Optimal testing designs under preferential sampling

Bren Case

2-14-2022

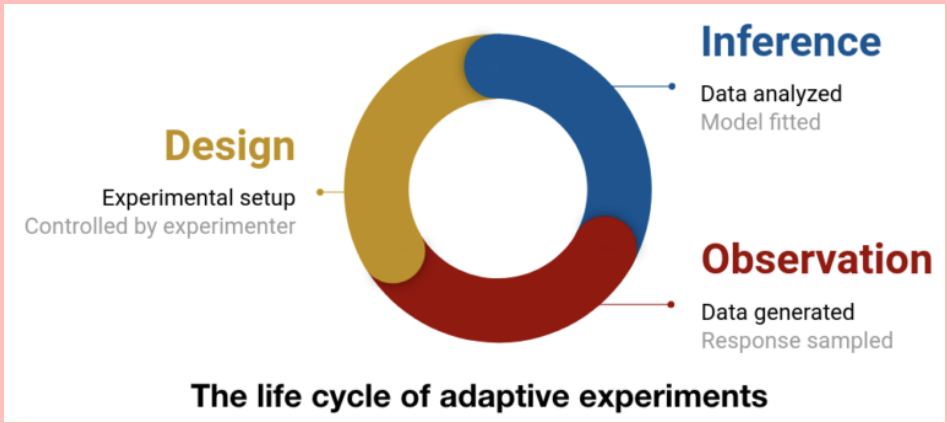
Outline

Conditional probabilities

- ❤️ $p(\boldsymbol{\theta})$ prior distribution. Parameters of the model
- ❤️ $p(\boldsymbol{y} \mid \boldsymbol{d}, \boldsymbol{\theta})$ likelihood. Model output with (noisy) observations determined by $\boldsymbol{d}$
- ❤️ $p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{d}) \propto p(\boldsymbol{y} \mid \boldsymbol{d}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$ posterior. Model parameters fit to data
- ❤️ $p(\boldsymbol{y} \mid \boldsymbol{d}) = \int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{d}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ model evidence. Probability that $\boldsymbol{y}$ came from model given prior beliefs.

What is (optimal) experimental design?

- ❤️ Specify how/what data should be collected, according to some goal (or *utility*)
  - Minimize response bias, reach certain power, maximize inference
- ❤️ Several standard definitions of optimal designs



eng.uber.com

Local vs. global optimality

For design $\boldsymbol{d}$, parameters $\boldsymbol{\theta}$, data $\boldsymbol{y}$, define utility U($\boldsymbol{d}$, $\boldsymbol{\theta}$, $\boldsymbol{y}$).
Local:

$$U(\boldsymbol{d};\boldsymbol{\theta}^*) = \int_{\mathcal{Y}} U(\boldsymbol{d},\boldsymbol{\theta}^*,\boldsymbol{y})p(\boldsymbol{y} \mid \boldsymbol{d},\boldsymbol{\theta}^*)d\boldsymbol{y}$$

Global:

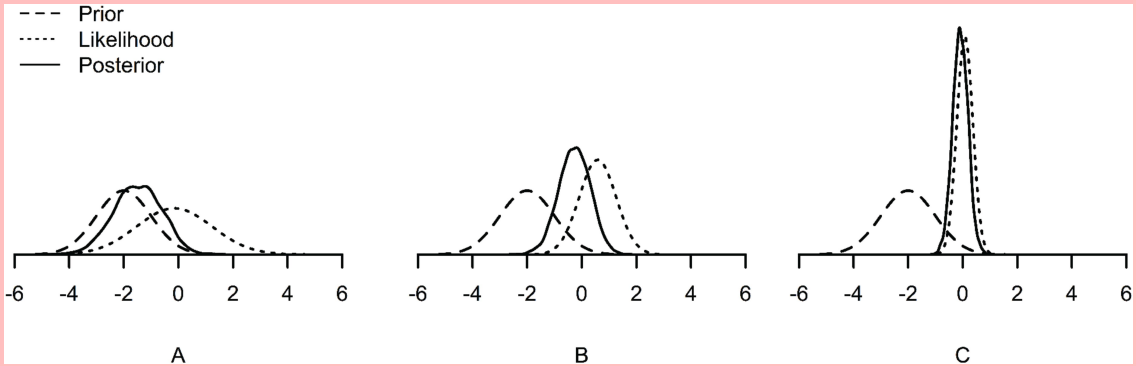
$$U(\boldsymbol{d}) = \int_{\Theta\times\mathcal{Y}} U(\boldsymbol{d},\boldsymbol{\theta},\boldsymbol{y})p(\boldsymbol{y} \mid \boldsymbol{d},\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{y}d\boldsymbol{\theta}$$

Shannon Information Gain:

$$U_{SIG}(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y}) = \log p(\boldsymbol{y} \mid \boldsymbol{d}, \boldsymbol{\theta}) - \log p(\boldsymbol{y} \mid \boldsymbol{d})$$

Negative sum squared error:

$$U_{NSSE}(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y}) = ||\boldsymbol{\theta} - \mathrm{E}[\theta \mid \boldsymbol{y}, \boldsymbol{d}]||_2^2$$



frontiersin.org

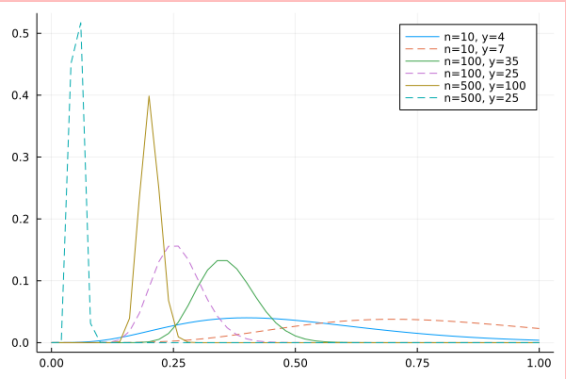
Design problems in epidemiology

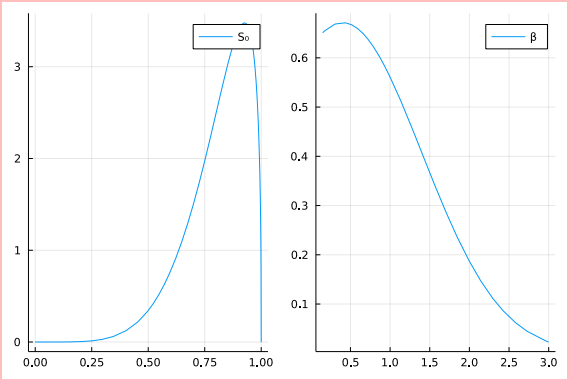
	Static (decision theory)	Sequential
Observation	$\boldsymbol{y}, \boldsymbol{\theta}$ not depend. on $\boldsymbol{d}$ d_t given prior to seeing $\boldsymbol{y}_{1:t-1}$	$\boldsymbol{y}, \boldsymbol{\theta}$ not depend. on $\boldsymbol{d}$ d_t depend. on $\boldsymbol{y}_{1:t-1}$
Control	$\boldsymbol{y}, \boldsymbol{\theta}$ depend. on $\boldsymbol{d}$ d_t given prior to seeing $\boldsymbol{y}_{1:t-1}$	$\boldsymbol{y}, \boldsymbol{\theta}$ depend. on $\boldsymbol{d}$ d_t depend. on $\boldsymbol{y}_{1:t-1}$

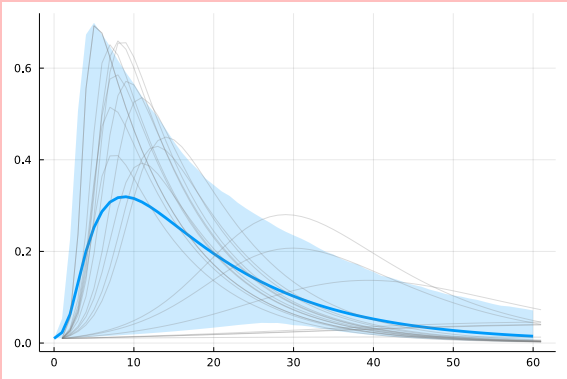
SIR model with Poisson testing rate:

$$\frac{dS}{dt} = -\beta I(t)S(t), \quad \frac{dI}{dt} = \beta I(t)S(t) - \alpha I(t)$$

$$y_t \mid \theta, n_t \sim \text{Poisson}(n_t I(t))$$

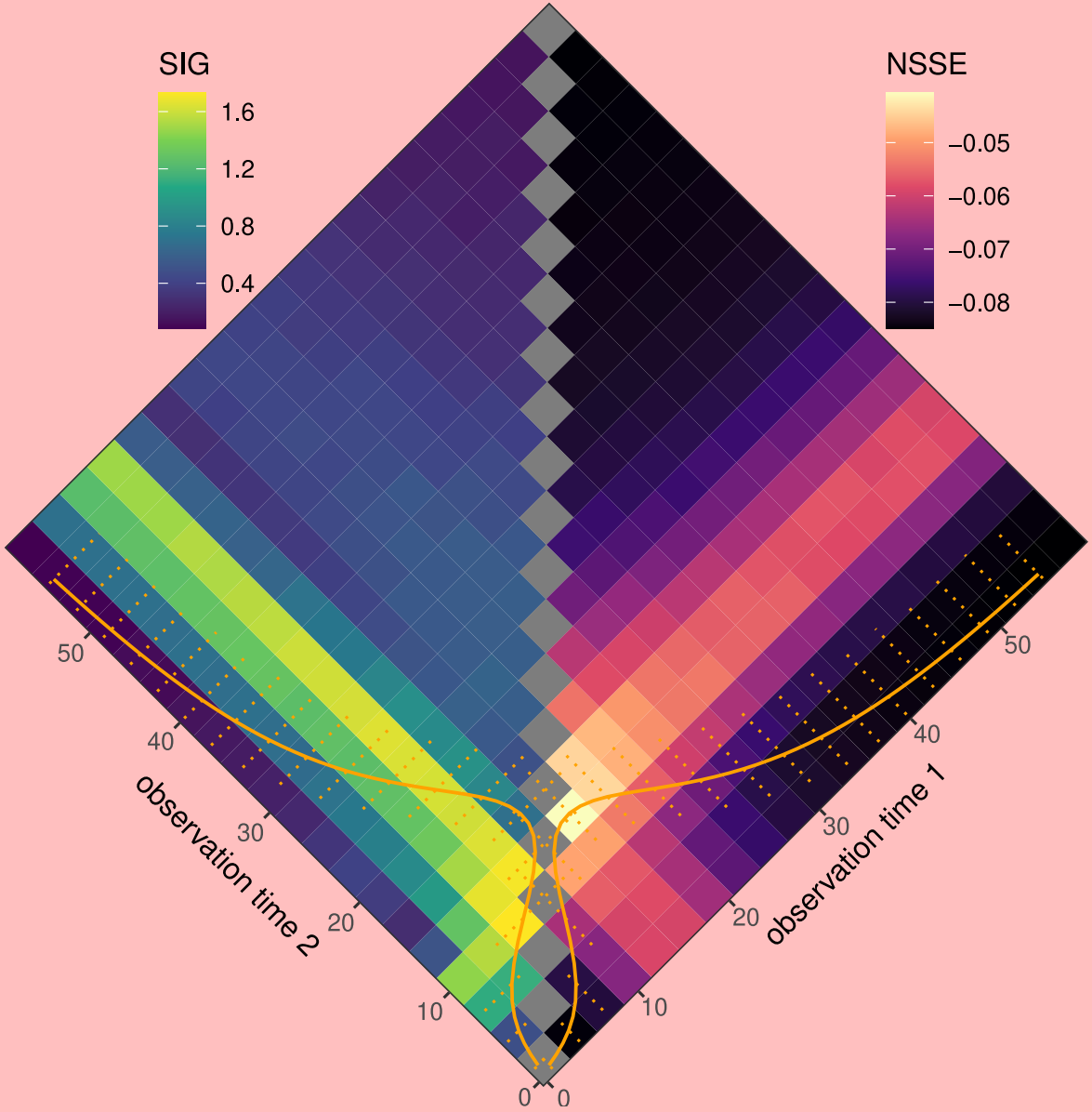






Some initial results

- Fixed true parameters $\boldsymbol{\theta}^*$
- $\boldsymbol{d} = (t_1, t_2)$, n_t fixed at 50
- Plot $U(\boldsymbol{d}; \boldsymbol{\theta}^*)$ for each pair



Introducing bias

- ❤️ Focused on features
  - tests applied to symptomatic individuals or approximately at random
  - tests applied less preferentially when more available
- ❤️ Ultimately want property that lots of tests needed for good inference ("feels reasonable" with COVID testing data)
  - Overdispersion?
  - Sensitivity/specificity?
  - Stochastic SIR?

Positive tests occur at (hidden) rate λ_+

$$\lambda_+ = n_1 p_1 + n_2 p_2$$

where (recall $I(t) = I \leq 1$)

$$p_1 = \frac{I}{G + I}$$
$$p_2 = I$$
$$n_1 = n^{-b}, \quad b \in (0, 1)$$
$$n_2 = n - n_1$$

Finding the true rate

- Fixed testing budget to predict I over T days

If $y \sim \text{Poisson}(\lambda_+)$, and $p_1 \approx z$. the MLE estimate for I is

$$\hat{I} = \frac{y - n_1 z}{n_2}$$

Let $U(n; x, b) = -\text{E}_y\left[(I - \hat{I})^2\right]$. Then utility associated with this particular day is

$$\frac{\frac{n_1}{n_2} z + I}{n_2}$$

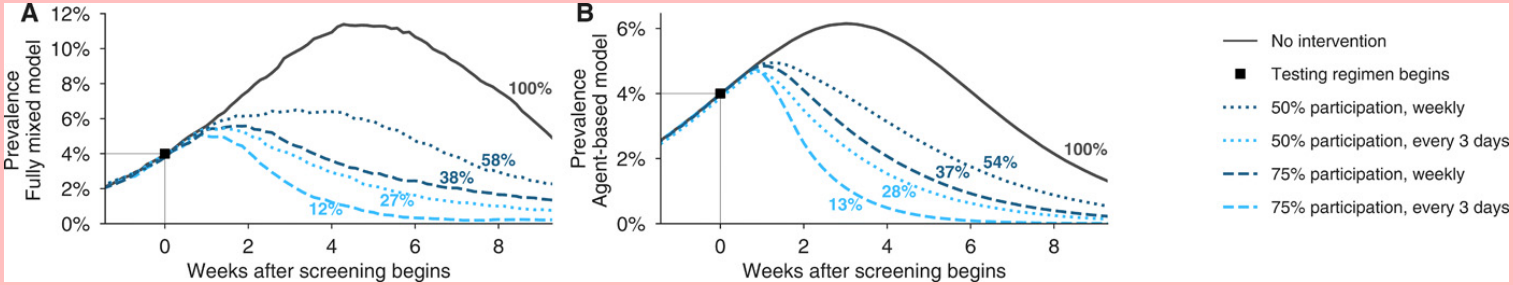
- Balance spreading tests to minimize bias and on high infection days

Current TODO list

- ❤️ Re-implementation with `MonteCarloMeasurements.jl`
- ❤️ Remake pair plot with some different assumptions
- ❤️ Apply current setup with biased testing
- ❤️ Explain wtf is going on

Apply biased testing framework in setting with asymptomatic cases, varying disease loads, and quarantine

- ❤️ Several cool models with simple testing strategies
- ❤️ Use ABC (and sequential Monte Carlo) to find robust (adaptive) strategies



Larremore, Daniel B., et al. "Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening." *Science advances*

Optimal testing schedules for model discrimination

- Lots of models use same collection of measurable parameters
- Priors on these params easy, but which model correct?
- Sequential design maximizing entropy of $p(\mathcal{M} \mid \boldsymbol{y}_{1:t}, \boldsymbol{d}_{1:t})$

Future work/side projects

Optimal testing schedules for observing and/or controlling EoN

- ❤️ Any hope with enormous search space?
- ❤️ Projection to lower dimensions
- ❤️ Using moment closures as proxies
- ❤️ Related to possible work with JGY and TGIR center