# Supplemental Information Appendix

# Modeling and replicating statistical topology, and evidence for CMB non-homogeneity

**Robert J. Adler, Sarit Agami and Pratyush Pranav**

Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion – Israel Institute of Technology

# Contents

A word of introduction: This 'Supplementary Information' began its life as a collection of a handful of tables containing additional details for the results in the body of the main paper, as well as some explanations for which there was not space there. It was primarily motivated as a response to a number of questions raised by two excellent referees of the first version of the paper.

As time went by, the document took on a life of its own, so that it is now more or less a paper in its own right. (Although it cannot be read without first reading the main paper, so you should read that first.) On the one hand, we apologize for its length. On the other hand, it contains a lot of useful information – and challenges – that seem to be important for the current and future development of RST.

So perhaps its detail and length are justified.

## SI 1. Homology and persistent homology

In this section we give a few pointers to the literature on both persistent homology, as the central tool of applied topology, as well as to the more recent literature looking at persistence diagrams from a statistical viewpoint. We make no attempt to give a complete, or even comprehensive, review, since there are already a number of excellent papers doing this. Rather, we shall give the briefest of overviews and then point the reader to the relevant papers.

One reference that is extremely relevant is a recently published, comprehensive and up to date review [43] (30 pages, and close to 100 references) by Larry Wasserman, on topological data analysis from the viewpoint of statistics. (See also [40] which also considers a number of interesting hypothesis testing issues for persistence diagrams.) Given the existence of Wasserman's review, it seems superfluous for us to attempt to cover the same material here, and so we shall often refer the reader to [43] for background rather than repeat what has already been written there.

Nevertheless, it is probably still worthwhile to give the briefest of introductions to some of the basic notions of algebraic topology and persistence, as a lead-in to the statistical analysis of persistence diagrams that we have proposed. It will also help set up notation and language for what we want to do.

### SI 1.1. Homology

While the classic book by Hatcher [24] is one of the best places to start learning about homology theory, more recent excellent and quite different books and reviews by Carlsson [8, 9], Edelsbrunner and Harer [15, 16, 17], Zomorodian [44], Oudot [31] and Ghrist [21] all give broad expositions of homology and are much closer to the spirit of our own work. More importantly, they also treat the much newer subject of persistent homology, which is not to be found in the classic texts. (For a description of the history of persistence, see the Introduction to [16].)

Algebraic topology focuses on studying topology by assigning algebraic, group theoretic, structures to topological spaces $\mathcal{X}$. Thus, homology, cohomology and homotopy groups can be used to classify objects into classes of 'similar shape'. We focus on homology. If $\mathcal{X}$ is of dimension $N$, then it has $N+1$ homology groups, each one of which is an abelian group. (Throughout, we take the coefficients from $\mathbb{Z}_2$, thereby making the groups into vector spaces.) The zero-th homology $H_0(\mathcal{X})$ is generated by elements that represent connected components of $\mathcal{X}$. For $k \geq 1$ the $k$-th homology group $H_k(\mathcal{X})$ is generated by elements representing $k$-dimensional 'loops' in $\mathcal{X}$. (Roughly speaking, a $k$-dimensional loop is a $k$-dimensional boundary of a $(k+1)$-dimensional set.) The rank of $H_k(\mathcal{X})$, denoted by $\beta_k$, is called the $k$-th *Betti number*. For $\mathcal{X}$ compact and $k \geq 1$, one can think of $\beta_k$

as counting the number of '$(k+1)$-dimensional holes in $\mathcal{X}$', while $\beta_0$ counts the number of connected components. The Euler characteristic, a central topological quantity and homotopy invariant, is then

$$\chi(X) \ = \ \sum_{k=0}^{N}(-1)^k\beta_k. \tag{1}$$

### SI 1.2. Persistent homology

The notion of persistent homology arises when one has a filtration of spaces; viz. a sequence (or continuum) of spaces $\mathcal{X}_1 \subseteq \mathcal{X}_2 \subset \dots$ (or $\mathcal{X}_t$, with $\mathcal{X}_s \subseteq \mathcal{X}_t$ whenever $s \leq t$) and one is interested in how homology changes as one moves along the sequence. The references above give full - and quite non-trivial - technical definitions of how this is done, but we shall suffice with an example that is all that is needed for this paper, and which also introduces Gaussian random fields. The example comes from an earlier paper, [2].

Suppose that $\mathcal{X}$ is a nice space, that $f : \mathcal{X} \to \mathbb{R}$ is smooth, and consider the filtration of excursion, or super-level, sets

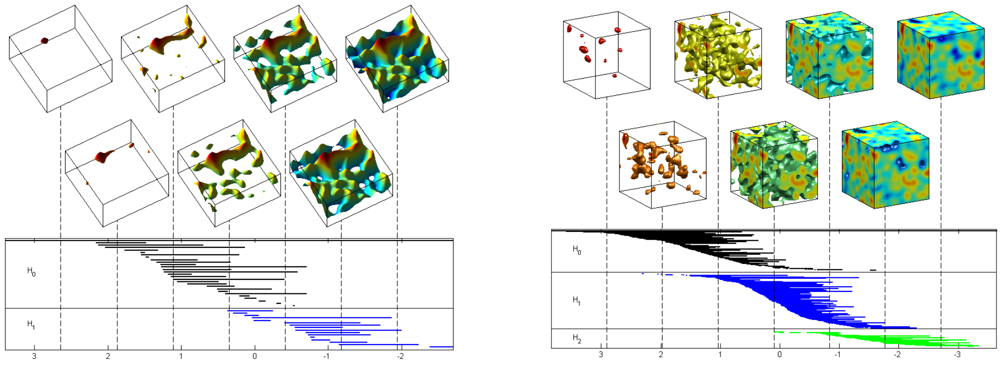
$$\mathcal{X}_u \ \stackrel{\Delta}{=} \{x \in \mathcal{X} : \, f(x) \in [u, \infty)\} \ \equiv \ f^{-1}([u,\infty)). \tag{2}$$

Note that if $u \geq v$ then $\mathcal{X}_u \subseteq \mathcal{X}_v$. Descending from $u$ to $v$, components of $\mathcal{X}_u$ may merge, new components may be born, and may possibly later merge with one another or with the components of $\mathcal{X}_u$. (When two components merge, the first of these to have appeared is treated as if it is the one continuing its existence beyond the merge level.) Similarly, the topology of these components may change, as holes and other structures form and disappear. Following the topology of these sets, as a function of $u$, by following their homology, is an example of persistent homology. The term 'persistence' comes from the fact that as the level $u$ changes there is no change in homology until reaching a level $u$ which is a critical point of $f$; i.e. the topology of the excursion sets remains static, or 'persists', between the heights of critical points. This, of course, is the basic observation of Morse theory, which links critical points to homology. However, the persistence of persistent homology goes further than regular Morse theory.

A useful way to describe persistent homology is via the notion of barcodes. Assuming that $\dim(\mathcal{X}) = N$, we also have, from the smoothness of $f$, that, if $\mathcal{X}_u$ is non-empty, then $\dim(\mathcal{X}_u)$ will typically also be $N$. A barcode for the excursion sets of $f$ is then a collection of $N+1$ diagrams, one for each collection of homology groups of common order. A bar in the $k$-th graph, starting at $u_1$ and ending at $u_2$ ($u_1 \geq u_2$) indicates the existence of a generator of $H_k(\mathcal{X}_u)$ that appeared at level $u_1$ and disappeared at level $u_2$. An example is given in Figure S1, in which the function $f$ is actually the realisation of a smooth, Gaussian random field (function) on the unit square (on the left) or on the unit cube (on the right). This is an example to which we shall return later when treating CMB data.

Figure S2 shows a different representation of the barcodes for the 2-dimensional example of Figure S1. Each bar has a 'birth time' $b$ and 'death time' $d$, where $d < b$ since, as described above, the filtration is for upper level sets, and we index these by levels descending from $+\infty$. The points $(d, b)$ corresponding to these bars appear in Figure S2, with the $H_0$ bars in the left hand diagram and the $H_1$ bars on the right. These diagrams, or, rather, the collection of points in them, are the persistence diagrams at the core of this paper.

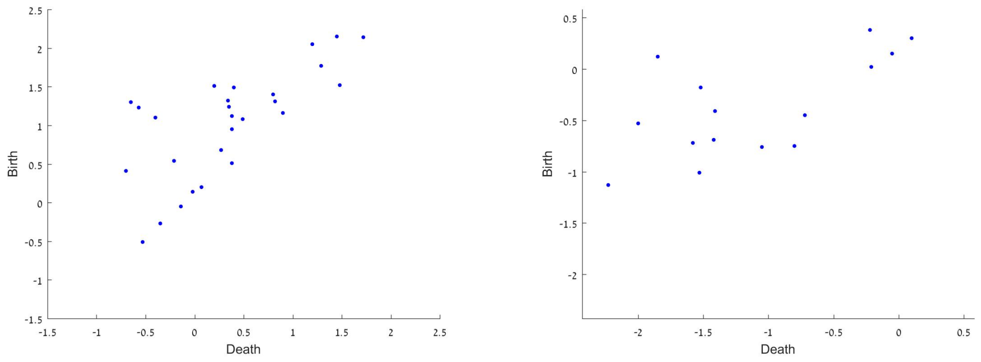
To get a feeling for the overall distribution of persistence diagrams, Figure S3 shows the points from 1,000 persistence diagrams of the kind used to generate Figure S2. Note



(a) (b)

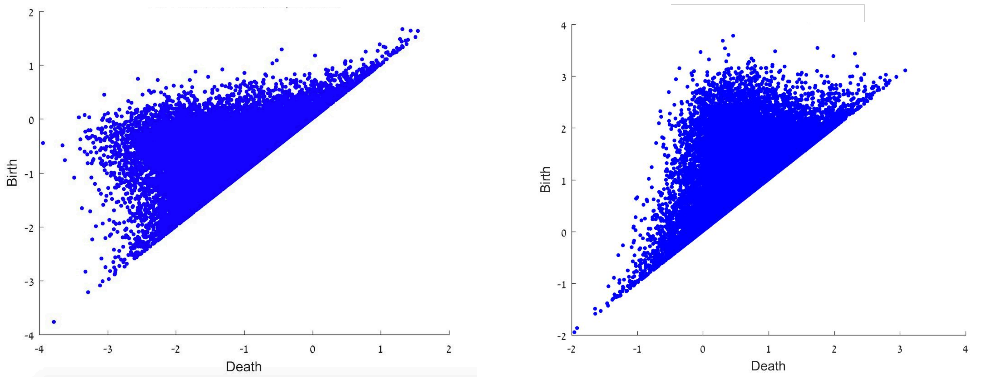
Figure S1: (a) Barcodes for the excursion sets of a Gaussian random field on $[0,1]^2$. The top seven boxes show the surfaces generated by a 2-dimensional random field above excursion sets $\mathcal{X}_u$ for different levels $u$. To determine the level for each figure, follow the vertical line down to the scale at the bottom of the barcode. As the vertical lines pass through the boxes labelled $H_0$ and $H_1$, the number of intersections with bars in the $H_0$ ($H_1$) box gives the number of connected components (resp. holes) in $A_u$. Thus, at $u \sim 1.9$, $\mathcal{X}_u$ has 4 connected components but no holes, while at $u \sim -1.2$, $\mathcal{X}_u$ has only 1 connected component, but 9 holes. The horizontal lengths of the bars indicate how long the different topological structures (generators of the homology groups) persist.
(b) Barcodes for the excursion sets of Gaussian random field on $[0,1]^3$. The barcode diagram is to be read as for (a) with two differences: The top 7 boxes now display the excursion sets themselves and the values of the field are colour coded. Furthermore, there are now three homology-groups/barcode-boxes, representing connected components, handles, and holes.
Computation of the barcodes, in both cases, was carried out in Matlab using Plex (Persistent Homology Computations) from Stanford [10].

that some care needs to be taken in interpreting this figure, since it *is not* a scatterplot of persistence diagrams. Each of the 1,000 persistence diagrams is a collection of points, and so should be thought of as a single point in $(\mathbb{R}^2)^N$, where $N$ is the largest number of points in any of the diagrams. However, since visualizing this is impossible, plotting all the points of the individual persistence diagrams in one two-dimensional plot still has value. In particular, two facts are immediately obvious from Figure S3, that cannot be seen in the individual persistence diagrams of Figure S2. One is that $H_0$ and $H_1$ 'distributions' are almost reflections of one another. This is a consequence of the fact that the underlying Gaussian functions are statistically symmetrical about zero, as well as the simple relationship between upper and lower level sets in two dimensions. The other is the fact that each 'distribution' is, itself, asymmetric. This will be an important fact to remember when we turn to the Gibbs' modeling below.



(a) (b)

Figure S2: Persistence diagrams corresponding to the 2-dimensional barcode diagram of Figure S1a. (a) The $H_0$ homology diagram (the upper set of bars in Figure S1a). (b) The $H_1$ homology diagram (the lower bars).



(a) (b)

Figure S3: Superpositioning of the points from 1,000 (random) persistence diagrams of the kind shown in Figure S2. (a) $H_0$ points (b) $H_1$ points.

### SI 1.3. Analysis of persistence diagrams: Current methodology

Persistence diagrams almost always arise as topological summaries of some underlying phenomenon, and, having been constructed, are typically subject to some kind of analysis. This can be thought of as a path

$$\textit{phenomenon} \;\rightarrow\; \textit{persistence diagram} \;\rightarrow\; \textit{analysis}. \tag{3}$$

The analysis can be of various forms. As have already noted, Wasserman's review [43] gives an excellent review of many of the statistical approaches to the analysis of persistence diagrams, making it rather superfluous for us to attempt a literature review here. There are also other approaches, many of which involve summarizing the diagram with either a low dimensional vector of numerical descriptors (typically involving various measures associated with metric measurements, bottleneck, Wasserstein, or other), a large dimensional vector (but of significantly lower dimension than the number of points in the diagram), or real valued function (e.g. the persistence landscape of [7], the persistence intensity functions of [13, 18], the persistence weighted Gaussian kernel approach of [27, 37, 39], and the persistence images of [1]). Many of these approaches (see also [26]) adopt techniques such as principal component analysis and support vector machines to analyse this summary data.

What is common to all approaches, however, is the need for multiple instances of the persistence diagram. When the analysis is essentially statistical, these are needed to compute means, averages, and so forth, as well as to justify the application of tools such as the central limit theorem. In the machine learning setting, multiple instances are required for the learning phase of the associated algorithms.

Providing multiple instances of the persistence diagram is not a trivial task. In some scenarios, multiple observations of the 'phenomenon' of (3) may be available. More typically, however, only one observation of the phenomenon is available, and so only one diagram. In those cases, the standard approach to effectively increase the number of instances is via resampling, either of the phenomenon or the diagram. Virtually all of the papers referred to above all have examples of this approach.

The contribution of the present paper is to enter the diagram (3) at the intermediate step, by providing a new approach to providing multiple instances of a persistence diagram when, perhaps, only one such original diagram is available. We do this via probabilistic modeling of persistence diagrams.

## SI 2. Probability models for persistence diagrams

The first point that needs to be made is that virtually nothing is known about the distribution of persistence diagrams for specific problems. On the one hand, the fundamental probability theory of persistence diagrams has been laid down in papers like [29, 30, 42]. These results establish that the space of persistence diagrams has properties that allow for the definition of probability measures which support expectations, variances, percentiles and conditional probabilities, and that the space is complete and separable when equipped with the Wasserstein metric. Despite the existence of this general theory, very few explicit models have been suggested which would assign explicit probabilities to a given, observed, persistence diagram. Indeed, even when given a simple model for the data underlying a persistence diagram (such as a filtration of Čech complexes defined over the points of a Poisson process) one is hard put to say much non-asymptotic about expectations, let alone distributions. (See, for example, the reviews [6, 25].)

The second point, which is a consequence of the first and already noted above, is that the lack of parametric models has meant that virtually all previous statistical analysis of persistence diagrams has relied on some sort of resampling. To replace this, we suggest a very specific, parametric, model for persistence diagrams. Thus, once the parameters are estimated – and this is possible from a single diagram, assuming that it is sufficiently rich in points – it is then possible to generate simulated diagrams from the hypothesised distribution to be used in statistical analysis.

### SI 2.1 Intensity functions and Poisson processes

Before turning to the Gibbs measures that we propose as a basis for RST, we want to discuss a seemingly much simpler, and quite natural approach (which was also, not surprisingly, raised by a referee).

Although persistence diagrams are collections of points, it is natural to smooth them with an appropriate kernel, and so obtain what [13, 18] refer to as 'persistence intensity functions'. One can then base statistical analysis of the diagram on statistical analysis of the intensity function, using any of a wide class of nonparametric statistical techniques for analyzing functional data, as, for example, in [13].

However, from our point of view, empirical persistence intensity functions could also be treated as if they were the estimated intensity of a non-homogeneous point process, such as a Poisson process, which could then be simulated in much the same spirit as described for the Hamiltonian model described in the paper, although the precise details would be different. (Nevertheless, some sort of MCMC approach would probably be the most appropriate.)

An important advantage of this model, over the one we have proposed, would be that it would be easier to allow, in the simulations, for changing numbers of points in the diagrams, allowing points near the diagonal to appear or disappear. This is actually a crucial aspect of the mathematical models of persistence diagrams developed in [29, 30, 42].

In our analysis we have kept $N$ fixed, for a number of reasons. The first is that allowing $N$ to vary in the Gibbs model requires additional modeling, and that one needs a good model for the distribution of $N$, effectively *external* to the Gibbs setting. (In the Poisson setting, whether homogeneous or not, $N$ would have a Possion distribution with parameter determined by the integral of the intensity measure.) We do not, at this point, have such a model.

The second reason is that modeling $N$ requires adding at least one additional dimension to the parameter space, and, at least for diagrams with a large number of points, we were already near the limits of available computing capacity with the five parameters over which we optimized models.

Both of these reasons are, obviously, choices of convenience. However, the main reason why we were content to keep $N$ fixed, and why we felt it was not a serious restriction, was that we applied (and recommend) RST primarily in situations in which $N$ is large. In these situations it is unreasonable to expect that small fluctuations in the numbers of points (particularly those in the neighborhood of the diagonal, which is where new points are 'born' and old ones 'die') would have more than a minor impact on the results.

In addition, we feel that there is a potential drawback of using intensity functions as a basis for generating (even non-homogeneous) Poisson processes as models for persistence diagrams, in that the Poisson process assumes the independence of numbers of points in disjoint regions, even if they are neighboring regions. This does not seem a reasonable assumption for persistence diagrams. Indeed, the strong significance of the interaction parameters that we found in the two examples treated in the paper indicate that such independence is unreasonable, and that there is, in fact, a tendency towards repulsion between the nearby points in persistence diagrams. Leaving the Poisson setting, but remaining in the setting of point processes in general, typically eventually leads to Gibbs measures, or similarly complicated models.

Another potential drawback is the essentially nonparametric nature of the intensity function, which makes it non-trivial to carry out parametric hypothesis testing of the kind we did for the CMB example. Of course, in the framework of nonparametric, and particularly Bayesian, statistics, this particular drawback might be considered a strength.

Nevertheless, alternative, non-homogeneous point processes based on persistence intensity functions present a serious alternative to the Hamiltonian models treated in this paper, and deserving of a more detailed study.

### SI 2.2 Gibbs measures

As described in the main paper, our choice of Gibbs measures to model the points of persistence diagrams was based on their long history as high quality parametric models, a rich literature on both their theory and application, and the ease with which they lead to simulations via Markov chain Monte Carlo.

In this section we want to address three main questions (along with a few lesser ones, en passant):

1. When fitting a Gibbs model to data, are the estimation techniques we suggested reliable?
2. When simulating from the fitted model, do the simulations reliably replicate the statistical properties of the persistence diagrams?
3. How long should the simulation be run, so as, on the one hand, to provide data consistent with the original persistence diagram while, on the other hand, providing (almost) independent replications of it?

Since we have no useful information on the true distributions of persistence diagrams, it is impossible to provide precise, theoretically justifiable, answers to either of these questions. Thus we will answer them with (typical) examples.

The first example that we shall consider comes from Gaussian random fields on the 2-sphere. We simulated 100 such fields, and, in preparation for the CMB example we study in the paper, used the simulation routines of NASA's HEALPix software[1]. The precise details of the simulation are not that important for our current purposes, and it suffices to note that they give 100 independent replications of Gaussian random fields, at 120 arcmin smoothings, that look like slightly smoothed versions of the CMB of Figure 2 of the main paper, as in Figure S4

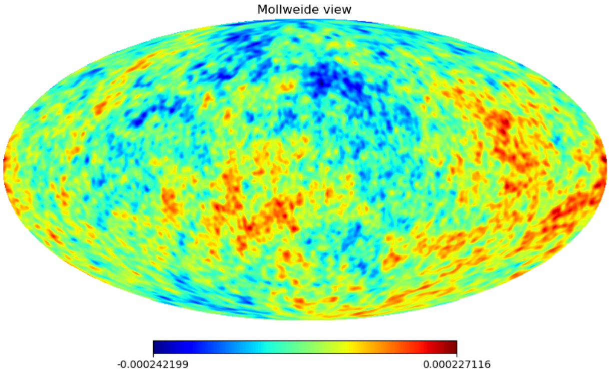


Figure S4: Simulation of a Gaussian random field on the sphere with a realistic CMB spectrum smoothed with a Gaussian kernel with full width half maximum 120 arcmin.

For each of these simulations we performed the following steps:

1. In preparation for the CMB analysis of Example 2, persistence diagrams for the $H_0$ and $H_1$ homologies for northern and southern spherical caps (60 degrees from the north and south poles) were calculated, using PHAT - Persistent Homology Algorithms Toolbox, [3]. The number of points in each diagram was around $N = 500$.

[1] HEALPix is an acronym for Hierarchical Equal Area isoLatitude Pixelization of a sphere. The pixelization produces a subdivision of a spherical surface in which each pixel covers the same surface area as every other pixel, and allows for fast simulation of homogenous, isotropic random fields on the sphere, with covariance functions appropriate for CMB. The simulation is based on representing the field as a sum of spherical harmonics with Gaussian coefficients; cf. [28] for the theory, and for details of the numerics see [22] and the HEALPix site healpix.jpl.nasa.gov. Section SI 4.1 has some more details.

2. For each diagram, we estimated the parameters in a Gibbs model with Hamiltonian

$$H^2_{\delta,\Theta}(\tilde{x}_N) = \theta_H \sigma_H^2 + \theta_V \sigma_V^2 + \sum_{k=1}^{3} \delta^{-2} \theta_k \mathcal{L}_{\delta,k}(\tilde{x}_N), \tag{4}$$

   where all terms are described in the main paper, along with the estimation procedure.

3. For each model, we produced simulations of the persistence diagrams, using the Metropolis-within-Gibbs procedure described in the main paper.

The results of the simulation are as follows:

1. The first two pages of Table A1 in the Appendix shows all 100 sets of parameter estimates for both homologies. Despite the fact that, a priori, there is no reason why a Gibbs model should be a good fit for the persistence diagrams, the consistency of the parameter estimates is obvious.

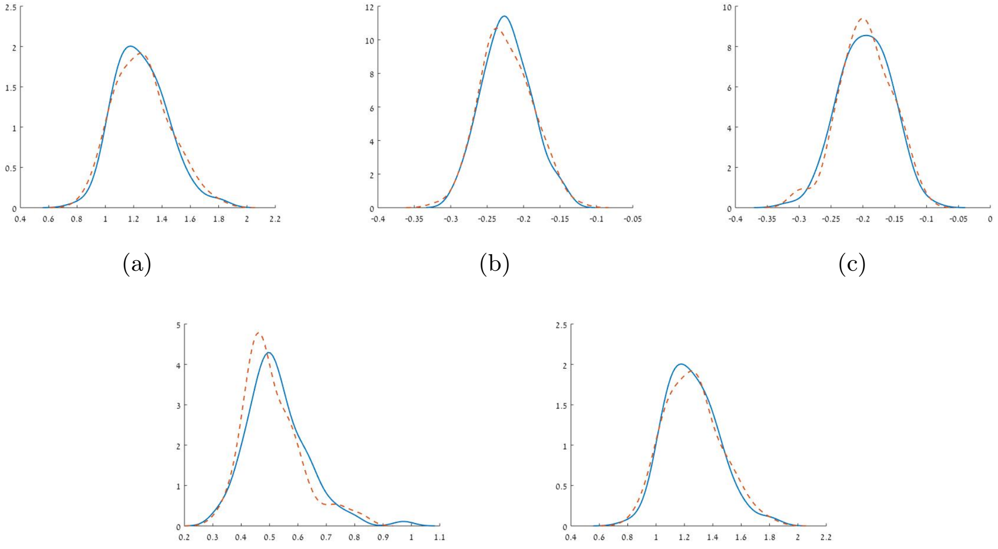
2. Figure S5 below shows the same information, but summarized as smoothed empirical densities.

3. Although the last two items indicate considerable stability in the parameter estimates over a number of realizations of the persistence diagrams, they do not confirm whether or not the Gibbs measure is, in fact, a good model. To check this, we ran the MCMC procedure on some of the estimated models, to see how they behaved as the samples became distanced from their initial states; viz. the persistence diagrams from which the parameters were estimated.

   Figure S6 shows typical results, for one of these cases, for which the estimated parameters were $(\delta, \theta_1, \theta_2, \theta_3, \sigma_H^2, \sigma_V^2) = (0.0518, -0.2480, -0.2038, -0.1712, 0.6074, 1.2532)$. While it is clear from the results that, if the number of MCMC steps is not large, then the simulated diagram is close to the initial one, it is also clear that there is a collapse of the diagram towards the diagonal as the MCMC proceeds. We discuss this, along with its implications for the practitioner, in detail below.

4. In order to further study the phenomenon of the previous item, we considered summary statistics of the persistence diagrams as the MCMC progressed. The blue (full line) graphs in Figure S7 show the empirical probability densities of average interaction strength within clusters, shown after 10, 50 and 1,000 MCMC steps, for each of the original 100 persistence diagrams. The red (dashed line) curves show the same thing, but for the original 100 diagrams. By average interaction strength we mean $\mathcal{L}_{\delta,k}/n_k$, where $n_k$ is the number of terms making up the sum $\mathcal{L}_{\delta,k}$, defined in the main paper. Consistent with the models fitted, we took $k = 1$, 2, or 3.

   As in the previous item, we see an excellent fit in the early stages of the MCMC (recall that there were only 100 replications of the MCMC runs) and less so later on. There is also a noticeable difference between the behavior of the first, second, and third order terms, with the deviations between the simulations and the MCMC results becoming more significant the higher the order. However, it is important to note that we always have $n_1 < n_2 < n_3$, and typically have $n_1 \ll n_2 \ll n_3$, (in the case of Figure S7 the numbers are of the order 400, 300, and 200, respectively) so that the actual impact of the higher order $\mathcal{L}_{\delta,k}$ on the Hamiltonian, and so the simulated persistence diagrams, is comparatively small.



(a)

(b)

(c)

(d)

(e)

Figure S5: Smoothed empirical densities for the five parameter estimates in the Hamiltonian (4) for the $H_0$ persistence diagram coming from 100 simulations of northern and southern caps of a Gaussian random field on the sphere, at a smoothing of 120 arcmin. Northern caps are in red (broken) lines, and southern caps are in blue (unbroken) lines. (a) $\theta_1$, (b) $\theta_2$, (c) $\theta_3$, (d) $\theta_V$, (e) $\theta_H$.

The data above is typical of a large number of tests that we performed, and indicate that the answer to the first question raised in this section - regarding the reliability of the estimation techniques - is, for those tests, positive.

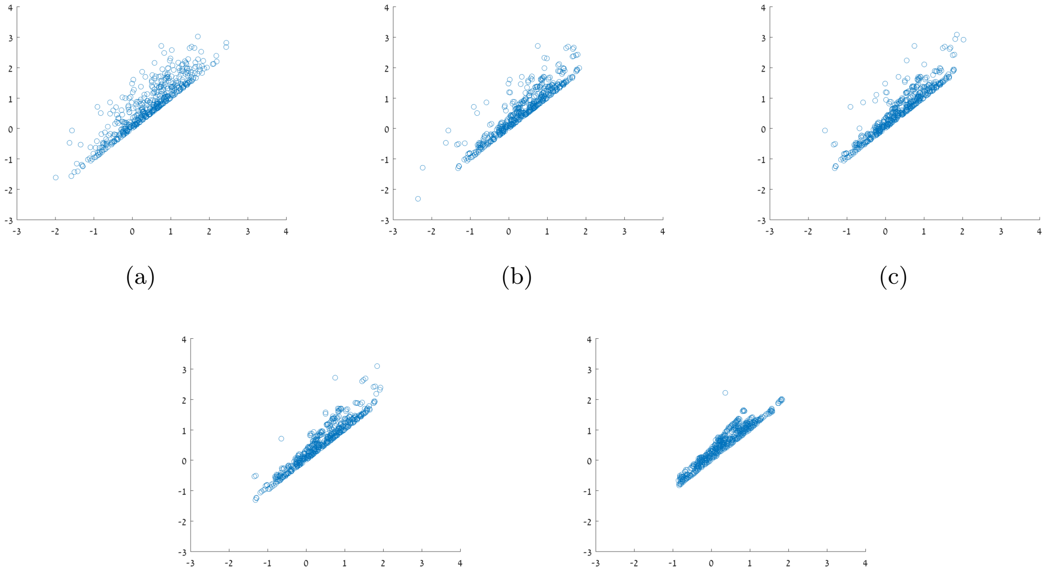
As an aside, we note that the choice of parametrization is an important part of achieving this stability. In our initial studies (and, indeed, in an earlier version of the main paper) we wrote the Hamiltonian (4) without the factor of $\delta^{-2}$ before the $\theta_k$'s. With this change the Hamiltonian becomes

$$H^2_{\delta,\Theta}(\tilde{x}_N) = \theta_H \sigma_H^2 + \theta_V \sigma_V^2 + \sum_{k=1}^{3} \theta_k^* \mathcal{L}_{\delta,k}(\tilde{x}_N), \tag{5}$$

so that the $\theta_k^*$ here is our original $\delta^{-2}\theta_k$.

While this makes no real difference to the model, it turned out to have a major effect on the stability of the estimates of the $\theta_k^*$. The reason behind this is that although $\delta$ is to a large extent a nuisance parameter, estimates of it vary considerably over diagrams. Consequently, adopting the Hamiltonian (5) rather than (4) leads to high variation in the estimates of the $\theta_k^*$. While the move to (4) was an a posteriori change of parametrization, motivated by a seeming lack of stability, the interpretation of the $\theta_k$ there as energy, or interaction, *intensities* also gives them an attractive physical meaning.

Regarding the second question - as to whether or not the simulations reliably replicate the statistical properties of the persistence diagrams - the evidence is less convincing. In fact, the simulations of the persistence diagrams, as the MCMC procedures progress, begin to exhibit characteristics that are not typical for the diagrams as a whole, and so it is indeed doubtful that the Gibbs models actually have stationary distributions which give realistic models for persistence diagrams. This is observable, for example, in the progression of diagrams in Figures S6 and S7.



(a) (b) (c)

(d) (e)

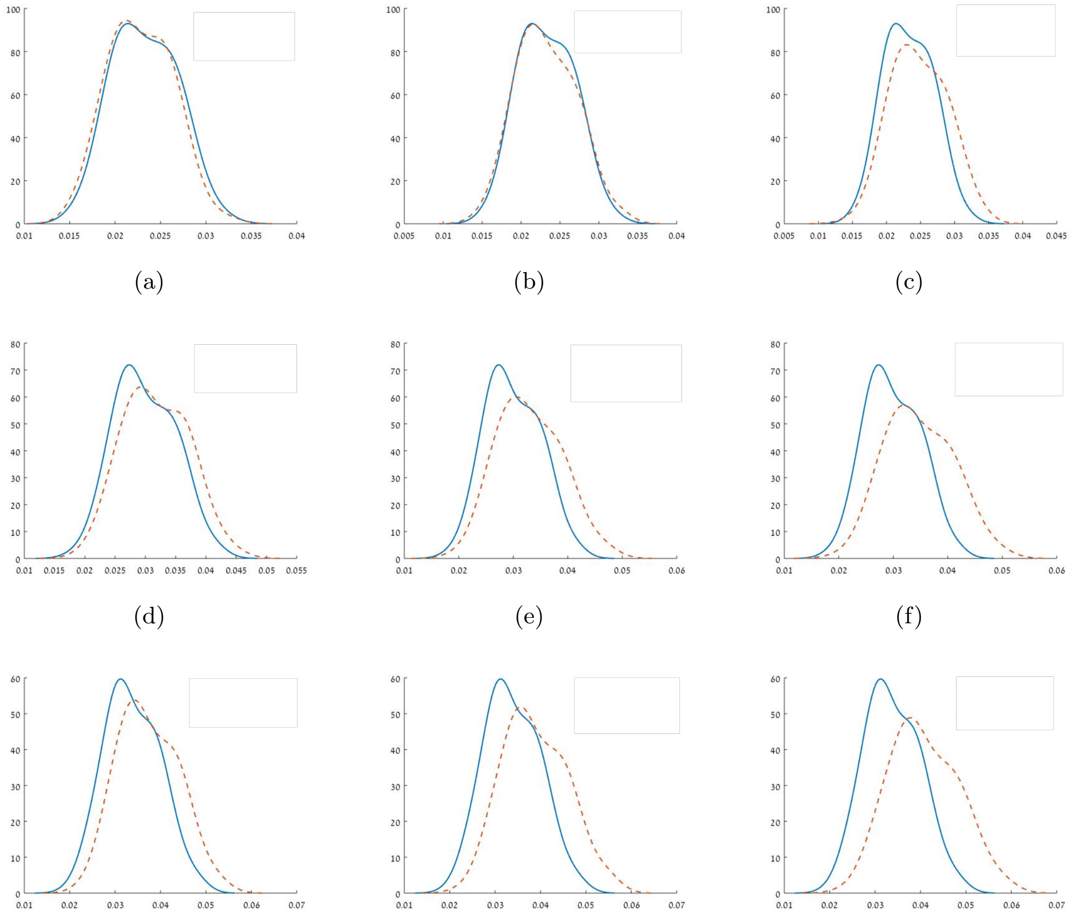
Figure S6: MCMC simulations of a persistence diagram arising from a Gaussian random field on the sphere. (a) The original diagram. (b)–(e) After 10, 25, 50, and 1,000 MCMC steps.

However, for the uses we make of the procedure, *this is not a major concern!* The underlying approach of RST is to provide multiple instances of persistence diagrams when the data only provides one, or a small number. In terms of the parameters $n_b$, $n_r$ and $n_R$ (number of steps in each MCMC block, number of block runs, and number of replications of the entire procedure) this can be achieved in a number of ways. A general methodological philosophy would be the following: When one believes that the model is good, it is most efficient (from a variance point of view) to take $n_b$ and $n_r$ large, and $n_R$ small, perhaps even $n_R = 1$. The cost for this is bias, which will typically be proportional to the lack of fit of the model. If one has less faith in the model itself, large $n_b$, small $n_r$, and large $n_R$ is a preferable route. In this case the procedure produces random perturbations of the original persistence diagram, rather than truly independent copies of it. Nevertheless, as we will see from the examples in the following sections, RST also works surprisingly well in this scenario.

This, of course, takes us to the third question raised at the beginning of this section; viz. When it comes to applying this philosophy, it is necessary to keep track of how far, and fast, the simulations diverge from the original data as the MCMC procedure progresses, so as to know how to best choose the various parameters. For this we tried the most common measures of distance between persistence diagrams; viz. the Wasserstein and bottleneck metrics. Recall that, for two diagrams $D_1$ and $D_2$, the Wasserstein $p$-distance, $W_p(D_1, D_2)$, $p > 0$, is defined as

$$W_p\left(D_1, D_2\right) \;=\; \inf_{\gamma} \Big( \sum_{u \in D_1} \|u - \gamma(u)\|_\infty^p \Big)^{1/p} \tag{6}$$

where $\gamma$ ranges over all matching between the points of $D_1$ and $D_2$, the latter having been augmented by adding all points on the diagonal. In the limit case of $p = \infty$ the Wasserstein distance is known as the bottleneck distance, which is the length of the longest edge in the best matching. From now on we shall use only $p = 2$, and so drop the explicit dependence



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure S7: Summary statistics of average interaction strengths for 100 persistence diagrams. From left to right: cluster sizes 2, 3, and 4. From top to bottom, after 10, 50, and 1,000 MCMC steps. See text for details.

on it.

Figure S8 gives additional insight into the distancing of the MCMC simulations from the original persistence diagrams shown in Figures S6 and S7, by showing how the bottleneck and Wasserstein distances change as the MCMC progresses. All figures here are based on averages taken over the same 100 simulated persistence diagrams used for Figure S7. The left hand panels show the bottleneck distances, while the right hand panels shows the $W_2$ differences. The first and second rows show the results of the first 50 steps of the MCMC algorithm, first on a linear scale and then on a logarithmic scale. The last row, again on a logarithmic scale, but based on only 20 simulations, goes out to 2,000 steps.

There are a number of interesting conclusions that can be drawn from these graphs. The first is that there seems to be far more sample variation for the bottleneck than for the Wasserstein distance. This is not unexpected, given that the former is highly dependent on one, or at most only a few, outlying observations, while the latter takes into account all the Euclidean distances in (6). This larger scale averaging should lead to additional stability. For this reason we recommend, in practice, relying more on Wasserstein than bottleneck distances, but the final recommendation that we shall make below turns out to be roughly the same in both cases.

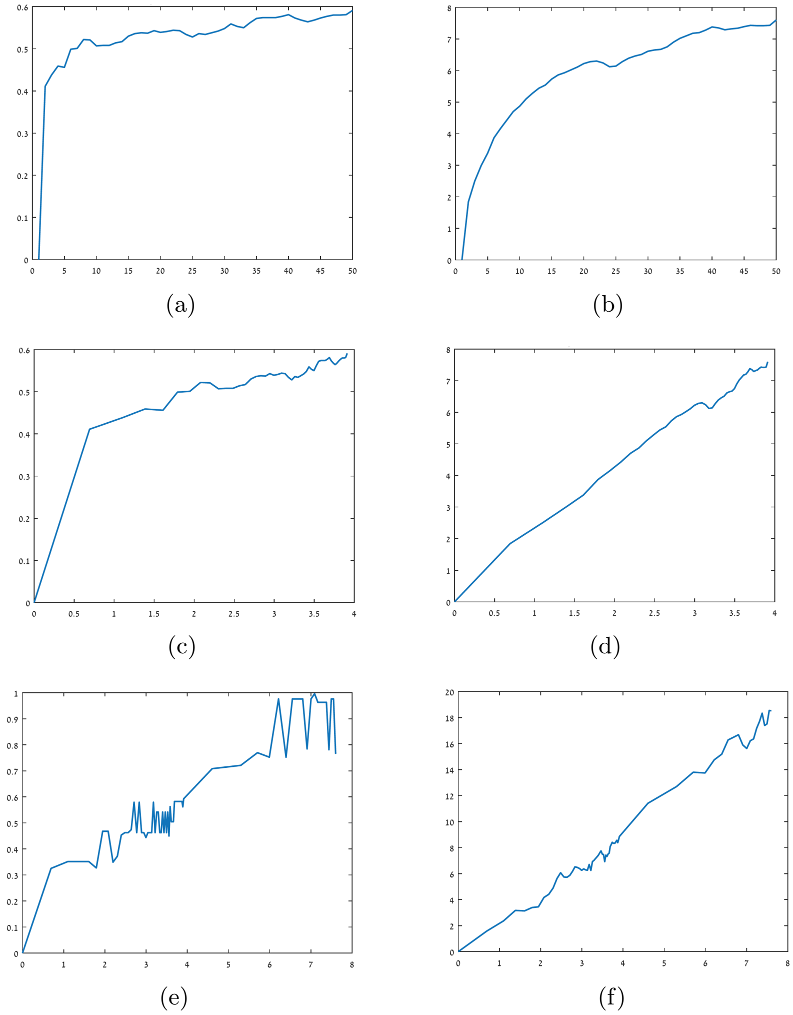


Figure S8: Growth of the bottleneck (a) and Wasserstein (b) differences of MCMC simulations from a specific persistence diagram (vertical axis), as a function of the number of steps $n_b$ (horizontal axis, $1 \leq n_b \leq 50$) averaged over 100 independent persistence diagrams, as in Figure S7. Panels (c) and (d) show the same data, but on a logarithmic scale, while (e) and (f), also on a logarithmic scale, take $1 \leq n_b \leq 2,000$.

The second conclusion is that while the initial growth of the distances is rapid, they approach asymptotes at exponential rates. The rapidity is clear in Panels (a) and (b), and the exponential rate is clear from the linear behavior of the curves in the logarithmic scales. The exponential approach to an asymptote is a standard consequence of the exponential convergence of any Markov process to its stationary state, and is classical. The rapid initial growth of the distance is also a common phenomenon for Markov processes, but is not classical, and is related to phenomena treated mathematically only relatively recently, as in the fundamental work of Diaconis and co-authors in the 1990's (cf. [4]) and a wide subsequent literature.

The newsworthy observation of [4] was that when shuffling a deck of 52 cards (with a 'riffle', or 'dovetail' shuffle) the effect of the initial ordering of pack was, to a large extent, lost after only 7 shuffles. Furthermore, the technical literature since [4] has shown that this phenomenon of rapid 'mixing' is common, although there is nothing sacrosanct about the number 7.

In our case, the evidence is that the 'magic number' is found at the point where the initial rapid growth of the distance functions ceases, which is approximately 10 for the bottleneck distance and somewhere in the range [10, 20] in the Wasserstein case. At 10 steps, therefore, the results of Figures S7 and S8 indicate that the dependence of the MCMC on the initial persistence diagram has dropped significantly, while at the same time the MCMC has produced persistence diagrams remaining close to the true distribution. Interestingly, the same range of numbers arises in the example of two circles treated below; cf. Figure S11.

These numbers are used in the examples below. (Actually, at the time we studied them we had not done the analysis leading to Figures S8 and S11. Once again, kudos to a referee for asking good questions and, in this case, also suggesting that we use the Wasserstein metric in this setting.)

Thus, in terms of the third question raised at the beginning of this section, our advice to the practitioner is to produce graphs such as those in Figure S8 and rely on the information they contain for choosing the MCMC parameters $n_b$, $n_r$ and $n_R$.

As an aside, it is worth commenting that the subsampling methods in common usage, and described above, are not that different, in some aspects, from the perturbative scenario of RST. They, too, do not produce independent copies of the original diagram, or its derivative statistical summaries, but rather provide perturbations that contain information about statistical variability.

## SI 3. Example 1: Learning two circles

This section gives further details on the first (toy) example treated in the paper, that of investigating the persistence diagram generated from a random sample from two concentric circles in the plane.

There are a number of points that we should make, before describing the analysis in detail. The first is to emphasize that this is indeed a very simple situation, and using RST to analyze it is definitely overkill. In fact, the empirical density and persistence diagram of Figure 1 of the paper are so simple, that anything beyond just looking at them is probably overkill. Our aim here is simply to use this simple example as a test case.

The second point is that we make no attempt to analyze the $H_1$ diagram. With only three points in this diagram, any attempt to fit a parametric model would be foolish. Indeed, with the order of only 50 points in the $H_0$ diagram, and Gibbs measures which typically estimate 5 parameters (plus the nuisance parameter $\delta$) it is not clear a priori that RST will work even in this case. Perhaps surprisingly, and certainly reassuringly, it does.

The last point is to emphasize, once again, that we are primarily interested in replicating the 'persistence diagram' stage of (3). Although we will propose a particular test statistic for the 'analysis' stage, the fact that the reader might prefer a different statistic does not affect the replication stage, which can be applied, just as well, with alternative statistics.

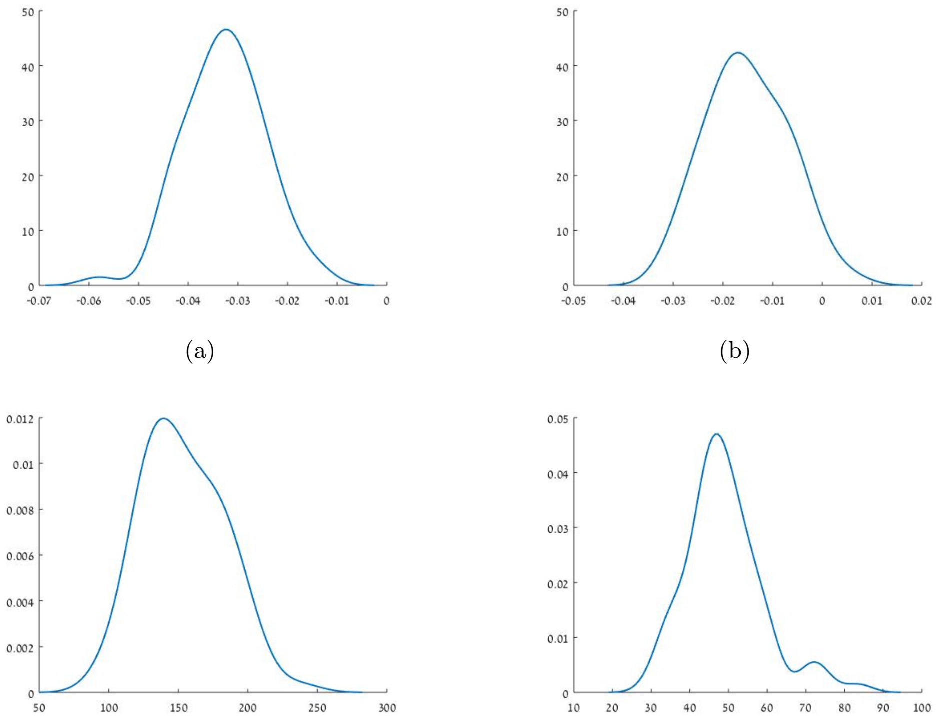
### SI 3.1 The Gibbs model

Following the procedure described in the paper, we fitted the Gibbs measure with Hamiltonian (4) to the $H_0$ persistence diagram, without the 'point at infinity'; viz. the point with the highest birth time. (We recommend removing this point in all small data sets, since it merely indicates the existence of an object and so is different in nature to all others.) This yielded the parameter estimates

$$(\delta, \theta_1, \theta_2, \theta_3, \theta_V, \theta_H) \;=\; (0.0038, -0.0225, -0.0104, 0, 144.7,\ 66.3). \tag{7}$$

Note that $\theta_3 = 0$, since the model selection criteria we used (both AIC and BIC) removed this parameter from the model. This is not surprising, given the small size of the data set (i.e. the persistence diagram).

Before attempting any further analysis, we check that the estimation procedure is reliable in this case, and whether or not the parameter set (7) 'makes sense'. Analogous to the investigation for random field generated diagrams in Section SI 2.2, we performed two checks. The first was to simulate the two-circle data 100 times, thus producing 100 persistence diagrams. For each of these we estimated the four parameters $\theta_1, \theta_2, \theta_V$ and $\theta_H$ (having, a priori, set $\theta_3 \equiv 0$). Empirical densities for each of the parameter estimates are shown in Figure S9, which indicates a mild spread in the estimates and reasonable stability.



(a) (b)

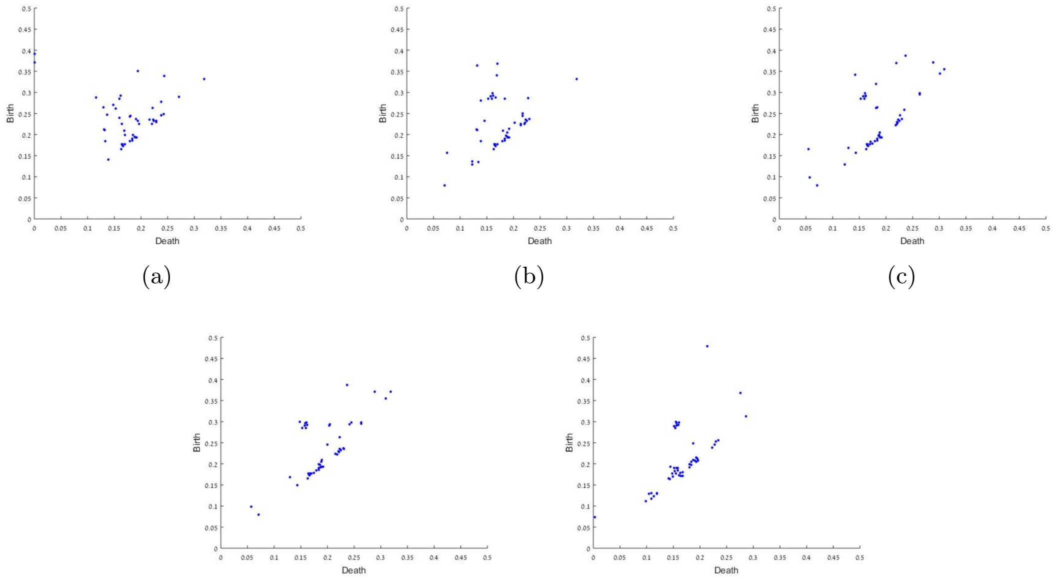
(c) (d)

Figure S9: Smoothed empirical densities for the four parameter estimates in the Hamiltonian (4) for the two circle data. (a) $\theta_1$, (b) $\theta_2$, (c) $\theta_V$, (e) $\theta_H$.

The second check involved running the MCMC procedure on the model with the estimated parameters, and comparing the simulations – considered, as before, as perturbations – with the original persistence diagram. Fig S10 shows some examples of this, at different stages. Overall, we observe a similar phenomenon to that observed in Section 2.2 for the Gaussian random field example, albeit considerably less marked; viz. (much slower) collapse of the persistence diagram towards the diagonal. Figure S11 shows what happens when running the MCMC routine, by studying the bottleneck and Wasserstein distances between the initial persistence diagram and those produced by the MCMC.

As in the Gaussian random field case that we studied in detail in Section 2.2, it once again seems that while the stationary distribution of the Gibbs measure is probably not an accurate model for the persistence diagram, treating the procedure as a mechanism for generating weakly dependent replications (perturbations) of the diagram is, nevertheless, appropriate, and following the behavior of the Wasserstein distance provides a useful practical tool for choosing MCMC parameters.

One fact that is worthwhile noting on comparing Figure S11 here and Figure S8 from the Gaussian example is the considerably higher (relative) variance of the estimated differences in the present case. At first thought, one might imagine that two circle case will be better behaved, since two circles create a much simpler topology than does a Gaussian field on a sphere. However, the fact that there are far more points in the persistence diagram in the Gaussian case leads to considerably more stability when studying properties of these diagrams.



(a) (b) (c)

(d) (e)

Figure S10: MCMC simulations of a persistence diagram arising from the two circle problem (a) The original diagram. (b)–(e) After 10, 25, 50, and 1,000 MCMC steps. See text for details.

## SI 3.2 Testing for homology

The statistics we chose to test for significant homology in the persistence diagram were the lifetime order statistics, $T_1, T_2, \ldots$ of all lifetimes than the 'infinite' one. Removing the infinite lifetime point, and given the remaining points $(d_i, b_i)$, of a persistence diagram, the $j$-th order statistic $T_j$ is the $j$-th largest among the differences $|d_i - b_i|$.

Having chosen statistics, we performed 1,000 MCMC simulations with the parameters (7) with a number of choices of the simulation parameters $n_b$, $n_r$ and $n_R$ defined in the paper. For each choice of these three parameters we computed both bootstrap style confidence intervals for the $T_j$ and $p$-values for the order statistics of the original persistence diagram. More formally, let $\mathbb{P}_j^*$ be the empirical distribution of $T_j$ over the 1,000 simulations, and $T_j^{org}$ the $j$-th order statistic of the original persistence diagram. Then a two-sided confidence interval at level $\alpha$ for the true $T_j$ is $[T_j^{org} - c_{j,1}, T_j^{org} + c_{j,2}]$ where

$$\begin{aligned} c_{j,1} &= \inf\left\{c \geq 0 : \mathbb{P}_j^*\left(T \leq T_j^{org} - c\right) \leq \alpha/2\right\}, \\ c_{j,2} &= \inf\left\{c \geq 0 : \mathbb{P}_j^*\left(T \geq T_j^{org} + c\right) \leq \alpha/2\right\}, \end{aligned}$$

and $T$ is a random variable with distribution $\mathbb{P}_j^*$. An alternative, one-sided confidence

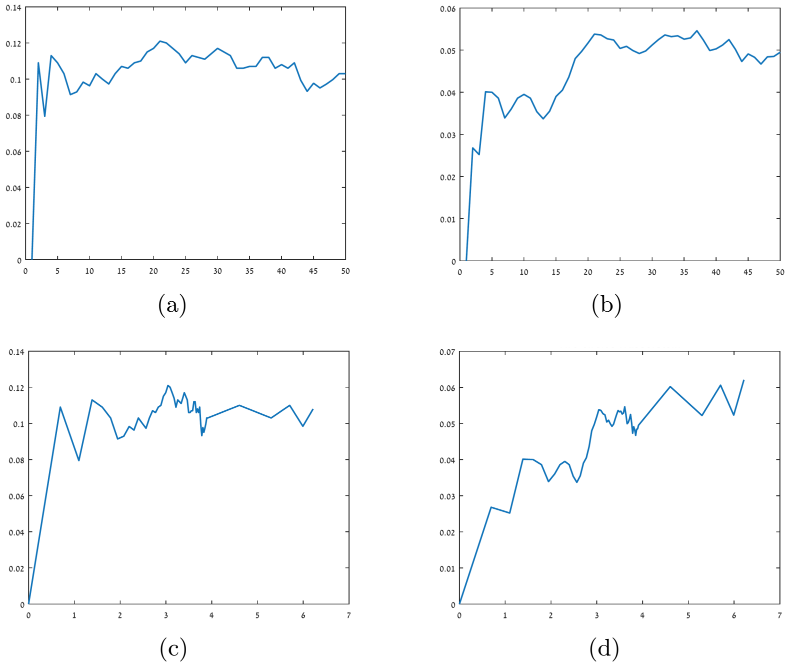


Figure S11: Growth of the bottleneck (a) and Wasserstein (b) differences of MCMC simulations from two-circle persistence diagrams (vertical axis), as a function of the number of steps $n_b$ (horizontal axis, $1 \leq n_b \leq 50$) averaged over 100 simulations of the diagrams. Panels (c) and (d) are on a logarithmic scale, for $1 \leq n_b \leq 500$.

interval $[0, T^{org} + c_j]$ can be defined by taking $c_j = c_{j,2}$ as above, but with $\alpha$ replacing $\alpha/2$ in the definition.

The $p$-value for $T_j^{org}$ is, similarly, defined as

$$p_j \;=\; \mathbb{P}_j^* \left( T > T^{org} \right).$$

On the basis that large values of the order statistics $T_j$ correspond to significant points in the $H_0$ persistence diagram – i.e. to connected components of the underlying object – we tested them sequentially, as follows. Firstly, we tested $T_1$ for significance with a one-sided test. If it was significant, only then did we test $T_2$, and only if it was significant did we test $T_3$, and so on.

Table S1 summaries the results of this procedure, for three different MCMC scenarios, each with a burn in period of 10 iterations and with $(n_b, n_r, n_R)$ given by (500,20,50), (500,40,25), and (500,100,10).

In view of the fact that it is unlikely that the stationary distribution of the Gibbs measure behind the MCMC is actually the right model for the persistence diagram, the agreement between the three scenarios is remarkable. In all cases, using either one-sided, 5% confidence intervals, or by considering $p$-values, the results indicated that $T_1$ and $T_2$ were highly significant, leading to the conclusion that the persistence diagram summarized an object with three components; viz. the component with infinite lifetime, and two more. Given that there are quite a few points in the persistence diagram close the the one corresponding to $T_2$, it is impressive that RST, which is most natural to apply in scenarios in which the diagram has a large number of points, worked so well in this small data situation, claiming only one superfluous component.

### SI 3.3 Alternative approaches

| Order Stat. | $T_j^{org}$ | $(n_b, n_r, n_R)$ | Conf. Interval | $p$-value |
|---|---|---|---|---|
| $T_1$ | 0.3161 | (500,20,50) | [0, 0.2430] | 0.0000 |
| | | (500,40,25) | [0, 0.2317] | 0.0010 |
| | | (500,100,20) | [0, 0.2271] | 0.0000 |
| $T_2$ | 0.2047 | (500,20,50) | [0, 0.1872] | 0.0160 |
| | | (500,40,25) | [0, 0.1887] | 0.0090 |
| | | (500,100,20) | [0, 0.1772] | 0.0110 |
| $T_3$ | 0.1535 | (500,20,50) | [0, 0.1669] | 0.1080 |
| | | (500,40,25) | [0, 0.1753] | 0.1220 |
| | | (500,100,20) | [0, 0.1548] | 0.0560 |

Table S1: Testing order statistics in the $H_0$ persistence diagram for two circles, showing one-sided 5% confidence intervals and $p$-values. See text for details.

Although the two-circle example is meant to be primarily illustrative of how RST works, and, because of the small number of points in the persistence diagram is not its natural scenario, it is almost obligatory that we compare the results to those from some existing methods.

Consequently, we also undertook an analysis using the bootstrap cum confidence band approaches of [12] and [20]. (As usual, see [43] for more detailed references.) In brief, this approach produces replicates of the persistence diagram by subsampling either the data or the diagram, and then produces confidence intervals based the bottleneck norm of the diagrams. This yields '$100(1-\alpha)$% acceptance regions', as in Figures S12 and S13, which are parallel to the diagonal, and within which all but $100(1-\alpha)$% of the subsampled persistence diagrams fall. All points outside these regions are then considered 'significant' at the $100\alpha$% level, and indicative of true homology.

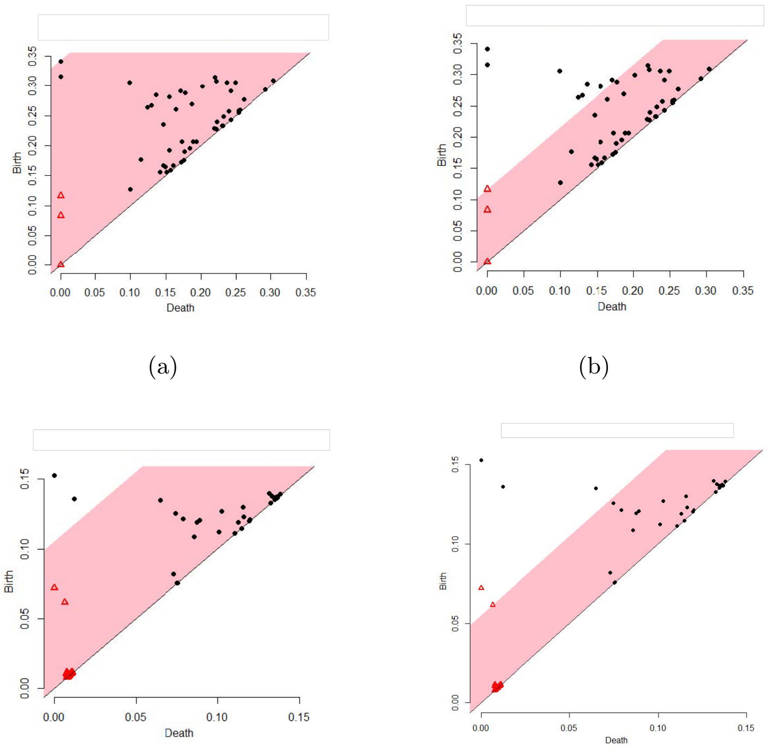
Applying the techniques of [12] and [20] to our persistence diagram, with $\alpha = 0.05$, gave the results in (a) and (b) of Figure S12 and (a) of Figure S13, respectively.

Figure S12 (a) shows that only one $H_0$ component is identified, and Figure S13 (a) identifies none. RST did much better than this, identifying two components correctly and providing marginal (but incorrect) for a third. On the other hand, Figure S12 (b) (marginally) identifies one $H_0$ component, whereas RST was not applicable here at all due to the small number of points in the $H_0$ diagram. The procedure of [20] leads to a single confidence interval for both homologies, and so Figure S13 (a) is also relevant for $H_1$. Again, it fails to identify any significant points.

Thus, the evidence at this point is that RST is doing very well against existing methods, at least as far as $H_0$ is concerned.

However, a careful referee decided to check out calculations. One parameter value that we have not defined yet is the bandwidth $h$ in equation (1) of the main paper. This was the bandwidth used to compute the empirical density shown in Figure 1 there, the upper level sets of which produced the persistence diagrams with which we have worked throughout. The value we took, somewhat arbitrarily, was $h = 0.1$, but in that version of the paper we neglected to mention this, so the referee carried out his/her checks with another value, in this case $h = 0.2$. The results are shown in Figures S12 (c) and (d) and S13 (b). As is clear from there, with this bandwidth the methods of [12] correctly identify the number of points in both the $H_0$ and $H_1$ diagrams, while those of [20] are still outperformed by RST.

Overall, however, it is important to re-emphasise that RST is a large data technique. Thus we now turn to a real data scenario, where the persistence diagrams contain large numbers of points, and where we believe RST comes into its own as a promising alternative, and complement, to existing methods.



(a)

(b)

(c)

(d)

Figure S12: Confidence bands for the persistence diagram of the two circles, using the bootstrap based techniques of [12], (a) $H_0,\ h = 0.1$. (b) $H_1,\ h = 0.1$ (c) $H_0,\ h = 0.2$. (d) $H_1,\ h = 0.2$ .

## SI 4. Example 2: Analyzing CMB data

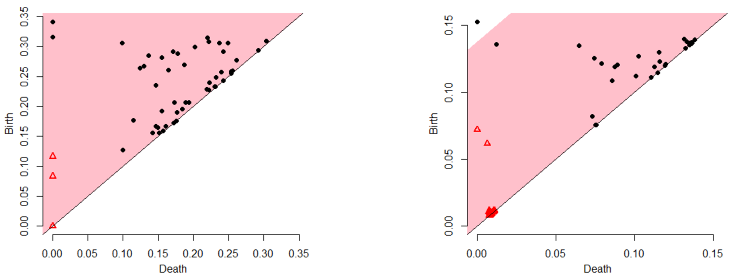
In the main paper, Figure 2 showed a reconstructed version of the CMB data from the Plank experiment, created using the Commander-Rule technique, without saying too much about what this was.

Although it would take far too much space to carefully define precisely what 'reconstructed' actually means, it is nevertheless worthwhile to devote a paragraph or two to various technical aspects of CMB data before we describe the RST analysis.

### SI 4.1 The data

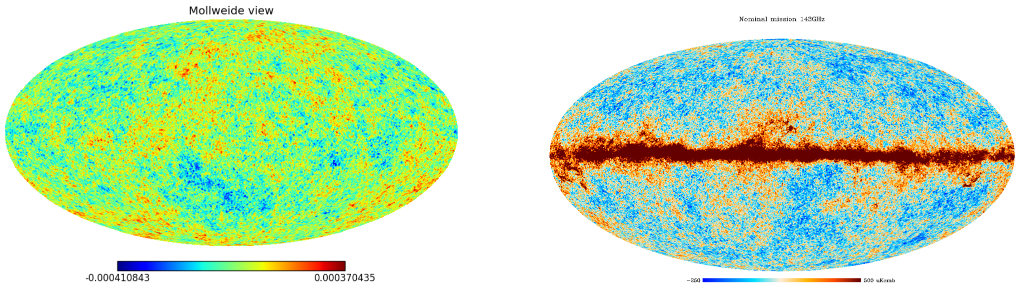
As described in the main paper, CMB is real-valued data on a sphere. More precisely, the CMB sky maps are presented in the HealPix [22] format, already mentioned in Section SI 2.2, which is based on a recursive equal-area pixelization of the sphere. It starts, using the faces of a rhombic dodecahedron, by decomposing the sphere into twelve patches of equal area. Further resolution is achieved by dividing these 12 base patches into $N^2$ equal area patches, so that the total number of patches at this resolution is $12N^2$. Maximum resolution is at $N = 2048$. Figure S14 (a) shows a simulated Gaussian random field using this scheme, with $N = 128$.

While the random function in (a) of Figure S14 looks not unlike the reconstructed CMB of Figure 2 of the main paper, unfortunately neither of them look much like 'raw' CMB data, or so-called 'fiducial sky' in (b). The map here is based on measurements which include instrument noise, astrophysical foregrounds, and various lensed scalar, tensor, and



(a) (b)

Figure S13: Confidence bands for the persistence diagram of the two circles, using the bootstrap based techniques of [20]. (a) $H_0, H_1,\ h = 0.1$. (b) $H_0, H_1,\ h = 0.2$.



(a) (b)

Figure S14: (a) Illustration of a 2-dimensional Gaussian random field simulated on the surface of the sphere, at the same resolution of the CMB in (b). (b) The fiducial sky as observed by the Planck satellite at frequency band 143GHz. Figure courtesy the Planck consortium.

non-Gaussian CMB components, and are taken at a particular frequency (in this case, 143GHz). The most obvious problem with the raw data of Figure S14(b) is the large 'belt' around the 'equator'. This is a region of the sky blocked by the Milky Way, and so there are no reliable measurements of the CMB in this region. Similarly problematic regions, of various sizes, appear throughout the sky, due to a variety of cosmological phenomena. Careful study of CMB must therefore take all of this into account.

In a masterpiece of data cleaning, involving classical physics, cosmological knowledge, and a lot of modeling, all these factors are (as much as possible) accounted for. Furthermore, the data from the eight observed frequency bands are combined, using what is known as a 'component separation technique', to produce the final 'optimized', data, known as a 'reconstructed map'. One aspect of this cleaning, of specific importance for us, is that data from the poorly observed regions are replaced with what are, effectively, interpolations based on simulations of isotropic, Gaussian, random fields. The Planck team produces these reconstructed maps through four different techniques: Commander-Ruler (C-R), NILC, SEVEM, and SMICA. A detailed description of these procedures can be found in [33]. Throughout this paper we work with the C-R maps, which is what produced the CMB of Figure 2 of the main paper.

It is accepted practice that, for many cosmological purposes, it is sufficient to work with reconstructed maps, which is what we do.

Recall that our aim is to exploit RST to test for inhomogeneity in the reconstructed

CMB, and we do this by fitting Gibbs measures to the persistence diagrams generated by the CMB in different parts of the sky, comparing the parameters in the corresponding Hamiltonians. From the brief description of CMB above, it is clear that in doing so we should avoid using data from near the equator. All such data is based on a reconstruction which *assumes* both Gaussianess and homogeneity, and so is unreliable from our point of view.

Consequently, we concentrate on two regions, which we will call the the 'northern and southern caps' of the CMB data, and which correspond to the top and bottom 60 degrees of data. This cuts out 30 degrees on either side of the equator, and so takes us into regions in which we expect reconstruction will have only minimal effect on the topology of the upper level sets which we use as filtrations for persistence diagrams.

### SI 4.2 The Gibbs model

As described in the main paper, we took 5 smoothed versions of the C-R reconstructed CMB, restricted to the northern and southern caps, each with different Gaussian smoothing kernels, at full width half maximum 300, 180, 120, 90, and 60 arcminutes. The highest level of smoothing (300') suppresses most of the fine scale variation, while the 60' level is closest to the actual CMB.

For each such smoothing, we produce persistence diagrams generated by the upper level set filtration, for both $H_0$ and $H_1$, leading to a total of $20 = 5 \times 2 \times 2$ diagrams. Although the aims there are different, details of the numerical procedure can be found in [36], and examples of two persistence diagrams are given in Figure 3 of the main paper. The parameter estimates for all 20 cases are given in Table S2 for the $H_0$ persistence diagrams, and in Table S3 for the $H_1$ diagrams. At this point, only the first two lines for each smoothing parameter are relevant. These are the parameter estimates.

As before, the questions of the stability of the parameter estimates, and of the effectiveness of the Gibbs measure as a model for the persistence diagrams, need to be addressed. However, we can assume that, at least in a null hypothesis scenario, CMB behaves like the realization of an isotropic Gaussian random field. Consequently, the discussion of the Gaussian case, in Section SI 2.2, along with extensive simulations discussed there, also serve for justification in the current situation.

### SI 4.3 Testing for inhomogeneity

In order to test for CMB inhomogenity, as described in the main paper, we need to compare the Gibbs measure parameter estimates for the two caps given in Tables S2 and S3.

(There are, of course, many other methods that one could use to test for differences between the two caps, even when restricting the methodology to that of TDA. These could include measuring the differences between their upper level set persistence diagrams with a standard TDA metric, such as the bottleneck metric, or comparing their Euler characteristic curves, which is a common diagnostic tool in cosmology (e.g. [23, 38]), or measuring differences between persistence landscapes. However, at the risk of belaboring the by-now obvious, since for each cap, and for each smoothing, there is only one observation, and additional universes that might allow us to obtain repeated observations are not readily available, all of the alternative methods will require some form of artificial replication. Planning for RST requires estimating the parameters of Tables S2 and S3 and so, with these at hand, working with them for hypothesis testing is a natural choice.)

Thus, given the parameter estimates in Tables S2 and S3, we now adopt a rather simple approach, that of pairwise tests of significance. However, in order to carry these out, we need estimates of variance for the parameter estimates themselves. There are at least four

| Smoothing | | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_3$ | $\hat{\theta}_H$ | $\hat{\theta}_V$ |
|---|---|---|---|---|---|---|
| 300' | North | -0.3582 | -0.2187 | -0.2258 | 1.0156 | 0.7685 |
| | South | -0.6140 | -0.4573 | -0.3941 | 0.6287 | 0.5120 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0283 | 0.0357 | 0.0381 | 0.2027 | 0.1622 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0656 | 0.0682 | 0.0662 | 0.1619 | 0.1500 |
| | $p$-value | 0.0003 | 0.0019 | 0.0276 | 0.1357 | 0.2455 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0827 | 0.1177 | 0.1572 | 0.1738 | 1.8287 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0785 | 0.1204 | 0.1521 | 0.2231 | 2.0230 |
| | $p$-value | 0.0249 | 0.1564 | 0.4418 | 0.1713 | 0.9251 |
| 180' | North | -0.2939 | -0.1869 | -0.1073 | 0.8262 | 0.8259 |
| | South | -0.2631 | -0.3757 | -0.2556 | 0.6410 | 0.8151 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0144 | 0.0169 | 0.0182 | 0.1015 | 0.1043 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0271 | 0.0268 | 0.0261 | 0.0790 | 0.1036 |
| | $p$-value | 0.3155 | 0.0000 | 0.0000 | 0.1500 | 0.9414 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0424 | 0.0557 | 0.0663 | 0.1269 | 0.6037 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0490 | 0.0624 | 0.0686 | 0.1591 | 0.7189 |
| | $p$-value | 0.6344 | 0.0240 | 0.1198 | 0.3627 | 0.9908 |
| 120' | North | -0.2039 | -0.2033 | -0.1719 | 0.8333 | 0.8623 |
| | South | -0.2784 | -0.2320 | -0.2072 | 0.5185 | 0.7703 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0105 | 0.0105 | 0.0102 | 0.0672 | 0.0695 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0095 | 0.0100 | 0.0103 | 0.0454 | 0.0665 |
| | $p$-value | 0.0000 | 0.0478 | 0.0145 | 0.0001 | 0.3386 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0257 | 0.0298 | 0.0332 | 0.1091 | 0.3027 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0260 | 0.0332 | 0.0440 | 0.1390 | 0.3113 |
| | $p$-value | 0.0415 | 0.5196 | 0.5214 | 0.0748 | 0.8322 |
| 90' | North | -0.1891 | -0.1980 | -0.1664 | 0.7993 | 0.7730 |
| | South | -0.2446 | -0.1857 | -0.1785 | 0.5753 | 0.7771 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0062 | 0.0060 | 0.0061 | 0.0488 | 0.0477 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0040 | 0.0055 | 0.0058 | 0.0357 | 0.0475 |
| | $p$-value | 0.0000 | 0.1327 | 0.1525 | 0.0000 | 0.9515 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0207 | 0.0239 | 0.0277 | 0.0975 | 0.2010 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0219 | 0.0259 | 0.0293 | 0.1253 | 0.2007 |
| | $p$-value | 0.0658 | 0.7276 | 0.7647 | 0.1582 | 0.9885 |
| 60' | North | -0.2363 | -0.2047 | -0.1456 | 0.7301 | 0.5638 |
| | South | -0.2361 | -0.2027 | -0.1847 | 0.6121 | 0.6294 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0027 | 0.0027 | 0.0003 | 0.0355 | 0.0272 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0024 | 6.5e-06 | 5.9e-06 | 0.0277 | 0.0275 |
| | $p$-value | 0.9459 | 0.4670 | 0.0000 | 0.0087 | 0.0901 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0301 | 0.0164 | 0.0181 | 0.0892 | 0.1156 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0530 | 0.0270 | 0.0244 | 0.1239 | 0.1329 |
| | $p$-value | 0.9968 | 0.9506 | 0.1984 | 0.4396 | 0.7096 |

Table S2: $H_0$ persistence diagram parameter estimates, along with estimates of variance and $p$-values for tests of North vs. South. See text for details.

| Smoothing | | $\hat{\theta}_1$ | $\hat{\theta}_2$ | $\hat{\theta}_3$ | $\hat{\theta}_H$ | $\hat{\theta}_V$ |
|---|---|---|---|---|---|---|
| 300' | North | -0.2994 | -0.4487 | -0.3156 | 0.7410 | 1.2152 |
| | South | -0.4545 | -0.4416 | -0.4362 | 0.3474 | 0.7910 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0022 | 0.0022 | 0.0023 | 0.1944 | 0.3435 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0043 | 0.0042 | 0.0041 | 0.0877 | 0.1975 |
| | $p$-value | 0.0536 | 0.9286 | 0.1318 | 0.0652 | 0.2844 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.1301 | 0.1423 | 0.1914 | 0.1701 | 0.7001 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.1228 | 0.1513 | 0.2010 | 0.2002 | 0.5961 |
| | $p$-value | 0.3861 | 0.9725 | 0.6638 | 0.1341 | 0.6446 |
| 180' | North | -0.3175 | -0.1709 | -0.0824 | 0.6763 | 1.0586 |
| | South | -0.2635 | -0.4110 | -0.4796 | 0.4220 | 0.8218 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0140 | 0.0172 | 0.0198 | 0.0816 | 0.1315 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0402 | 0.0402 | 0.0379 | 0.0618 | 0.1197 |
| | $p$-value | 0.2048 | 0.0000 | 0.0000 | 0.0130 | 0.1829 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0520 | 0.0686 | 0.0819 | 0.1095 | 0.4294 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0520 | 0.0592 | 0.0735 | 0.1288 | 0.3872 |
| | $p$-value | 0.4627 | 0.0080 | 0.0000 | 0.1327 | 0.6821 |
| 120' | North | -0.2254 | -0.2436 | -0.1845 | 0.5986 | 0.8074 |
| | South | -0.2871 | -0.1934 | -0.1734 | 0.4229 | 0.9385 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0123 | 0.0122 | 0.0119 | 0.0501 | 0.0721 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0092 | 0.0104 | 0.0112 | 0.0353 | 0.0767 |
| | $p$-value | 0.0001 | 0.0017 | 0.4996 | 0.0042 | 0.2128 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0300 | 0.0332 | 0.0387 | 0.0849 | 0.2661 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0332 | 0.0346 | 0.0436 | 0.1054 | 0.2735 |
| | $p$-value | 0.1675 | 0.2944 | 0.8498 | 0.1940 | 0.7312 |
| 90' | North | -0.2176 | -0.1896 | -0.2125 | 0.5579 | 0.7206 |
| | South | -0.2475 | -0.1942 | -0.1922 | 0.4705 | 0.8633 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0067 | 0.0068 | 0.0066 | 0.0367 | 0.0499 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0045 | 0.0051 | 0.0056 | 0.0292 | 0.0535 |
| | $p$-value | 0.0000 | 0.5939 | 0.0193 | 0.0625 | 0.0511 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.0257 | 0.0211 | 0.0273 | 0.0678 | 0.1828 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.0214 | 0.0255 | 0.0290 | 0.0854 | 0.1811 |
| | $p$-value | 0.3712 | 0.8906 | 0.6107 | 0.4230 | 0.5792 |
| 60' | North | -0.1956 | -0.1806 | -0.1646 | 0.5132 | 0.6361 |
| | South | -0.1710 | -0.2273 | -0.2420 | 0.4330 | 0.6130 |
| | $\hat{\sigma}_{N,\text{Info}}$ | 0.0031 | 0.0031 | 0.0031 | 0.0238 | 0.0298 |
| | $\hat{\sigma}_{S,\text{Info}}$ | 0.0065 | 0.0053 | 0.0053 | 0.0201 | 0.0280 |
| | $p$-value | 0.0006 | 0.0000 | 0.0000 | 0.0101 | 0.5720 |
| | $\hat{\sigma}_{N,\text{Sim}}$ | 0.5626 | 0.2799 | 0.1279 | 0.0856 | 0.1373 |
| | $\hat{\sigma}_{S,\text{Sim}}$ | 0.5291 | 0.2052 | 0.0740 | 0.0770 | 0.2277 |
| | $p$-value | 0.9746 | 0.8931 | 0.6005 | 0.4861 | 0.9308 |

Table S3: $H_1$ persistence diagram parameter estimates, along with estimates of variance and $p$-values for tests of North vs. South. See text for details.

ways to obtain these:

(i) Subsampling: Subsample the persistence diagrams, produce estimates of the $\theta_k$ for each subsample, and use these to estimate their variances. Note that the subsampling must be done on the diagrams themselves, and not on the original CMB data, since there are no natural replications of the CMB.

(ii) Fisher information matrices: In estimating the $\theta_k$, the Hessian of the log (pseudo) likelihood at its maximum can be used to produce estimates of not only the required variances, but also of the full covariance matrix of the $\hat{\theta}_k$.

(iii) MCMC resampling: Having estimated the $\theta_k$, we have a model for the persistence diagram, which can then be simulated by MCMC as described in the main paper. Re-estimating the parameters from such simulations would give information on the variance (and, indeed, full distribution) of the original set of estimates.

(iv) The HEALPix software that we have already mentioned produces not only simulations of purely Gaussian random fields, but also produces simulations of the 'true, reconstructed, CMB'. The estimation process can be repeated on these simulations to obtain distributional information on the $\hat{\theta}_k$.

We did not try (i), since we were primarily interested in seeing how RST worked. Nevertheless it is an obvious route to take in the topological investigation of CMB.

Approach (iii) is natural within the framework of RST, and we experimented with it. In the end, we will not report on detailed results, since they were similar to those from (iv), which is the approach we adopted in the end. One problem with (iii) was the amount of time it took to produce simulations under the MCMC scheme, whereas the HEALPix simulations of (iv) were relatively quick and easy to generate. Another, more conceptual problem, was that it is unclear what impact estimation error in the original set of estimates has on variance estimates from the MCMC. We plan to study this phenomenon in the future on smaller scale problems.

This leaves us with (ii) and (iv). Results in an earlier version of the main paper were based purely on (ii), but we now add (iv), which we believe to be more reliable. However it is also more time and computer consuming. These methods lead to the additional rows in Tables S2 and S3, which we now explain.

For each smoothing level in these tables, we have already noted that the lines 'North' and 'South' give the estimates of the $\theta_k$ via the pseudo-likelihood approach described in the main paper. The two lines following give estimates of the standard deviations $\hat{\sigma}_{N,\mathrm{Info}}$ and $\hat{\sigma}_{S,\mathrm{Info}}$ of these estimates (for north and south, respectively) based on the sample Fisher information matrix.

The row following the standard deviations gives $p$-values for simple pairwise tests of the differences

$$\Delta_k \;=\; \frac{\hat{\theta}_k^N - \hat{\theta}_k^S}{\sqrt{\hat{\sigma}^2_{N,\mathrm{Info}} + \hat{\sigma}^2_{S,\mathrm{Info}}}}, \tag{8}$$

where the $N$ and $S$ superscript on the $\hat{\theta}_k$ indicate north or south. Inference for the $\Delta_k$, based on their approximate normality, with the empirical Fisher information matrix as the asymptotic covariance matrix, is as justified in our Gibbs, pseudo-likelihood scenario much as it is in the classical maximum likelihood scenario (cf. [11, 14]). The $p$-values that are significant at the 5% level are highlighted.

The three lines following, for each smoothing level, are similar, except that the estimates of the standard deviations, $\hat{\sigma}_{N,\mathrm{Sim}}$ and $\hat{\sigma}_{S,\mathrm{Sim}}$, are now taken from the simulation procedure described in (iv) above. Again, $p$-values that are significant at the 5% level are highlighted.

Since there are multiple tests being carried out here, and it is unlikely that they are independent, it is difficult to draw definitive conclusions from Tables S2 and S3. (Since the $p$-values are given explicitly, the reader is invited to perform any of the standard techniques for multiple testing, such as the classical Bonferroni correction, the Benjamini-Hochberg FDR procedure [5], or a chi-squared goodness of fit test. We actually did so, but did not find the results particularly illuminating.) Nevertheless, the highlighting in the tables leads to two distinct visual impressions:

(i) The tests based on the variances computed via the estimated information matrix yield more statistically significant results (19) than do the tests based on simulation derived variances (5). It is hard to know whether this means that this former approach is more powerful, or whether it simply leads to false positives. Experimentation with simulations leads us to believe that the latter situation is the case, and so we have more faith in the simulation based results. However, since all simulations are based on imperfect models, that, among other things, actually assume isotropy, it is hard to be completely definite about this.

(ii) Most of the significant north/south differences occur at the middle levels of smoothing. This is particularly noticeable in Table S3, for the $H_1$ parameters, but also, to some extent, in Table S2. Table S2 also shows differences at the level of 300 arcmins of smoothing.

It turns out that there are good cosmological explanations for these impressions, and we will see in Section SI 4.4 below.

The result of all this analysis is that adopting a parametric approach via Gibbs measures, at least as applied to the complex CMB data set, leads to some interesting patterns of results. That, for this data, the patterns are indicative, but not definitive, is not surprising. CMB is famously difficult data to analyse, and even hints of new phenomena are of considerable interest to cosmologists.

In the following, final, section we give the briefest of cosmological interpretations of what the above analysis has shown us.

### SI 4.4 Cosmological interpretations of the statistical analysis

Our aim in this final section is not to delve deeply into Cosmology, but rather to put the statistical findings of the previous section – of a pattern of topological differences between the northern and southern CMB, as seen via Gibbs measure parameter estimates – into perspective for the non-cosmologist.

The hypothesis of asymmetry in the temperature distribution between the northern and the southern hemispheres of the CMB has been the focus of many studies in the past decade and a half, with pioneering works by Eriksen et al [19], who studied it via power spectrum estimation, and Park [32], who studied it using topological statistics based on the genus (Euler characteristic) of upper level sets. In both cases they worked with WMAP data, but they worked at different 'angular scales', a term which requires some explanation.

An important issue in CMB analysis is the issue of smoothing. This is related to the physical fact that points in the CMB sphere that are separated by a degree or larger, are not causally connected, in the sense that they are so far away from one another that they are less susceptible to influences from late time effects that may alter the CMB photon characteristics. (Recall that the measured CMB is coming from sources some 13 billion light years away, so that a one degree of angular separation translates into a distance of the order of hundreds of megaparsecs between the sources.)

For example, it is known that, at small scales of the order of arcminutes, the frequency of the CMB photons is Doppler-shifted due to the influence of the gravitational potential generated by the intermediate matter distribution between the CMB sphere and the observer [41]. Such effects are negligible at larger scales.

Returning now to the literature, we note that while [19] focuses on large angular scales (typically $2° - 5°$), while [32] focuses on sub-degree scales. Since the appearance of these papers, studying the north-south asymmetry in the CMB has become a standard practice, culminating, most recently, in the studies using data from the Planck satellite [34]. Regardless of the issue of angular scales, the general consensus (as opposed to 'agreement in the face of overwhelming evidence') is that the properties of the temperature distribution in the north appear to be different from those in the south.

As in [19], we, throughout, have concentrated on scales larger than one degree. Thus, if the smoothing is performed, as in our examples, with Gaussian kernels with 'full width, half maximum' parameter $\alpha$-arcmin (roughly twice the standard deviation $\sigma$) with

$$\alpha \;=\; 60, 90, 120, 180, 300,$$

this translates into angular scales of between 2–5 angular degrees.

Consequently, at least at the higher levels of smoothing, most our results should be free of most late time effects, and any differences that the above analysis shows should be true CMB phenomena, unaffected by late time effects.

With this in mind, let us look back at the results in Tables S2 and S3. Both tables show that many parameter estimates differ significantly between the north and the south caps, for both $H_0$ and $H_1$ parameter sets, and for scales approximately $2° - 5°$. Overall, the differences are more marked at the higher smoothing scales. This is consistent with the findings of [19]. This is also roughly the scale at which we found additional anomalous behaviour in the CMB with respect to the standard cosmological model; cf. [35].

In closing, we note that the indications of non-homogeneity at these scales (and even possible non-Gaussianity, which would be consistent with our results, although we did not explicitly consider this as an explanation for them) which we found in the previous section are studied, in far more detail, in [35]. The methodology of [35] is also topological in nature, but quite different to what we have developed in the current paper. Here we have been motivated by adopting a vanilla approach, based on a generic analysis of persistence diagrams, that will work for a wide variety of cases. [35] develops more powerful topological methods, specifically tailored to CMB analysis.

## Appendix

**Table A1**. Parameter estimates for 100 simulations of a Gaussian process on a sphere. (a) Southern cap. (b) Northern cap. See text for details.

| (a) Parameter estimates: Southern cap | | | | | |
|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_H$ | $\theta_V$ |
| **1** | -0.2092 | -0.2239 | -0.1883 | 0.5594 | 1.4782 |
| **2** | -0.1500 | -0.2616 | -0.2495 | 0.5436 | 1.3001 |
| **3** | -0.2204 | -0.1641 | -0.1491 | 0.7893 | 1.2806 |
| **4** | -0.2355 | -0.2310 | -0.2269 | 0.5089 | 1.2028 |
| **5** | -0.2721 | -0.2066 | -0.2028 | 0.7364 | 0.8027 |
| **6** | -0.2188 | -0.1837 | -0.1936 | 0.4859 | 1.2384 |
| **7** | -0.2386 | -0.2149 | -0.1994 | 0.5397 | 1.0558 |
| **8** | -0.2463 | -0.2083 | -0.1686 | 0.5081 | 1.3957 |
| **9** | -0.2327 | -0.1939 | -0.1412 | 0.6217 | 1.2619 |
| **10** | -0.2316 | -0.1875 | -0.1378 | 0.5072 | 1.4256 |
| | | | | | |
| **11** | -0.2615 | -0.2008 | -0.1703 | 0.5536 | 1.0217 |
| **12** | -0.2747 | -0.2061 | -0.1399 | 0.5137 | 1.3893 |
| **13** | -0.2563 | -0.2002 | -0.1721 | 0.4996 | 1.1939 |
| **14** | -0.2149 | -0.2230 | -0.1854 | 0.5906 | 1.0959 |
| **15** | -0.2145 | -0.2283 | -0.1845 | 0.6051 | 1.1344 |
| **16** | -0.2072 | -0.2709 | -0.2245 | 0.4982 | 1.0845 |
| **17** | -0.2703 | -0.2016 | -0.1728 | 0.3851 | 1.4760 |
| **18** | -0.2514 | -0.1844 | -0.1682 | 0.5681 | 1.0986 |
| **19** | -0.2535 | -0.1536 | -0.1858 | 0.6543 | 1.2848 |
| **20** | -0.2405 | -0.2724 | -0.2390 | 0.4728 | 1.0321 |
| | | | | | |
| **21** | -0.2371 | -0.1691 | -0.1654 | 0.3964 | 1.3287 |
| **22** | -0.2591 | -0.2019 | -0.1907 | 0.4524 | 1.6082 |
| **23** | -0.2471 | -0.1879 | -0.1411 | 0.4297 | 1.8069 |
| **24** | -0.2417 | -0.2210 | -0.1371 | 0.5668 | 1.1758 |
| **25** | -0.1898 | -0.2742 | -0.2700 | 0.4817 | 1.2852 |
| **26** | -0.2651 | -0.2183 | -0.1630 | 0.5554 | 1.1705 |
| **27** | -0.2655 | -0.2278 | -0.1551 | 0.5741 | 1.1028 |
| **28** | -0.2053 | -0.2878 | -0.2669 | 0.4904 | 1.2583 |
| **29** | -0.1963 | -0.1840 | -0.1607 | 0.6254 | 1.7888 |
| **30** | -0.1941 | -0.2592 | -0.2448 | 0.4580 | 1.2140 |
| | | | | | |
| **31** | -0.2134 | -0.1500 | -0.0947 | 0.6826 | 1.4140 |
| **32** | -0.2345 | -0.1527 | -0.1407 | 0.5984 | 1.3556 |
| **33** | -0.2906 | -0.2274 | -0.1986 | 0.3093 | 1.2103 |
| **34** | -0.2527 | -0.2111 | -0.2003 | 0.4527 | 1.6455 |
| **35** | -0.2506 | -0.1920 | -0.1814 | 0.5144 | 1.0174 |
| **36** | -0.2171 | -0.2467 | -0.2183 | 0.4760 | 1.0795 |
| **37** | -0.2515 | -0.2507 | -0.2074 | 0.4379 | 1.1064 |
| **38** | -0.2471 | -0.2412 | -0.2168 | 0.4081 | 1.3466 |
| **39** | -0.2236 | -0.2664 | -0.2308 | 0.4046 | 1.1544 |
| **40** | -0.2377 | -0.2489 | -0.2270 | 0.4403 | 1.2640 |
| | | | | | |
| **41** | -0.2680 | -0.2289 | -0.1843 | 0.3855 | 1.1621 |
| **42** | -0.2471 | -0.2519 | -0.2515 | 0.5195 | 0.9502 |
| **43** | -0.1992 | -0.2109 | -0.2316 | 0.5558 | 1.2306 |
| **44** | -0.2503 | -0.2320 | -0.2438 | 0.4569 | 1.0905 |
| **45** | -0.2334 | -0.2059 | -0.1631 | 0.7309 | 1.2434 |
| **46** | -0.2462 | -0.2152 | -0.1632 | 0.4059 | 1.5510 |
| **47** | -0.2141 | -0.2525 | -0.2283 | 0.4486 | 1.3947 |
| **48** | -0.2470 | -0.2698 | -0.2316 | 0.4167 | 1.1914 |
| **49** | -0.2181 | -0.1846 | -0.1589 | 0.6608 | 1.1253 |
| **50** | -0.2555 | -0.2603 | -0.2289 | 0.3548 | 1.4489 |
| **51** | -0.2598 | -0.2503 | -0.2055 | 0.3877 | 1.1845 |
| **52** | -0.2472 | -0.2469 | -0.2410 | 0.4471 | 1.1756 |
| **53** | -0.2464 | -0.2671 | -0.2773 | 0.3979 | 1.1802 |
| **54** | -0.2179 | -0.2448 | -0.2155 | 0.6909 | 1.1448 |
| **55** | -0.2212 | -0.2403 | -0.1455 | 0.4822 | 1.4229 |
| **56** | -0.2503 | -0.1931 | -0.1464 | 0.5515 | 1.0624 |
| **57** | -0.2190 | -0.2334 | -0.1935 | 0.4729 | 1.3464 |
| **58** | -0.2697 | -0.2250 | -0.1769 | 0.4601 | 1.4426 |
| **59** | -0.1880 | -0.2533 | -0.3154 | 0.4562 | 1.1015 |
| **60** | -0.2120 | -0.2395 | -0.1897 | 0.5695 | 1.1943 |
| | | | | | |
| **61** | -0.1970 | -0.2359 | -0.1664 | 0.6421 | 1.4194 |
| **62** | -0.2229 | -0.2372 | -0.2375 | 0.5679 | 1.1138 |
| **63** | -0.2706 | -0.2152 | -0.1657 | 0.4865 | 1.2152 |
| **64** | -0.2589 | -0.1827 | -0.1373 | 0.5156 | 1.3990 |
| **65** | -0.2119 | -0.1959 | -0.1702 | 0.6085 | 1.3614 |
| **66** | -0.2568 | -0.2081 | -0.1460 | 0.4760 | 1.1390 |
| **67** | -0.1659 | -0.2883 | -0.2568 | 0.4460 | 1.3428 |
| **68** | -0.2711 | -0.2389 | -0.2072 | 0.4093 | 1.4778 |
| **69** | -0.2071 | -0.2510 | -0.2008 | 0.4833 | 1.5442 |
| **70** | -0.2462 | -0.2160 | -0.2112 | 0.5090 | 1.4840 |
| | | | | | |
| **71** | -0.2200 | -0.2315 | -0.2199 | 0.4662 | 1.3638 |
| **72** | -0.2340 | -0.2208 | -0.1871 | 0.5009 | 1.3085 |
| **73** | -0.2214 | -0.1882 | -0.1888 | 0.6735 | 1.2476 |
| **74** | -0.2307 | -0.2555 | -0.2081 | 0.4888 | 1.1051 |
| **75** | -0.2410 | -0.2326 | -0.2083 | 0.4569 | 1.0430 |
| **76** | -0.1836 | -0.2207 | -0.2143 | 0.6246 | 1.6236 |
| **77** | -0.2539 | -0.2245 | -0.2205 | 0.5389 | 1.0371 |
| **78** | -0.2544 | -0.1809 | -0.1677 | 0.6192 | 1.0170 |
| **79** | -0.2656 | -0.2614 | -0.2571 | 0.3339 | 1.3311 |
| **80** | -0.2287 | -0.2292 | -0.1883 | 0.5226 | 1.0944 |
| | | | | | |
| **81** | -0.2254 | -0.2898 | -0.2546 | 0.4140 | 1.3739 |
| **82** | -0.2290 | -0.1460 | -0.1202 | 0.7240 | 1.3505 |
| **83** | -0.1973 | -0.2155 | -0.1683 | 0.9722 | 1.3143 |
| **84** | -0.2330 | -0.2263 | -0.2187 | 0.5738 | 0.9325 |
| **85** | -0.1889 | -0.2420 | -0.2160 | 0.5123 | 1.4870 |
| **86** | -0.2488 | -0.1883 | -0.1904 | 0.5216 | 1.1991 |
| **87** | -0.1976 | -0.2452 | -0.2022 | 0.5332 | 1.0335 |
| **88** | -0.2061 | -0.2222 | -0.1978 | 0.6464 | 1.2531 |
| **89** | -0.2442 | -0.2338 | -0.2266 | 0.4994 | 1.0302 |
| **90** | -0.2369 | -0.2268 | -0.1812 | 0.5597 | 1.0745 |
| | | | | | |
| **91** | -0.2332 | -0.2108 | -0.2126 | 0.5089 | 1.3182 |
| **92** | -0.2142 | -0.2867 | -0.2716 | 0.6230 | 0.9782 |
| **93** | -0.2320 | -0.1894 | -0.1754 | 0.6236 | 1.3597 |
| **94** | -0.2275 | -0.2540 | -0.2140 | 0.5309 | 1.1414 |
| **95** | -0.2064 | -0.2229 | -0.2341 | 0.4669 | 1.5399 |
| **96** | -0.2395 | -0.1955 | -0.1605 | 0.5228 | 1.2067 |
| **97** | -0.2445 | -0.2298 | -0.1706 | 0.6380 | 1.1417 |
| **98** | -0.2513 | -0.2173 | -0.2283 | 0.5110 | 1.0886 |
| **99** | -0.2473 | -0.2628 | -0.2382 | 0.3310 | 1.2741 |
| **100** | -0.2244 | -0.1577 | -0.1218 | 0.7995 | 1.2935 |

| **(b) Parameter estimates: Northern cap** | | | | | |
|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_H$ | $\theta_V$ |
| **1** | -0.2961 | -0.2401 | -0.2175 | 0.5008 | 1.1089 |
| **2** | -0.2698 | -0.2302 | -0.1761 | 0.4428 | 1.3066 |
| **3** | -0.2958 | -0.2491 | -0.1866 | 0.4061 | 1.2390 |
| **4** | -0.2474 | -0.2331 | -0.1807 | 0.4730 | 1.2733 |
| **5** | -0.2743 | -0.2090 | -0.2102 | 0.4544 | 1.2451 |
| **6** | -0.2485 | -0.2399 | -0.1923 | 0.5516 | 0.9871 |
| **7** | -0.2857 | -0.2526 | -0.2448 | 0.4525 | 0.9761 |
| **8** | -0.2472 | -0.2456 | -0.1956 | 0.4559 | 1.3274 |
| **9** | -0.1943 | -0.2394 | -0.2378 | 0.4474 | 1.1109 |
| **10** | -0.2031 | -0.1977 | -0.1997 | 0.5923 | 1.4493 |
| | | | | | |
| **11** | -0.2315 | -0.2427 | -0.1787 | 0.5995 | 1.2011 |
| **12** | -0.2282 | -0.2427 | -0.2349 | 0.4838 | 1.1874 |
| **13** | -0.2684 | -0.2532 | -0.2207 | 0.4434 | 1.2951 |
| **14** | -0.2411 | -0.2695 | -0.2173 | 0.4288 | 1.0232 |
| **15** | -0.2115 | -0.1639 | -0.1182 | 0.4683 | 1.5538 |
| **16** | -0.2079 | -0.3166 | -0.2976 | 0.5774 | 0.8619 |
| **17** | -0.2332 | -0.2230 | -0.1973 | 0.4111 | 1.5700 |
| **18** | -0.2478 | -0.2401 | -0.2603 | 0.4334 | 1.1420 |
| **19** | -0.2481 | -0.2137 | -0.1852 | 0.4365 | 1.2622 |
| **20** | -0.2495 | -0.2196 | -0.1845 | 0.5453 | 1.1024 |
| | | | | | |
| **21** | -0.2273 | -0.2227 | -0.2133 | 0.6195 | 1.2773 |
| **22** | -0.2592 | -0.2225 | -0.1858 | 0.4146 | 1.5094 |
| **23** | -0.1913 | -0.2338 | -0.1831 | 0.4861 | 1.6797 |
| **24** | -0.2062 | -0.2871 | -0.2839 | 0.4587 | 1.1413 |
| **25** | -0.2428 | -0.2119 | -0.1531 | 0.5099 | 1.3123 |
| **26** | -0.2635 | -0.2048 | -0.1937 | 0.4389 | 1.4183 |
| **27** | -0.2277 | -0.1566 | -0.1365 | 0.4785 | 1.2866 |
| **28** | -0.2702 | -0.1886 | -0.1791 | 0.5167 | 1.2494 |
| **29** | -0.2821 | -0.1943 | -0.1920 | 0.4522 | 1.1222 |
| **30** | -0.2193 | -0.2437 | -0.2290 | 0.5535 | 1.2458 |
| | | | | | |
| **31** | -0.2138 | -0.1963 | -0.1577 | 0.5567 | 1.2599 |
| **32** | -0.2092 | -0.2826 | -0.3039 | 0.4651 | 0.9712 |
| **33** | -0.2065 | -0.1616 | -0.1375 | 0.7779 | 1.2095 |
| **34** | -0.2431 | -0.2066 | -0.1727 | 0.4743 | 1.8006 |
| **35** | -0.2195 | -0.1505 | -0.1171 | 0.6149 | 1.3557 |
| **36** | -0.2487 | -0.2017 | -0.1625 | 0.7291 | 1.0344 |
| **37** | -0.2236 | -0.1932 | -0.1383 | 0.8256 | 1.0296 |
| **38** | -0.2466 | -0.2008 | -0.1878 | 0.5939 | 1.1464 |
| **39** | -0.2506 | -0.2115 | -0.1395 | 0.5614 | 1.3477 |
| **40** | -0.2141 | -0.2441 | -0.2206 | 0.4274 | 1.3171 |
| | | | | | |
| **41** | -0.2375 | -0.2609 | -0.2402 | 0.4531 | 1.1184 |
| **42** | -0.2082 | -0.2185 | -0.2178 | 0.5647 | 1.4260 |
| **43** | -0.2246 | -0.2044 | -0.2043 | 0.5415 | 1.3351 |
| **44** | -0.2900 | -0.2580 | -0.2446 | 0.4457 | 0.9774 |
| **45** | -0.2462 | -0.2372 | -0.1940 | 0.4009 | 1.6377 |
| **46** | -0.2292 | -0.2734 | -0.2319 | 0.3572 | 1.4269 |
| **47** | -0.2350 | -0.1913 | -0.1614 | 0.5477 | 1.4842 |
| **48** | -0.2223 | -0.2803 | -0.2227 | 0.3987 | 1.3048 |
| **49** | -0.3054 | -0.2437 | -0.2340 | 0.3554 | 1.0429 |
| **50** | -0.1905 | -0.1680 | -0.1588 | 0.7190 | 1.7280 |

| **(b) Parameter estimates: Northern cap** | | | | | |
|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_H$ | $\theta_V$ |
| **51** | -0.2073 | -0.1709 | -0.1434 | 0.6106 | 1.4891 |
| **52** | -0.2626 | -0.2221 | -0.1993 | 0.4276 | 1.1451 |
| **53** | -0.2384 | -0.1643 | -0.1482 | 0.4713 | 1.4859 |
| **54** | -0.2609 | -0.2035 | -0.1841 | 0.4426 | 1.0780 |
| **55** | -0.2736 | -0.2392 | -0.1646 | 0.3826 | 1.2882 |
| **56** | -0.2496 | -0.2303 | -0.1956 | 0.4907 | 1.0544 |
| **57** | -0.2273 | -0.2369 | -0.2249 | 0.4902 | 1.0219 |
| **58** | -0.2322 | -0.1633 | -0.1338 | 0.5833 | 1.3810 |
| **59** | -0.2380 | -0.2811 | -0.2441 | 0.4892 | 0.9193 |
| **60** | -0.2334 | -0.2421 | -0.2368 | 0.5111 | 1.2001 |
| | | | | | |
| **61** | -0.2476 | -0.2255 | -0.2084 | 0.4402 | 1.4944 |
| **62** | -0.1817 | -0.2850 | -0.2346 | 0.6710 | 1.2087 |
| **63** | -0.1968 | -0.1869 | -0.1554 | 0.7343 | 1.2720 |
| **64** | -0.2477 | -0.2394 | -0.2119 | 0.4282 | 1.2773 |
| **65** | -0.2555 | -0.1853 | -0.1381 | 0.5345 | 1.4186 |
| **66** | -0.2855 | -0.2642 | -0.2462 | 0.4218 | 1.0581 |
| **67** | -0.2084 | -0.2526 | -0.2453 | 0.6753 | 1.0514 |
| **68** | -0.2240 | -0.2009 | -0.2010 | 0.4759 | 1.3247 |
| **69** | -0.2576 | -0.2575 | -0.1542 | 0.3871 | 1.5953 |
| **70** | -0.2389 | -0.2448 | -0.2139 | 0.5148 | 1.3030 |
| | | | | | |
| **71** | -0.2510 | -0.1963 | -0.1655 | 0.5654 | 1.4345 |
| **72** | -0.1973 | -0.2497 | -0.2003 | 0.4961 | 1.3257 |
| **73** | -0.2279 | -0.2118 | -0.2227 | 0.5458 | 1.0982 |
| **74** | -0.2006 | -0.2308 | -0.1953 | 0.5804 | 1.2942 |
| **75** | -0.2274 | -0.2680 | -0.2125 | 0.4675 | 1.0949 |
| **76** | -0.2225 | -0.2104 | -0.1913 | 0.4715 | 1.5988 |
| **77** | -0.2266 | -0.1812 | -0.1522 | 0.4818 | 1.1669 |
| **78** | -0.2068 | -0.2321 | -0.2525 | 0.6159 | 1.1225 |
| **79** | -0.2256 | -0.1879 | -0.1443 | 0.5702 | 1.2820 |
| **80** | -0.2571 | -0.2578 | -0.2519 | 0.4349 | 1.0929 |
| | | | | | |
| **81** | -0.2344 | -0.2259 | -0.2090 | 0.4767 | 1.3229 |
| **82** | -0.2490 | -0.2131 | -0.2072 | 0.4456 | 1.5248 |
| **83** | -0.2461 | -0.2646 | -0.3005 | 0.3109 | 1.2982 |
| **84** | -0.2739 | -0.2196 | -0.1953 | 0.4527 | 0.9043 |
| **85** | -0.2262 | -0.2234 | -0.1702 | 0.7560 | 1.2102 |
| **86** | -0.2084 | -0.1728 | -0.1256 | 0.8231 | 1.2475 |
| **87** | -0.2496 | -0.2388 | -0.1849 | 0.3480 | 1.1309 |
| **88** | -0.2618 | -0.1773 | -0.1455 | 0.4246 | 1.5500 |
| **89** | -0.2355 | -0.2696 | -0.2478 | 0.4182 | 1.0503 |
| **90** | -0.2348 | -0.2579 | -0.2134 | 0.4985 | 1.0906 |
| | | | | | |
| **91** | -0.2143 | -0.2089 | -0.2062 | 0.6423 | 1.1569 |
| **92** | -0.2534 | -0.2403 | -0.1634 | 0.4926 | 1.1837 |
| **93** | -0.2867 | -0.2022 | -0.1720 | 0.3718 | 1.2329 |
| **94** | -0.2794 | -0.1852 | -0.1409 | 0.5628 | 1.0147 |
| **95** | -0.2123 | -0.2473 | -0.2188 | 0.4758 | 1.4412 |
| **96** | -0.2663 | -0.1278 | -0.1758 | 0.5434 | 1.3269 |
| **97** | -0.2490 | -0.2529 | -0.1941 | 0.5582 | 1.1384 |
| **98** | -0.2594 | -0.1782 | -0.1066 | 0.4955 | 1.3202 |
| **99** | -0.1880 | -0.2399 | -0.2172 | 0.4987 | 1.5291 |
| **100** | -0.2208 | -0.2154 | -0.2091 | 0.6120 | 1.0823 |

## References

[1] H. Adams, S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta, R. Neville, C. Peterson, P. Shipman, and L. Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *ArXiv e-prints*, July 2015.

[2] R.J. Adler, O. Bobrowski, M.S. Borman, E. Subag, and S. Weinberger. Persistent homology for random fields and complexes. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*, pages 124–143. Institute of Mathematical Statistics, 2010.

[3] U. Bauer, M. Kerber, and J. Reininghaus. Clear and compress: Computing persistent homology in chunks. *ArXiv e-prints*, March 2013.

[4] Dave Bayer and Persi Diaconis. Trailing the dovetail shuffle to its lair. *Ann. Appl. Probab.*, 2(2):294–313, 1992.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[6] O. Bobrowski and M. Kahle. Topology of random geometric complexes: a survey. *Preprint at http://arxiv. org/abs/1409.4734*, 2014.

[7] P. Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015.

[8] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

[9] G. Carlsson. Topological pattern recognition for point cloud data. *Acta Numer.*, 23:289–368, 2014.

[10] G. Carlsson and V. de Silva. Plex: MATLAB software for computing persistent homology of finite simplicial complexes. comptop.stanford.edu/programs/∼plex.

[11] B. Chalmond. *Modeling and Inverse Problems in Imaging Analysis*, volume 155 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2003. Translated from the French, With a foreword by Henri Maître.

[12] F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Robust topological inference: Distance to a measure and kernel distance. *ArXiv e-prints*, December 2014.

[13] Y.-C. Chen, D. Wang, A. Rinaldo, and L. Wasserman. Statistical analysis of persistence intensity functions. *ArXiv e-prints*, October 2015.

[14] J-F. Coeurjolly and R. Drouilhe. Asymptotic properties of the maximum pseudo-likelihood estimator for stationary Gibbs point processes including the Lennard-Jones model. *Electron. J. Statist.*, 4:677–706, 2010.

[15] H. Edelsbrunner. *A Short Course in Computational Geometry and Topology*. Springer Briefs in Applied Sciences and Technology. Springer, Cham, 2014.

[16] H. Edelsbrunner and J. Harer. Persistent homology—a survey. In *Surveys on discrete and computational geometry*, volume 453 of *Contemp. Math.*, pages 257–282. Amer. Math. Soc., Providence, RI, 2008.

[17] H. Edelsbrunner and J.L. Harer. *Computational Topology.* American Mathematical Society, Providence, RI, 2010. An introduction.

[18] H. Edelsbrunner, A. Ivanov, and R. Karasev. Current open problems in discrete and computational geometry. *Modelirovanie i Analiz Informats. Sistem*, 19:5–17, 2012.

[19] H. K. Eriksen, F. K. Hansen, A. J. Banday, K. M. Gorski, and P. B. Lilje. Asymmetries in the cosmic microwave background anisotropy field. *The Astrophysical Journal*, 605(1):14–20, 2004.

[20] B.T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *Ann. Statist.*, 42(6):2301–2339, 2014.

[21] R. Ghrist. *Elementary Applied Topology.* Createspace, 2014.

[22] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *Astrophysical Journal*, 622:759–771, April 2005.

[23] J. R. Gott, III, M. Dickinson, and A. L. Melott. The sponge-like topology of large-scale structure in the universe. *Astrophysical Journal*, 306:341–357, July 1986.

[24] A. Hatcher. *Algebraic Topology.* Cambridge University Press, Cambridge, 2002.

[25] M. Kahle. Topology of random simplicial complexes: a survey. In *Algebraic topology: applications and new directions*, volume 620 of *Contemp. Math.*, pages 201–221. Amer. Math. Soc., Providence, RI, 2014.

[26] S. Kalisnik-Verovsek. Tropical coordinates on the space of persistence barcodes. *ArXiv e-prints*, March 2016.

[27] G. Kusano, K. Fukumizu, and Y. Hiraoka. Kernel method for persistence diagrams via kernel embedding and weight factor. *ArXiv e-prints*, June 2017.

[28] D. Marinucci and G. Peccati. *Random Fields on the Sphere*, volume 389 of *London Mathematical Society Lecture Note Series.* Cambridge University Press, Cambridge, 2011. Representation, limit theorems and cosmological applications.

[29] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 22, 2011.

[30] E. Munch, K. Turner, P. Bendich, S. Mukherjee, J. Mattingly, and J. Harer. Probabilistic Fréchet means for time varying persistence diagrams. *Electron. J. Stat.*, 9(1):1173–1204, 2015.

[31] S.Y. Oudot. *Persistence Theory: From Quiver Representations to Data Analysis*, volume 209 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2015.

[32] Chan-Gyung Park. Non-Gaussian signatures in the temperature fluctuation observed by the Wilkinson Microwave Anisotropy Probe. *Monthly Notices of the Royal Astronomical Society*, 349(1):313–320, 2004.

[33] Planck Collaboration, R. Adam, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, and et al. Planck 2015 results. IX. Diffuse component separation: CMB maps. *Astron. & Astrophysics*, 594:A9, September 2016.

[34] Planck Collaboration, P. A. R. Ade, N. Aghanim, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, C. Baccigalupi, A. J. Banday, and et al. Planck 2013 results. xxiii. isotropy and statistics of the cmb. *ArXiv e-prints*, March 2013.

[35] P. Pranav, R.J. Adler, H. Edelsbrunner, H. Wagner, T. Buchert, A. Schwartzman, B.J.T. Jones, and R. van de Weygaert. Loops abound in the cosmic microwave background. 2017. In preparation.

[36] P. Pranav, H. Edelsbrunner, R. van de Weygaert, G. Vegter, M. Kerber, B. J. T. Jones, and M. Wintraecken. The topology of the cosmic web in terms of persistent Betti numbers. *Monthly Notices of the Royal Astronomical Society*, 465:4281–4310, March 2017.

[37] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4741–4748, 2015.

[38] J. E. Rhoads, J. R. Gott, III, and M. Postman. The genus curve of the Abell clusters. *Astrophysical Journal*, 421:1–8, January 1994.

[39] V. Robins and K. Turner. Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D Nonlinear Phenomena*, 334:99–117, November 2016.

[40] A. Robinson and K. Turner. Hypothesis testing for topological data analysis. *ArXiv e-prints*, October 2013.

[41] R. K. Sachs and A. M. Wolfe. Perturbations of a cosmological model and angular variations of the microwave background. *Astrophysical Journal*, 147:73, January 1967.

[42] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet means for distributions of persistence diagrams. *Discrete Comput. Geom.*, 52(1):44–70, 2014.

[43] L. Wasserman. Topological data analysis. *Annual Reviews in Statistics*, 5, 2018.

[44] A.J. Zomorodian. *Topology for Computing*, volume 16 of *Cambridge Monographs on Applied and Computational Mathematics.* Cambridge University Press, Cambridge, 2005.