UNIVERSITY OF OXFORD
Mathematical Institute

Parameter identifiability and model selection for PDE models of cell invasion

YUE LIU
Supervisors: Ruth Baker, Philip Maini
Mathematical Institute
University of Oxford

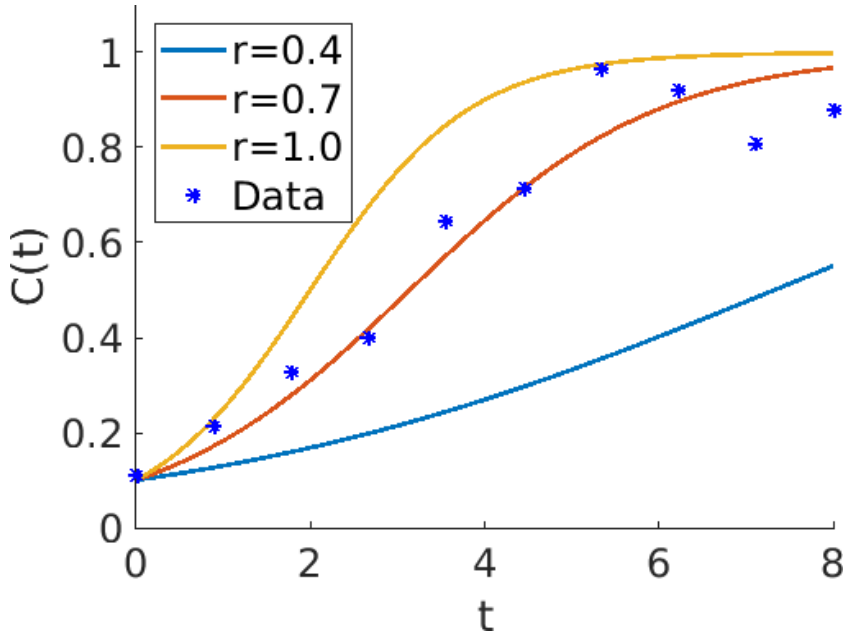
European Conference on Mathematical and Theoretical Biology
Heidelberg, 2022

Oxford
Mathematics

We all have done model fitting...

Example: what's the best value of r?

$$\frac{\partial C}{\partial t} = rC(1 - C)$$



... but how sure are we?

- Structural identifiability: can the true parameter be recovered, given theoretically *infinite* amount of data? (well-defined, objective property of the model itself)
- Practical identifiability: can the true parameter be recovered, given a *finite, realistic* amount of data? (subjective, depending on both model and available data)

... but how sure are we?

- Structural identifiability: can the true parameter be recovered, given theoretically *infinite* amount of data? (well-defined, objective property of the model itself)
- Practical identifiability: can the true parameter be recovered, given a *finite, realistic* amount of data? (subjective, depending on both model and available data)

How do we choose between models?

What are we looking for in a good model?

1. Ability to reproduce data (goodness of fit)
2. Simplicity (fewer free parameters)
3. Parameter identifiability

How do we choose between models?

What are we looking for in a good model?

1. Ability to reproduce data (goodness of fit)
2. Simplicity (fewer free parameters)
3. Parameter identifiability

How do we choose between models?

What are we looking for in a good model?

1. Ability to reproduce data (goodness of fit)
2. Simplicity (fewer free parameters)
3. Parameter identifiability

How do we choose between models?

UNIVERSITY OF OXFORD
Mathematical Institute

What are we looking for in a good model?

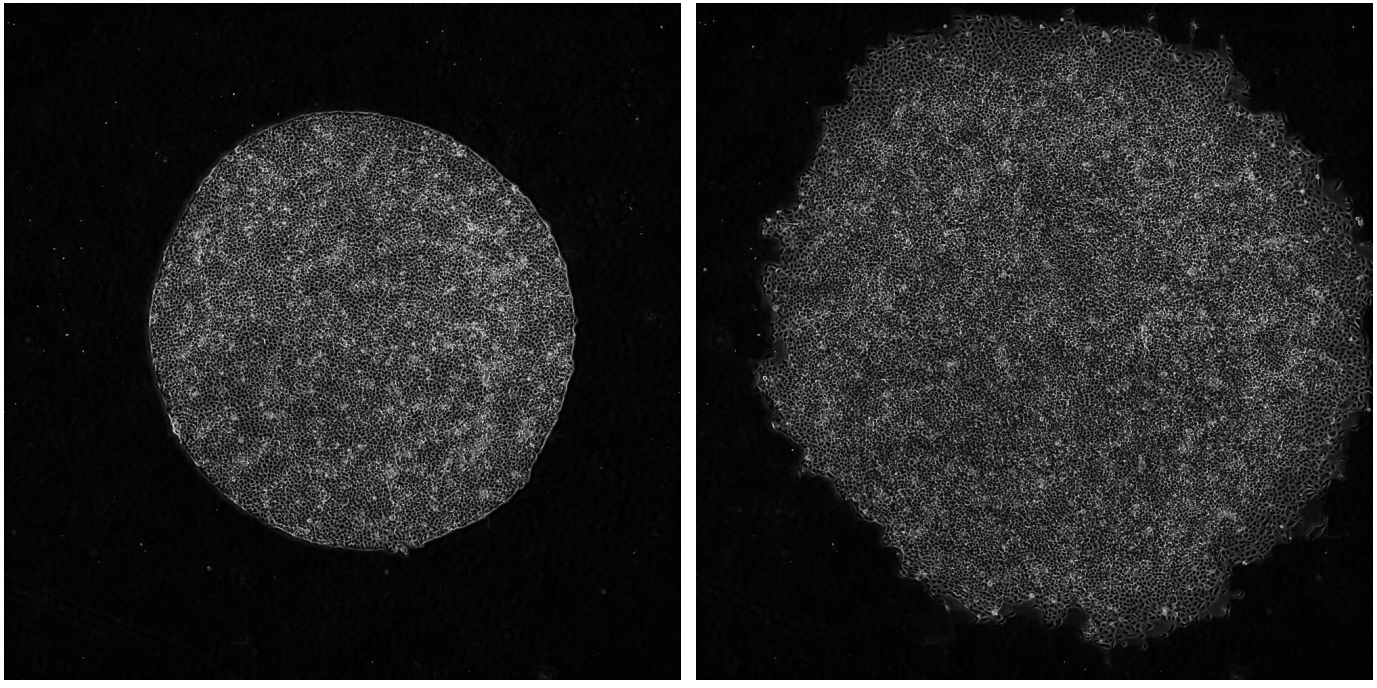
1. Ability to reproduce data (goodness of fit)
2. Simplicity (fewer free parameters)
3. Parameter identifiability

$$AIC = -2\log(p(C_{\text{data}}|\boldsymbol{\theta})) + 2m$$
$$BIC = -2\log(p(C_{\text{data}}|\boldsymbol{\theta})) + \log(N)m$$
(lower = better)

$m = \#\text{free params}$
$N = \#\text{data pts}$
$p(C_{\text{data}}|\boldsymbol{\theta}) = \text{likelihood}$

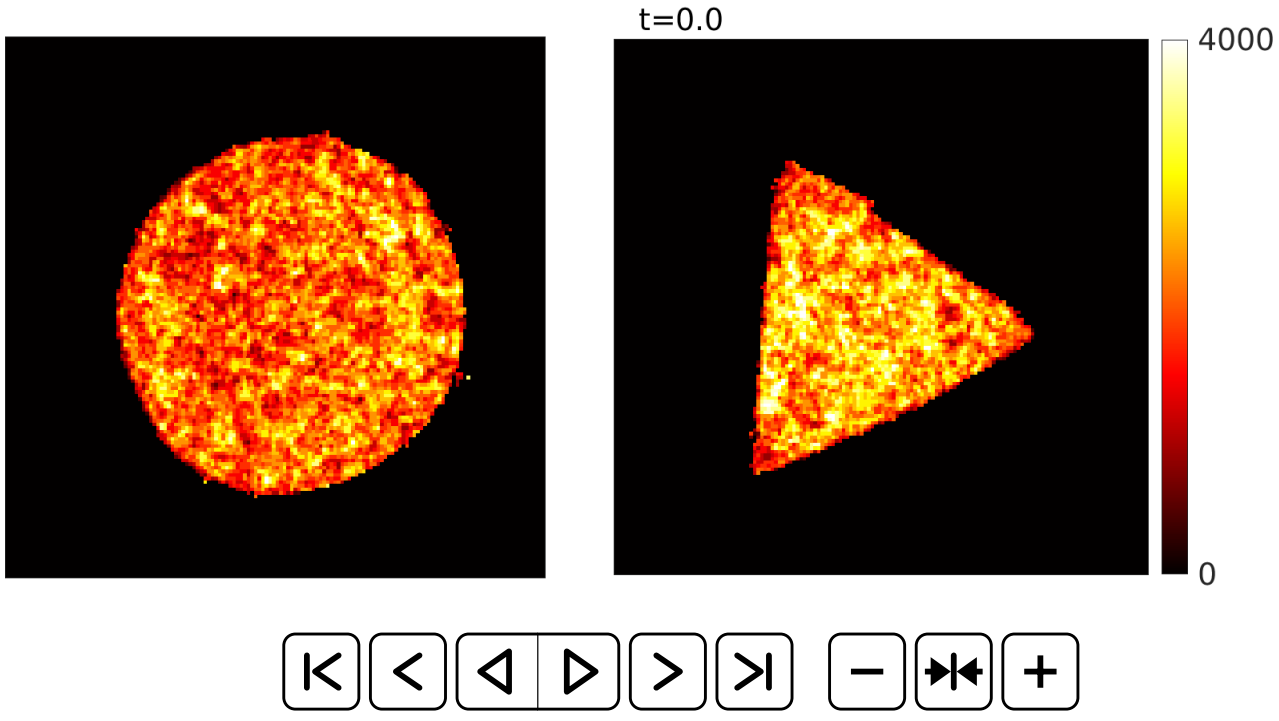
Cell invasion: an interesting problem

Barrier assay experiments



8 experiments, $n_t = 77$ images, Δt =20 min, $n_x = n_y = 150$

How to best model cell invasion?



Spatially-discretized cell density from 2 experiments with different initial conditions
(link if movie doesn't work)

What do we want to know?

- Are the models practically identifiable?
- How much data do we need to make the models identifiable?
- Which model is the “best”?
- Are parameter estimates consistent across experimental replicates?
- What’s a computationally efficient method to answer these?

What do we want to know?

- Are the models practically identifiable?
- How much data do we need to make the models identifiable?
- Which model is the “best”?
- Are parameter estimates consistent across experimental replicates?
- What’s a computationally efficient method to answer these?

What do we want to know?

- Are the models practically identifiable?
- How much data do we need to make the models identifiable?
- Which model is the “best”?
- Are parameter estimates consistent across experimental replicates?
- What’s a computationally efficient method to answer these?

What do we want to know?

- Are the models practically identifiable?
- How much data do we need to make the models identifiable?
- Which model is the “best”?
- Are parameter estimates consistent across experimental replicates?
- What’s a computationally efficient method to answer these?

What do we want to know?

- Are the models practically identifiable?
- How much data do we need to make the models identifiable?
- Which model is the “best”?
- Are parameter estimates consistent across experimental replicates?
- What’s a computationally efficient method to answer these?

Mathematical Model

UNIVERSITY OF OXFORD
Mathematical Institute

4 candidate models in the competition: $C = C(x, y, t)$

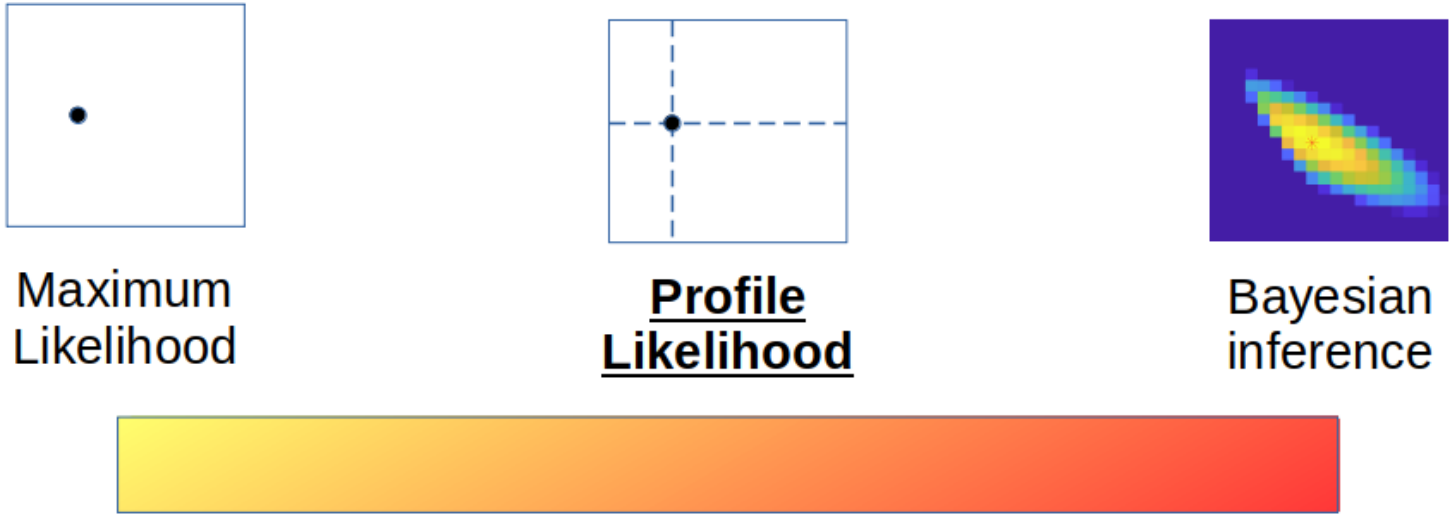
$$\text{Std. Fisher } (m = 3): \quad \frac{\partial C}{\partial t} = D_0 \nabla^2 C + rC(1 - C/K)$$

$$\text{Porous Fisher } (m = 4): \quad \frac{\partial C}{\partial t} = \nabla \cdot (D_0 (C/K)^{\eta} \nabla C) + rC(1 - C/K)$$

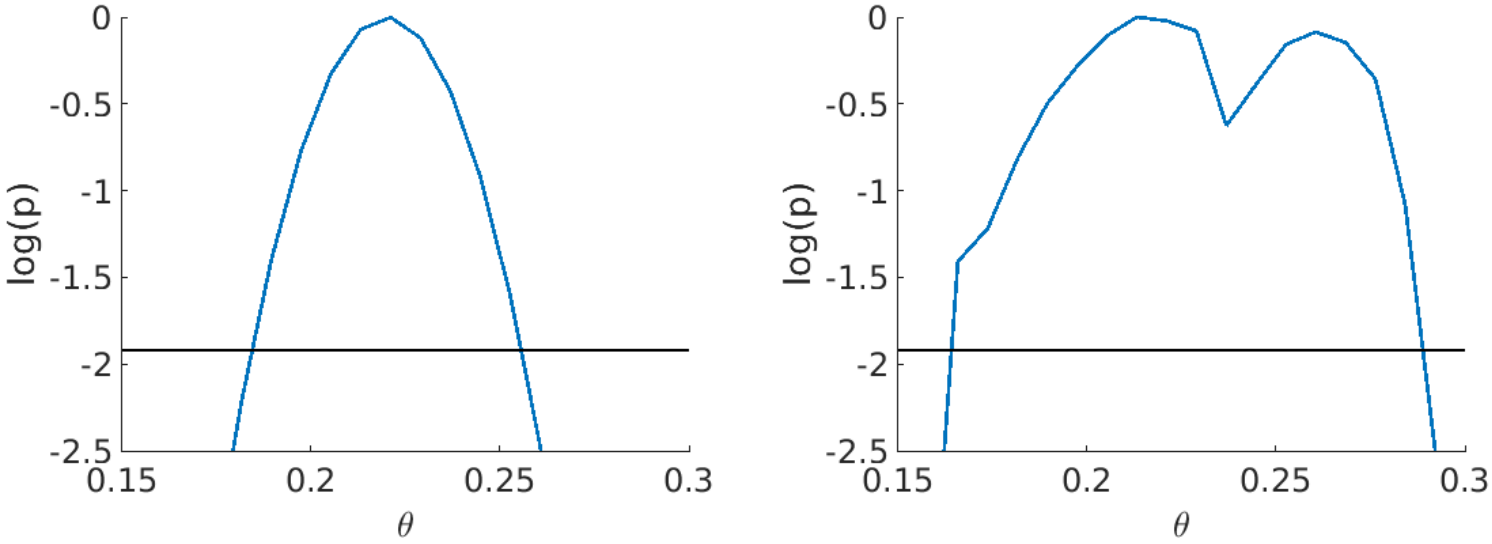
$$\text{Richards } (m = 4): \quad \frac{\partial C}{\partial t} = D_0 \nabla^2 C + rC(1 - (C/K)^{\gamma})$$

$$\text{Gen. Fisher } (m = 5): \quad \frac{\partial C}{\partial t} = D_0 \nabla^2 C + rC^{\alpha}(1 - C/K)^{\beta}$$

What method to use?



What does a profile likelihood curve looks like?



Line at $log(p) = -1.92$: cutoff for 95% confidence interval
Left: identifiable case: smooth, narrow, $\sim$ parabolic.
Right: non-identifiable case: broader, flat top, multimodal, jagged

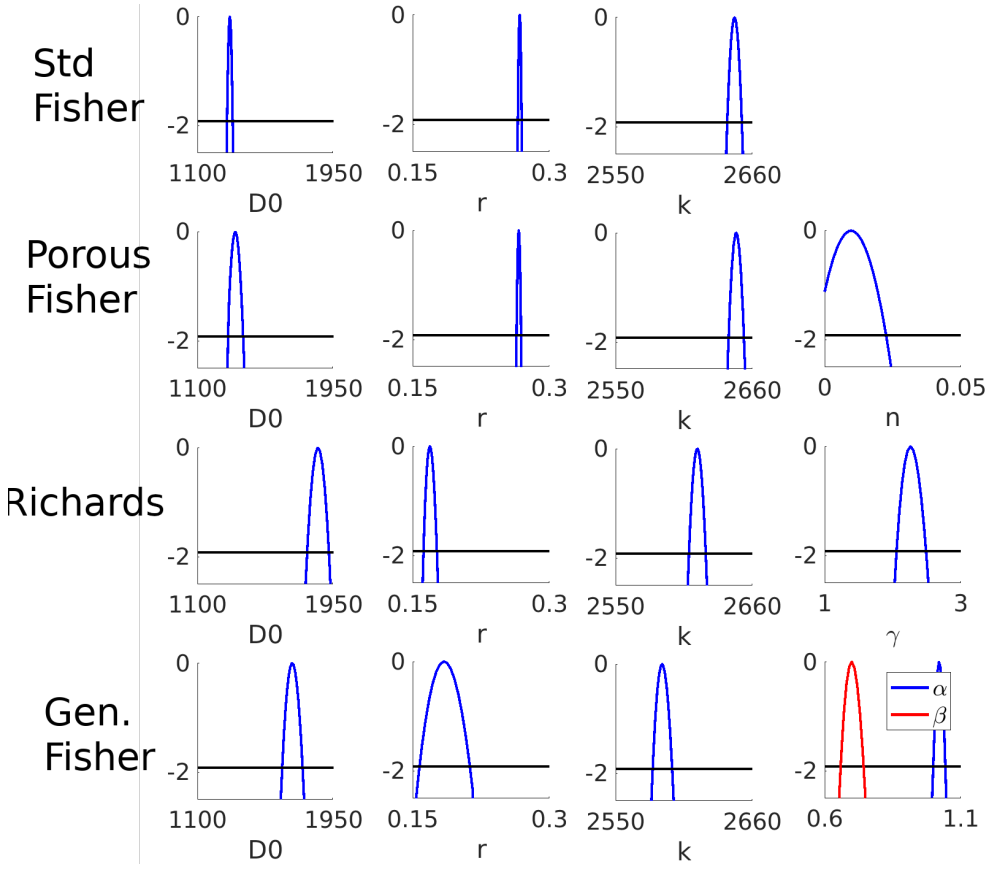
Are the models practically identifiable?

Yes (but only because we have high resolution data)

To emphasize: these are real, not synthetic, data

Are the models practically identifiable?

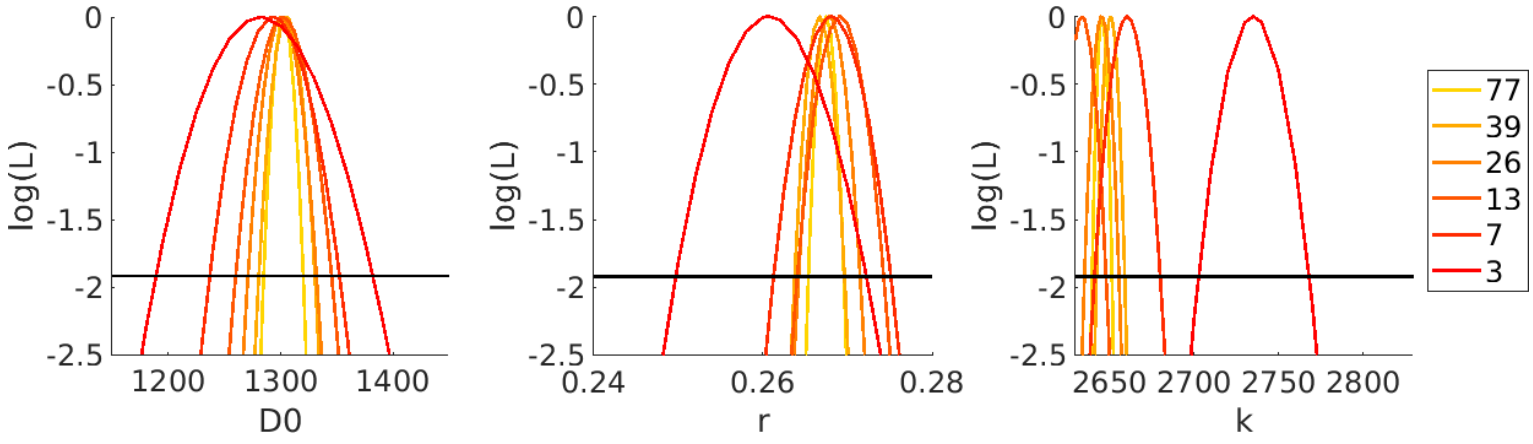
Yes (but only because we have high resolution data)



To emphasize: these are real, not synthetic, data

How much data do we need?

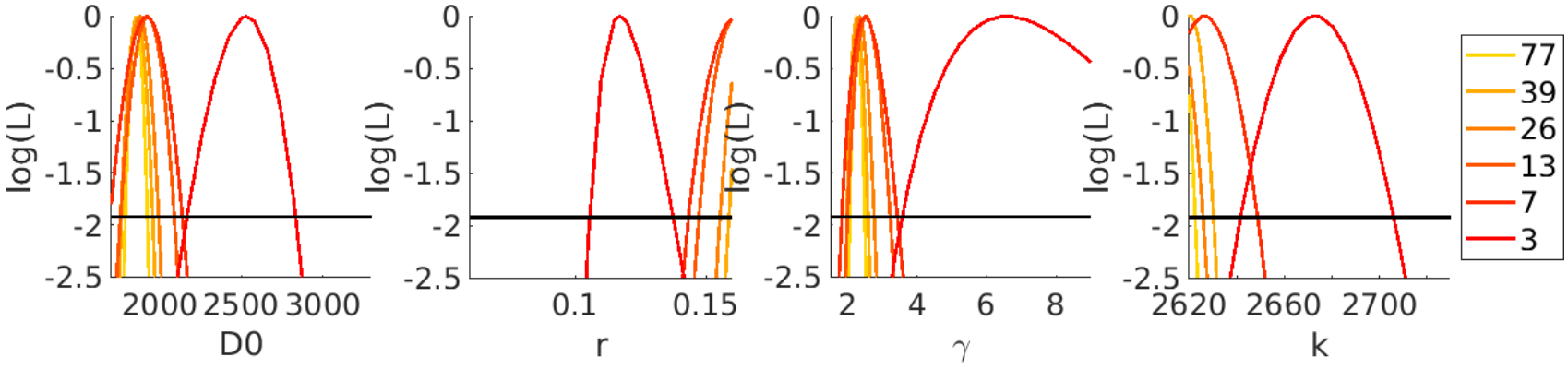
Repeat profile likelihood calculations with temporally down-sampled data (lower n_t/higher Δt)



The Standard Fisher model remains identifiable even when we down-sample the data to $n_t = 3$

How much data do we need?

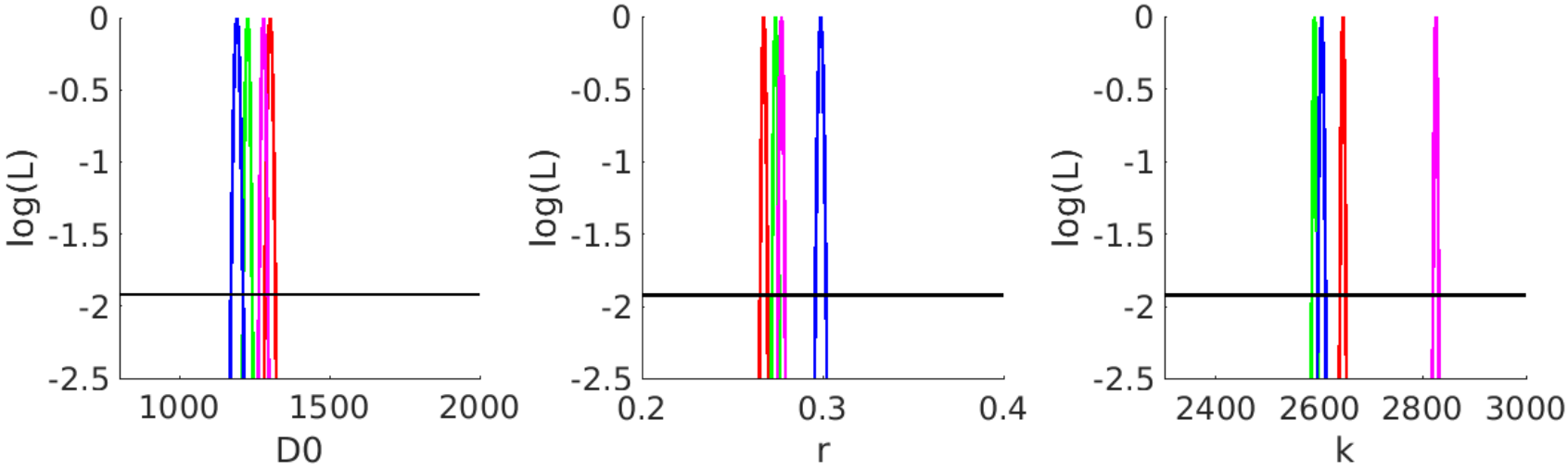
Repeat profile likelihood calculations with temporally down-sampled data (lower n_t/higher Δt)



The Richards model cease to be identifiable when the data resolution is sufficiently low

Are parameter estimates consistent across experimental replicates?

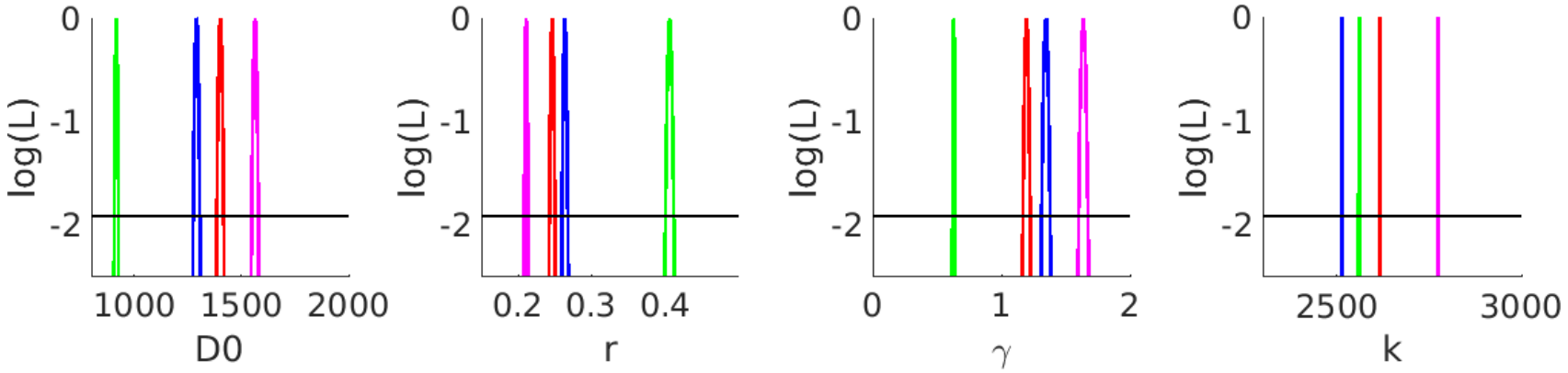
Depends on the model.



Std. Fisher (shown), and Porous Fisher: mostly consistent

Are parameter estimates consistent across experimental replicates?

Depends on the model.



Richards (shown), Generalised Fisher: much less consistent

Conclusions

- The 4 models (Std Fisher, Porous Fisher, Richards, Gen Fisher) are all identifiable given high resolution data
- Richards and Gen Fisher becomes non-identifiable if data resolution is low
- These two models also show inconsistencies across experiment replicates even if they appears identifiable
- **Inconsistency reflects sensitivity of the model to process noise, a symptom of non-identifiability**
- Experimental initial conditions can have a major impact on parameter identifiability
- Computational cost of profile likelihood compares favourably to MCMC, while still being informative

Conclusions

- The 4 models (Std Fisher, Porous Fisher, Richards, Gen Fisher) are all identifiable given high resolution data
- Richards and Gen Fisher becomes non-identifiable if data resolution is low
- These two models also show inconsistencies across experiment replicates even if they appears identifiable
- **Inconsistency reflects sensitivity of the model to process noise, a symptom of non-identifiability**
- Experimental initial conditions can have a major impact on parameter identifiability
- Computational cost of profile likelihood compares favourably to MCMC, while still being informative

Conclusions

- The 4 models (Std Fisher, Porous Fisher, Richards, Gen Fisher) are all identifiable given high resolution data
- Richards and Gen Fisher becomes non-identifiable if data resolution is low
- These two models also show inconsistencies across experiment replicates even if they appears identifiable
- **Inconsistency reflects sensitivity of the model to process noise, a symptom of non-identifiability**
- Experimental initial conditions can have a major impact on parameter identifiability
- Computational cost of profile likelihood compares favourably to MCMC, while still being informative

Conclusions

- The 4 models (Std Fisher, Porous Fisher, Richards, Gen Fisher) are all identifiable given high resolution data
- Richards and Gen Fisher becomes non-identifiable if data resolution is low
- These two models also show inconsistencies across experiment replicates even if they appears identifiable
- **Inconsistency reflects sensitivity of the model to process noise, a symptom of non-identifiability**
- Experimental initial conditions can have a major impact on parameter identifiability
- Computational cost of profile likelihood compares favourably to MCMC, while still being informative

Conclusions

- The 4 models (Std Fisher, Porous Fisher, Richards, Gen Fisher) are all identifiable given high resolution data
- Richards and Gen Fisher becomes non-identifiable if data resolution is low
- These two models also show inconsistencies across experiment replicates even if they appears identifiable
- **Inconsistency reflects sensitivity of the model to process noise, a symptom of non-identifiability**
- Experimental initial conditions can have a major impact on parameter identifiability
- Computational cost of profile likelihood compares favourably to MCMC, while still being informative

Acknowledgement

Prof. Ruth Baker
Oxford
(supervisor)

Prof. Philip Maini
Oxford
(supervisor)

Kevin Suh
Princeton
(experiments)

Prof. Daniel Cohen
Princeton
(Kevin's supervisor)

Wolfson Centre for Mathematical Biology
MATHEMATICAL INSTITUTE

NSERC CRSNG

Canadian Centennial Scholarship Fund

What are these methods?

$\boldsymbol{\theta}$: model parameters, $\boldsymbol{\theta}_{-i}$: parameters except θ_i

Maximum likelihood estimate (MLE):

$$\boldsymbol{\theta}^* = \operatorname*{argsup}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|C_{\text{data}})$$

Bayesian inference:

$$p(\boldsymbol{\theta}|C_{\text{data}}) \sim p(C_{\text{data}}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

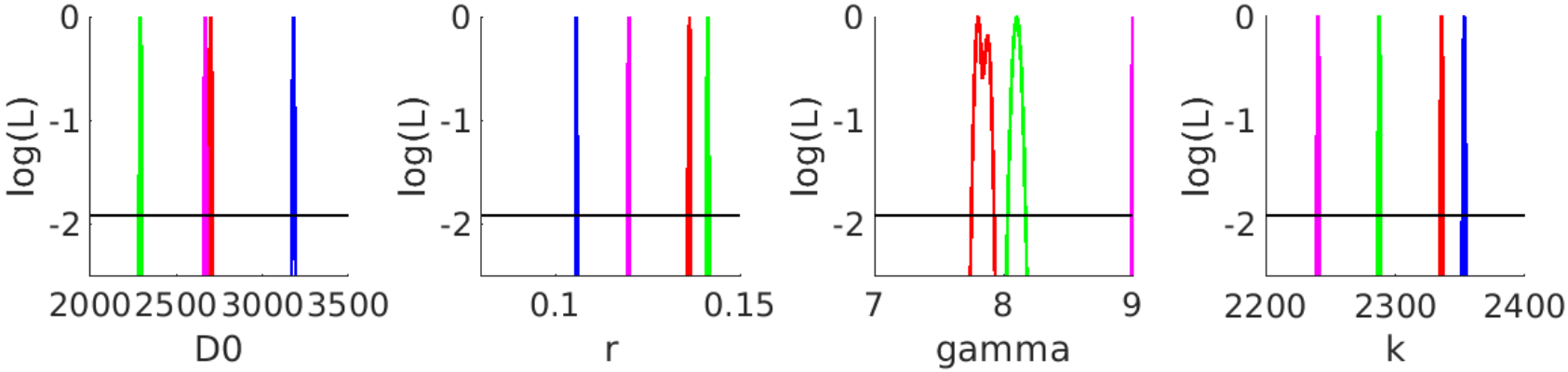
Profile likelihood:

$$p(\theta_i = \theta_i'|C_{\text{data}}) \sim \max_{\boldsymbol{\theta}_{-i}} p(C_{\text{data}}|\boldsymbol{\theta}_{-i}, \theta_i = \theta_i')$$

Result: Does the initial condition of the experiment have any impact?

Mathematical Institute

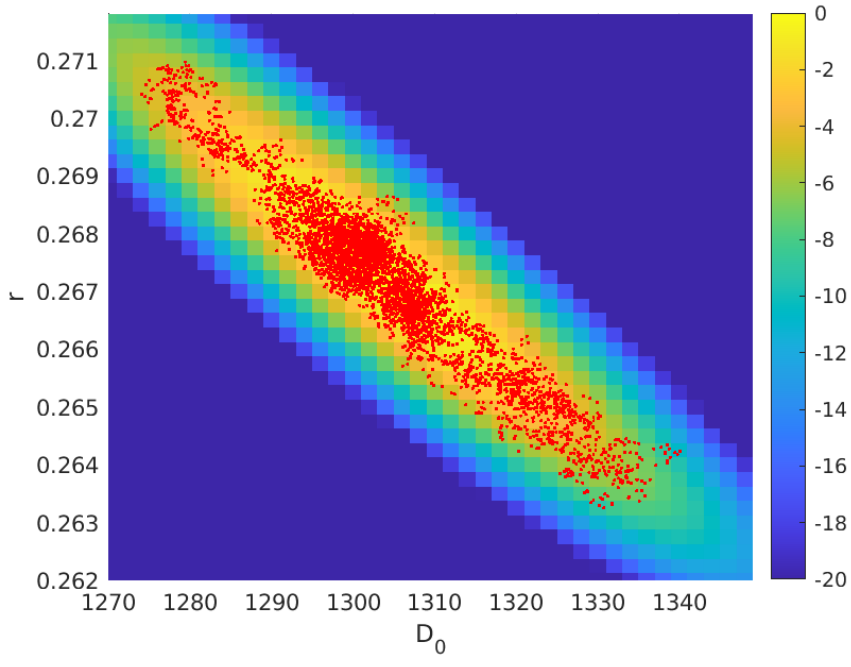
Yes



Triangular initial condition makes the parameter less consistent. Parameter estimates in Richards and Gen Fisher models can be far from the estimates with circular initial conditions

Result: Does profile likelihood agree with MCMC?

Yes



The samples generated from Metropolis-Hastings MCMC closely matches the contours of the two-parameter profile likelihood function

Computational cost of profile likelihood?

We measure computational cost by the number of model simulations required to compute the profile likelihood curves (all other costs negligible)
Total cost $\approx$ # free parameters * # points per curve (we chose 10) * average # model simulations needed for optimization

- 3 f.p. (Standard Fisher): 3 * 10 * (40 – 60) $\approx 1200 - 1800$
- 4 f.p. (Porous Fisher, Richards): 4 * 10 * (60 – 100) $\approx 2400 - 4000$
- 5 f.p. (Gen. Fisher): 5 * 10 * (140 – 250 $\approx 7000 - 12500$

Optimization may fail with off-the-shelf methods with $\geq$ 5 f.p.