Parameter Inference with Topological Approximate Bayesian Computation

Tom Thorne

23rd May 2022

Topological Approximate Bayesian Computation

Topological approximate Bayesian computation for parameter inference of an angiogenesis model

Thomas Thorne ✉, Paul D W Kirk, Heather A Harrington ✉

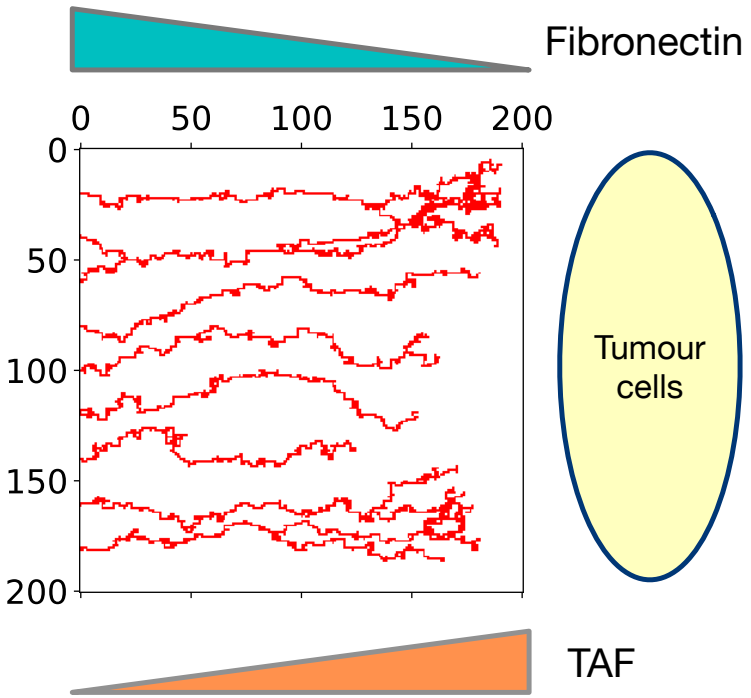
Bioinformatics, Volume 38, Issue 9, 1 May 2022, Pages 2529–2535,
https://doi.org/10.1093/bioinformatics/btac118

Outline:

- The Anderson—Chaplain angiogesis model
- Approximate Bayesian parameter inference
- Topological Data Analysis (TDA) for ABC

The Anderson–Chaplain model of angiogenesis

- Models the growth of vasculature from endothelial tip cells.
- Growth depends on gradients of *fibronectin* and *tumour angiogenic factors* (TAF).
- The model is simulated on a 2D lattice with linear chemoattractant distribution increasing along the x axis.
- Focus on two parameters χ and ρ, coefficients for chemotaxis and haptotaxis.



Anderson A.R., Chaplain M.A. (1998) Continuous and discrete mathematical models of tumor-induced angiogenesis. Bull. Math. Biol., 60, 857–899.

Nardini, J.T. et al. (2021) Topological data analysis distinguishes parameter regimes in the Anderson-Chaplain model of angiogenesis. PLoS Comput. Biol., 17, e1009094.

Bayesian inference

Why Bayesian inference?

Applying Bayesian inference:

- We can build interpretable mechanistic models
- Gives an idea of uncertainty in parameter estimates
- Allows us to draw predictive samples

The target is to learn the posterior distribution,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

But what to do if we can't evaluate $p(y|\theta)$?

Approximate Bayesian Computation

For expensive models where it would be prohibitively expensive to evaluate the likelihood $p(x|\theta)$, we can approximate the likelihood using simulations of the model $x \sim p(x|\theta)$

$$p(y|\theta) = \int_x \mathbb{1}[x = y]p(x|\theta)dx$$

In practice we use a kernel, for example a uniform kernel giving a cutoff at distance ϵ

$$p(y|\theta) \approx \int_x \mathbb{1}[D(x, y) < \epsilon]p(x|\theta)dx$$

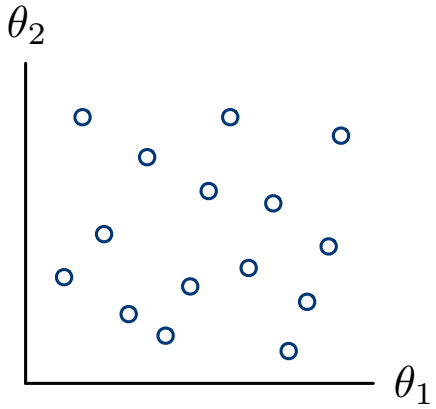
Marin, JM. et al. Approximate Bayesian computational methods. Stat Comput 22, 1167–1180 (2012).

Summary statistics

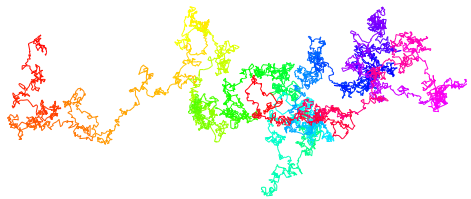
Typically rather than using a distance on the data themselves, ABC uses summary statistics of the data.

- In the case of estimating the mean of a normal distribution, we could use the sample mean of simulated data.
- Needs some care – summary statistics may not be sufficient. Robert C.P. et al. (2011) Lack of confidence in approximate Bayesian computation model choice. PNAS, 108, 15112–15117.
- Often domain specific summaries are used.

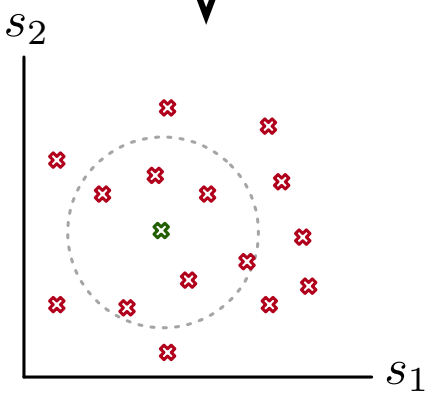
ABC rejection sampler



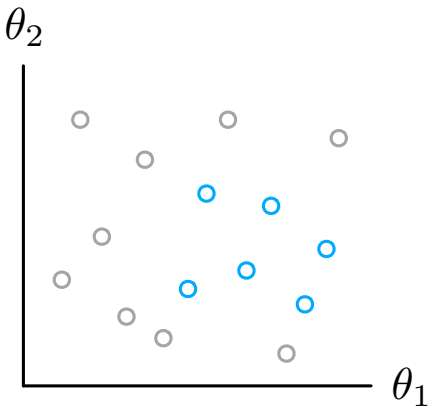
Sample $\theta_i \sim p(\theta)$



Simulate $x_i \sim f_{\theta_i}(\cdot)$



Summarise x_i as $s_i = g(x_i)$
Select s_i s.t. $D(g(y), s_i) < \epsilon$

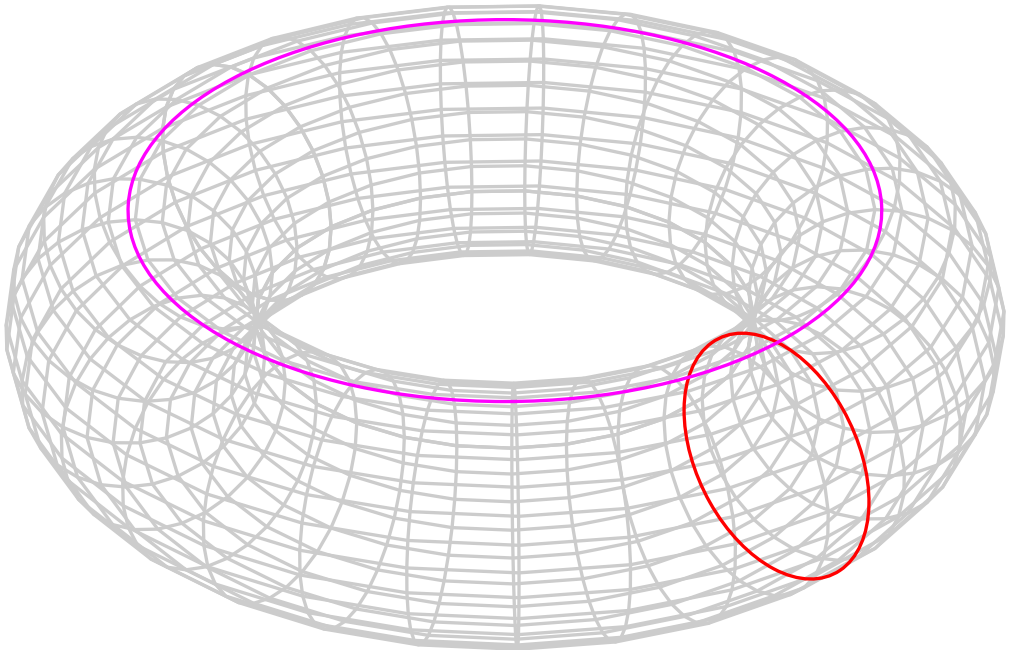


Corresponding θ_i

Topological Data Analysis

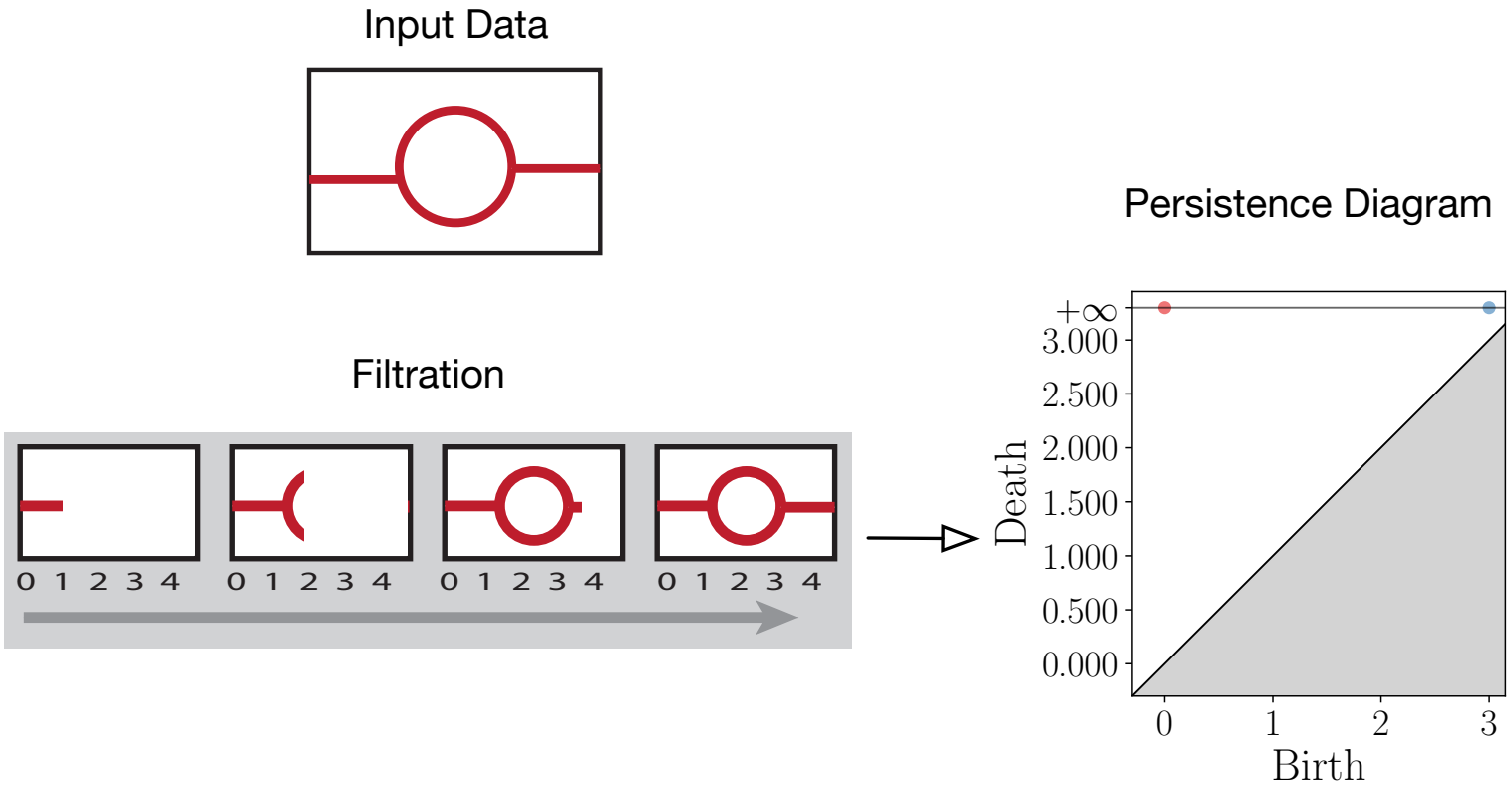
Topological Data Analysis considers the shape of data by computing topological properties of the data.

- Persistent homology characterises the features of a topological space and their persistence across scales in the data.



Persistent homology

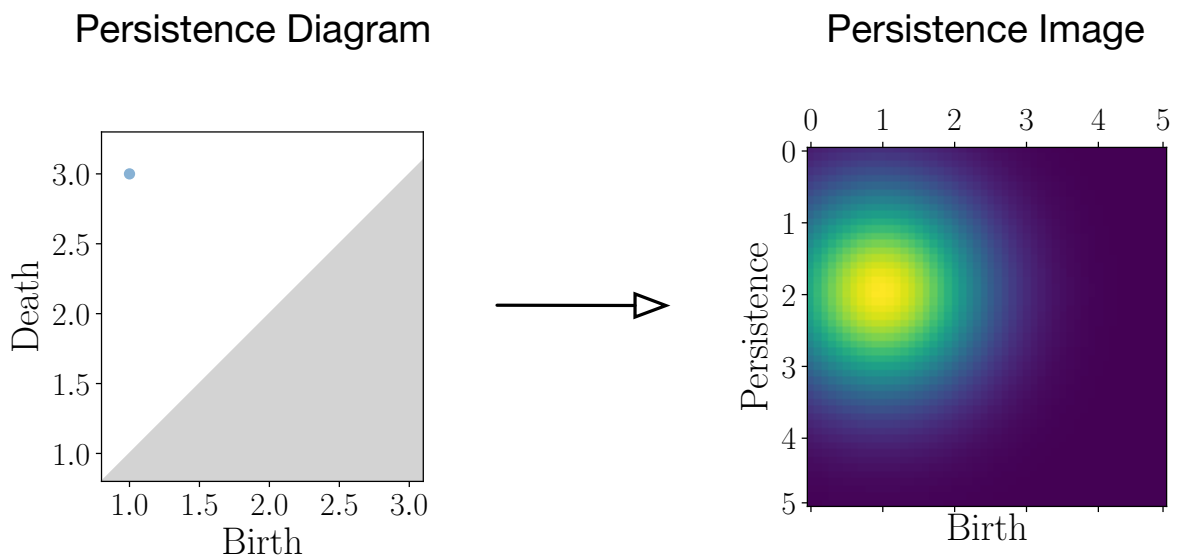
Persistent homology considers the birth and death of elements of the homology groups $H_k(X)$ in dimension k over a filtration on the data, considering sublevel sets $f^{-1}(-\infty, a]$ with $a_0 < a_1 < \ldots < a_n$:



Persistence images

- Persistence images allow us to vectorise persistence diagrams by applying a kernel and discretising them on a grid. Adams H. et al. (2017) Persistence images: a stable vector representation of persistent homology. J. Mach. Learn. Res., 18, 1–35.

$$f(x,y) = \sum_{(b,p)\in B} g(b,p)h(x,y;b,p),$$



Filtrations and Extended Persistence

Shown in **Nardini et al.** that left to right filtration alone is not enough to capture behaviour of the model.

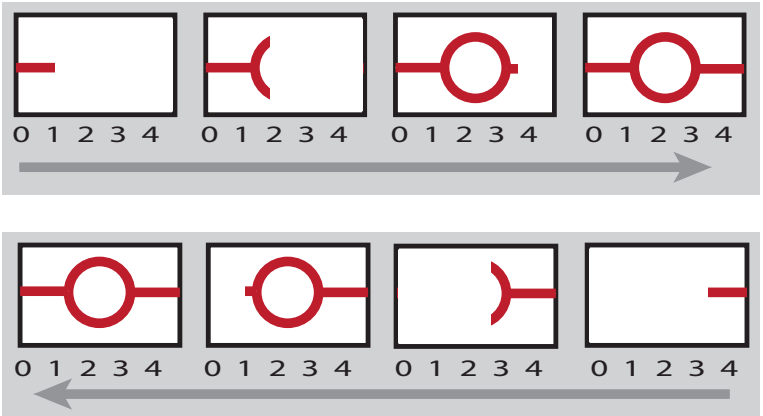
In the left to right filtration:

- Some topological features persist to end of filtration.
- Information on the size of loops is missing.

Using extended persistence:

- Features have finite death.
- We can also capture the size of loops.

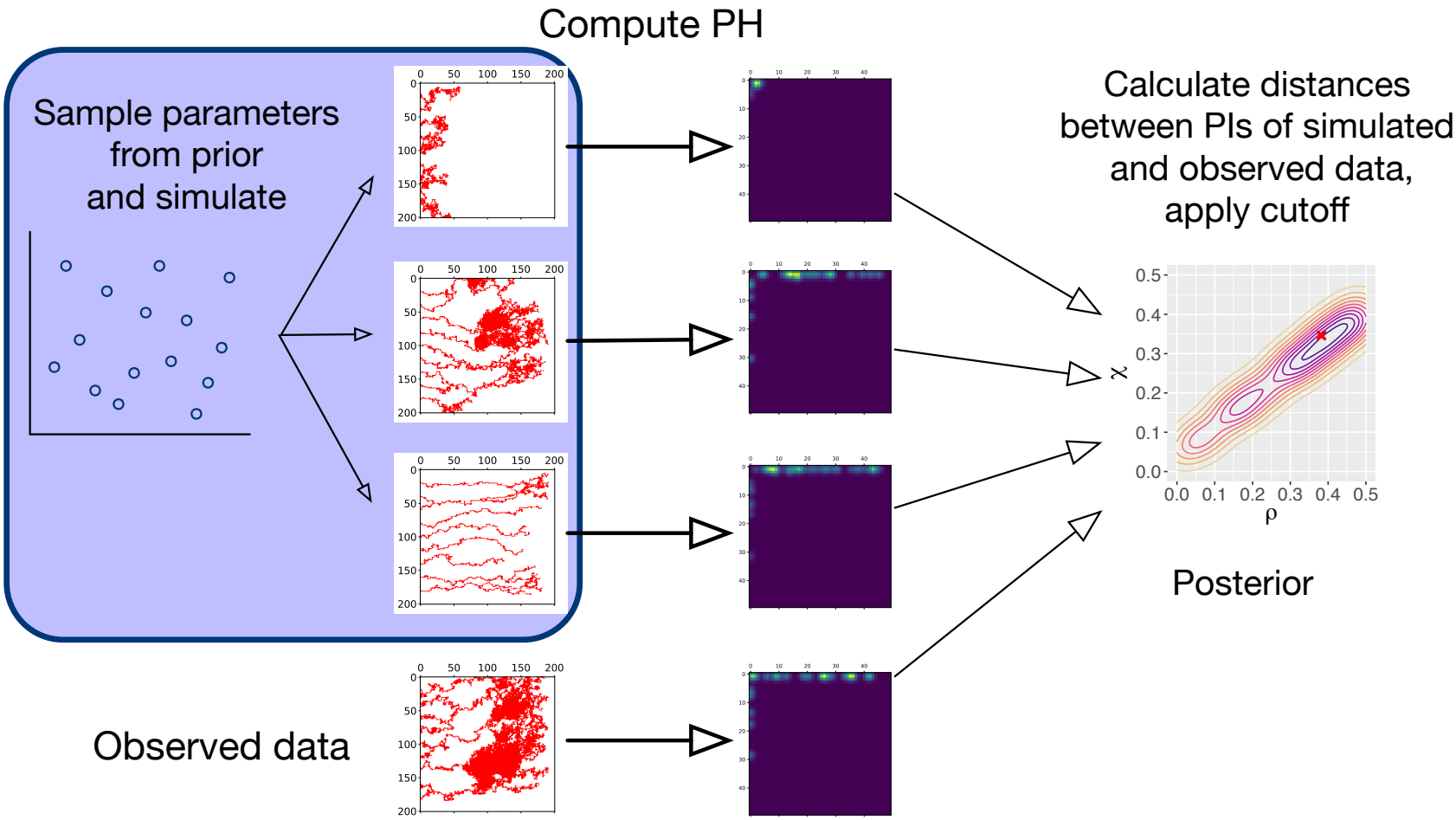
Extended persistence considers relative homology with respect to superlevel sets, as well as sublevel sets:



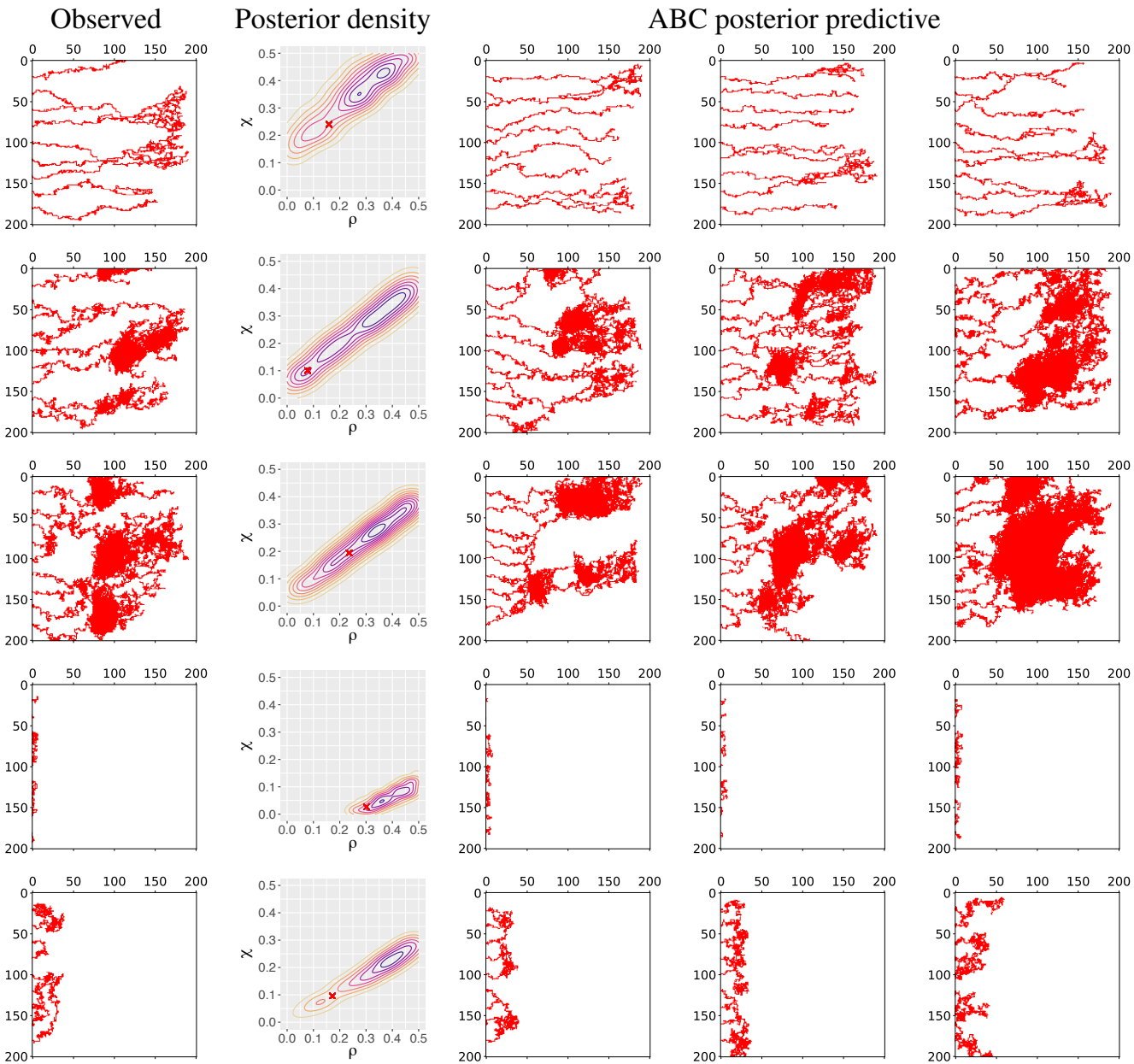
Nardini, J.T. et al. (2021) Topological data analysis distinguishes parameter regimes in the Anderson-Chaplain model of angiogenesis. PLoS Comput. Biol., 17, e1009094.

Cohen-Steiner D. et al. (2009) Extending persistence using Poincaré and Lefschetz duality. Found. Comput. Math., 9, 79–103.

Topological Approximate Bayesian Computation



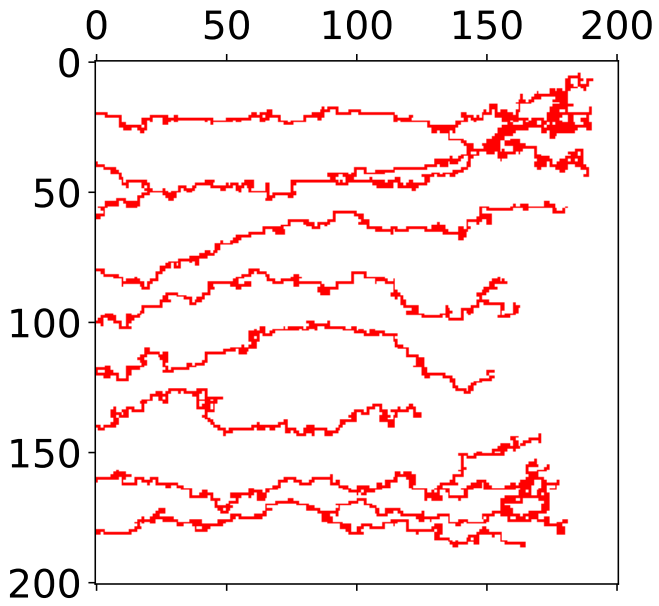
Inference results



Inference results

Comparison with some simple summary statistics:

- Mean X coordinate of occupied pixels
- Mean Y coordinate of occupied pixels
- Maximum X coordinate of occupied pixels
- Fraction of occupied pixels



For 100 parameter values sampled from the prior, we simulated data and calculated the mean RSSE of the inferred posterior samples from the known parameter value, and the entropy of the posterior distributions:

Statistics	Mean RSSE	$2\sigma_{\overline{x}}$ RSSE	Mean Entropy	$2\sigma_{\overline{x}}$ Entropy
Simple	4.30	0.25	-2.86	0.12
Topological	3.61	0.27	-3.31	0.12

Summary

- We can use topological features of the data to perform parameter inference.
- Including topological summary statistics improves parameter inference in the Anderson–Chaplain angiogenesis model.
- The TABC framework extends naturally to higher dimensional data, and other classes of data.

Future directions

- How to choose the filtration?
- How to calculate distances between persistence diagrams?
- More advanced Monte Carlo samplers (e.g. ABC-SMC) to work in higher dimensional parameter spaces.

Thanks for listening!

Acknowledgements

University of Surrey

- Duong Vo
- Tom Wilson
- NICE group

University of Oxford

- Heather Harrington
- Algebraic Systems Biology Group

MRC Biostatistics Unit

- Paul Kirk

tom.thorne@surrey.ac.uk
tt104.gitlab.io