure 1, together with a scale space family of KDEs shown as blue curves. Note that the range of smooths starts with even less smoothing than the blue curve in Figure 1 and ranges through more smoothing than the red curve, in particular including a uni-modal member. The bandwidths of this family are logarithmically equally spaced, which gives a good visual impression, because bandwidths work in a multiplicative way. The second major contribution of SiZer is to focus the statistical inference on bump hunting, through the observation that a bump is determined by a region of increase, followed by a region of decrease. Hence, the inference is based on statistical SIgnificance of the ZERo crossings of the derivative of the density estimate.

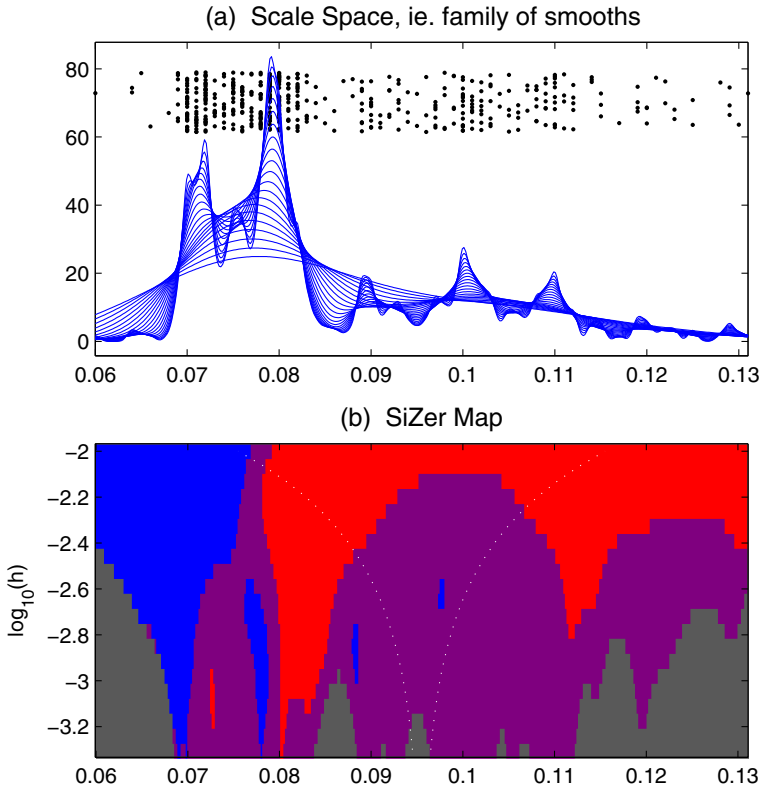


Figure 2. Sizer analysis of the Hidalgo stamp data. Relatively few of the modes are found to be statistically significant.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.167

The ISI's Journal for the Rapid Dissemination of Statistics Research

In particular, at each location, a hypothesis test is performed, and where the slope is significantly positive (negative), the colour blue (red, respectively) is shown. That original colour choice was based on what is natural for economic data (where the terminology "in the red" is associated with decrease). In other applications, for example, in climatology (e.g. Holmström & Erästö (2002) and Erästö & Holmström (2007, 2012)), it is natural to reverse the colours (as red is thought of as "hotter," with blue corresponding to "cooler"). At locations where the test is inconclusive, the intermediate colour of purple is used, and grey appears where the data are too sparse for reliable inference. This is performed across a scale, with the results summarized in the SiZer map shown in the bottom panel of Figure 2. The horizontal axis is the same as that for the top panel, for direct comparison. The vertical axis indexes the $\log_{10}$scale, that is, bandwidth. This shows that the two largest peaks of the green smooth in Figure 1 are strongly statistically significant (flanked by both blue and red patches). Of the next three largest peaks, only the first and second receive some SiZer support in the form of blue patches to the left, but these are not called as significant modes, because there are no corresponding red regions to the right. Specific insights into how much smoothing is represented at each scale comes from the white dots, which show ± 2 standard deviations of the Gaussian kernel window. The top rows of the SiZer map show that at the coarsest scale, the kernel smooth is uni-modal, while the colours around $\log_{10} h = -2.6$ show that the stamp data set is bimodal at that level of resolution.

The third major contribution of the SiZer idea is to skirt the traditional bias problems of kernel smoothing by ignoring them. This makes sense from the scale space viewpoint by changing the target of the statistical inference from the underlying density to the *density at the given scale*, that is, to the convolution of the density with the kernel. See Hannig & Marron (2006) for an improved version of SiZer (currently implemented in the examples shown here), where some early distributional approximations have been made precise. A two-dimensional version is available by Godtliebsen et al. (2002).

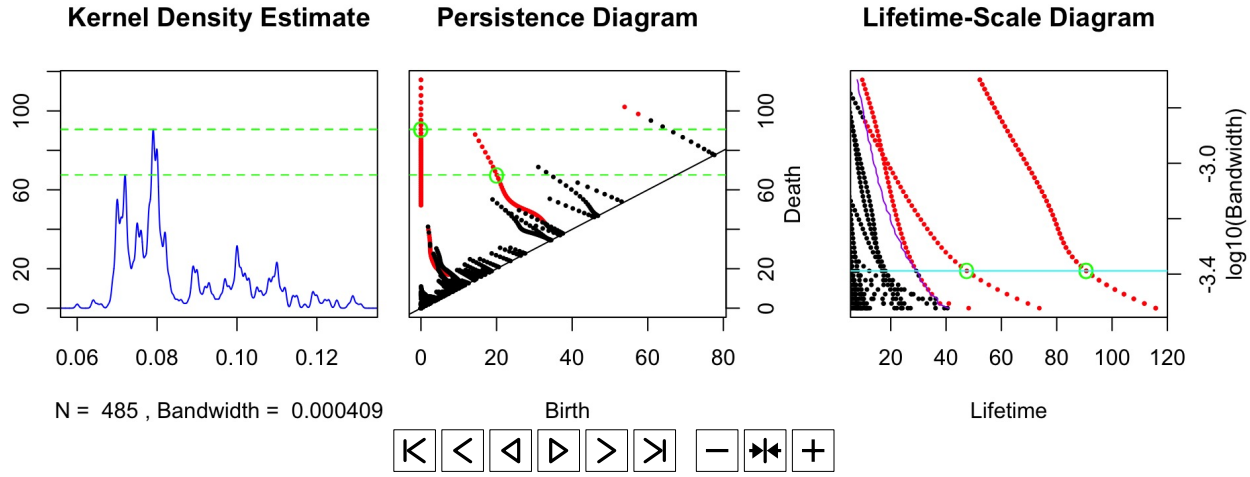
2 Methods

2.1 Persistent homology

For a sufficiently regular function $f : \mathbb{R}^d \to \mathbb{R}$, *persistent homology* (Edelsbrunner & Harer, 2008) tracks the change of topological *features* such as connected components, holes and voids of *sub-level sets* $L_\epsilon := \{x \in \mathbb{R}^d : f(x) \leq \epsilon\}$ as the level parameter $\epsilon \in \mathbb{R}$ changes. We give a brief review of the main ideas here and refer to the growing literature for details (Bubenik & Kim, 2007; Ghrist, 2008; Carlsson & Zomorodian, 2009; Carlsson, 2009; Chazal et al., 2011).

As sub-level sets are nested, that is, $L_{\epsilon_1} \subset L_{\epsilon_2}$ for $\epsilon_1 \leq \epsilon_2$, there is a natural correspondence between the topological features of L_{ϵ_1} and those of L_{ϵ_2}. As the level parameter ϵ grows, new features can appear and existing features disappear as, for example, connected components merge or holes are filled in. Each feature can be assigned a smallest ϵ for which it appears and a largest ϵ after which it is no longer present—these levels are called the *birth* and *death* times of the feature, respectively. Naturally, the difference between death and birth time of a feature is called its *lifetime*. This is typically depicted in a *persistence diagram* where each feature is represented by a point with its birth and death times as its horizontal and vertical coordinates, respectively.

Bump hunting with persistent homology Persistent homology can be used for hunting bumps of a univariate continuous probability density function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$. To this end, we track the connected components of the sub-level sets of f. New connected components will be born when the level ϵ reaches the height of a local minimum of f; likewise, the death times of features (as above meaning the open intervals that comprise the complement of L_ϵ) will correspond to the height of local maxima. As f is a density function, it approaches zero for $x \to \pm\infty$. Therefore, every sub-level set for positive ϵ will have usually two unbounded connected components. These two components will always have birth time $\epsilon = 0$, and one of them dies when ϵ is the global maximum of f. All other connected components are born and



Movie 1. Topological data analysis-based mode hunting for the Hidalgo stamp data.

die between those extremes. We may ignore one of these unbounded components because it will always be born at $\epsilon = 0$, lives for all positive ϵ, and hence carries no information about f. With this modification, each feature in the persistence diagram corresponds to a bump in the density f—namely, the local maximum at which it dies. Other work on using persistence in bump hunting can be found in Xia et al. (2015a, 2015b).

2.2 Significance

Stability theorem The stability theorem (Cohen-Steiner et al., 2007) provides the foundation for statistical inference with persistent homology. It asserts that if a function f has k features with lifetime greater or equal to $C > 0$, then any function g with $\sup_x |f(x) - g(x)| < C$ has at least k features (actually, the stability theorem says more than that, but this simplified version suffices for our purpose).

Significance of bumps Assume now we are hunting for bumps of a density f of a univariate random variable X based on a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$. We propose to consider the persistence diagrams of the KDE $\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$ as an estimator for the persistence diagram of the smoothed true density $f_h = K_h * f$.

The stability theorem allows us to decide whether a feature found in the persistence diagram of $\hat{f}_h$ is significant or not. Indeed, let $q_h(\alpha)$ be the $1 - \alpha$ quantile of the random variable

$$\sup_{x \in \mathbb{R}} |\hat{f}_h(x) - f_h(x)|.$$

In practice, we determine $q_h(\alpha)$ from the data by bootstrapping. It follows from the stability theorem that, with probability $1 - \alpha$, the number of features of $f_h(x)$ is at least the number of features of $\hat{f}_h$ whose lifetime exceeds $q_h(\alpha)$.

This approach to assessing the significance of features in persistence diagrams has been first proposed by Fasy et al. (2014) but has not been used for mode hunting yet.

Scale-lifetime diagrams An important contribution of this paper is the *dynamic scale-lifetime diagram* shown in Movie 1. Time in this movie represents scale, indexed by the bandwidth h. It is useful to both watch the full movie and to stop it at particularly interesting values of h, with manual advancement through frames.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.167

The ISI's Journal for the Rapid Dissemination of Statistics Research

For each scale h, the left panel shows the corresponding KDE as a blue curve. Also shown are horizontal green lines that correspond to peaks, that are statistically significant in the sense that the corresponding feature (recall this is the gap between connected components in L_ϵ) has a lifetime that is $> q_h(\alpha)$. Behaviour across scales is displayed in the centre and left panels.

The centre panel shows an aggregation of conventional persistence diagrams across scales. The distinct patterns are a result of the strong connections between diagrams across scales. At each scale h, each peak generates a dot. The vertical coordinate of each dot is the height of the feature corresponding to the peak, because that is when the peak dies in the ϵ filtration. The horizontal coordinate of each point is the birth time of the feature (gap) beneath each peak. Points are coloured red at peaks that are statistically significant in the aforementioned sense. At the smallest bandwidth of $h = 0.0003$, there are three peaks that touch horizontal green lines in the left panel. Each of these persist through a wide range of the ε filtration. These appear as the three red dots circled in green in the centre panel, and note that the heights are the same as that in the left panel, because the height of the dot is its death time. All other peaks appear as black dots whose vertical coordinate is the height of the peak. As can be understood by slowly changing the scale using the right arrow key (>), the nearly linear patterns show the motion of each dot (corresponding to peaks) over bandwidth. The upper right set of points stops being significant already at the third value of h because it corresponds to the very thin second peak, which completely disappears at the bandwidth $h = 0.0045$. The tallest peak, at stamp thickness around 0.08, generates the set of points on the far left, because this component is born at $\epsilon = 0$. All of these features are statistically significant over all shown scales h. The second bandwidth $h = 0.0031$ is also interesting, because that is the finest scale where the peak near stamp thickness 0.10 is statistically significant. Following the pattern downwards through scale shows that this peak is intermittently significant.

The right panel provides additional insights, especially in regard to the issue of statistical significance. This shows roughly the same set of dots, but now displayed with log bandwidth (scale) on the vertical axis, and lifetime (which is the difference of vertical and horizontal coordinates in the centre panel) on the horizontal axis. To help make the needed visual connection, the points with green circles in the centre panel also have green circles in the right. Because scale is on the vertical axis, these green circle all have the same height, which increases with bandwidth h. Looking along each row, one sees the number of both significant (red) and insignificant (black) peaks at that scale, except that very small peaks with a lifetime of less than about 5 (i.e. within five units of the diagonal in the centre panel) are not shown. Good insight into how the statistical significance works across scale is given by the purple curve, which traces out $q_h(\alpha)$ as a function of h. This gives a good impression as to why the significance of the peak near thickness 0.01 is rather intermittent while the two major peaks appear over a wide range of scales.

Our final contribution to the oft considered issue of how many modes are in the stamp data is that at the level of resolution $h = 0.0031$ there are four statistically significant modes, which are substantially different from those found in previous analyses.

3 Power comparison with SiZer

To demonstrate improved statistical power, the Hidalgo stamp data from Figures 1 and 2 are analysed here using persistent homology. Movie 1 shows the results of this analysis. Using the same range of bandwidths as the SiZer, the topological data analysis (TDA) method finds up to four significant modes for small bandwidths, and three bumps are marked as significant over nearly the entire bandwidth range. The TDA method thus finds more significant modes than does SiZer and finds them to be significant over a larger range of bandwidths.

Another approach demonstrating the improved statistical power is to compare performance on interesting simulated data sets. Figure 3 shows a SiZer analysis of $n = 450$ data points simulated from the Marron & Wand (1992)

(wileyonlinelibrary.com) DOI: 10.1002/sta4.167

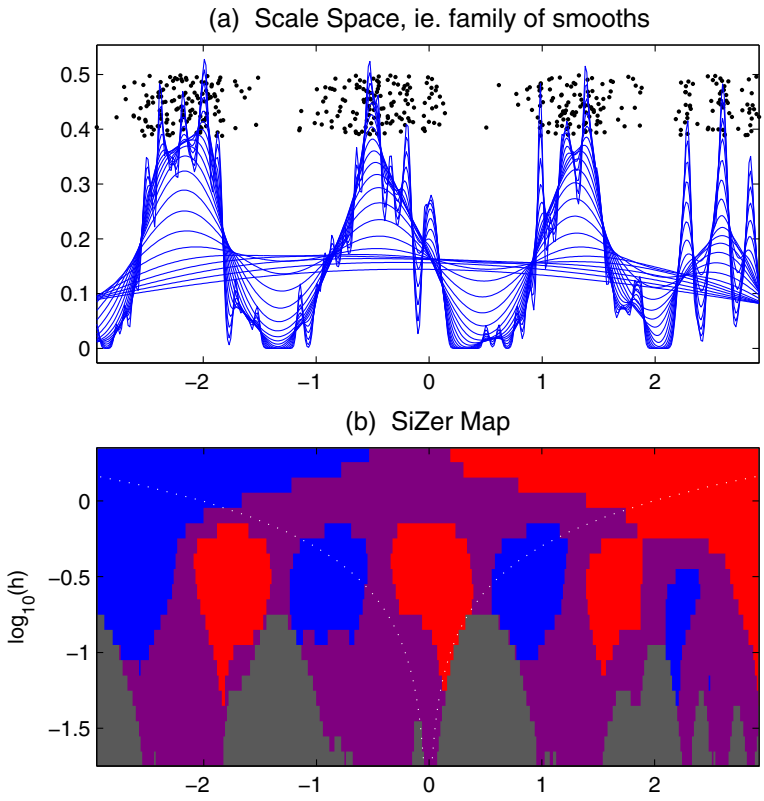
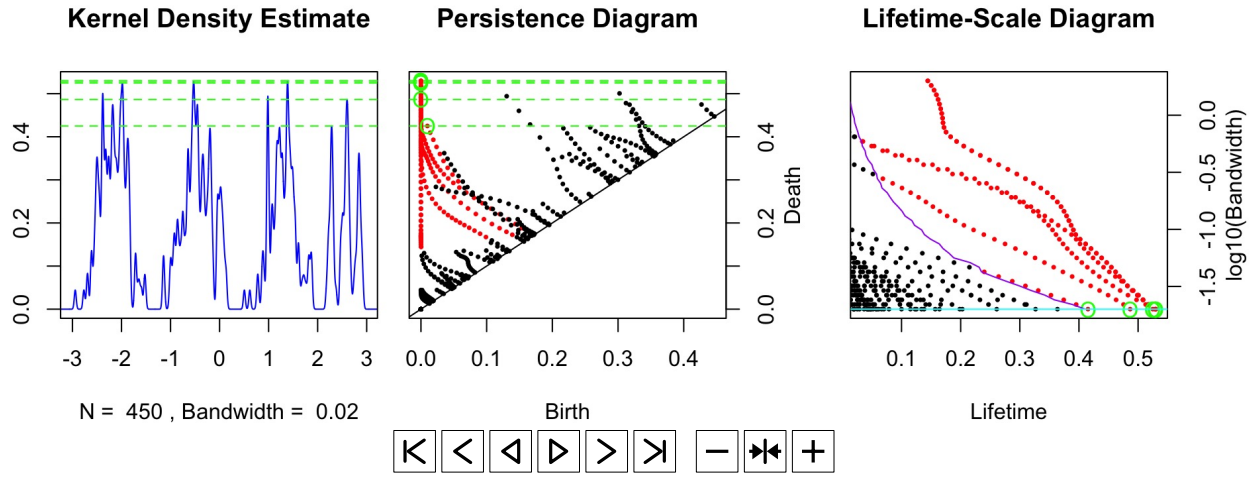


Figure 3. SiZer analysis of 450 points simulated from the discrete comb density. Only the three strongest of the six modes, and one of the thin peaks, are significant.



Movie 2. Topological data analysis-based mode hunting of 450 points simulated from the discrete comb density.

Density 15, called the *discrete comb density*. As one might guess from a careful inspection of the jitter (black dots) and scale space family (blue curves) plots, this density is a mixture of three relatively broad Gaussian components and three much narrower Gaussian peaks.

The SiZer analysis in the bottom panel of Figure 3 shows that the three big peaks on the left are flagged as significant at scales around $\log_{10} h = -0.05$, but the three thin peaks on the right are all combined into a single significant peak. The middle of the three narrow peaks is significant at the scale $\log_{10} h = -1.2$, but the other two are not flagged. This behaviour is an artefact of the deliberately chosen sample size of $n = 450$, aimed at highlighting the comparison between SiZer and persistent homology. For larger sample sizes, such as $n = 1000$, all six modes are typically found by SiZer.

Movie 2 shows a TDA analysis of the same data set. The three big peaks and one of the thin peaks are marked as significant for a large range of bandwidths. Additionally, a fifth peak is found to be significant for bandwidths from 0.02 to approximately 0.04 with a brief interruption around bandwidth 0.022. The lifetime of the sixth mode can be seen to be just barely below the threshold for significance for bandwidths up to around 0.035.

4 Conclusion

We have demonstrated how persistent homology offers a novel approach to the classical problem of bump hunting for probability densities. At the heart of this approach is the combination of persistence diagrams of KDEs over a range of bandwidths—as well as a further reduction of this information into a lifetime-scale diagram. These representations allow us to track the size and statistical significance of modes of the density over different levels of smoothing.

In a direct comparison with SiZer on one real data set and one simulated data set, the presented method shows greater power; that is, it finds more modes significant over larger ranges of bandwidths. This shows that persistent homology is a competitive method for bump hunting in density estimation.

Acknowledgements

A significant part of the presented research was conducted during the SAMSI 2013–2014 programme on Low-dimensional Structure in High-dimensional Systems. M. S. acknowledges support by the Studienstiftung des Deutschen Volkes. G. H. was supported by grant Natural Sciences and Engineering Research Council of Canada DG 293180 and the McIntyre Memorial Fund. J. S. M. was partially supported by the US National Science Foundation grant IIS-1633074.

References

Basford, KE, McLachlan, GJ & York, MG (1997), 'Modelling the distribution of stamp paper thickness via finite normal mixtures: the 1872 Hidalgo stamp issue of Mexico revisited', *Journal of Applied Statistics*, **24**(2), 169–180.

Bubenik, P & Kim, PT (2007), 'A statistical approach to persistent homology', *Homology, Homotopy and Applications*, **9**(2), 337–362.

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

Carlsson, G & Zomorodian, A (2009), 'The theory of multidimensional persistence', *Discrete & Computational Geometry*, **42**(1), 71–93.

Chaudhuri, P & Marron, J (1999), 'SiZer for exploration of structures in curves', *Journal of the American Statistical Association*, **94**(447), 807–823.

Chaudhuri, P & Marron, J (2000), 'Scale space view of curve estimation', *Annals of Statistics*, **28**(2), 408–428.

Chazal, F, Cohen-Steiner, D & Mérigot, Q (2011), 'Geometric inference for probability measures', *Foundations of Computational Mathematics*, **11**(6), 733–751.

Cohen-Steiner, D, Edelsbrunner, H & Harer, J (2007), 'Stability of persistence diagrams', *Discrete & Computational Geometry*, **37**(1), 103–120.

Devroye, L & Gyorfi, L (1985), *Nonparametric Density Estimation: The L1 View*, Vol. 119, *John Wiley & Sons Incorporated*.

Edelsbrunner, H & Harer, J (2008), 'Persistent homology—a survey', *Contemporary Mathematics*, **453**, 257–282.

Efron, B & Tibshirani, RJ (1994), *An Introduction to the Bootstrap*, *CRC press*.

Erästö, P & Holmström, L (2007), 'Bayesian analysis of features in a scatter plot with dependent observations and errors in predictors', *Journal of Statistical Computation and Simulation*, **77**(5), 421–431.

Erästö, P & Holmström, L (2012), 'Bayesian multiscale smoothing for making inferences about features in scatterplots', *Journal of Computational and Graphical Statistics*, **14**(3), 569–589.

Fasy, BT, Lecci, F, Rinaldo, A, Wasserman, L, Balakrishnan, S & Singh, A (2014), 'Confidence sets for persistence diagrams', *The Annals of Statistics*, **42**(6), 2301–2339.

Fisher, N & Marron, JS (2001), 'Mode testing via the excess mass estimate', *Biometrika*, **88**(2), 499–517.

Ghrist, R (2008), 'Barcodes: the persistent topology of data', *Bulletin of the American Mathematical Society*, **45**(1), 61–75.

Godtliebsen, F, Marron, J & Chaudhuri, P (2002), 'Significance in scale space for bivariate density estimation', *Journal of Computational and Graphical Statistics*, **11**(1), 1–21.

Good, I & Gaskins, R (1980), 'Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data', *Journal of the American Statistical Association*, **75**(369), 42–56.

Hannig, J & Marron, J (2006), 'Advanced distribution theory for SiZer', *Journal of the American Statistical Association*, **101**(474), 484–499.

Holmström, L & Erästö, P (2002), 'Making inferences about past environmental change using smoothing in multiple time scales', *Computational Statistics & Data Analysis*, **41**(2), 289–309.

Izenman, AJ & Sommer, CJ (1988), 'Philatelic mixtures and multimodal densities', *Journal of the American Statistical association*, **83**(404), 941–953.

Jones, M, Marron, JS & Sheather, S (1996a), 'Progress in data-based bandwidth selection for kernel density estimation', *Computational Statistics*, **11**(3), 337–381.

Jones, MC, Marron, JS & Sheather, SJ (1996b), 'A brief survey of bandwidth selection for density estimation', *Journal of the American Statistical Association*, **91**(433), 401–407.

Marron, JS & Wand, MP (1992), 'Exact mean integrated squared error', *The Annals of Statistics*, **20**, 712–736.

Minnotte, MC (2010), 'Mode testing via higher-order density estimation', *Computational Statistics*, **25**(3), 391–407.

Minnotte, MC & Scott, DW (1993), 'The mode tree: a tool for visualization of nonparametric density features', *Journal of Computational and Graphical Statistics*, **2**(1), 51–68.

Scott, DW (2015), *Multivariate Density Estimation: Theory, Practice, and Visualization*, *John Wiley & Sons*.

Silverman, BW (1986), *Density Estimation for Statistics and Data Analysis*, Vol. 26, *CRC Press*.

Simonoff, JS (2012), *Smoothing Methods in Statistics*, *Springer Science & Business Media*.

Walther, G (2002), 'Detecting the presence of mixing with multiscale maximum likelihood', *Journal of the American Statistical Association*, **97**(458), 508–513.

Wand, MP & Jones, MC (1994), *Kernel Smoothing*, *CRC Press*.

Xia, K, Zhao, Z & Wei, GW (2015a), 'Multiresolution persistent homology for excessively large biomolecular datasets', *The Journal of Chemical Physics*, **143**(13), 134103.

Xia, K, Zhao, Z & Wei, GW (2015b), 'Multiresolution topological simplification', *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **22**(9), 887–891.