Supplemental Web Materials

MODELING OF PERSISTENT HOMOLOGY

Sarit Agami and Robert J. Adler

Andrew and Erna Viterbi Faculty of Electrical Engineering

Technion – Israel Institute of Technology

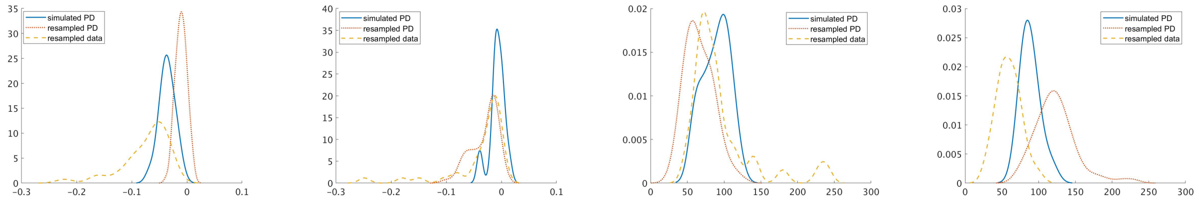
Contents

S.1. The two dimensional sphere - H_1 persistence diagram

S.1.1. Fitting the model

We present here the analysis of the H_1 persistence diagram for the two dimensional sphere that is described in Section 4.1 in the paper. Again estimating the parameters for the Gibbs pseudolikelihood (7) in the paper, taking $K = 3$, the estimate of δ was 0.0047. For this δ, the estimates of Θ were $\theta_1 = -0.0331$, $\theta_2 = 0$, $\theta_3 = 3.3842$, $\theta_H = 60.00$, and $\theta_V = 110.00$.

To check the match between the estimated model and the H_1 persistence diagram, we used the same 100 simulated sets of the 2-sphere used for the H_0 diagram, following the same procedure that we adopted then, this time restricting to a model with only θ_1, θ_3, θ_H, and θ_V non-zero. The blue plot in Figure 1 shows the smoothed empirical densities for the parameters estimates generated by these simulations. As for the H_0 case, the results indicate that the estimation procedure is stable.



(a)

(b)

(c)

(d)

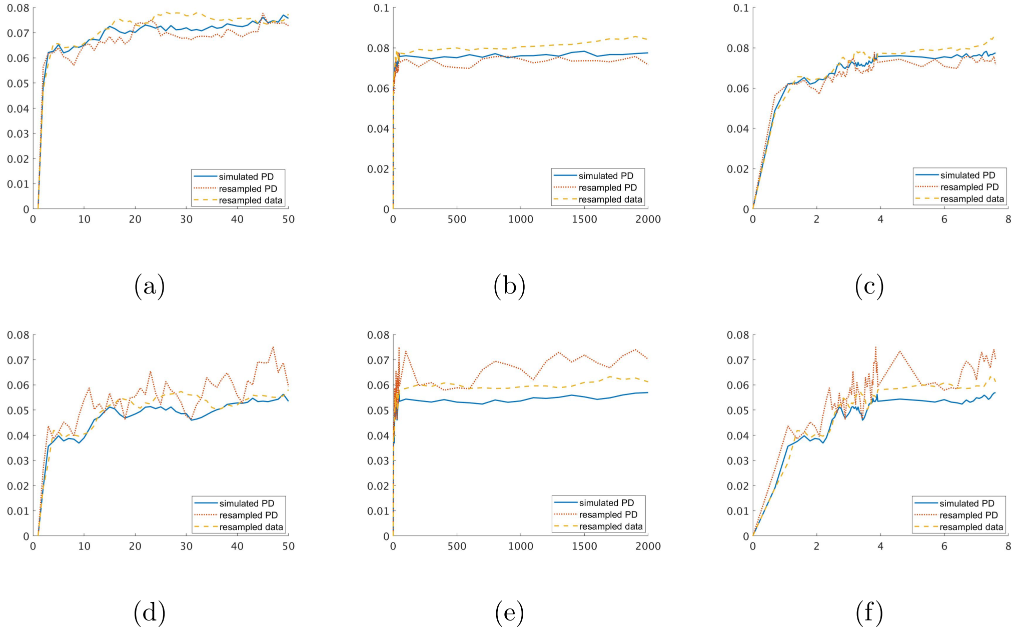
Figure 1: Smoothed empirical densities for the four parameter estimates of H_1 persistence diagram coming from the simulations of 2-sphere, see text for details. (a) θ_1, (b) θ_3, (c) θ_H, (d) θ_V.

.

S.1.2. Replicating the H_1 persistence diagram

As for the analysis of the H_0 diagram, we calculated bottleneck and the Wasserstein distances between the original persistence diagrams and the corresponding 100 MCMC simulated diagrams of the previous section. The results are shown by the blue plots in Figure 2.

The first row in Figure 2 shows the bottleneck distances, while the second row shows the W_2 differences. The first column shows the results of the first 50 steps of the MCMC algorithm on a linear scale. The second and third columns go out to 2,000 steps, first on a linear scale and then on a logarithmic scale. The point where the initial rapid growth of the distance functions ceases, is approximately 44 for the bottleneck distance and 47 in the Wasserstein case.



(a) (b) (c)

(d) (e) (f)

Figure 2: Growth of the bottleneck (a) and Wasserstein (d) 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. Panels (b) and (e) take $1 \leq n_b \leq 2,000$, while (c) and (f) show he same data but on a logarithmic scale.

In addition we considered summary statistics of the 100 simulated persistence diagrams as the MCMC progressed, to ensure that the simulations reliably replicate the statistical properties of the persistence diagrams. Here the best fits were for a burn in of 50, which is consistent with the results of Figure 2.

S.1.3. Resampling H_1

As for the H_0 case, we again examine the performance of resampling from the original persistence diagram (Setting I) and from the original data (Setting II), repeating each procedure 100 times. The results are summarised in Figure 1. The red (dot dashed) plots are the smoothed empirical densities for the parameter estimates in Setting I, while the yellow (dashed) plot correspond to Setting II.

In order to assess the fit of the simulated data to the original, we computed, as previously, the bottleneck and the Wasserstein distances between the MCMC simulations and the data itself. The results are presented in Figure 2, in addition to the results based on the 100 simulated persistence diagrams. The red (dot dashed) plot shows the results for the 100 resampled sets from the original persistence diagram, and the yellow (dashed) plot shows the same thing, but for the 100 resampled sets from the original data. The point where the initial rapid growth of the distance functions ceases, is approximately 22 and 46 in Setting I and Setting II, respectively, for the bottleneck distance, and approximately 20 and 48 in the Wasserstein case. This suggests taking a burn in period of 50 for generating the replicated persistence diagrams for H_1.

S.1.4. Statistical inference

We are now finally in a position to carry out a simulation study to test how well we can identify the homology of 2-sphere, using the methodology described earlier. To do so, we generated 1,000 persistence diagrams from the fitted model, via MCMC, with a burn in period of 50 iterations and with (n_b, n_r, n_R) given by (500,10,100), (500,20,50), (500,40,25), or (500,100,10). Using these four sets of simulations, we computed the maximum statistic $T_1 = \max_i |d_i - b_i|$, its confidence interval and p-value, for both the H_0 and H_1 persistence diagrams. Table 1 summarizes the results.

The results for the H_0 persistence diagram show that T_1, in two first scenarios, was statistically insignificant, and in the two other scenarios was significant. In other words, the evidence is split between one connected component (represented by the ‘point at infinity’ not included in the analysis) and two components. The fact that the correct result occurs in the cases of a larger number of shorter MCMC runs is consistent with earlier findings in Adler et al. (2017).

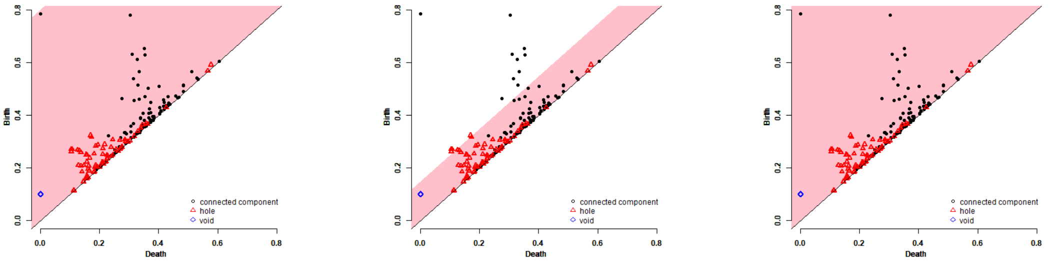
As for the H_1 topology, all four scenarios showed that T_1 was insignificant for all MCMC parameter, implying, correctly, a trivial H_1 homology.

In order to appreciate the power of the above inferences, we now carry out inference

Homology	Statistic	Real PD	(n_b, n_r, n_R)	CI	p-value	Significance
H_0	T_1	0.4769	(500,10,100)	[0, 0.4769]	0.0990	no
			(500,20,50)	[0, 0.4769]	0.0520	no
			(500,40,25)	[0, 0.3273]	0.0320	yes
			(500,100,10)	[0, 0.2616]	0.0100	yes
H_1	T_1	0.1673	(500,10,100)	[0, 0.2140]	0.4060	no
			(500,20,50)	[0, 0.2069]	0.3780	no
			(500,40,25)	[0, 0.2065]	0.3550	no
			(500,100,10)	[0, 0.1995]	0.3270	no

Table 1: Maximum statistic T_1 for the real H_0 and H_1 persistence diagram and the simulated H_0 and H_1 persistence diagrams of the 2-sphere. The CI is a one-sided confidence interval at a 95% confidence level. The p-value is also one-sided. Both the CI and the p-value are based on 1,000 simulated persistence diagrams.

based on two existing, bootstrap based, methods. Figure 3 presents 95% confidence sets for the persistence diagram using the bootstrap with 1000 bootstrap samples. The pink region describes the confidence set, and the number of bootstrap samples was 1000. There are two approaches for bootstrapping, either via bootstrapping the original data data (Fasy et al. (2014)), or bootstrapping from the original persistence diagram (Chazal et al. (2014)). Figure 3 (a) shows that bootstrapping from the original persistence diagram fails to detect any connected components at all (including the one corresponding to the 'point at infinity', which appears in this particular diagram!), while (b) shows that it (incorrectly) identified several significant holes. Bootstrapping from the original sphere, as in (c), failed to detect either connected components or holes.



(a)

(b)

(c)

Figure 3: 95% confidence sets for persistence diagrams using the bootstrap: See text for more details. (a) A 95% confidence set for the H_0 persistence points of the sphere, using bootstrapping from the original persistence diagram. (b) A 95% confidence set for the H_1 persistence points of the sphere, using bootstrapping from the original persistence diagram. (c) A 95% confidence set for the persistence diagram, using bootstrapping from the original sphere. Black circles are H_0 persistence points, red triangle is H_1 point. Birth times are on the vertical axis.

S.2. 3-torus

S.2.1. Replicating the persistence diagram

The determination of the burn in period in this example, for both H_0 and H_1, was only heuristic. Figure 4 presents the original persistence diagrams of H_0 and H_1 and their MCMC with burn in periods of 10, 25, 50 and 1000. The best fits for both H_0 and H_1 occur for burn in periods in the range $[10, 50]$.

S.2.2. Statistical inference

We generated 1,000 replicated persistence diagrams from the fitted model with a burn in period of 10 iterations. Table 2 summarizes the results.

The results for the H_0 diagram, for all scenarios, showed that T_1 was insignificant (the lowest p-value reached in any of the six cases was 0.235). Thus, adding the 'point at infinity' back into the diagram, we have evidence for exactly one connected component, as we hoped to find.

For the H_1 diagram, the results for all scenarios showed that $T_1 - T_3$ were all significant

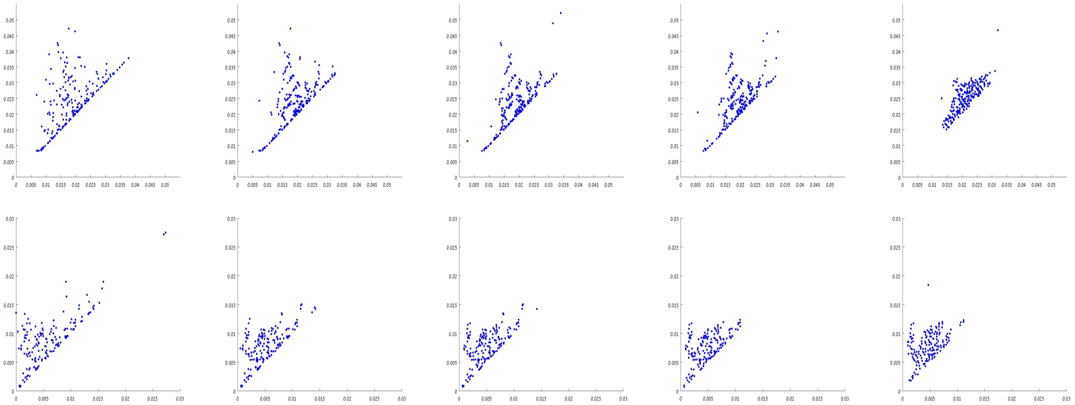


Figure 4: The first row shows the persistence diagrams of H_0 of the 3-torus, and the second row shows the persistence diagrams of H_1 of the 3-torus. At each row, the left plot is the original persistence diagram, and the four other plots are simulated persistence diagrams based on an MCMC simulation with burn in of 10, 25, 50 and 1000.

Homology	Statistic	Real PD	(n_b, n_r, n_R)	CI	p-value	Significance
H_0	T_1	0.0295	(500,10,100)	[0, 0.0371]	0.3360	no
			(500,20,50)	[0, 0.0362]	0.2770	no
			(500,40,25)	[0, 0.0359]	0.2350	no
			(500,100,10)	[0, 0.0322]	0.2720	no
H_1	T_1	0.0136	(500,10,100)	[0, 0.0123]	0.0340	yes
			(500,20,50)	[0, 0.0118]	0.0320	yes
			(500,40,25)	[0, 0.0107]	0.0220	yes
			(500,100,10)	[0, 0.0134]	0.0490	yes
H_1	T_2	0.0118	(500,10,100)	[0, 0.0103]	0.0030	yes
			(500,20,50)	[0, 0.0102]	0.0060	yes
			(500,40,25)	[0, 0.0102]	0	yes
			(500,100,10)	[0, 0.0101]	0.0020	yes
H_1	T_3	0.0103	(500,10,100)	[0, 0.0100]	0.0360	yes
			(500,20,50)	[0, 0.0098]	0.0060	yes
			(500,40,25)	[0, 0.0099]	0.0060	yes
			(500,100,10)	[0, 0.0099]	0.0180	yes

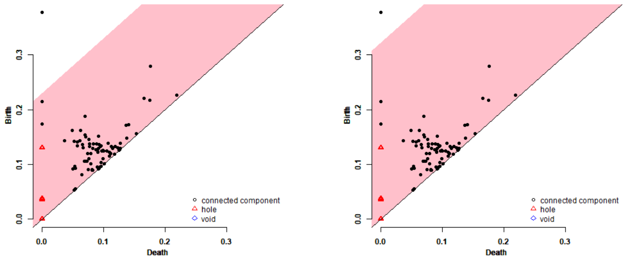
Table 2: Order statistics T_1, T_2, T_3 for the real H_0 and H_1 persistence diagrams and the simulated H_0 and H_1 persistence diagrams for the 3-torus example. The CI is a one-sided confidence interval at a 95% confidence level. The p-value is also a one-sided. Both the CI and the p-value are based on 1000 simulated persistence diagrams.

(the highest p-value reached in any of the 8 cases was 0.049). That is, three significant ‘holes’, as we hoped to find.

However, as mentioned in the main paper, $T_4, ..., T_{10}$ were also statistically significant, leading to a significant over-estimation of the complexity of the H_1 homology. Some possible explanations, and ways to correct, for this are described there.

S.3. Three circles

The results of the bootstrap method are presented in Figure 5: we have that bootstrapping from the original persistence diagram and from the original three circles recognize one connected component.



(a) (b)

Figure 5: 95% confidence sets for the persistence diagrams of the three circles, using the bootstrap. (a) A 95% confidence set for the H_0 persistence points of the three circles, using bootstrapping from the original persistence diagram. (b) A 95% confidence set for the persistence diagram of the three circles, using bootstrapping from the original sphere. Black circles are H_0 persistence points, red triangle is H_1 point. Birth times are on the vertical axis.

S.4. Noisy circle

S.4.1. Fitting the model

In order to compare the estimates distributions over noisy circle and a circle without noise, we generated 100 collections of samples from the noisy circle and from the circle without noise, according to the same procedure that generated the original data; for each sample we fitted the model that includes all the five parameters of Θ. The top plots in Figure 6 show the (smoothed) empirical densities of the resulting parameter estimates for the noisy circle[1]. The bottom plots in Figure 6 show the (smoothed) empirical densities of the resulting parameter estimates for the circle without noise[2]. We see that the estimates of δ, θ_1, θ_2, θ_3, θ_H, are typically smaller for the noisy circle than the corresponding estimates under a circle without noise, whereas for θ_V the opposite is true. In addition, we see that the behaviour of the distributions is the same over the two cases except for θ_1: the distributions of δ and θ_V are symmetric, whereas the distributions for the other estimates are asymmetric to the right. That is, the noise in general made the estimates to be with smaller values, except θ_V which had the opposite direction of sign.

Most noticeable, however, are the facts that while the centres (as measured by modes) of the distributions are not seriously affected by the noise, the means and variances do change. In other words, estimation in the presence of noise involves both mean bias (not expected) and (typically) increased variance (expected). Table 3 gives some summary statistics supporting these claims.

[1] Some of these estimates included outliers, probably since the model with the all five parameters is not the best model for the specific sample. We omitted these cases (7 cases), and the plots are based on the samples without outliers in the estimates.

[2] Also here some of the estimates included outliers. We omitted these cases (15 cases), and the plots are based on the samples without outliers in the estimates.

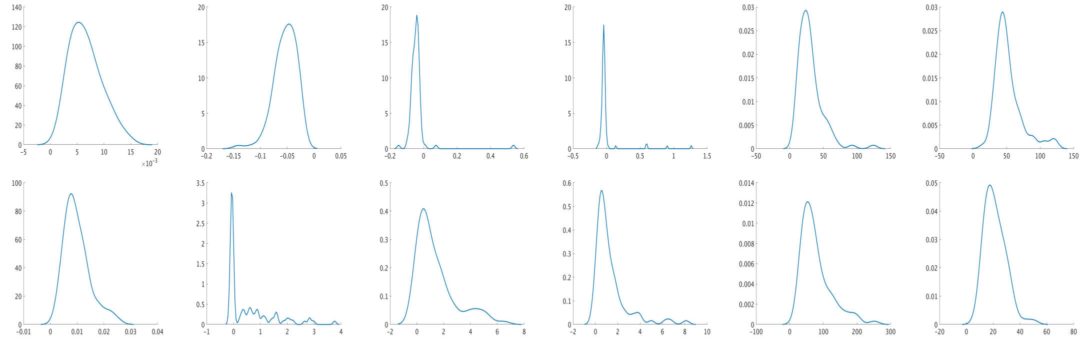


Figure 6: Smoothed empirical densities for δ and the five parameter estimates for the H_0 persistence diagram. Top: 100 simulations of noisy circle. Bottom: 100 simulations of circle without noise. See text for details. From left to right: δ, θ_1, θ_2, θ_3, θ_H, θ_V.

.

	Without noise			With noise		
Statistic	Mode	Average	Relative std^a	Mode	Mean	Relative std^a
θ_1	-0.1412	-0.0539	0.0307	-0.1214	0.5749	1.2914
θ_2	-0.1489	-0.0411	0.0523	-0.0977	1.3950	1.1879
θ_3	-0.1322	-0.0056	0.1551	-0.1240	1.3201	1.2133
θ_4	7.8188	30.4283	0.4383	21.5787	78.1979	1.0950
θ_5	14.4942	52.3556	0.9348	7.1351	21.3760	0.3574
δ	0.0016	0.0066	0.6896	0.0020	0.0099	1.1279

Table 3: Comparison of the center location (mode and mean) and the spread (relative standard deviation) over the circle without noise and the circle with noise. ^{a}Relative standard deviation is the fraction of the relevant standard deviation from the total standard deviation. The total standard deviation was calculated over the two cases of the circle with and without noise for the specific parameter.

S.4.2. The determination of the burn in period

The determination of the burn in period in this example, for both the noisy circle and the circle without noise, was only heuristic. Figure 7 presents the original persistence diagrams of the noisy circle and the circle without noise and their MCMC with burn in periods of 10, 25, 50 and 1000. The best fits for both noisy circle and circle without noise occur for burn in periods in the range $[10, 50]$.

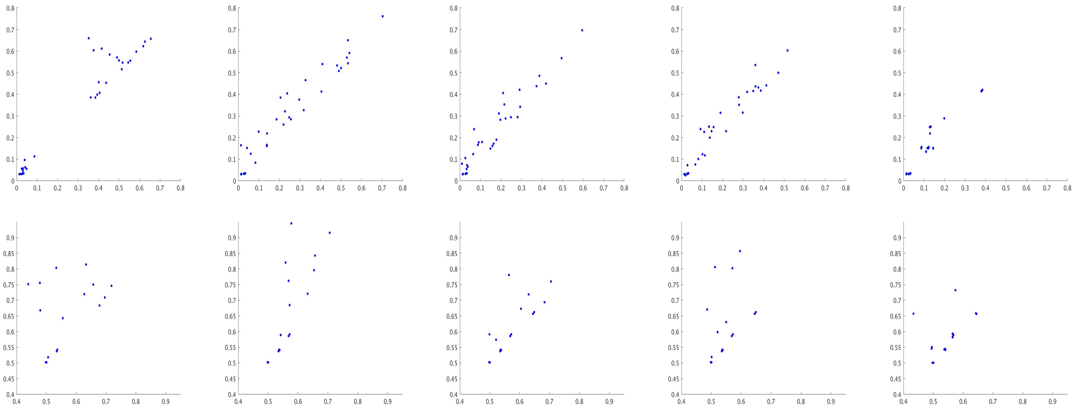


Figure 7: The first row shows the persistence diagrams of the noisy circle, and the second row shows the persistence diagrams of circle without noise. At each row, the left plot is the original persistence diagram, and the four other plots are simulated persistence diagrams based on an MCMC simulation with burn in of 10, 25, 50 and 1000.

S.4.3. Statistical inference

We generated 1,000 replicated persistence diagrams from the fitted model of the both cases with a burn in period of 10 iterations. Table 4 summarizes the results for the noisy circle, and Table 5 summarizes the results for the circle without noise.

Homology	Statistic	Real PD	(n_b, n_r, n_R)	CI	p-value	Significance
H_0	T_1	0.3080	(500,10,100)	[0, 0.2165]	0.0210	yes
			(500,20,50)	[0, 0.1843]	0.0070	yes
			(500,40,25)	[0, 0.1562]	0.0030	yes
			(500,100,10)	[0, 0.1734]	1.00E-03	yes

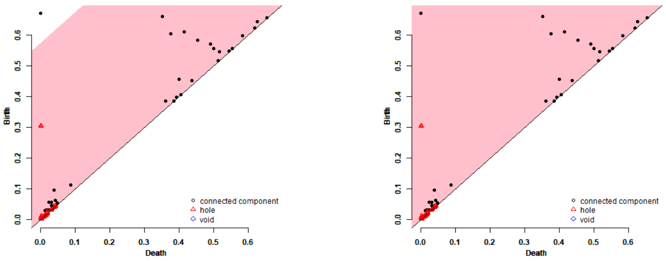
Table 4: Maximum statistics T_1, T_2, T_3 for the real H_0 persistence diagrams and the simulated H_0 persistence diagram of the noisy circle. The CI is a one-sided confidence interval at a 95% confidence level. The p-value is also a one-sided. Both the CI and the p-value are based on 1000 simulated persistence diagrams.

Comparing these results with the bootstrap method described in Figure 8, we have that bootstrapping from the original persistence diagram in the noisy circle finds a single, significant, connected component, but bootstrapping from the original noisy circle indentifies

Homology	Statistic	Real PD	(n_b, n_r, n_R)	CI	p-value	Significance
H_0	T_1	0.3122	(500,10,100)	[0, 0.3419]	0.0790	no
			(500,20,50)	[0, 0.3298]	0.0770	no
			(500,40,25)	[0, 0.3285]	0.0650	no
			(500,100,10)	[0, 0.3138]	0.0540	no

Table 5: Maximum statistics T_1, T_2, T_3 for the real H_0 persistence diagram and the simulated H_0 persistence diagram of the circle without noise. The CI is a one-sided confidence interval at a 95% confidence level. The p-value is also a one-sided. Both the CI and the p-value are based on 1000 simulated persistence diagrams.

none.



(a)

(b)

Figure 8: 95% confidence sets for persistence diagrams of the noisy circle using the bootstrap. (a) A 95% confidence set for the persistence diagram of the noisy circle using bootstrapping from the original persistence diagram. (b) A 95% confidence set for the persistence diagram of the noisy circle, using bootstrapping from the original noisy circle. Black circles are H_0 persistence points, red triangle is H_1 point. Birth times are on the vertical axis.

BIBLIOGRAPHY

Adler, R. J., and S. Agami, and P. Pranav. 2017. Modeling and replicating statistical topology and evidence for CMB nonhomogeneity. *Proceedings of the National Academy of Sciences* 114:11878–11883.

Chazal, F., B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman (2014, December). Robust Topological Inference: Distance To a Measure and Kernel Distance. ArXiv e-prints.

Fasy, B. T., F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh (2014). Confidence sets for persistence diagrams. Ann. Statist. 42(6), 23012339.