Robust Bayesian experimental design through flexible modelling structures

Speaker: James McGree (james.mcgree@qut.edu.au)

Collaborators: Najimuddin, N., Warne, D.J. and Thompson, H

Affiliation: School of Mathematical Sciences
Queensland University of Technology
Brisbane, Queensland, Australia

QUT

Design problem

Motivation

- With an Australian producer, we aim to increase our understanding of avocado growth in Australian conditions

- This is to ensure the production of high quality, robust fruit

- Indicator of quality: Dry matter (also indicator of flavour)

- Propose to undertake sampling throughout growing season

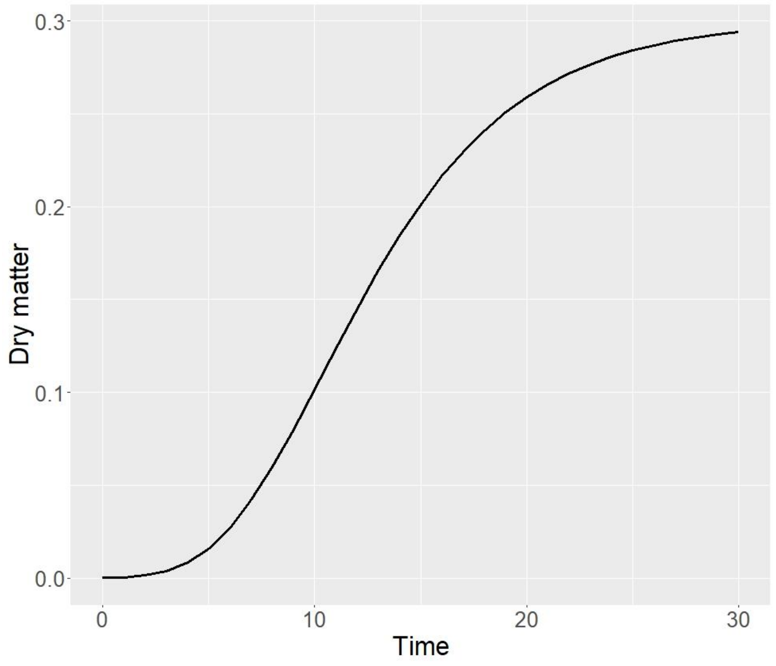
Problem

- Need an efficient sampling design. Methods available but depend on prior information

- No previous studies were found in the literature that described avocado growth in Australian conditions but found studies conducted overseas

- Can use this as a source of prior information but need to account for potential misspecification

Typical Bayesian design solution



- Seek prior information e.g. model and parameters

- Based on previous research, sigmoidal growth curves have been proposed to describe dry matter for avocados and many other fruit.

- E.g. Gompertz model:

$$\frac{dy}{dt} = r\, y \log\left(\frac{\lambda}{y}\right); \;\; y \sim \mathrm{N}(\mathrm{E}[y|t,r,\lambda], {\sigma_e}^2)$$

y – response (dry matter)

t – time

r – growth rate

λ – carrying capacity

$\theta = (r, \lambda, {\sigma_e}^2)$

- Propose a goal of data collection e.g. learn about model parameters

- Form an expected utility function then maximise through choice of design (here time points) e.g.

$$\begin{aligned} U(d) &= E_{y,\theta}[u(d,y,\theta)] \\ &= \int_Y \int_\Theta u(d,y,\theta)\, p(y,\theta \mid d)\, \mathrm{d}\theta\, \mathrm{d}y \\ &= \int_Y \int_\Theta u(d,y,\theta)\, p(y \mid \theta, d)\, p(\theta \mid d)\, \mathrm{d}\theta\, \mathrm{d}y \end{aligned}$$

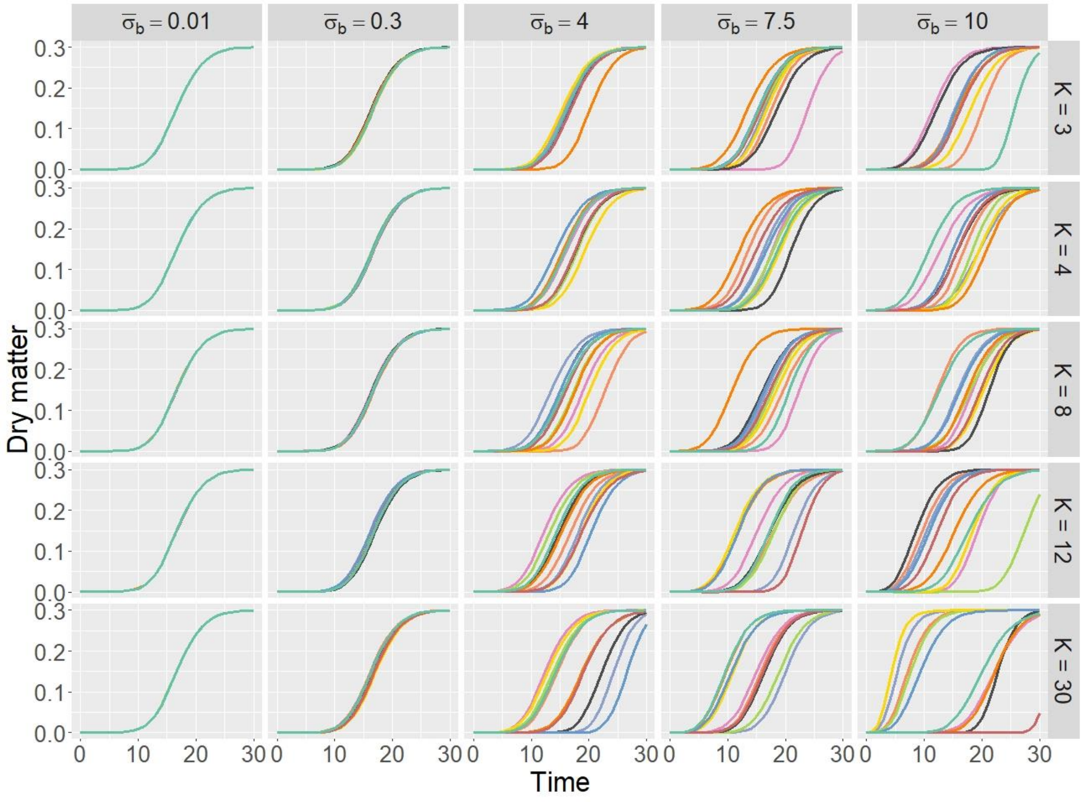
$$d^* = \arg\max_{\mathrm{d} \in \mathrm{D}} U(d)$$

Proposed solution

- Form designs based on flexible models
- Exploit flexibility to provide robust designs
- E.g. a flexible Gompertz model

$$\frac{dy}{dt} = ry\log\left(\frac{\lambda}{y}\right) B'(t)\ ;\ B'(t) = \left(\beta_0 + \beta_1 t + \sum_{k=1}^{K} b\left(t^k - \tau_k\right)\right)$$

K = no. of knots, b = additional parameters, τ = knots, $\left(t^k - \tau_k\right)$ = spline basis fn

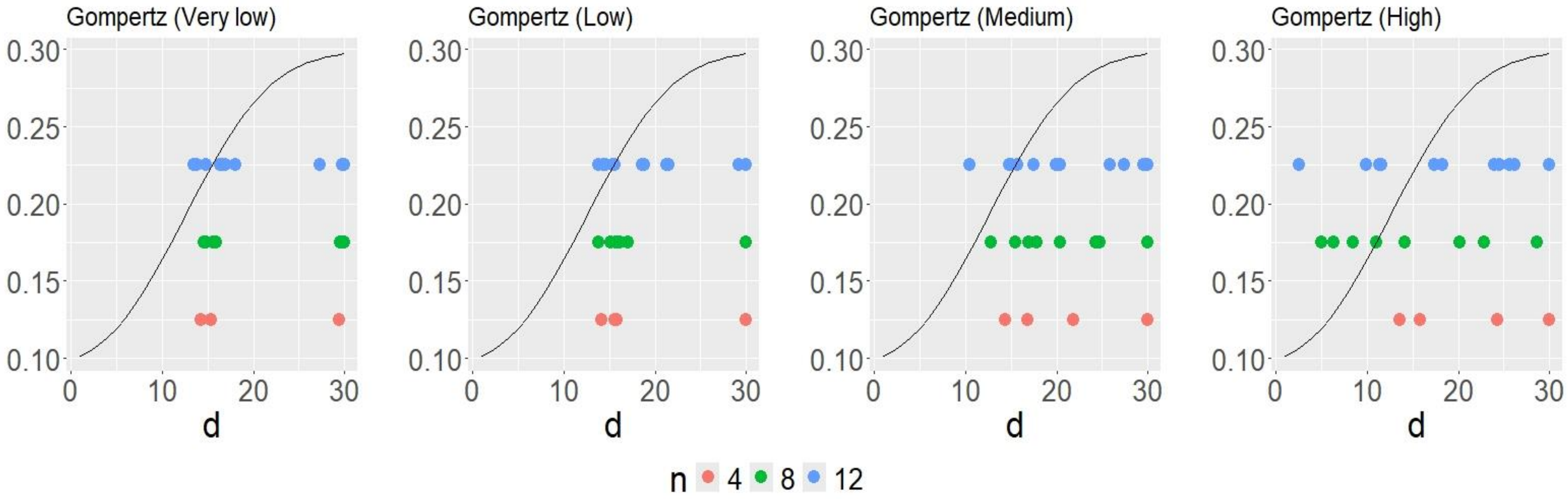


- Then, extend expected utility function:

$$\begin{aligned} U(d) &= E_{y,\theta,b}[u(d,y,\theta,b)] \\ &= \int_Y \int_\Theta \int_B u(d,y,\theta,b)\, p(\,y,\theta,b \mid d\,)\, \mathrm{d}b\, \mathrm{d}\theta\, \mathrm{d}y \\ &= \int_Y \int_\Theta \int_B u(d,y,\theta,b)\, p(\,y \mid \theta,b,d\,)\, p(\,\theta,b \mid d\,)\, \mathrm{d}b\, \mathrm{d}\theta\, \mathrm{d}y \end{aligned}$$

- Found designs under a range of flexible models i.e. “very low”, “low”, “medium” and “high”, and evaluated robustness properties of these designs.

Resulting designs



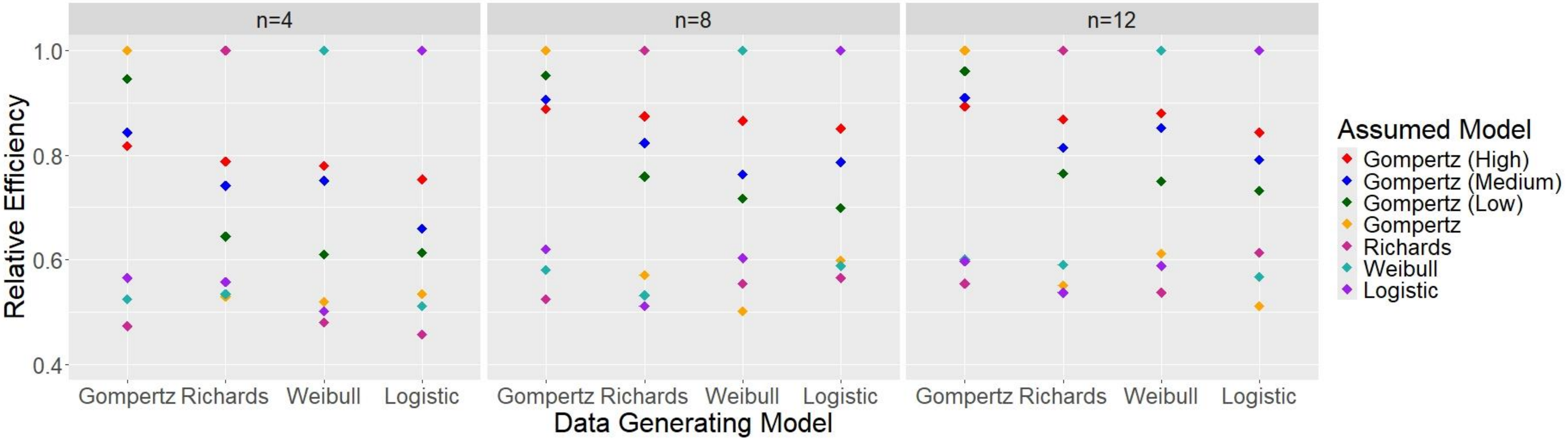
Robustness properties of designs

- Evaluate design efficiency: Measures expected information gain of design d relative to d^*

$$Eff(d, d^*) = U(d)/U(d^*)$$

- E.g. Suppose d based on Gompertz model. Find d^* based on assuming alternative Logistic growth.

- Evaluate efficiency assuming Logistic growth then (e.g.) 0.5 would suggest twice as much sampling would need to be undertaken with d to obtain as much information as d^*



Concluding remarks

- Proposed an approach to find Bayesian designs where prior information is potentially misspecified

- Needed for motivating design problem as prior information was based on studies conducted overseas

- Demonstrated robustness properties of resulting designs e.g. these remain efficient despite alternative growth models potentially being more appropriate to describe the data

- Increased flexibility led to increased robustness to alternative models (in our scenario)

- A range of models considered but only presented results for Gompertz model

- **Limitation:** Only considered one approach to allow the ODE to be more flexible. Other options available e.g. alternative inclusion/formulation of the spline term, Gaussian processes, etc.

- **Limitation:** Still assumption dependent e.g. assumed data are normally distributed.