Variance vs. mean: distinguishing heterogeneity from noise

Xuefei Wang
supervised by Dr. Alexander Browning and Dr. Age Zanca
The University of Melbourne

Introduction

- **Context:** Standard Ordinary Differential Equation (ODE) models often neglect biological heterogeneity, treating population variability as simple measurement noise.
- **Method:** We assess statistical identifiability by comparing a Random Parameter (RP) Model against a Fixed Parameter (FP) Model.
- **Result:** Distinguishing between RP and FP models is unachievable without sufficient data.

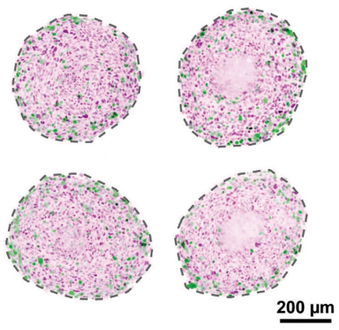

Fig1. *Heterogeneity in tumour spheroids grown from melanoma cell line WM983b (Browning et al., 2022, p. 3)*

Methods

1. Model Formulation

We focus on a standard logistic growth model,

$$\frac{dr}{dt} = \frac{\lambda}{3} r\left(1 - \frac{r}{R}\right),\ r(0) = r_0 \qquad (1)$$

where the model parameters $\theta = (\lambda, R, r_0)$ represent the "intrinsic growth rate", "carrying capacity", and "initial spheroid radius" (Browning et al., 2022, p. 10).

2. The Random Parameter Model

To incorporate biological heterogeneity, we assume:

$$\lambda \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2)$$
$$R \sim \mathcal{N}(\mu_R, \sigma_R^2)$$
$$r_0 \sim \mathcal{N}(\mu_{r_0}, \sigma_{r_0}^2);$$

and they are independent.

Then, we utilise the Surrogate Likelihood method which approximates the distribution of the solution to the ODE model $r(t;\theta)$ using a shifted gamma distribution (Browning et al., 2022, p. 7, 8).

3. The Equivalent Fixed Parameter Model

Our first experiment is to distinguish between heterogeneity (RP model) and a simplified deterministic model (FP model) designed to share the exact same mean trajectory. To construct such FP model, we follow the steps:

- **Step 1:** We compute the mean trajectory of the stochastic approximation, denoted as $\mu(t) = \mathbb{E}[r(t)]$.
- **Step 2:** We define the FP solution $\tilde{r}(t)$ such that $\tilde{r}(t) = \mu(t)$. Consequently, their derivatives must also match: $\tilde{r}'(t) = \mu'(t)$.
- **Step 3:** We numerically evaluate pairs of $(\tilde{r}(t), \tilde{r}'(t))$ for a time sequence $t = 1, \ldots, 50$ to reconstruct the function $G(\tilde{r})$, a FP ODE $\frac{d\tilde{r}}{dt} = G(\tilde{r})$ that is indistinguishable from the random model in terms of mean growth. (See ***Fig2.*** for plot of $G(\tilde{r})$)

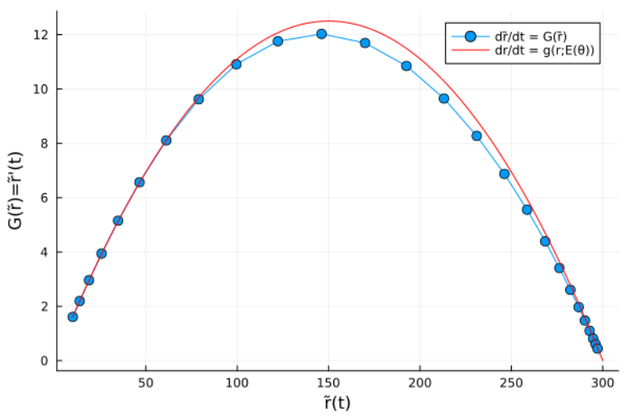

Fig2. *Blue curve is a plot of $G(\tilde{r})$; Red curve is plotted with parameters set to the exact mean ($\lambda_0 = \mathbb{E}(\lambda), R_0 = \mathbb{E}(R)$), and they are different since $G(\tilde{r}) = \mathbb{E}[g(r;\theta)] \neq g(r;\mathbb{E}[\theta])$ in general.*

Experiment A

1. Goal and Methodology

In this experiment, we simulate the case where our data has intrinsic heterogeneity and answer the question: can we distinguish heterogeneity from pure measurement noise if such data is provided? To address this, we follow the procedures:

- **Data Generation:** We generate synthetic data using the RP model overlaid with measurement noise:
  $$r^{obs}(t) = r(t;\theta_0) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \eta^2)$$
- **Model Fitting:** We fit the data back to the RP model using Maximum Likelihood Estimation (MLE) to infer 7 parameters: $(\mu_\lambda, \sigma_\lambda, \mu_R, \sigma_R, \mu_{r_0}, \sigma_{r_0}, \sigma_\epsilon = \eta)$. Using the inferred parameters, we construct the corresponding "Equivalent Fixed Parameter Model" which shares the exact same mean trajectory, and we only optimise for the noise parameter η.
- **Distinguishability Test:** We compute $\Delta AIC = AIC_{RP} - AIC_{FP}$ (AIC stands for Akaike information criterion) to evaluate distinguishable capability in different scenarios.
- **Scenarios:** We compare performance under two distinct conditions: Sufficient Data / Low Noise vs. Sparse Data / High Noise to assess the degradation of identifiability.

2. Result:

- **Sufficient Data (High Identifiability):** The AIC strongly favour the RP model ($\Delta AIC \approx -53.0$). In ***Fig4.*** The red confidence ribbon captures the data distribution accurately, confirming that the method successfully identifies that the heterogeneity exists.
- **Sparse Data (Loss of Identifiability):** The models become statistically indistinguishable (see ***Fig5.***, $\Delta AIC \approx 0.35$), which demonstrates that without sufficient data density, we cannot distinguish intrinsic biological heterogeneity from pure measurement noise.

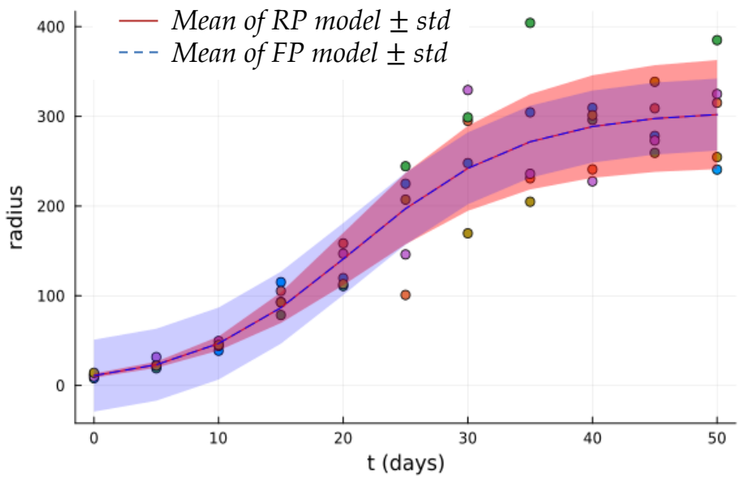

Fig3. *Comparison of RP and FP models with enough data and less noise. $|\Delta AIC| \approx 53.0 \gg 2$ (Statistically distinguishable). While the mean trajectories are identical, the variance structures differ: the blue ribbon reflects only constant measurement noise.*

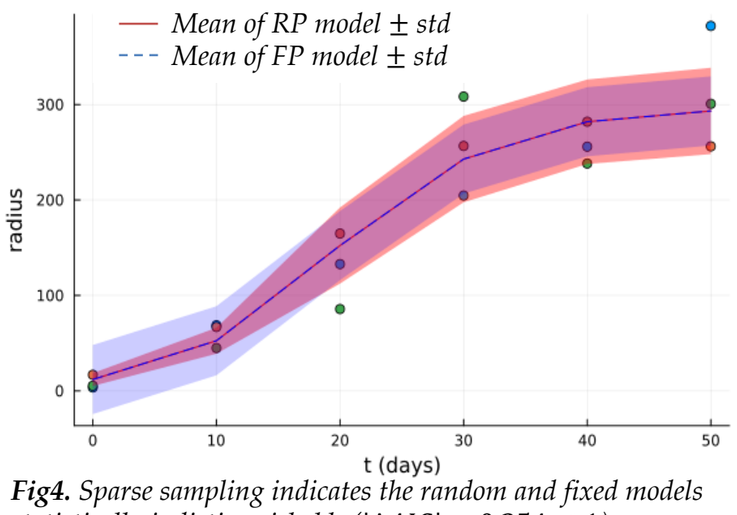

Fig4. *Sparse sampling indicates the random and fixed models statistically indistinguishable ($|\Delta AIC| \approx 0.354 < 1$)*

Experiment B

1. Goal and Methodology

In this experiment, we aim to distinguish model misspecification from intrinsic heterogeneity. We generated data from a completely unrelated FP model (with a distinct mean trajectory) and fit both the original RP model and this alternative FP model to assess whether they can be statistically distinguished. We employ the following methods:

- **Data Generation:** We generate synthetic data from an unrelated FP ODE model (different from equation ***(1)***) with measurement noise:
  $$\frac{dr}{dt} = \frac{\lambda}{3} r\left(1 - \left(\frac{r}{R}\right)^2\right),\ r(0) = r_0 \quad (2)$$
- **Distinguishability Test:** We evaluate the Difference in Log-Likelihood ($\Delta\ln(\mathcal{L})$) rather than AIC. This strictly assesses fit quality without the confounding influence of parameter penalties.

2. Result:

- **Indistinguishability**: Across all scenarios—regardless of noise level or data density—the two models remain statistically indistinguishable, with log-likelihood differences consistently negligible ($|\Delta\ln(\mathcal{L})| < 2$).
- **Implication:** We cannot distinguish model misspecification from heterogeneity even with sufficient data is given.

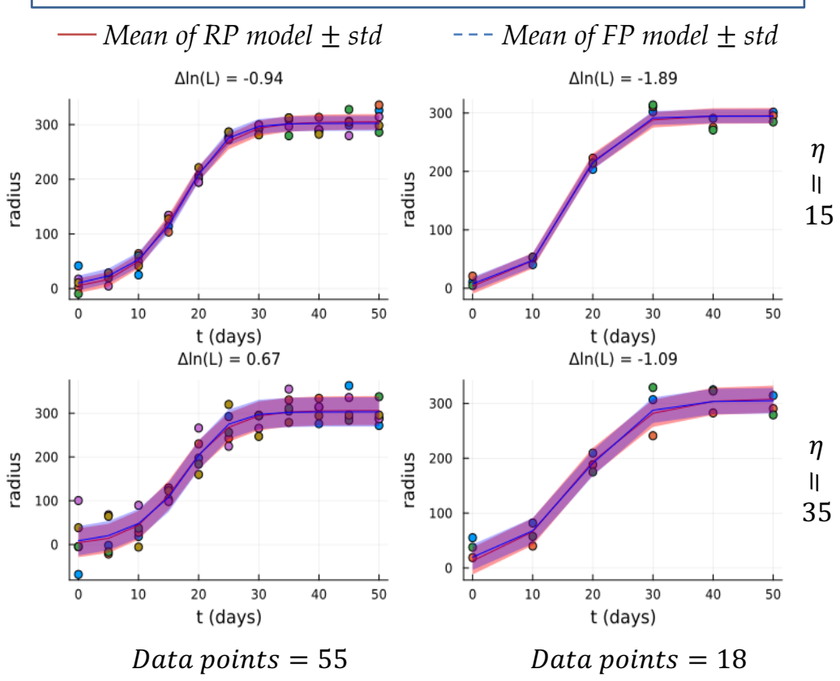

Fig5. *Comparison of RP model and FP model fits to deterministic data across varying noise levels (η) and data densities.*

Conclusion and Discussions

Conclusion:

- We successfully construct a FP model sharing the same mean as the RP model, demonstrating that variance structure, rather than the mean, is the critical factor for identifying biological heterogeneity.
- Although sufficient data allows for the identification of true heterogeneity, we find that it is often impossible to distinguish model misspecification from heterogeneity even under ideal conditions.
- Even with a known, simple ODE structure, correct model identification is not guaranteed. This suggests the necessity of extreme caution when applying any inference techniques if heterogeneity may be present.

Limitations and Future Work:

- Future research should extend validation to clinical measurements and utilise machine learning or non-parametric approaches to investigate scenarios where the model structure itself is unknown.

Acknowledgement:
I am deeply grateful to Dr. Alexander Browning and Dr. Age Zanca for their patient guidance and for introducing me to the world of research.

Reference:

Browning, A. P., Drovandi, C., Turner, I. W., Jenner, A. L., & Simpson, M. J. (2022). Efficient inference and identifiability analysis for differential equation models with random parameters. *PLOS Computational Biology*, *18*(11), e1010734. https://doi.org/10.1371/journal.pcbi.1010734