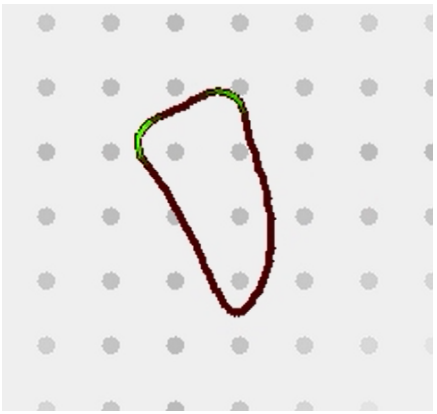
GP-enhanced semi-automatic ABC for inference in an NSPDE system for chemotaxis

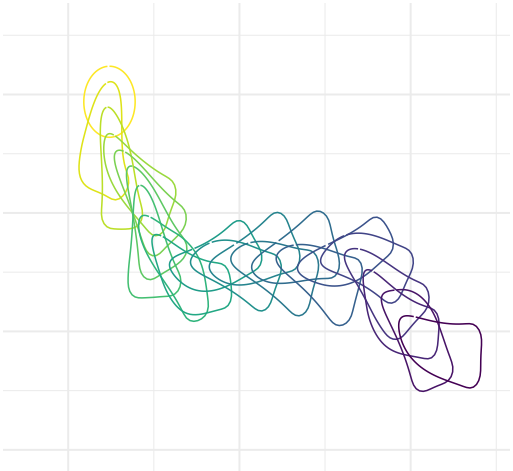
Agnieszka Borowska

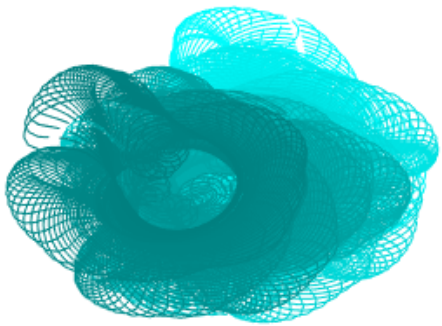
Joint with: Diana Giurghita, Dirk Husmeier

School of Mathematics and Statistics, University of Glasgow

29.10.2020







Research problem

Overall goal: to estimate the parameters of an NSPDE (*nonlinear stochastic partial differential equations*) system modelling chemotaxis.

Context:

- Chemotaxis: a type of directed cell movement in which cells are steered by chemical signals.
- Cell migration is a complex phenomenon, frequently analysed as a biophysical system, e.g. with a reaction-diffusion model.
- Plays a key role in a range of critical processes, in particular cancer metastasis.

Challenge: high-dimensional and complex outputs from the model simulator.

Outline

NSPDE system

NSPDE system

NSPDE models of chemotaxis: Neilson et al. (2011), Tweedy et al. (2013), Tweedy (2018).

The NSPDE system from Tweedy (2018):

$$\partial_t a(\mathbf{x},t) = D_a \Delta_\Gamma a + \frac{s(a(\mathbf{x},t)^2/c(t) + b_a)}{(k_M + b(\mathbf{x},t))(1+s_a^2)} - d_a a(\mathbf{x},t), \tag{1}$$

$$\partial_t b(\mathbf{x},t) = D_b \Delta_\Gamma b(\mathbf{x},t) + k_b a(\mathbf{x},t) - d_b b(\mathbf{x},t), \tag{2}$$

$$\partial_t c(t) = \frac{r_c}{|\Gamma(t)|} \oint_\Gamma a(\mathbf{x},t) d\mathbf{x} - r_c a(\mathbf{x},t), \tag{3}$$

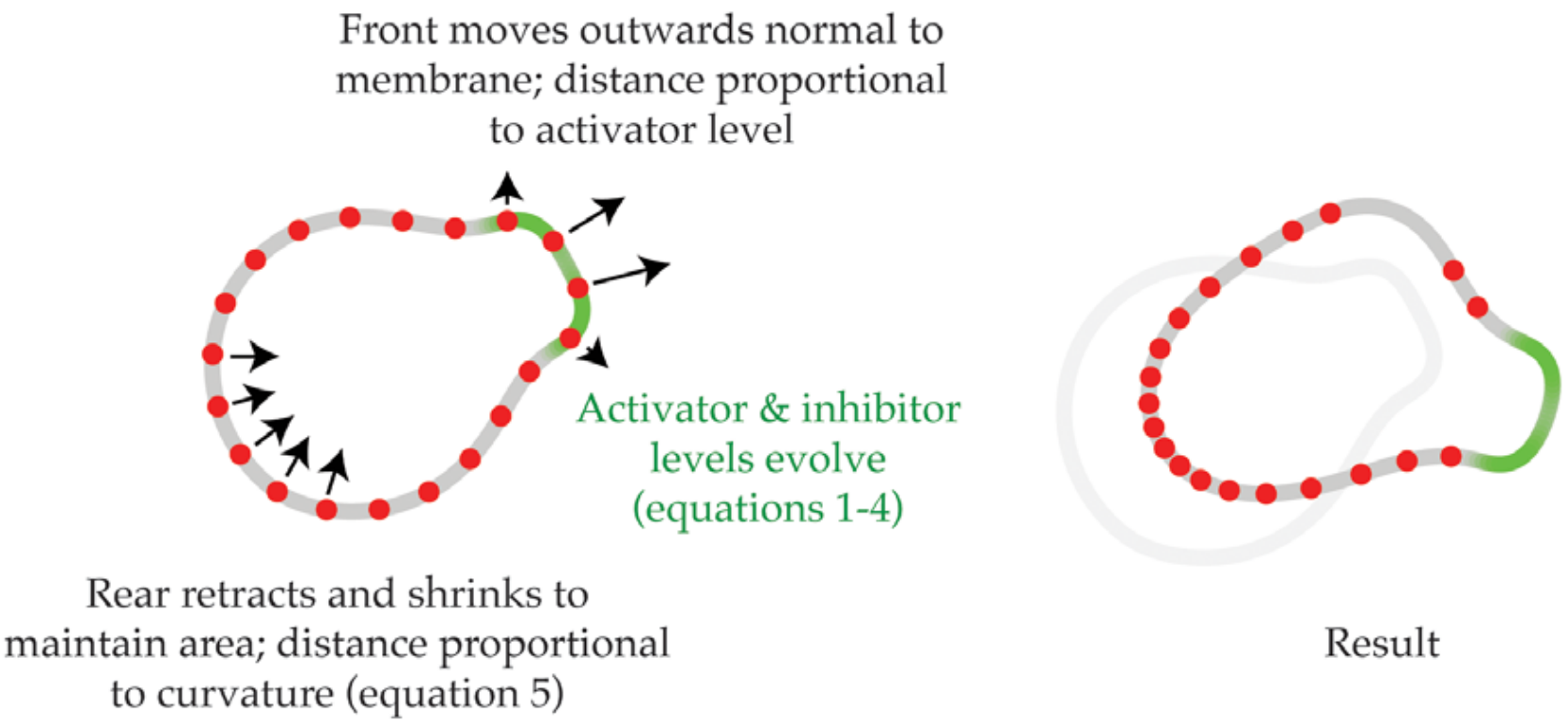
$$s(\mathbf{x},t) = 1 + d_r \mathrm{RND} + \left(\frac{C(\mathbf{x},t)}{C(\mathbf{x},t) + k_d} \right) (1 + d_r \mathrm{RND}), \tag{4}$$

$$\partial_t \Gamma(t) = (f_a a(\mathbf{x},t) - \lambda(A - A_0)) \hat{\Gamma}(\mathbf{x},t), \tag{5}$$

a – local activator (LA); b – local inhibitor (LI); c – global inhibitor (GI);
s – stimulus strength in terms of external concentration C; RND $\sim \mathcal{U}(0,1)$ – a noise term for symmetry-breaking for pseudopod formation;
Γ – the cell boundary; $\hat{\Gamma}(\mathbf{x},t)$ – the (unit) outward normal to Γ at $\mathbf{x}$.

Idea behind the NSPDE model

An evolving cell boundary coupled with the three reaction-diffusion equations of the Meinhardt (1999) model of dynamic pattern (pseudopod) formation.



From Neilson et al. (2011)

NSPDE system

The NSPDE system stated by Tweedy (2018)

$$\partial_t a(\mathbf{x},t) = D_a \Delta_\Gamma a(\mathbf{x},t) + \frac{s(a(\mathbf{x},t)^2/c(t) + b_a)}{(k_M + b(\mathbf{x},t))(1 + {s_a}^2)} - d_a a(\mathbf{x},t), \quad (1)$$

$$\partial_t b(\mathbf{x},t) = D_b \Delta_\Gamma b(\mathbf{x},t) + k_b a(\mathbf{x},t) - d_b b(\mathbf{x},t), \quad (2)$$

$$\partial_t c(t) = \frac{r_c}{|\Gamma(t)|} \oint_\Gamma a(\mathbf{x},t) d\mathbf{x} - r_c a, \quad (3)$$

$$s(\mathbf{x},t) = 1 + d_r \mathrm{RND} + \left(\frac{C(\mathbf{x},t)}{C(\mathbf{x},t) + k_d} \right) (1 + d_r \mathrm{RND}), \quad (4)$$

$$\partial_t \Gamma(t) = (f_a a(\mathbf{x},t) - \lambda(A - A_0)) \hat{\Gamma}(\mathbf{x},t), \quad (5)$$

Parameters

10 parameters to be inferred:

$$\theta = (f_a, r_c, k_b, d_b, D_b, k_M, s_a, b_a, D_a, d_a)^T$$

(the remaining parameters treated as fixed in the simulator).

Parameter	Meaning	Default value $\tilde{\theta}$
f_a	rate of the outward force from LA	0.0015
r_c	response speed of GI	0.07
k_b	birth rate of LI	0.0028
d_b	death rate of LI	0.013
D_b	diffusivity of LI	0.045
k_M	Michaelis-Menten-like constant for LI	0.16
s_a	variable controlling saturation of LA	7.0E-5
b_a	basal production level of LA	0.1
D_a	diffusivity of LA	0.025
d_a	the death rate of LA	0.02

Data observability

Two observational settings:

- fully observed data: all the outputs, including chemical signals, assumed to be measurable
- partially observed data: based on visible outputs only (cell contours)

Simulator outputs

Simulator outputs

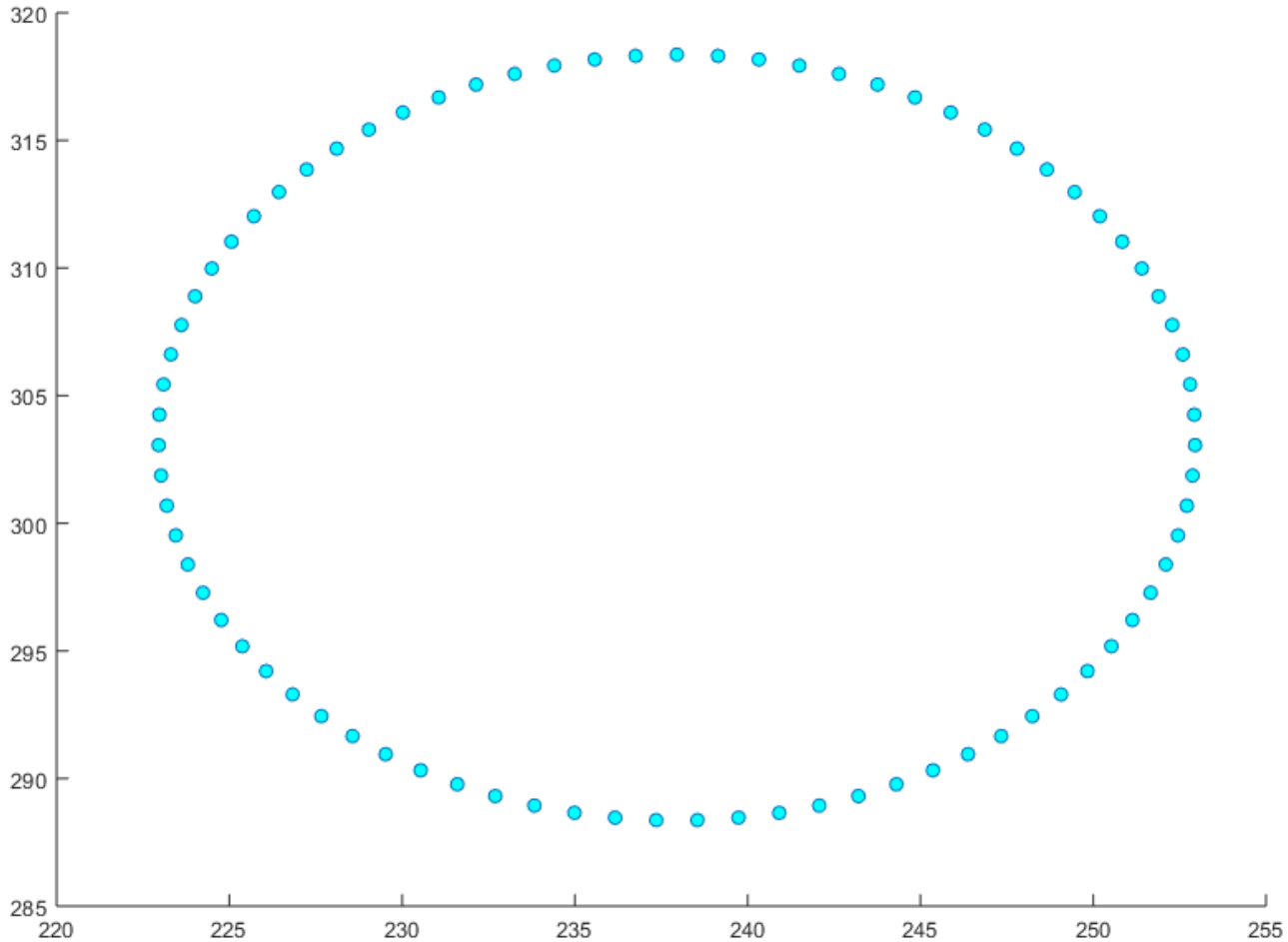
Time series of length $T = 1000$ of:

- XY coordinates of finite element (FE) nodes
- GI signal (the only univariate output)
- S signal at each node
- LA signal at each node
- LI signal at each node

← the ONLY outputs for partially observed data

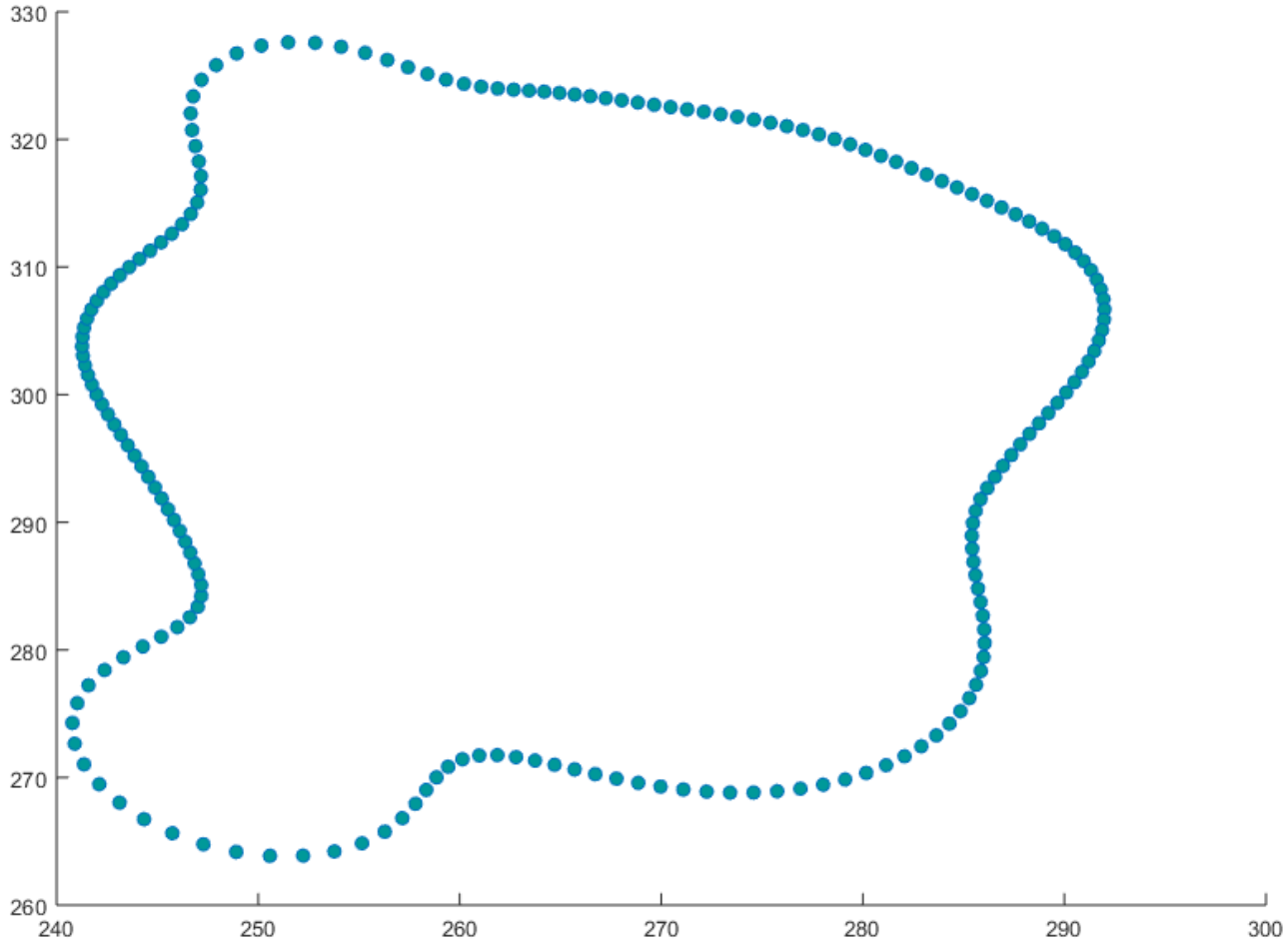
Output examples: XY coordinates (FE nodes) [single θ]

$t = 1$



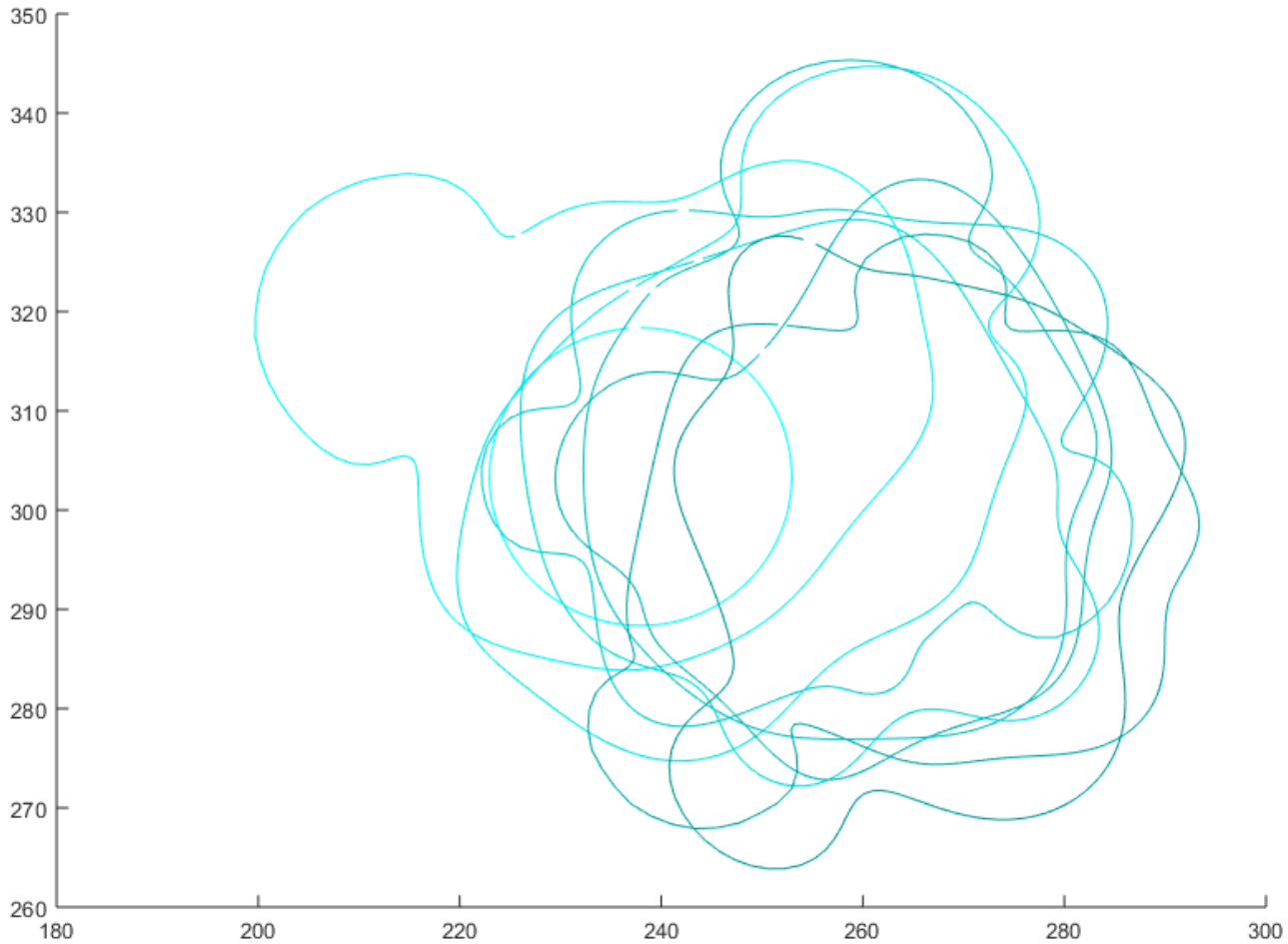
Output examples: XY coordinates (FE nodes) [single θ]

$$t = 1 + 8 \times 100$$

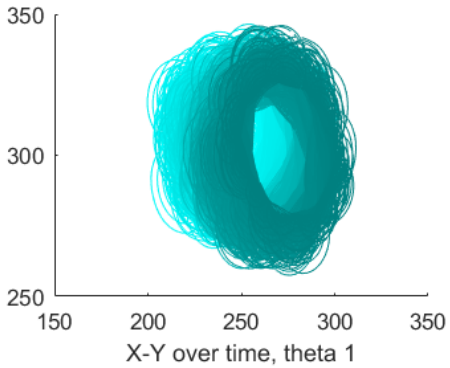


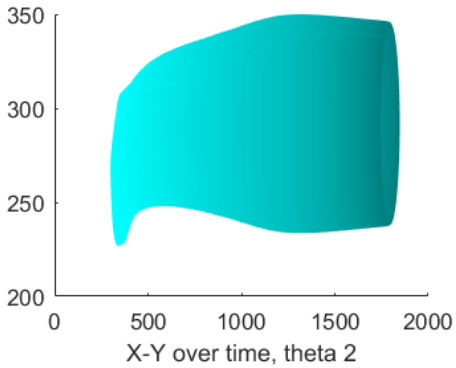
Output examples: XY coordinates (contours) [single θ]

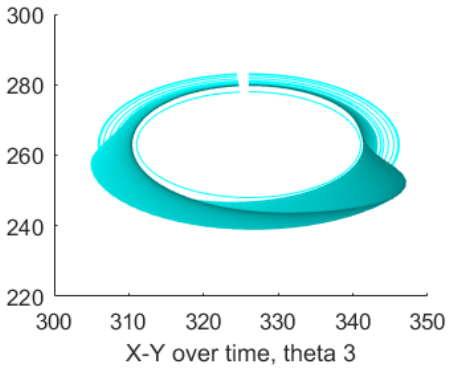
$$t \in \{1 + \bar{t} \times 100, \bar{t} = 1, \ldots, 9\}$$

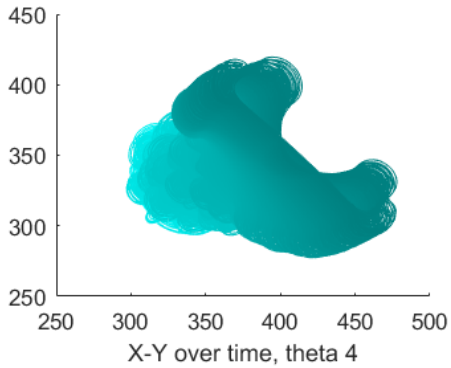


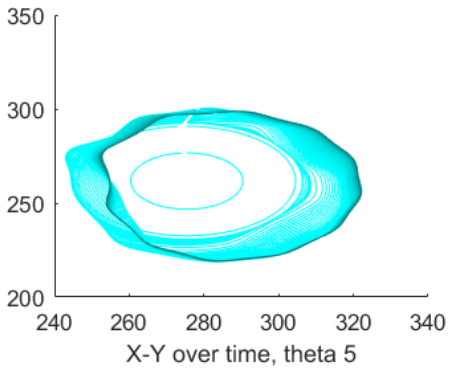
Output examples: time series of XY coordinates

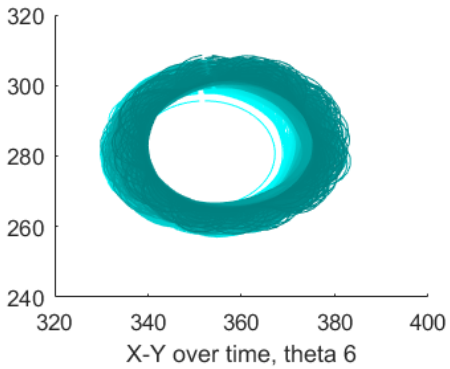


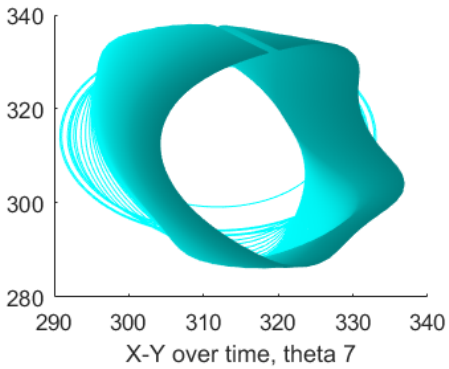


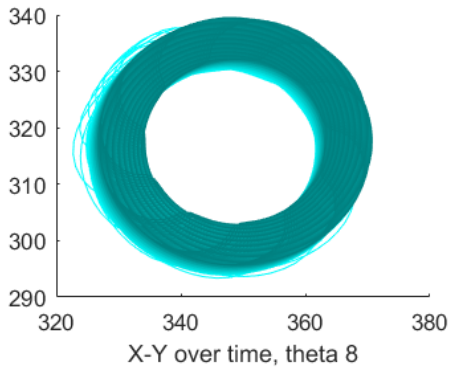


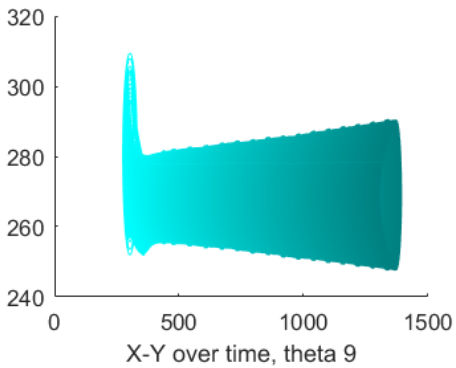


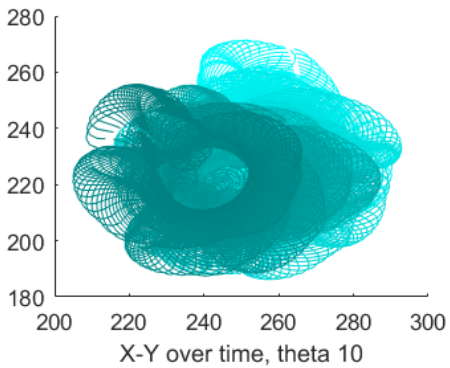


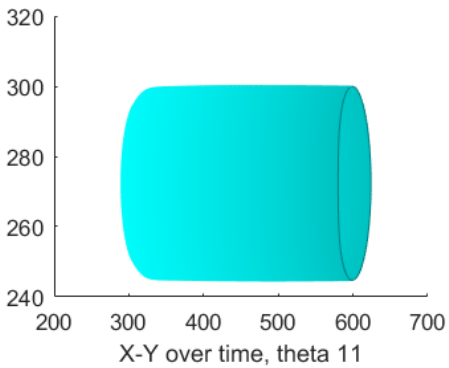


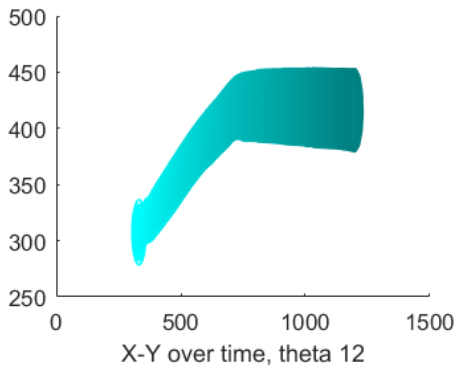




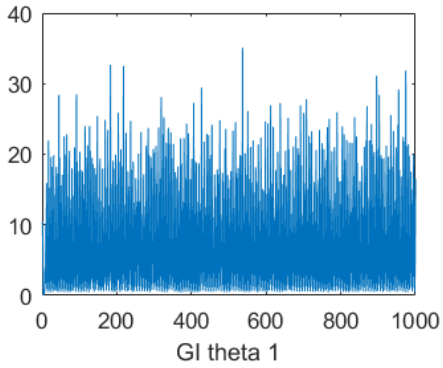


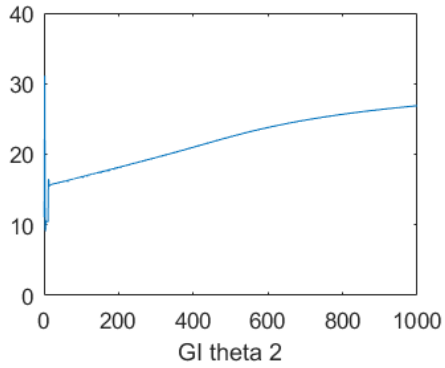


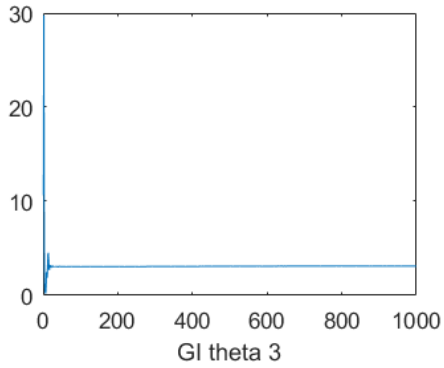


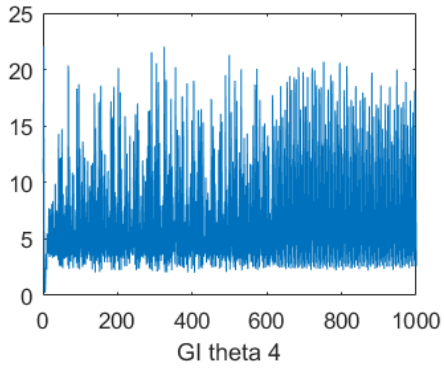


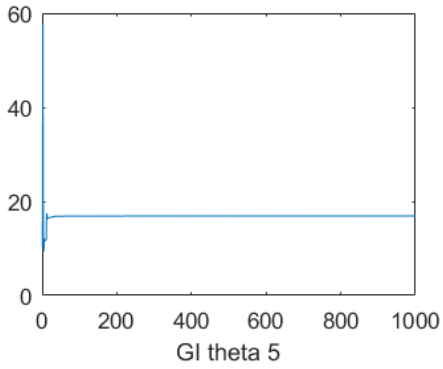
Output examples: GI (univariate series)

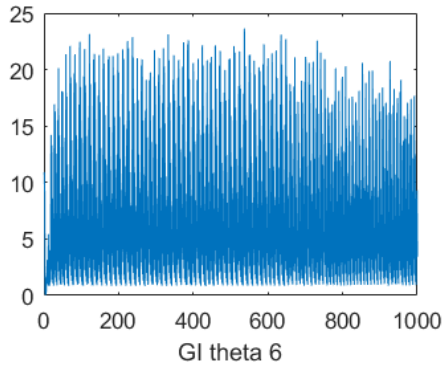


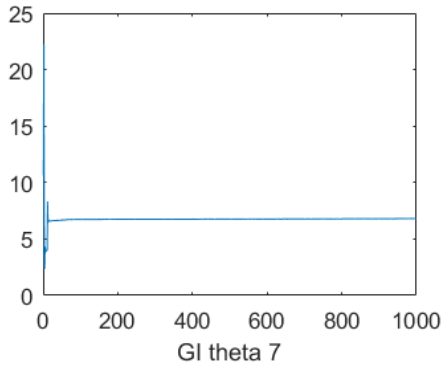


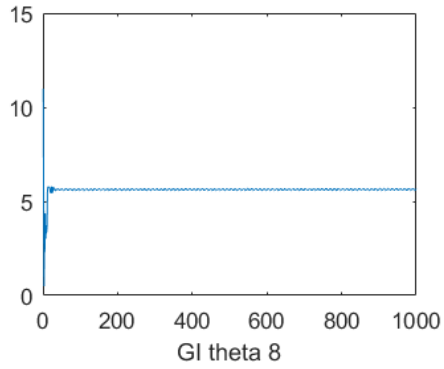


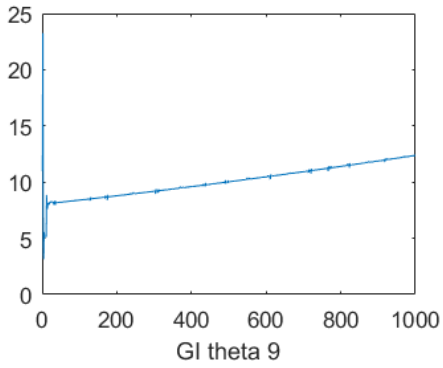


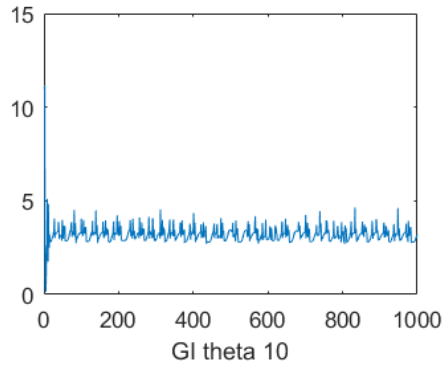


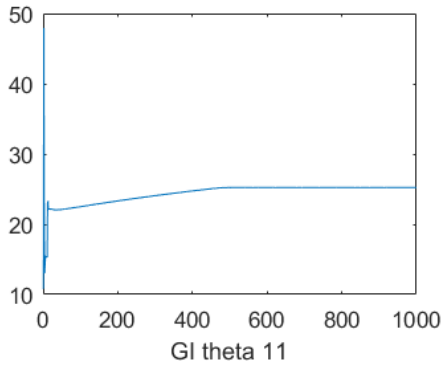


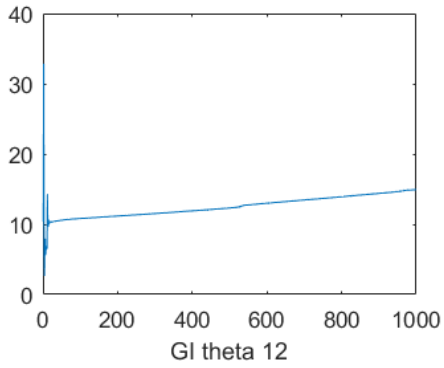






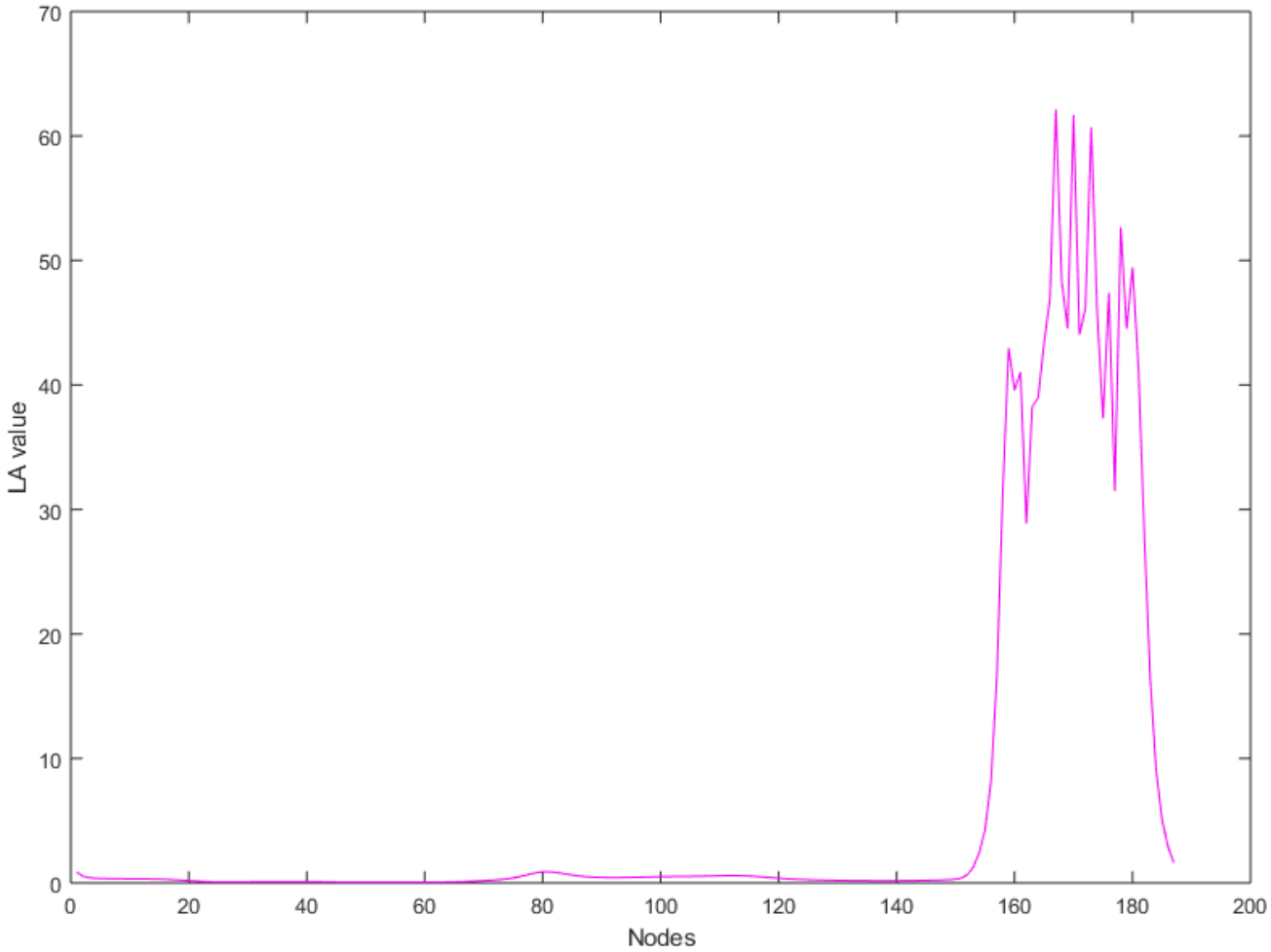




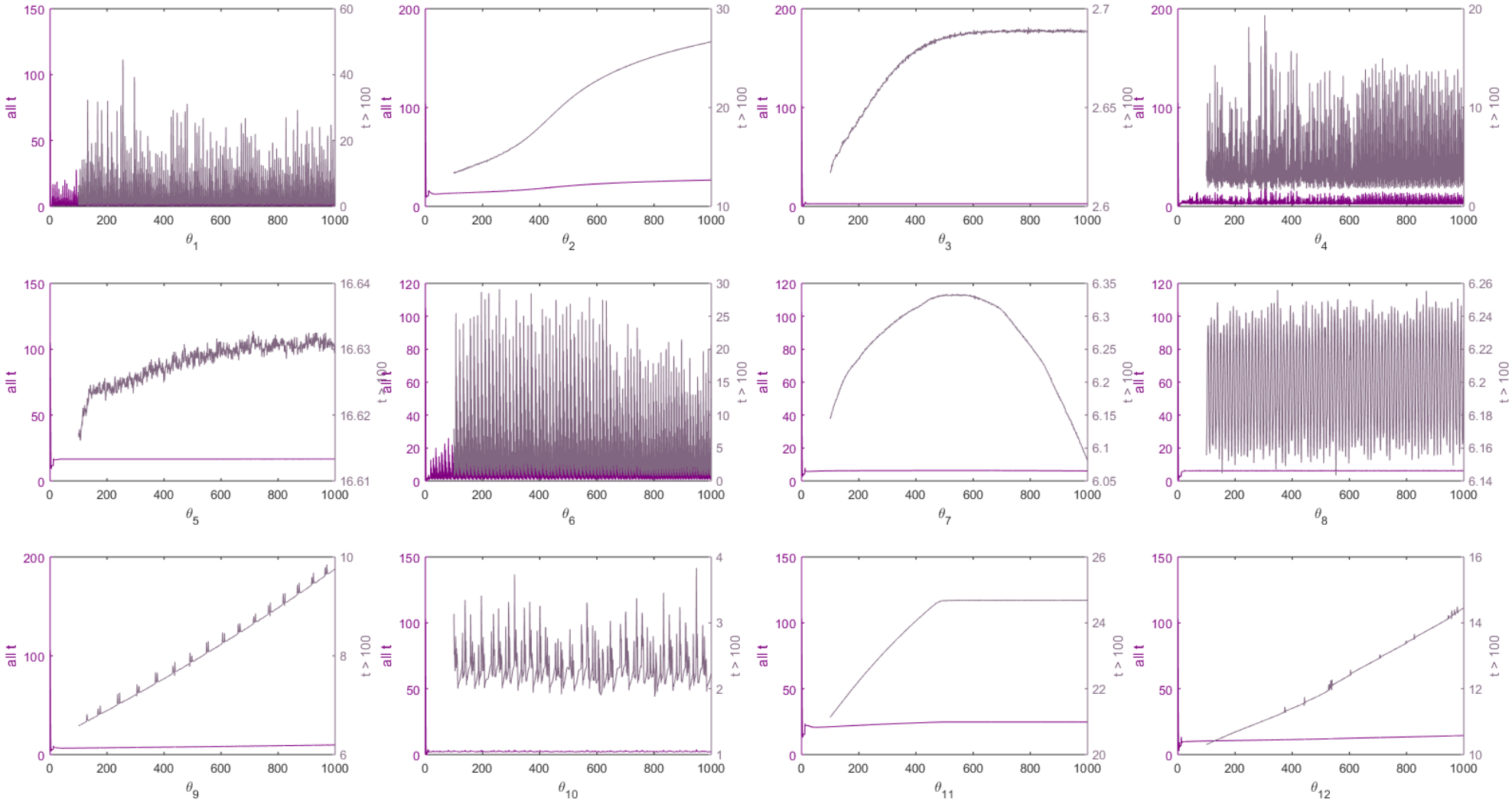


Output examples: LA over nodes [single θ]

$$t = 1 + 1 \times 100$$



Output examples: time series of LA means [over space]



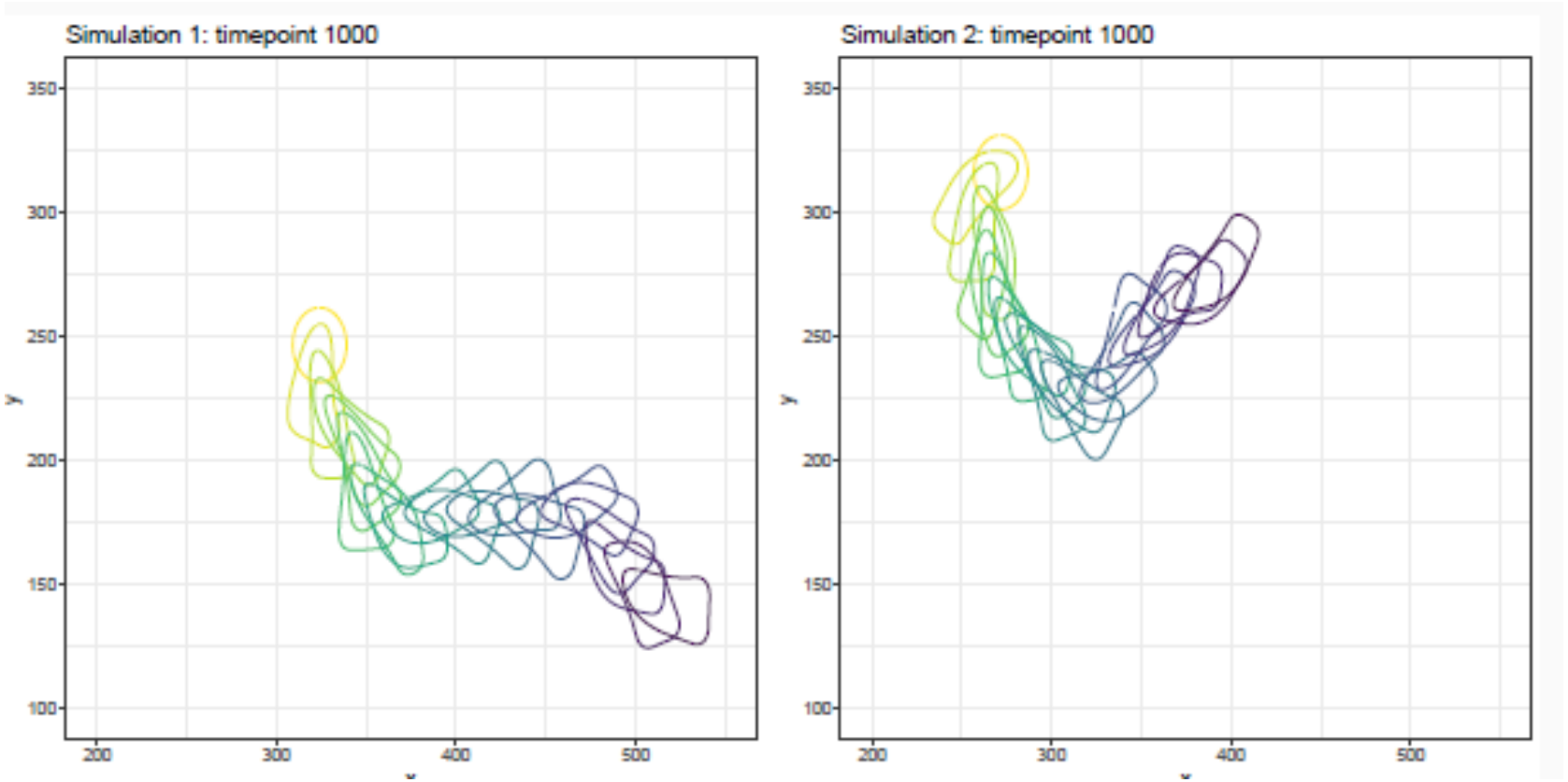
Output examples: conclusions

- **Very** heterogeneous.
- No clear clustering (in terms of θ).
- Both stable and dynamic patters.

Challenges (problems?)

Stochastic system

Each time a different realisation for the same θ.



Challenges (problems?) [cont'd]

- **Varying dimensions of the data**
  In ∼35% of cases the dimensions of the generated "matrices" were constant over time and equal to the initial size of 79 points.
  For other parameters cells were growing (up to 407 points) or shrinking (to 58 points).
- **Not time-aligned points**
  The points observed in subsequent time instances were not the same points: in each time instance the finite element nodes were uniformly drawn from the cell membrane.
  ⇒ Time-series analysis for the membrane points impossible.
- **No clear/known boundaries for the parameters**
  Working assumption needed: $\Rightarrow \theta \in [0.5\tilde{\theta}, 2\tilde{\theta}]$,
  where $\tilde{\theta}$ – the default values from Tweedy (2018).

Methodology: GP-ABC

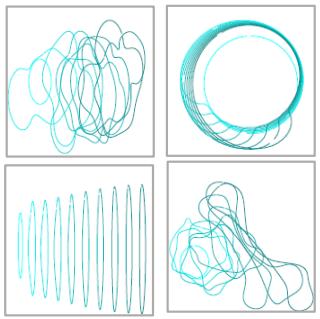
Overview

Pre-data phase: model based

Post-data phase: data based

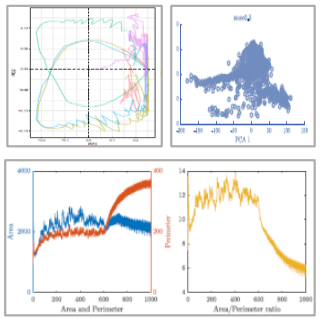
Problem specific

Simulating from the NSPDE model



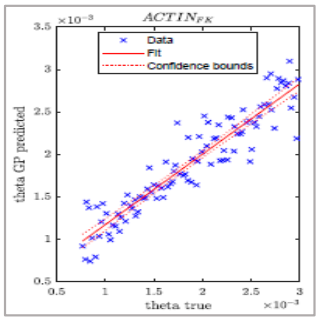
Feature extraction or construction

Methodology: Heuristic (domain knowledge)



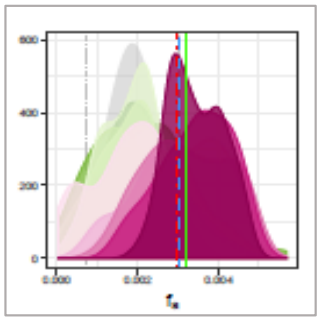
Feature combination and weighting

Methodology: GP regression



Estimation and uncertainty quantification

Methodology: ABC-SMC



General method

Step 2: features

In total:

56 variables for the fully observed data

31 variables for the partially observed data

Simple statistics: total distance travelled, mean area (over time), mean radius (over time), etc.

Label of the model best fitting the time series of area–perimeter ratios (polynomials, exponentials, Fouriers)

Fourier transform + PCA: e.g. scores of first 2 components, percentiles of PCA components for signals and cell contours

Fourier shape descriptors + PCA: e.g. number of crossings of the x and y axis in the space of the first 2 components

Step 3: Gaussian process regression

- Gaussian process (GP) regression (Rasmussen and Williams, 2006) instead of linear regression in semi-automatic ABC of Fearnhead and Prangle (2012).
- Separate regression for each parameter θ_i, $i = 1, \ldots, 10$
- Initial design: 2000 points from Sobol sequences (for training) and 100 points from Latin hypercube (for validation).

Step 3: kernel sensitivity analysis

Kernels (covariance functions) considered:

1. squared exponential (se),
2. Matérn 3/2 (m32),
3. Matérn 5/2 (m52),
4. rational quadratic (rq),
5. neural network (nn),

all with automatic relevance determination (ARD).

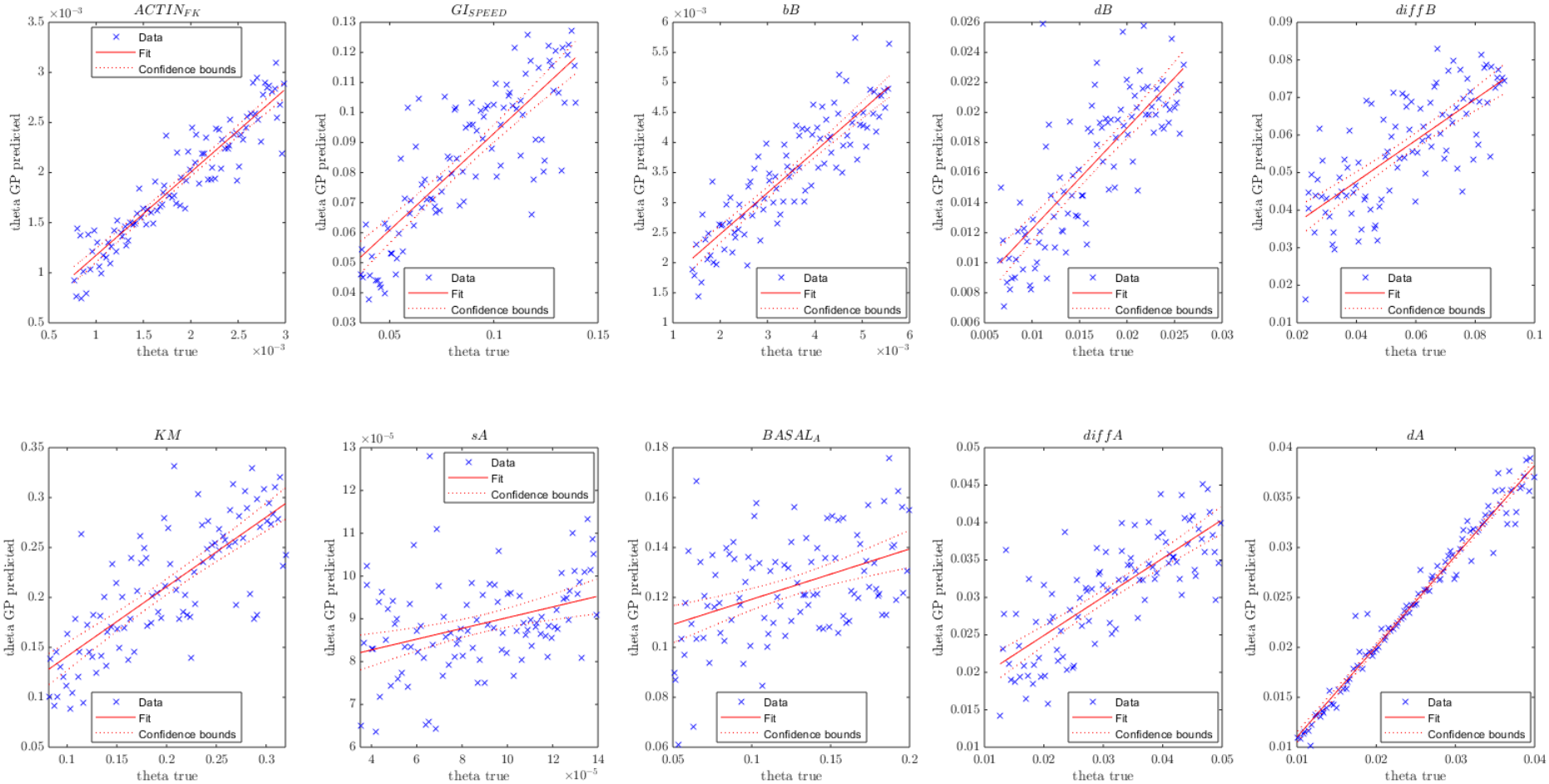
The final configuration: based on the out-of-sample performance (lowest RMSE) of 15 different kernels (for each of 10 parameters).

Step 3: final kernel configuration

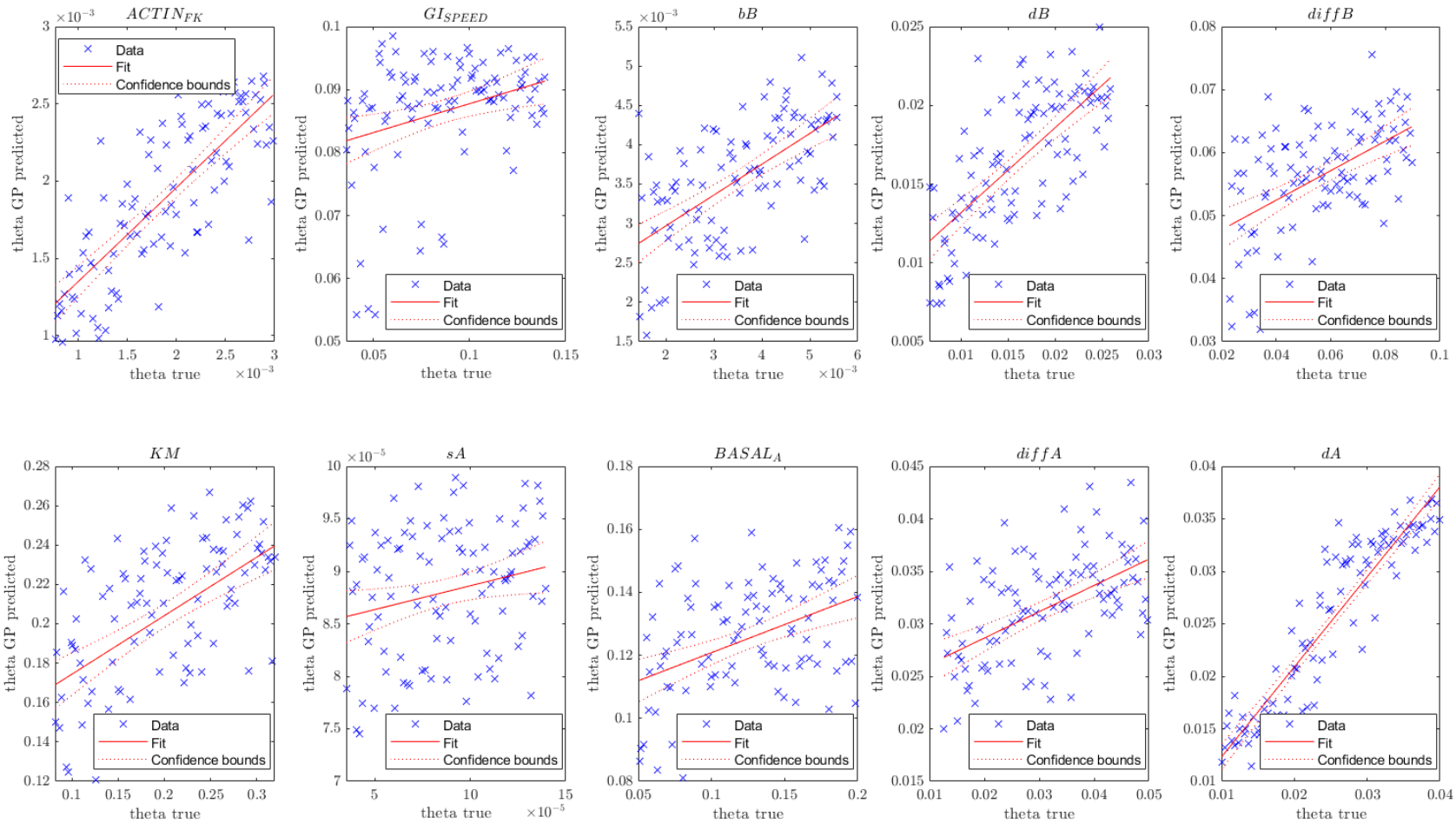
Parameter	Fully observed		Partially observed	
f_a	seARD	NR	m32ARDrstr	R
r_c	m32	NR	m32ARD	R
k_b	m32	NR	qrARDrstr	R
d_b	m32ARD	R	nn	R
D_b	rqARD	R	m52ARDrstr	R
k_M	rq	R	nn	R
s_a	nn	NR	nn	R
b_a	rq	R	m52ARDrstr	R
D_a	rqARD	R	nnARD	R
d_a	rq	R	qrARD	R

(R/NR refers to 'refined'/'not refined' i.e. estimated or not on 2000 calibration points starting from the estimates from 1000 calibration points)

Step 3: out-of-sample fit (fully observed data)



Step 3: out-of-sample fit (partially observed data)



Step 4: approximate Bayesian computations

Why ABC?

- Likelihood free methods
- Based on simulations from the model.
- (Typically) use summary statistics to assess whether the generated dataset is "similar" to the observed dataset.

.

Semi-automatic ABC of Fearnhead and Prangle (2012): use predictions from linear regression as a summary statistics.

We extend the approach of Fearnhead and Prangle (2012): base summary statistics on predictions from GP regression.

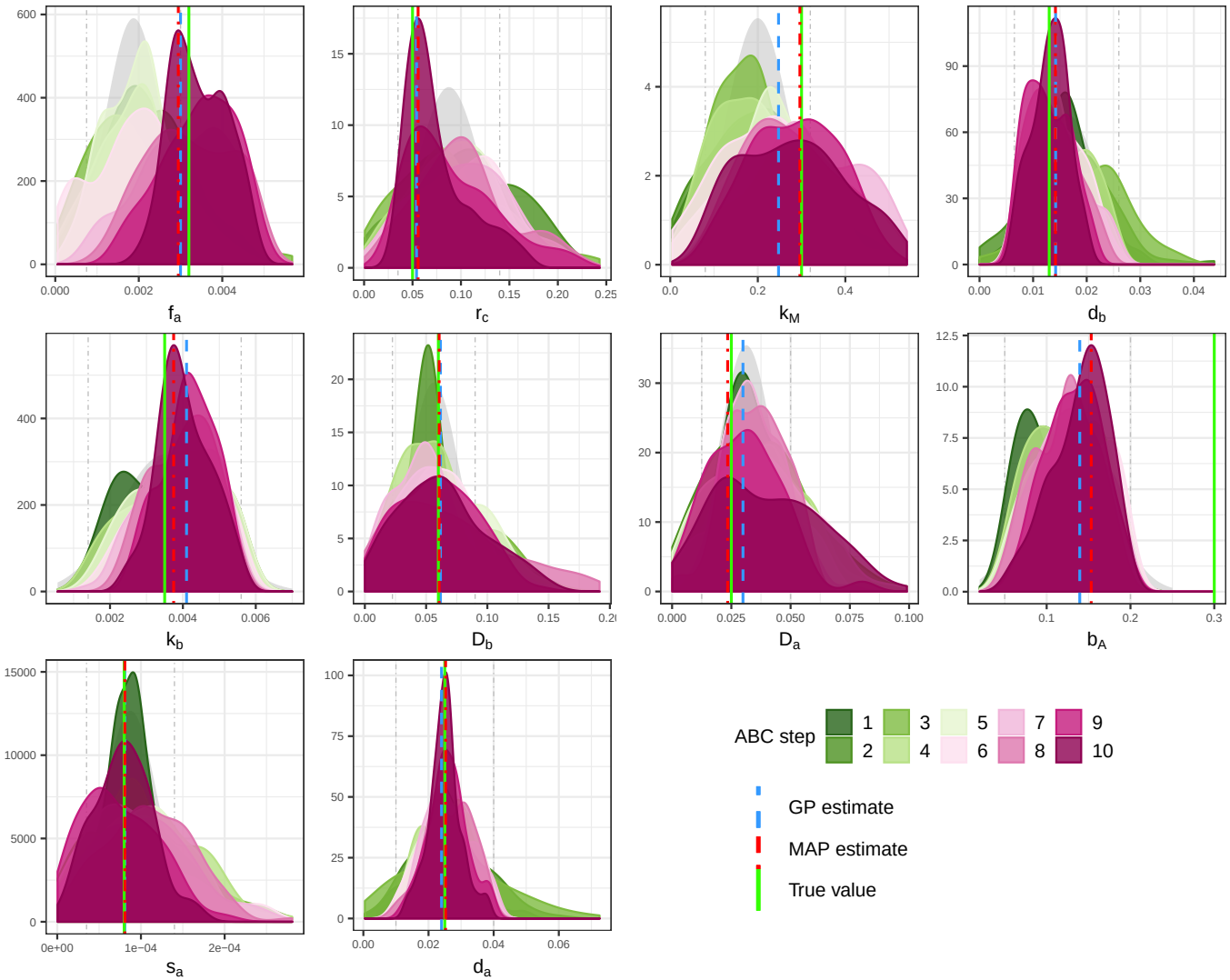
Step 4: ABC-SMC

ABC-SMC (Beaumont et al., 2002) more efficient than rejection ABC or ABC-MCMC for high dimensional, stochastic models.

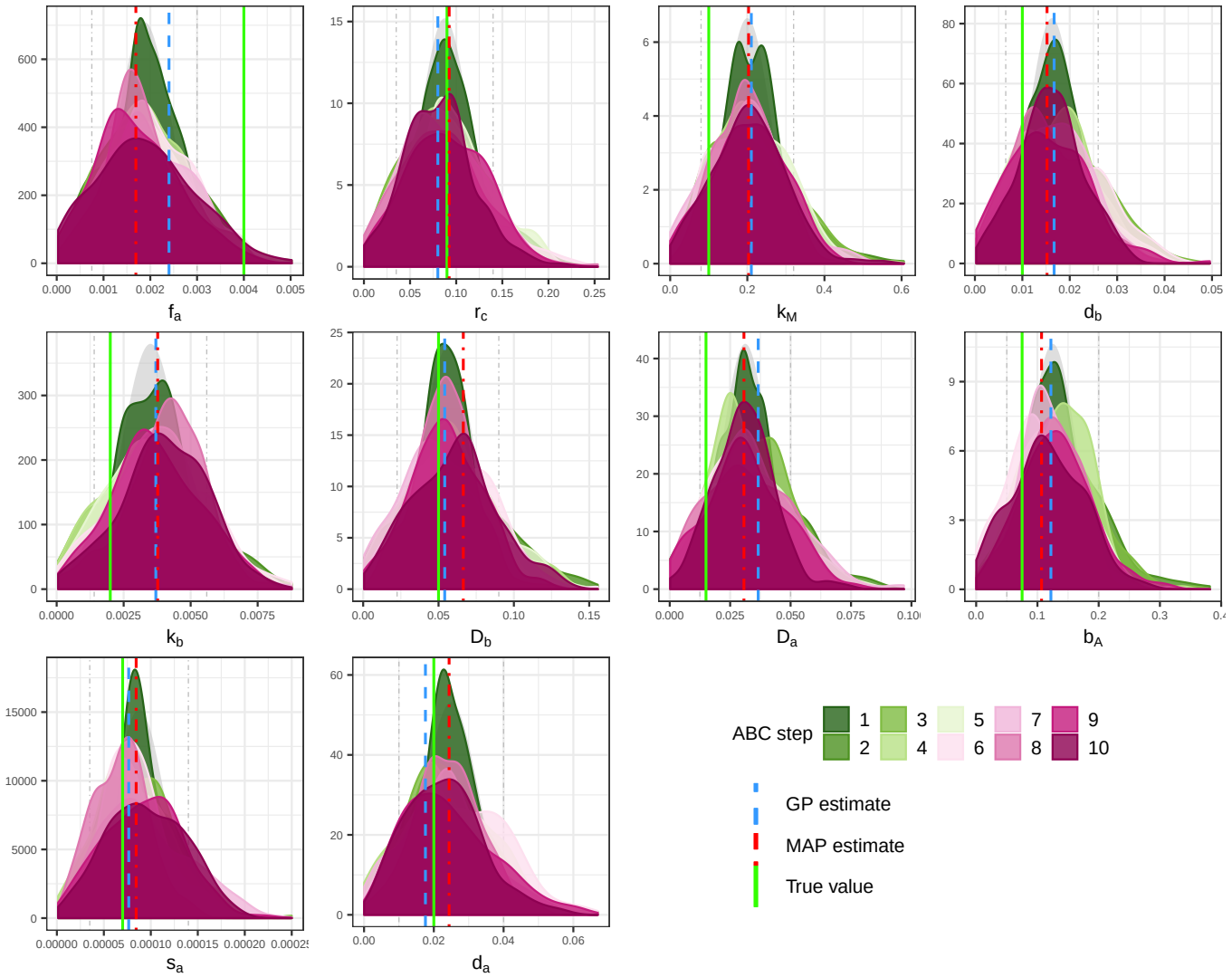
Simplified ABC-SMC steps:

1. sample $\boldsymbol{\theta}^*$ from prior $\pi(\boldsymbol{\theta})$ $[\mathcal{N}(1.25\tilde{\theta}_i, (0.375\tilde{\theta}_i)^2)]$
2. simulate from the model with $\boldsymbol{\theta}^*$ [2–3 minutes]
3. compute distance D between real and simulated data [Euclidean]
4. if $D \leq \epsilon$ accept $\boldsymbol{\theta}^*$, otherwise reject
5. repeat until P particles are accepted $[P = 100]$
6. construct a new prior by perturbing the P accepted particles
7. decrease the tolerance level ϵ [linearly]
8. repeat from step 1 [10 times]

Step 4: posterior densities (fully observed data)



Step 4: posterior densities (partially observed data)



Step 4: benefits from ABC correction

1. More accurate predictions (compared to using only the predictions from GP regression):
   - for most of the NSPDE parameters for fully observed data,
   - for half of the parameters for partially observed data.
2. ABC can help with parameters for which the initial compact calibration domain was too narrow
3. Flexible uncertainty quantification: GP regression implies Gaussian posterior, ABC density plots very non-Gaussian.

Discussion

Limitations and further research

1. Ranges for the initial designs.
2. Bayesian estimation of GP hyperparameters.
3. Multi–output GPs instead of univariate regressions.
4. Search for further informative features.

Applications to real data

Ultimate goal: to apply the proposed modelling and inference framework to real cell migration data (e.g. obtained by high-resolution microscopy).

Such applications would be of particular interest to cancer research, so as to shed more light on the mechanisms underlying metastasis.

Methodological challenges:

1. Video data ⇒ segmentation and tracking (manual or automatic).
2. Dimensionless quantities of the NSPDE model of Tweedy et al. (2013) ⇒ variable transformations to link the equations to real observations
3. No ground truth ⇒ difficult to assess inference accuracy.

References

Beaumont, M. A., W. Zhang, and D. J. Balding (2002), "Approximate Bayesian Computation in Population Genetics." *Genetics*, 162, 2025–2035.

Fearnhead, P. and D. Prangle (2012), "Constructing Summary Statistics for Approximate Bayesian Computation: Semi-Automatic Approximate Bayesian Computation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 419–474.

Meinhardt, H. (1999), "Orientation of Chemotactic Cells and Growth Cones: Models and Mechanisms." *Journal of Cell Science*, 112, 2867–2874.

Neilson, M. P., D. M. Veltman, P. J. M. van Haastert, S. D. Webb, J. A. Mackenzie, and R. H. Insall (2011), "Chemotaxis: a Feedback-Based Computational Model Robustly Predicts Multiple Aspects of Real Cell Behaviour." *PLoS biology*, 9, 1–11.

Rasmussen, C. E. and C. K. I Williams (2006), *Gaussian Processes for Machine Learning*. MIT Press.

Toni, Tina and Michael PH Stumpf (2009), "Tutorial on abc rejection and abc smc for parameter estimation and model selection." *arXiv preprint arXiv:0910.4472*.

Tweedy, L. (2018), "Meinhart Simulations on an Evolving Line." Technical report, Cside 2018 Supplementary Notes.

Tweedy, L., B. Meier, J. Stephan, D. Heinrich, and R. G. Endres (2013), "Distinct cell shapes determine accurate chemotaxis." *Scientific Reports*, 3, 2606.

Appendix

Simulator

Methodology:

Equations (1)–(3): approximated on the evolving cell perimeter using an Arbitrary Lagrangian Eulerian surface finite element method using piecewise linear elements.

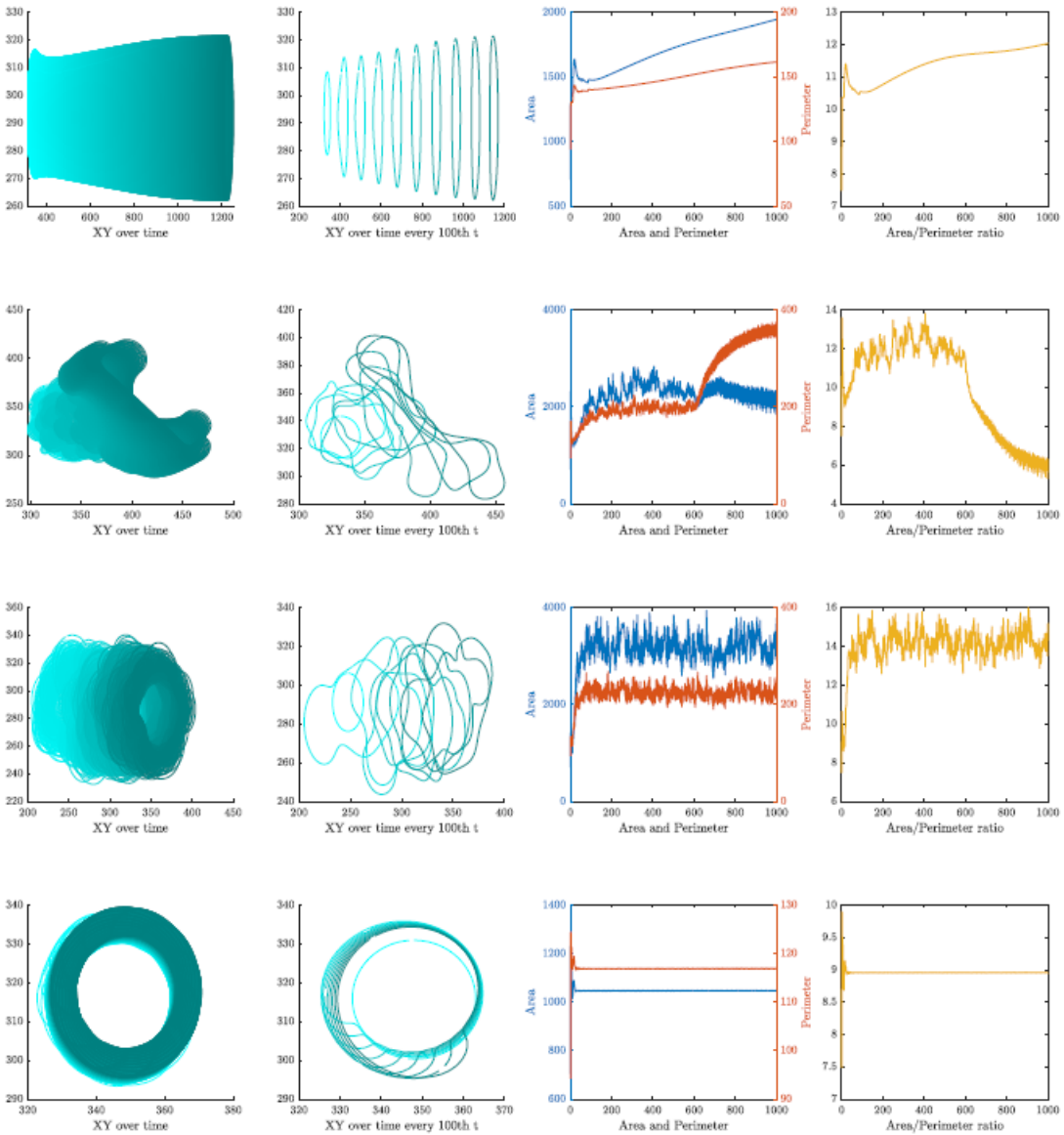
Time integration: achieved using a semi-implicit approach.
The computed activator profile is used to drive a mechanical model of the protrusive and retractive forces exerted on the cell membrane.

Movement of the cell: obtained using a level set method and a moving Cartesian mesh. Calculations are performed using the level set toolbox in MATLAB

Default settings:

- simulation of duration of 100000;
- arbitrary units u;
- outputs samples every $100u$ (evenly spaced);
- output of time dimension $T = 1000$.

Step 2: features (area, perimeter, area-perimeter ratios)



Step 2: features (zero-crossings in PCA space)