Sequential Monte Carlo for estimating parameters of differential equations

Liangliang Wang

Dept. of Statistics and Actuarial Science
Simon Fraser University (SFU)
Burnaby, BC, Canada

Oct 16, 2020

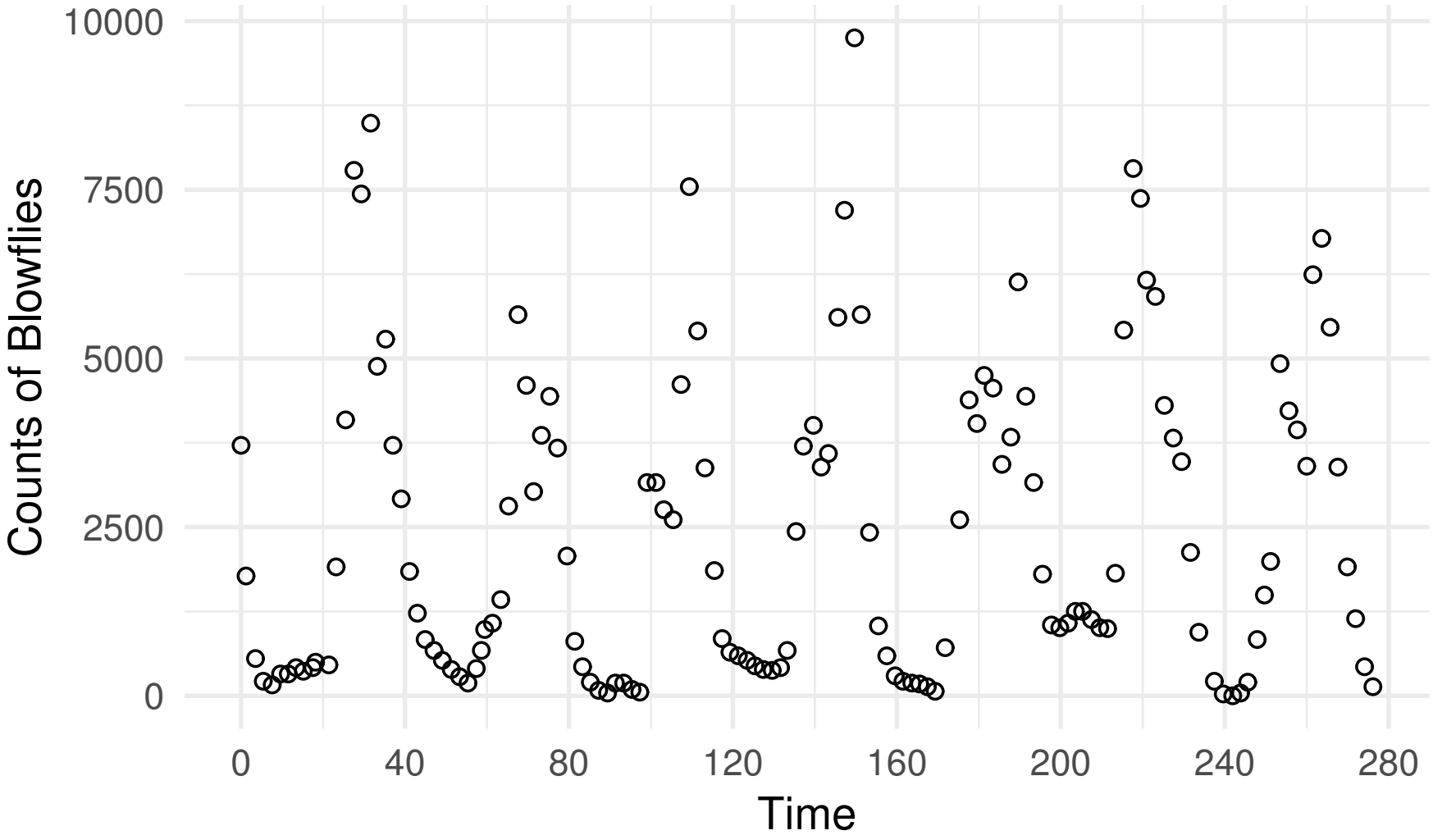
Outline

Outline

Population of Blowflies

- Nicholson (1954, 1957) published his classic studies on the competition for food in laboratory populations of Australian sheep blowflies.
- The blowflies were cultured in a room maintained at 25 $^{\circ}\mathrm{C}$.
- Nicholson measured the population of blowflies every day for approximately one year.

Number of Blowflies VS Days



Population of Blowflies

$$\frac{dx(t)}{dt} = rx(t)\left(1 - \frac{x(t-\tau)}{1000 \cdot P}\right)$$

- $x(t)$: the blowfly population
- r: the rate of increase of the blowfly population
- P: a resource limitation parameter set by the supply of food
- τ: the time delay, roughly equal to the time for a larva to grow up to an adult

Delay Differential Equations (DDEs)

A typical form for DDEs:

$$\begin{aligned} \frac{dx_i(t)}{dt} &= g_i(\mathbf{x}(t), \mathbf{x}(t-\tau)|\boldsymbol{\theta}), \\ x_i(0) &= x_{i0}, \end{aligned}$$

- $x(t)$: the dynamic process on $[t_1, t_J]$.
- τ, a constant delay parameter.
- $\boldsymbol{\theta}$, a vector of parameters describing the system.
- The parametric form of $g_i(\cdot)$ is often known.
- Ordinary Differential Equations (ODE) is a special case of DDE when $\tau = 0$.

Likelihood function

- Let y_{ij} be the j-th observation of the i-th DE $x_i(\cdot)$ at time t_{ij}.
- Likelihood function:

$$L(y_{ij}|x_i(t_{ij}), \boldsymbol{\theta}_y), j = 1, \ldots, J,$$

where $\boldsymbol{\theta}_y$ is the vector of the parameters in the observation model.

- For example,

$$y_{ij} \sim N(x_i(t_{ij}), \sigma_i^2), j = 1, \ldots, J,$$

where $\boldsymbol{\theta}_y = (\sigma_1^2, \ldots, \sigma_I^2)^T$.

Challenges

Question: Estimate θ, τ, and x_{i0} from noisy and/or partially observed data?

Challenges

- The DDEs usually have no analytic solutions and can only be solved numerically.
- The DDE solutions will not only depend on the parameters θ and τ, but will also rely on the history of the dynamic process $\mathcal{H}_\tau = \{x(t), t \in [t_1 - \tau, t_1]\}$ which is an infinite-dimensional set.
- DDE solutions are extremely sensitive to delay parameters.

Outline

Maximum Likelihood Method

The DE parameters θ can be estimated by maximizing the log likelihood function

$$H^*(\theta) = \sum_{i=1}^{n} \log L(y_i | X(t_i|\theta)) \,.$$

Difficulties with MLE

- The DE solution $X(t|\theta)$ is an implicit function of θ.
- High computation load associated with the numeric solver.
- Many local optima in the log likelihood function.
- We have to specify the unknown history of the dynamic process $\mathcal{H}_\tau$ to solve the DDE numerically.

A two-step procedure

Ellner et al. (1997)

Step 1: Nonparametric Smoothing

- $y_i = X(t_i) + \epsilon_i$
- Estimate $X(t)$ and $X'(t)$ from noisy data using nonparametric smoothing methods

Step 2: Nonlinear Regression

$$\min_{\boldsymbol{\theta},\tau} \sum_{i=1}^{n} \|\hat{X}'(t_i) - g(\hat{X}(t), \hat{X}(t-\tau)|\boldsymbol{\theta})\|^2$$

Pros and cons of the two-step procedure

Pros

- No need to solve DDE
- Fast computation

Cons

- $X'(t)$ may not be well estimated, especially from sparse data.
- Thus DE parameter estimation is not accurate.

Semiparametric Method

Wang and Cao (2012)

Advantages

- No need to solve DDE numerically
- No need to specify DDE history
- Estimation is more accurate than the two-step method.

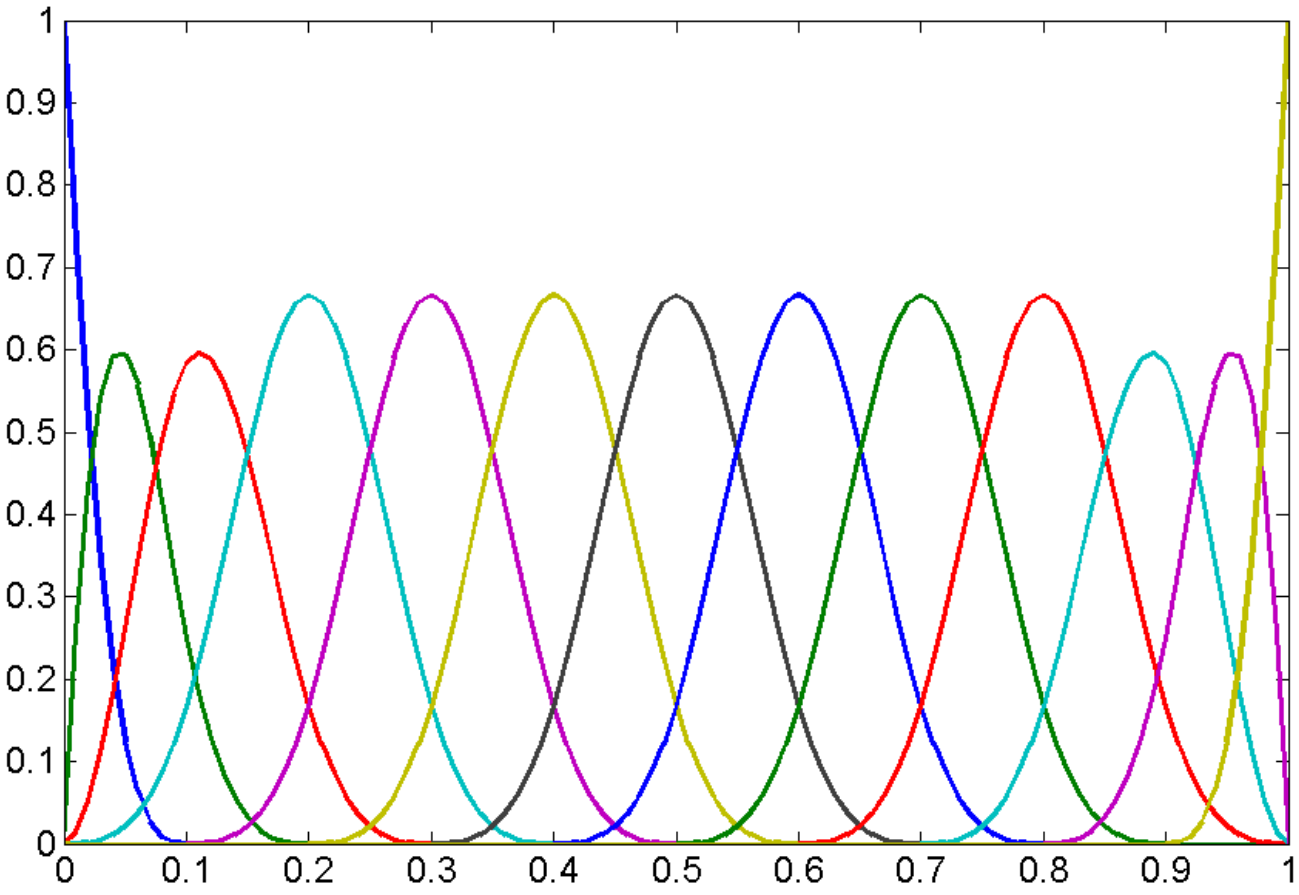
Semiparametric Method

Estimate a nonparametric function to approximate DDE solutions by a linear combination of basis functions:

$$x(t) = \sum_{k=1}^{K} \phi_k(t) c_k = \phi^T(t)\mathbf{c}\,,$$

- $\phi(t) = (\phi_1(t), \ldots, \phi_K(t))^T$ is the vector of basis functions at time t, for example, B-spline (Fixed and Known).
- $\mathbf{c} = (c_1, \ldots, c_K)^T$ are the basis function coefficients.

Cubic B-spline Basis



Estimating Spline Coefficients

Fitting to data

- Observations: y_i
- Nonparametric function: $x(t) = \phi^T(t)\mathbf{c}$
- Fitting to data: $C_1 = \sum_{i=1}^{n} \log L(y_i|x(t_i))$

Fidelity to DDE $\frac{dx(t)}{dt} = g(x(t), x(t-\tau)|\theta)$

- **Difference between two sides of DDE**: $\frac{dx(t)}{dt} - g(x(t), x(t-\tau)|\theta)$
- **Fidelity to DDE**: $C_2 = \int_{t_1+\tau}^{t_n} \left\{ \frac{dx(t)}{dt} - g(x(t), x(t-\tau)|\theta) \right\}^2 dt$

Criterion: $J(\mathbf{c}|\theta) = C_1 + \lambda C_2$

Estimating Spline Coefficients

Fitting to data

- Observations: y_i
- Nonparametric function: $x(t) = \phi^T(t)\mathbf{c}$
- Fitting to data: $C_1 = \sum_{i=1}^{n} \log L(y_i|x(t_i))$

Fidelity to DDE $\frac{dx(t)}{dt} = g(x(t), x(t-\tau)|\theta)$

- **Difference between two sides of DDE**: $\frac{dx(t)}{dt} - g(x(t), x(t-\tau)|\theta)$
- **Fidelity to DDE**: $C_2 = \int_{t_1+\tau}^{t_n} \left\{ \frac{dx(t)}{dt} - g(x(t), x(t-\tau)|\theta) \right\}^2 dt$

Criterion: $J(\mathbf{c}|\theta) = C_1 + \lambda C_2$

Estimating Spline Coefficients

Fitting to data

- Observations: y_i
- Nonparametric function: $x(t) = \phi^T(t)\mathbf{c}$
- Fitting to data: $C_1 = \sum_{i=1}^{n} \log L(y_i|x(t_i))$

Fidelity to DDE $\frac{dx(t)}{dt} = g(x(t), x(t-\tau)|\theta)$

- **Difference between two sides of DDE**: $\frac{dx(t)}{dt} - g(x(t), x(t-\tau)|\theta)$
- **Fidelity to DDE**: $C_2 = \int_{t_1+\tau}^{t_n} \left\{ \frac{dx(t)}{dt} - g(x(t), x(t-\tau)|\theta) \right\}^2 dt$

Criterion: $J(\mathbf{c}|\theta) = C_1 + \lambda C_2$

Estimating DDE parameter θ

- Criterion:

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L(y_i | \hat{x}(t_i | \boldsymbol{\theta})) \,,$$

where $\hat{x}(t|\boldsymbol{\theta}) = \phi^T(t)\hat{\mathbf{c}}(\boldsymbol{\theta})$ and we obtain the estimate $\hat{\mathbf{c}}$ by maximizing the penalized log likelihood function.

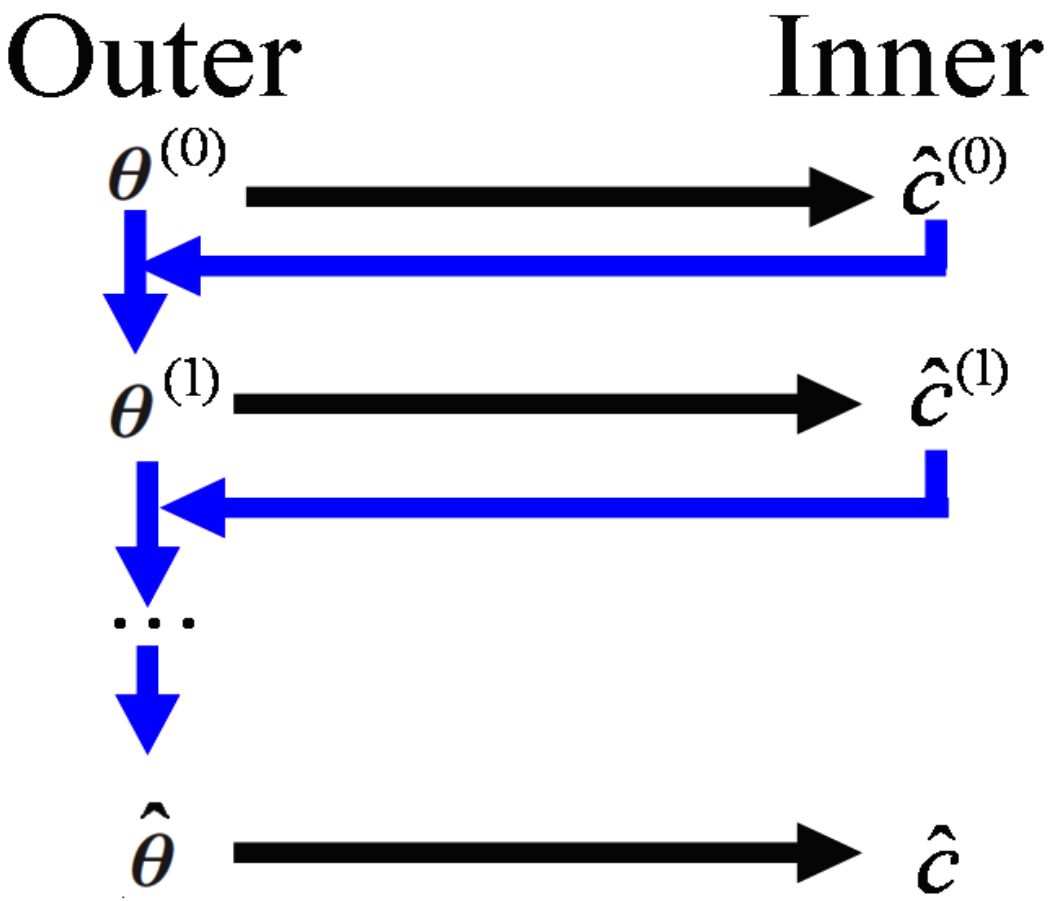
Two nested levels of optimization

- **Inner level:** $J(\mathbf{c}|\theta)$
  $\Rightarrow$**$\hat{\mathbf{c}}$ is a function of θ:** $\hat{\mathbf{c}}(\theta)$**.**
- **Outer level:** $H(\hat{\mathbf{c}}(\theta), \theta)$

c: basis coefficients;

θ: DDE parameter;

Diagram for our semiparametric method



Smoothing Parameter Selection

The K-fold cross validation:

$$CV = \frac{1}{K} \sum_{j=1}^{K} \sum_{i \in \mathcal{A}^{(-j)}} \log L(y_i | \hat{x}(t_i | \theta^{(-j)})).$$

Outline

Bayesian inference

Posterior distribution

$$\begin{aligned}\pi(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda) &\propto \gamma(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda) \\ &= L(\mathbf{y}|\boldsymbol{\theta}_y, \mathbf{c})\pi_0(\mathbf{c}|\boldsymbol{\theta}_{\mathrm{DE}}, \lambda)\pi_0(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \lambda),\end{aligned}$$

where γ is the unnormalized posterior distribution, and $\boldsymbol{\theta}_{\mathrm{DE}}$ includes $\boldsymbol{\theta}$ and τ.

Prior distribution for c

Given $\lambda, \boldsymbol{\theta}, \tau$, the prior distribution for $\mathbf{c}$ is

$$\begin{aligned}
&\pi_0(\mathbf{c}|\boldsymbol{\theta}, \tau, \lambda) \propto \\
&\exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{I} \int_{t_1+\tau}^{t_J} \left[\frac{dx_i(s)}{ds} - g_i(\mathbf{x}(s), \mathbf{x}(s-\tau)|\boldsymbol{\theta}) \right]^2 ds \right\}, \\
= \quad &\exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{I} \int_{t_1+\tau}^{t_J} \left[\frac{d\boldsymbol{\Phi}(s)'}{ds} \mathbf{c}_i - g_i(\boldsymbol{\Phi}(s)'\mathbf{c}, \boldsymbol{\Phi}(s-\tau)'\mathbf{c}|\boldsymbol{\theta}) \right]^2 ds \right\}.
\end{aligned}$$

This prior distribution measures how well the estimated DE variables $\hat{\mathbf{x}}(t)$ satisfy the DE system.

Prior distributions $\pi_0(\boldsymbol{\theta}_{\mathbf{DE}}, \boldsymbol{\theta}_y, \lambda)$

For example,

$$\begin{aligned}
\boldsymbol{\theta} &\sim MVN(\mathbf{0}_D, \sigma_\theta^2 \mathbf{I}_D), \\
\tau &\sim U(t_1, t_J), \\
\sigma_i^2 &\sim IG(g_0, h_0),\ i = 1, \ldots, I, \\
\lambda &\sim Gamma(a_\lambda, b_\lambda).
\end{aligned}$$

Likelihood function

For example,

$$L(\mathbf{y}|\mathbf{c},\boldsymbol{\sigma}) \propto$$
$$(\prod_{i=1}^{I}\prod_{j=1}^{J}\sigma_i^2)^{-1/2}\exp\left\{-\sum_{i=1}^{I}\left(\sum_{j=1}^{J}\frac{(y_{ij}-\mathbf{\Phi}(t_{ij})'\mathbf{c}_i)^2}{2\sigma_i^2}\right)\right\}$$

Intermediate distributions

The r-th intermediate distribution:

$$\begin{aligned} & \pi_r(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda) \propto \gamma_r(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda) \\ = \; & [L(\mathbf{y}|\boldsymbol{\theta}_y, \mathbf{c})\pi_0(\mathbf{c}|\boldsymbol{\theta}_{\mathrm{DE}}, \lambda)\pi_0(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \lambda)]^{\alpha_r}\rho(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda)^{1-\alpha_r}, \\ = \; & [L(\mathbf{y}|\boldsymbol{\beta})\pi_0(\boldsymbol{\beta})]^{\alpha_r}\rho(\boldsymbol{\beta})^{1-\alpha_r} \end{aligned}$$

- $\rho(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda)$ is a reference distribution.
- $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{R-1} < \alpha_R = 1$ is a sequence of annealing parameters.
- $\pi_0(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda) = \rho(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda)$.
- $\pi_R(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda) = \pi(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda)$.
- Denote $\boldsymbol{\beta} = (\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda)^T$.

Annealed SMC

(Del Moral et al. 2006, 2007)

sample $\tilde{\beta}_{0,k} \sim \rho(\cdot)$
set its unnormalized weight $w_{0,k} = 1$.
for $r \in 1, \ldots, R$ **do**
sample $\beta_{r,k} \sim T_r(\tilde{\beta}_{r-1,k}, \cdot)$
compute

$$w_{r,k} = w_{k,r-1} \cdot \frac{\gamma_r(\beta_{r,k})}{\gamma_{r-1}(\tilde{\beta}_{r-1,k})} \cdot \frac{L_{r-1}(\beta_{r,k}, \tilde{\beta}_{r-1,k})}{T_r(\tilde{\beta}_{r-1,k}, \beta_{r,k})}$$

normalize weights $W_{r,k} = w_{r,k} / \sum_{k=1}^{K} w_{r,k}$.
if ESS $<$ threshold, resample $\{\beta_{r,k}, W_{r,k}\}$ to obtain new particles denoted $\{\tilde{\beta}_{r,k}\}$, and set $w_{r,k} = 1$.
end for

Markov kernels $T_r(\tilde{\beta}_{r-1}, \cdot)$

- We propagate new particles $\{\beta_{r,k}\}$ via π_r-invariant MCMC moves, $\beta_{r,k} \sim T_r(\tilde{\beta}_{r-1,k}, \cdot)$.
- Sample from the full conditional distributions:
  - $\pi_r(\boldsymbol{\theta}_{\mathrm{DE}}|\cdot)$
  - $\pi_r(\boldsymbol{\theta}_y|\cdot)$
  - $\pi_r(\mathbf{c}_i|\cdot)$
  - $\pi_r(\lambda|\cdot)$
- If the full conditional distribution does not admit a closed form, we implement one step of the Metropolis-Hastings algorithm.

Backward Markov kernel

A convenient backward Markov kernel that allows an easy evaluation of the importance weight is

$$L_{r-1}(\beta_r, \tilde{\beta}_{r-1}) = \frac{\pi_r(\tilde{\beta}_{r-1}) T_r(\tilde{\beta}_{r-1}, \beta_r)}{\pi_r(\beta_r)}.$$

With this backward kernel, the incremental importance weight becomes

$$\begin{aligned} w_r &= w_{r-1} \cdot \frac{\gamma_r(\beta_r)}{\gamma_{r-1}(\tilde{\beta}_{r-1})} \cdot \frac{L_{r-1}(\beta_r, \tilde{\beta}_{r-1})}{T_r(\tilde{\beta}_{r-1}, \beta_r)} \\ &= w_{r-1} \cdot \frac{\gamma_r(\tilde{\beta}_{r-1})}{\gamma_{r-1}(\tilde{\beta}_{r-1})} \\ &= w_{r-1} \cdot \left(\frac{L(\mathbf{y}|\tilde{\beta}_{r-1}) \pi_0(\tilde{\beta}_{r-1})}{\rho(\tilde{\beta}_{r-1})} \right)^{\alpha_r - \alpha_{r-1}} \end{aligned}$$

Choose $\alpha_r - \alpha_{r-1}$ adaptively

Conditional ESS (CESS) (Zhou, Johansen, and Aston (2016))

$$\text{CESS}_r = \frac{K(\sum_{k=1}^{K} W_{r-1,k} w_{r,k})^2}{\sum_{k=1}^{K} W_{r-1,k} (w_{r,k})^2},$$

which is a function of $\alpha_r - \alpha_{r-1}$.

The next annealing parameter α_r can be determined by solving $\text{CESS}_r = \phi_r$, a user specified threshold.
(Wang, Wang, and Bouchard-Côté (2020))

Outline

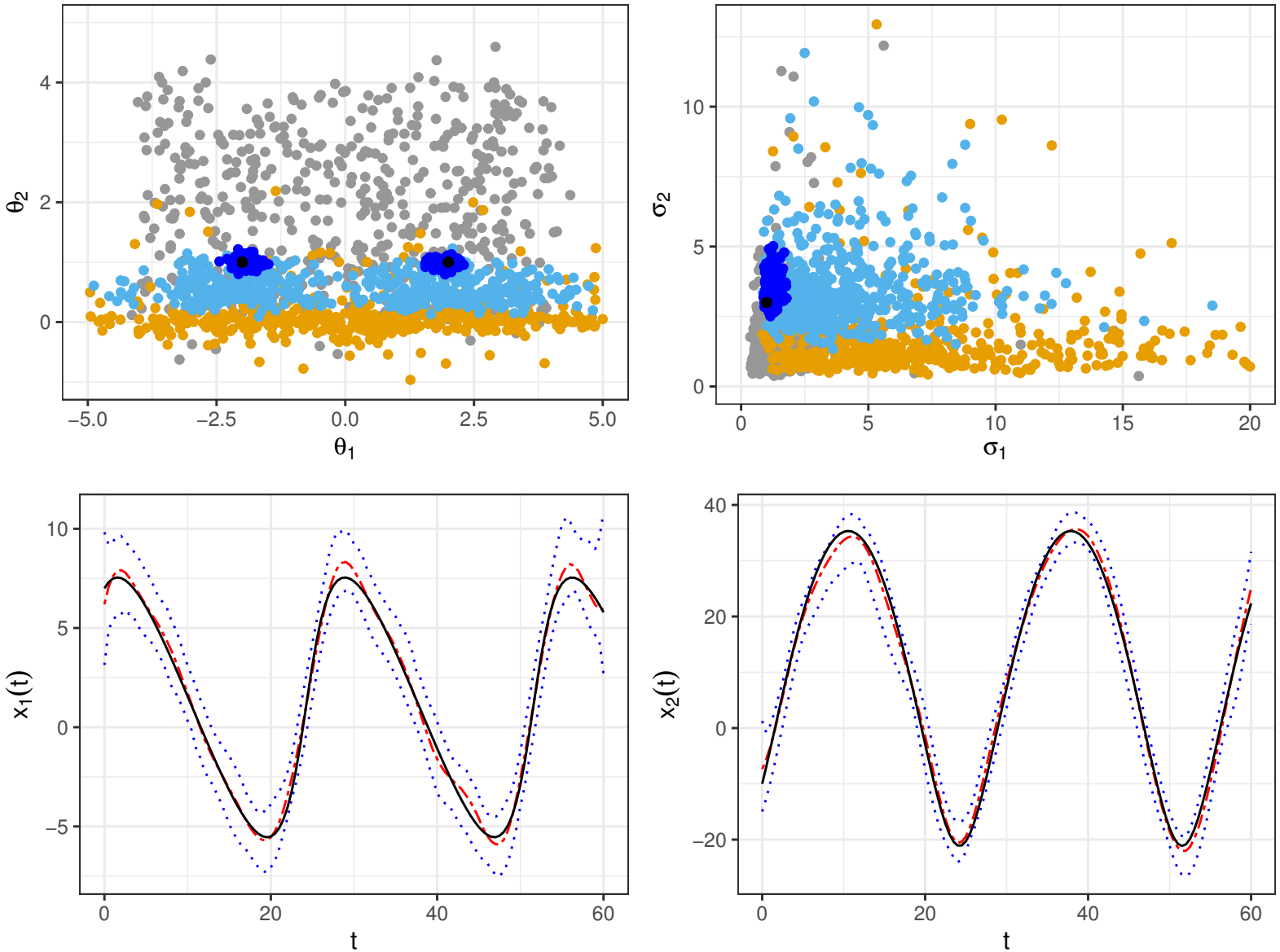
Simulation setting

- ODE

$$\begin{aligned}\frac{dx_1(t)}{dt} &= \frac{72}{36 + x_2(t)} - |\theta_1|,\\ \frac{dx_2(t)}{dt} &= \theta_2 x_1(t) - 1.\end{aligned}$$

- Observation model: $y_i \sim \texttt{Normal}(X(t_i|\boldsymbol{\theta}), \sigma^2)$.
- $|\theta_1| = 2$, $\theta_2 = 1$.
- $x_1(0) = 7$, and $x_2(0) = -10$.
- The standard deviation of the noise is set to $\sigma = 0.5$.

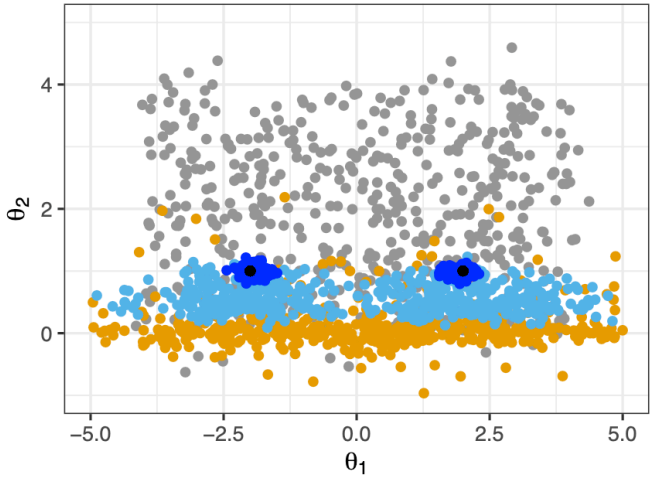
ASMC results



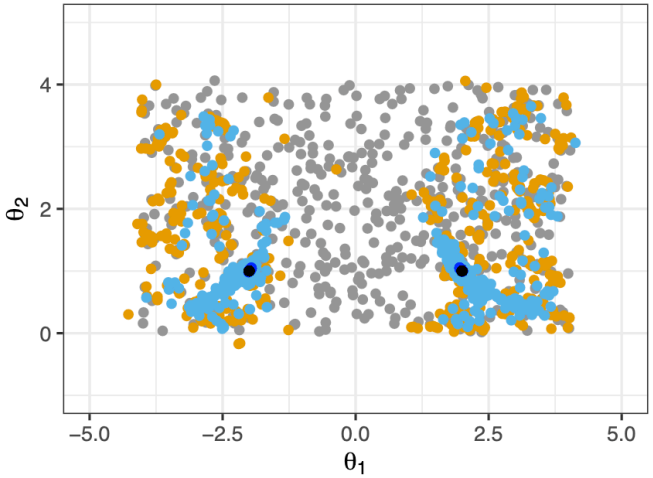
Comparison SMC-spline to alternative methods

	Target distribution	ODE solver	Algorithm
SMC-spline	$\pi(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda)$		ASMC
SMC-deSolve	$\pi(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{x}(0))$	"lsoda"	ASMC
MCMC-spline	$\pi(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{c}, \lambda)$		MCMC
MCMC-deSolve	$\pi(\boldsymbol{\theta}_{\mathrm{DE}}, \boldsymbol{\theta}_y, \mathbf{x}(0))$	"lsoda"	MCMC

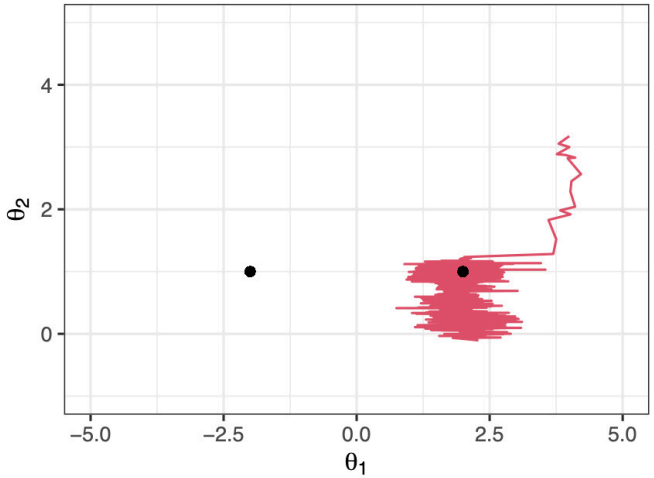
Comparison SMC-spline to alternative methods



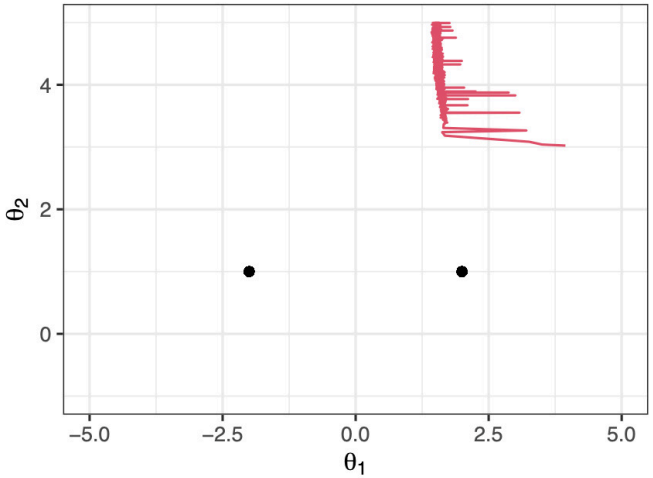
(a) SMC1



(b) SMC2



(c) MCMC1



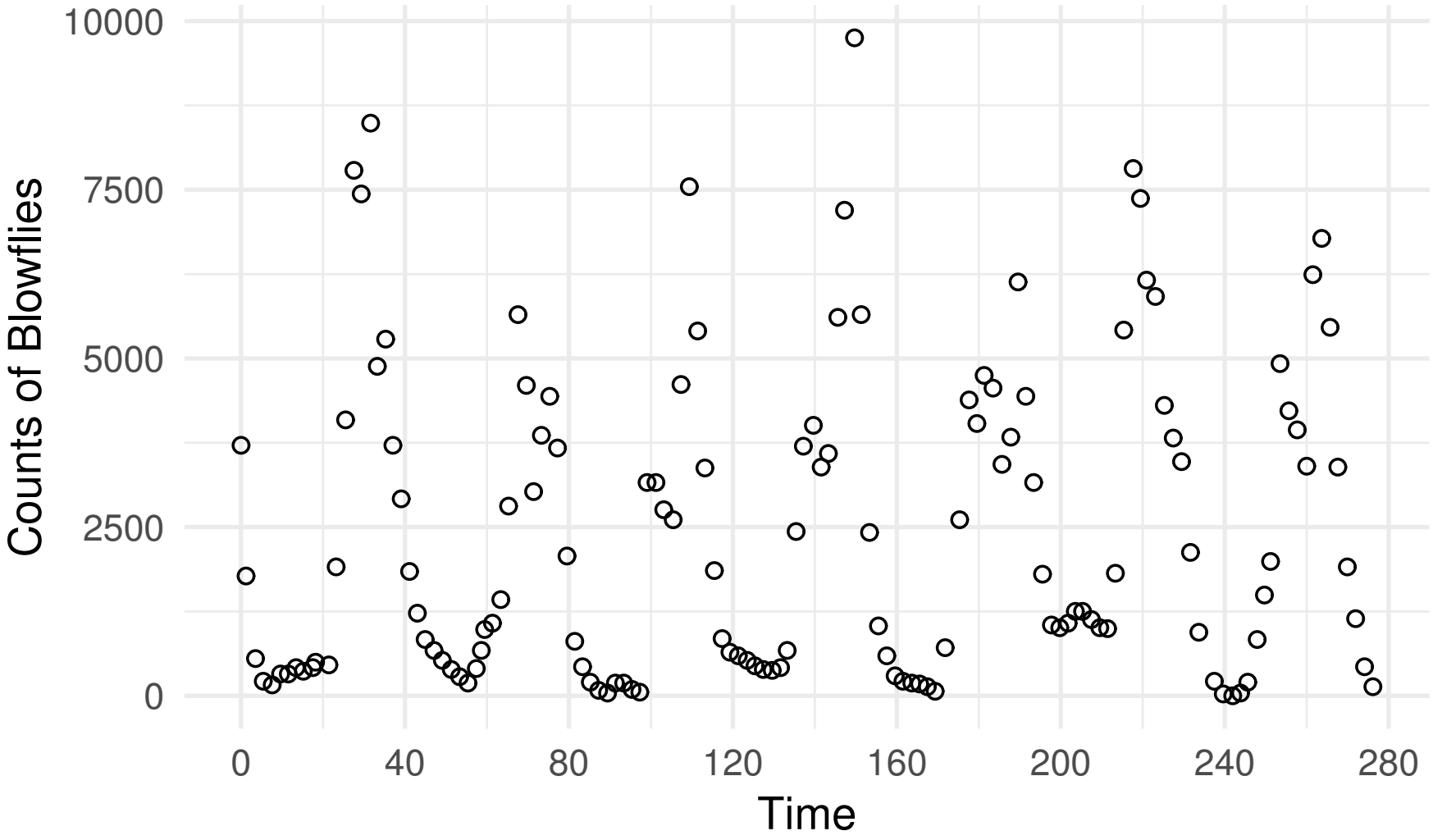
(d) MCMC2

Comparison SMC-spline to alternative methods

	True value	SMC-spline	SMC-deSolve	MCMC-spline	MCMC-deSolve
$\lvert\theta_1\rvert$	2	1.93 (1.68, 2.19)	1.84 (1.81, 1.88)	1.92 (1.48, 2.37)	5.37 (5.37, 5.37)
θ_2	1	0.99 (0.90, 1.09)	1.12 (1.08, 1.17)	0.97 (0.87, 1.08)	0.85 (0.85, 0.85)
$x_1(0)$	7	6.43 (2.56, 10.11)	3.94 (3.71, 4.22)	6.48 (4.01, 8.92)	4.55 (4.55, 4.55)
$x_2(0)$	-10	-10.66 (-17.86, -2.26)	-4.47 (-5.63, -3.23)	-10.28 (-14.18, -6.71)	0.68 (0.65, 0.70)

Outline

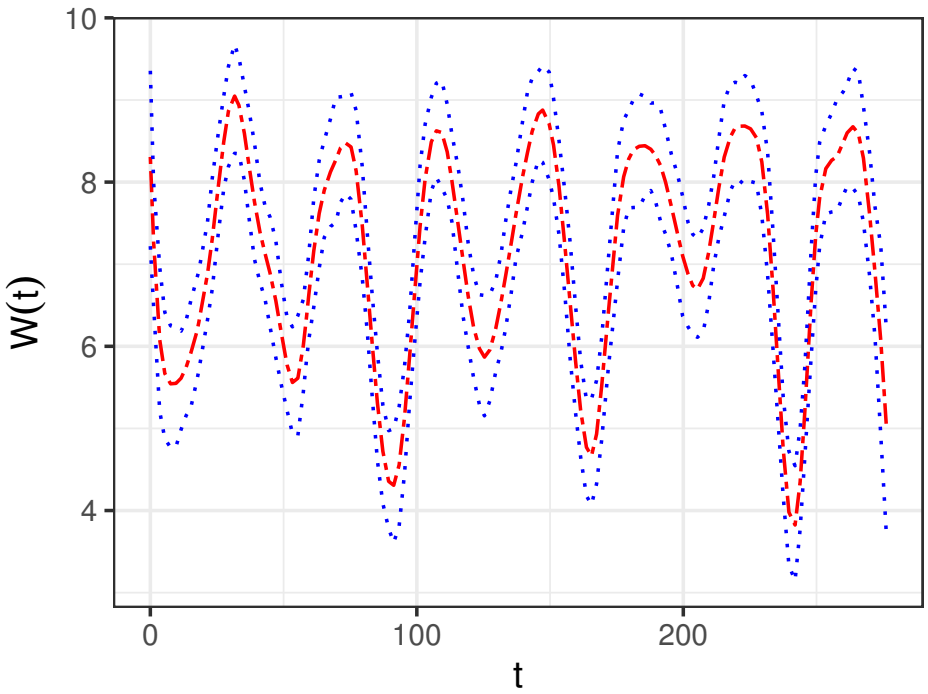
Blowflies

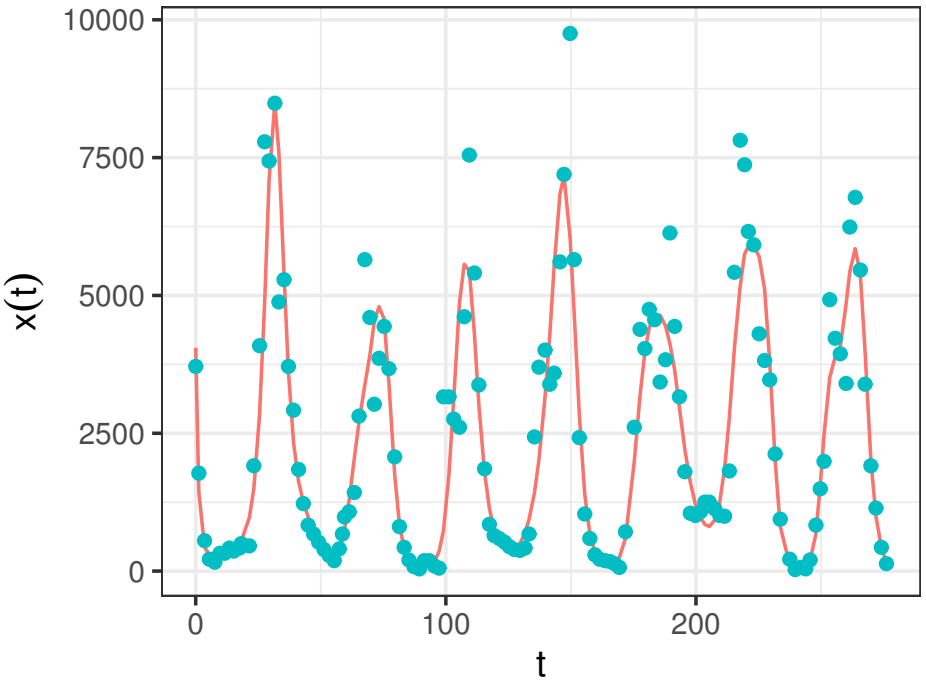

10000
7500
5000
2500
0
Counts of Blowflies
0
40
80
120
160
200
240
280
Time

DDE for Blowflies

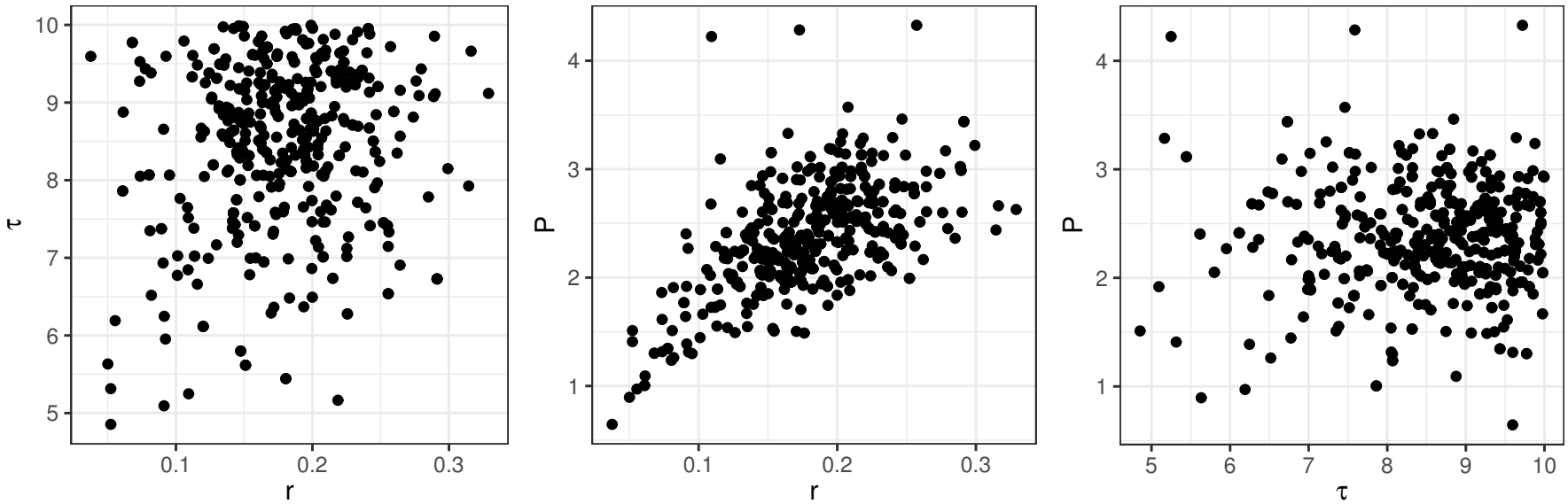
$$\frac{dX(t)}{dt} = rX(t)\left(1 - \frac{X(t-\tau)}{1000 \cdot P}\right)$$

- $X(t)$ is the blowfly population
- r is the rate of increase of the blowfly population
- P is a resource limitation parameter set by the supply of food
- τ is the time delay, roughly equal to the time for a larva to grow up to an adult





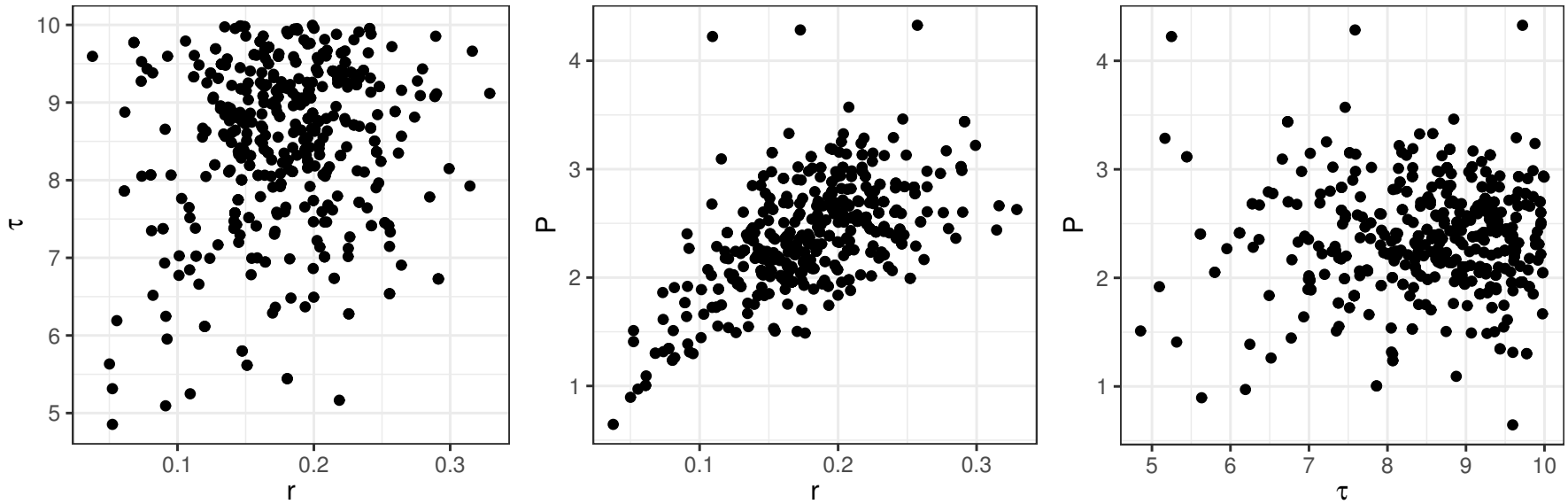
Estimated posterior mean trajectory and 95% credible bands for the delay differential equation modelling the population dynamics of blowflies. Here $X(t) = e^{W(t)}$.



Posterior samples of r, τ, P estimated via SMC. Correlation: $\texttt{corr}(r, \tau) = 0.139$, $\texttt{corr}(r, P) = 0.598$, $\texttt{corr}(P, \tau) = 0.008$.

Estimate for τ

- Annealed SMC:
  - Posterior mean and its 95% CI : 8.368 (5.656, 9.916)
- MLE in Wang and Cao (2012).
  - $\hat{\tau} = 8.78$, standard error 0.039



Posterior samples of r, τ, P estimated via SMC. Correlation: $\texttt{corr}(r, \tau) = 0.139$, $\texttt{corr}(r, P) = 0.598$, $\texttt{corr}(P, \tau) = 0.008$.

Estimate for τ

- Annealed SMC:
  - Posterior mean and its 95% CI : 8.368 $(5.656, 9.916)$
- MLE in Wang and Cao (2012).
  - $\hat{\tau}$ = 8.78, standard error 0.039

Outline

Summary

- Many statistical challenges for estimating ODE/DDE parameters.
- Our Bayesian semiparametric method via SMC has several advantages:
  - No need to solve DEs numerically
  - No need to specify DDE history
  - Provides uncertainty estimation for the parameters
  - Shows advantages over MCMC methods

Acknowledgements

Shijia Wang (Nankai University)

The Natural Sciences and Engineering Research Council of Canada (NSERC).

Thank you!

Bibliography

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B 68*(3), 411–436.

Del Moral, P., A. Doucet, and A. Jasra (2007). Sequential Monte Carlo for Bayesian computation. *Bayesian Statistics 8*, 1–34.

Wang, L. and J. Cao (2012). Estimating parameters in delay differential equation models. *Journal of Agricultural, Biological, and Environmental Statistics 17*(1), 68–83.

Wang, L., S. Wang, and A. Bouchard-Côté (2020). An annealed sequential Monte Carlo method for Bayesian phylogenetics. *Systematic Biology 69*(1), 155–183.

Zhou, Y., A. M. Johansen, and J. A. Aston (2016). Toward automatic model comparison: An adaptive sequential monte carlo approach. *Journal of Computational and Graphical Statistics 25*(3), 701–726.