Parameter identifiability for PDE models of cell invasion

Yue Liu[1*], Kevin Suh[2], Philip K. Maini[1], Daniel J. Cohen[2,3], Ruth E. Baker [1]

[1]Mathematical Institute, University of Oxford, [2]Department of Chemical and Biological Engineering, Princeton University, [3]Department of Mechanical and Aerospace Engineering, Princeton University

[*]yue.liu@maths.ox.ac.uk

Introduction

The Fisher-KPP model is one of the simplest partial differential equation models exhibiting travelling wave behaviour, and it has been widely used to model the growth and spread of populations in biology. When using the model to describe experimental data, it is often tempting to generalise the model with additional parameters to obtain a better fit. However, this increase in model complexity also increases the difficulty in obtaining accurate estimates of the parameter values, which is important for making predictions in a wider range of scenarios.

In this study, we use a profile likelihood approach to investigate parameter identifiability of extensions of the Fisher-KPP model on both simulated data, and experimental data from a cell invasion assay. We show that this approach allows us to determine which model is the most appropriate, which model best approximates the data, and the accuracy with which the parameter values in the model can be estimated. We also explore ways to design experiments to yield data more useful for parameter identification.

Experimental data

The experimental data are obtained from a barrier assay with MDCK cells. The assay is done over a period of 25 hours, with images taken once per 20 minutes. Initially, the cells are confined to a circular region in the middle of the domain by a barrier, which is removed at $t = 0$. The images are analyzed using convolutional neural networks [1] to obtain a density profile. Since the initial condition is radially symmetric, we can reduce the data to one dimension by computing the radially-averaged cell density.

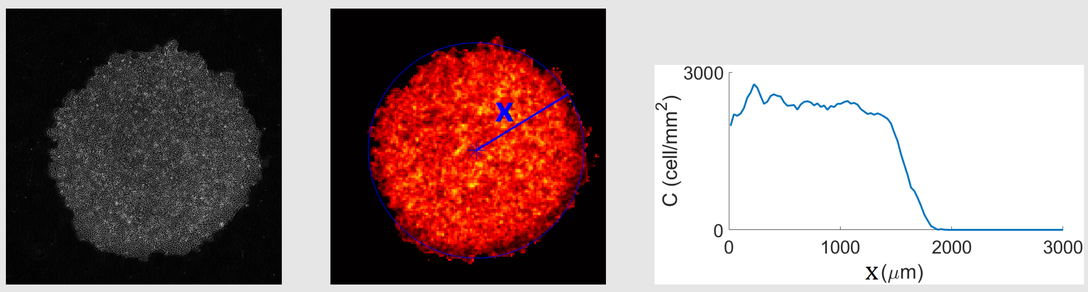

Figure 1. Left: Raw images of cells; Middle: Cell density profile in two dimensions; Right: Radially averaged cell density, where x is distance away from the centre.

Profile likelihood

Let $C_{\text{data}}(x,t)$ be either the cell density obtained from *in vitro* experiments, or *in silico* data generated by a known model with additive noise. Let $C_{\text{model}}(x,t;\boldsymbol{\theta})$ be the output from a proposed model for describing the data, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ are the free parameters in the model. Let p denote prior or posterior probability. The profile likelihood approach [2] is a compromise between maximum likelihood and Bayesian approaches, in terms of computational cost, and the amount of information it provides.

Maximum likelihood:

$$\boldsymbol{\theta}^* = \max_{\boldsymbol{\theta}} p(C_{\text{data}}|\boldsymbol{\theta}).$$

Bayesian inference:

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

Profile likelihood:

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

where $\boldsymbol{\theta_{-i}}$ denotes the parameters other than θ_i. The profile likelihood can be thought of as an approximation to the univariate marginal posterior distribution of parameter values, therefore it allows us to estimate confidence intervals in addition to the point estimates provided by maximum likelihood, while being computationally cheaper than Bayesian inference.

We assume that the experimental data can be written as a sum of the solution of the model, and a noise representing measurement error:

$$C_{\text{data}}(x_i, t_i) = C_{\text{model}}(x_i, t_i; \boldsymbol{\theta}) + \mathcal{N}(0, \sigma^2).$$

The log-likelihood function for parameters $\boldsymbol{\theta}$ is then

$$\begin{aligned} l(\boldsymbol{\theta}, \sigma) = \log p(C_{data}|\boldsymbol{\theta}, \sigma) &= \log \left[\prod_{\substack{1\le i\le n_x \\ 1\le j\le n_t}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[C_{\text{data}}(x_i,t_i) - C_{\text{model}}(x_i,t_i;\boldsymbol{\theta})]^2}{2\sigma^2} \right) \right] \\ &= -N\log(\sigma) - \frac{N}{2}\log(2\pi) - \frac{R}{2\sigma^2}, \end{aligned}$$

where $N = n_x n_t$ is the number of data points, and $R(\boldsymbol{\theta}) = \sum_{i,j}(C_{\text{data}}(x_i,t_j) - C_{\text{model}}(x_i,t_j))^2$, the squared residual. For each parameter θ_i, define the profile log-likelihood function

$$\begin{aligned} l_i(\theta_i') &= \max_{\boldsymbol{\theta_{-i}}, \sigma} l(\boldsymbol{\theta_{-i}}, \theta_i = \theta_i', \sigma), \\ &= -\frac{N}{2}\left[\log\left(\frac{2\pi R_{\min}}{N}\right) + 1\right], \\ \text{where } R_{\min} &= \min_{\boldsymbol{\theta_{-i}}} R(\boldsymbol{\theta_{-i}}, \theta_i = \theta_i'). \end{aligned}$$

Therefore, computing the profile likelihood amounts to finding the best-fitting model for the data by optimizing over all free parameters other than θ_i. The minimization of residual R is done using the interior point algorithm implemented with Matlab's *fmincon*.

The 95% confidence interval for parameter θ_i is the interval where $l_i(\theta_i') > -1.96$. We say a parameter can be better identified if the confidence interval is narrower.

Models

The model outputs C_{model} are generated by finite difference simulations of the following equation:

$$\frac{\partial C}{\partial t} = \nabla\cdot(D(C)\nabla C) + f(C), \quad f(C) = rC^\alpha\left[1 - \left(\frac{C}{K}\right)^\gamma\right]^\beta, \quad D(C) = D_0 C^n, \tag{1}$$

where $C(x,t)$ is cell density, $D(C) \ge 0$ is the diffusion coefficient, the function $f(C)$ models net cell proliferation, and $K > 0$ is the carrying capacity, and $D_0, r, \alpha, \beta, \gamma > 0, n \ge 0$ are constant parameters. The specific models we consider are:

- Standard Fisher-KPP: free parameters D_0, r, k, fixed $\alpha = \beta = \gamma = 1, n = 0$;
- Generalized porous Fisher: free parameters D_0, r, n, k, fixed $\alpha = \beta = \gamma = 1$; (standard porous Fisher takes $n = 1$);
- Richards growth: free parameters D_0, r, γ, k, fixed $\alpha = \beta = 1, n = 0$;
- Generalized Fisher: free parameters D_0, r, α, β, k, fixed $\gamma = 1, n = 0$;

Results

We present the profile log-likelihood functions, using the cell density from *in vitro* experiment as C_{data}.

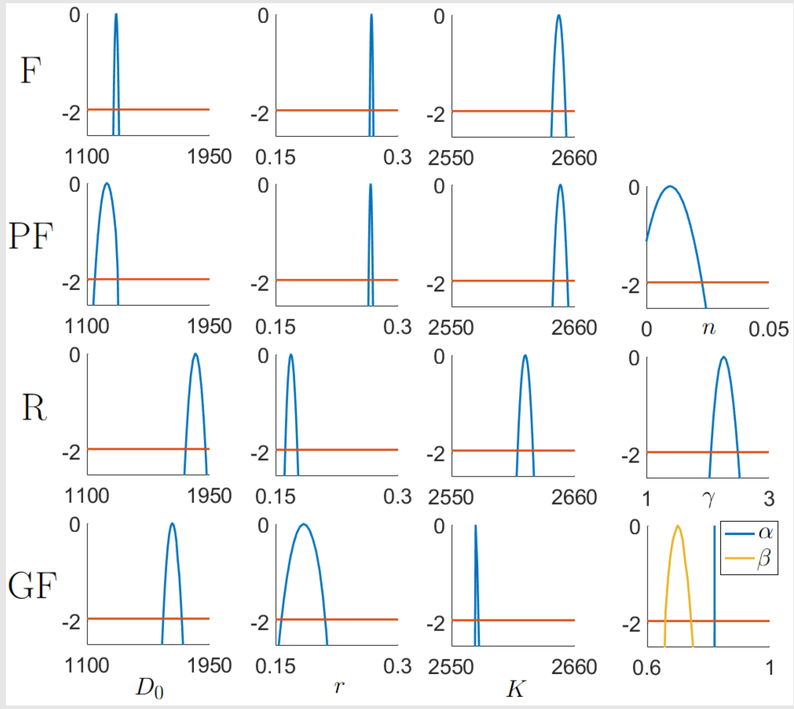

Figure 2. Profile likelihood function for each parameter under the four different models. F - Standard Fisher-KPP; PF - Generalized porous Fisher; R - Richards growth; GF - Generalized Fisher. The estimated confidence interval for each parameter is bounded by the points where the log-likelihood curve (blue) intersect with the constant line at -1.96.

Model	AIC	BIC
F	213821.313	213842.706
PF	213821.064	213849.589
R	213531.509	213560.034
GF	213595.768	213631.425

Table 1. AIC and BIC of the maximum likelihood parameter estimator for each model, with the same set of experimental data.

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are widely used as guides to model selection. They are defined as:

$$\begin{aligned} AIC &= -2l(\boldsymbol{\theta}^*, \sigma^*) + 2m, \\ BIC &= -2l(\boldsymbol{\theta}^*, \sigma^*) + \log(N)m, \end{aligned}$$

where $\boldsymbol{\theta}^*, \sigma^*$ are the maximum likelihood estimates, N the number of data points, and m the number of free parameters. Both criteria seek to find a maximum likelihood estimator while penalising the model for having more free parameters to avoid over-fitting. A lower score indicates a better model. Our main findings are:

- increasing the number of free parameters in general widens the confidence interval of each of the parameters, thus reducing parameter identifiability;
- decreasing the amount of data points available (eg. by down-sampling the data) reduces parameter identifiability;
- the peak and the confidence intervals of the parameters shared by all models, (D_0, r, K), differ depending on model.

Discussions

- The profile likelihood method provides a computationally efficient way to explore parameter identifiability
- In the context of cell invasion, where a number of models are capable of replicating the qualitative behaviour and are practically identifiable, the use of information criterion can be used to select a model
- Out of the four specific models we considered, the Richards growth model has the lowest AIC and BIC scores, therefore it is considered a better fit to data after penalising for over-fitting. However, its parameters are less identifiable compared to other models with a similar number of free parameters.
- Our results highlight that parameter interpretations and estimates are model dependent, therefore we must be cautious when taking parameter values estimates from one model and using them in related models.

References

J. LaChance and D. J. Cohen. "Practical fluorescence reconstruction microscopy for large samples and low-magnification imaging". In: *PLOS Computational Biology* 16 (2020), e1008443.

M. J. Simpson, A. P. Browning, D. J. Warne, O. J. Maclaren, and R. E. Baker. "Parameter identifiability and model selection for sigmoid population growth models". In: *Journal of Theoretical Biology* 535 (2022), p. 110998.

Acknowledgement

YL acknowledges support by the Natural Sciences and Engineering Research Council of Canada and the Canadian Centennial Scholarship Fund.