# Model parameter estimation

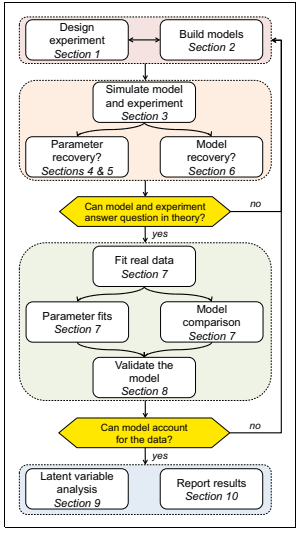
Carsten Murawski

6 October 2021

# Overview

In today's lecture, we will be covering:

- Experiment design
- Building computational models
- **Parameter estimation**
- Simulating model(s) (first pass)
- Parameter/model recoverability analysis (first pass)



# Overview

Please refer http://github.com/bmmlab/emds for resources, materials and references.

Our main reference for this lecture will be Robert C Wilson and Anne GE Collins. “Ten simple rules for the computational modeling of behavioral data”. In: *eLife* 8 (Nov. 2019), e49547.

# Experiment design

# Experiment design

Some issues to consider:

- Which scientific (research) question are you asking?
- What is the theoretical foundation of your question?
- Does your experiment engage the targeted cognitive (or neural) processes?
- Will ‘signatures’ of the targeted cognitive (or neural) processes be evident from simple statistics of your data?

# Let’s design a toy experiment

**Research question:** How do people trade off reward and delay to receipt of reward?

# Theoretical background: Intertemporal choice

We assume that participant's preferences can be represented by a utility function $u$, to be specified, and that discounting is exponential.

We expect a participant to be indifferent between two income options $M_t$ and $M_{t+\tau}$ if

$$u(\omega + M_t) + \frac{1}{(1+\delta)^\tau} u(\omega) = u(\omega) + \frac{1}{(1+\delta)^\tau} u(\omega + M_{t+\tau}), \quad (1)$$

where $u(\omega + M_t)$ is the utility of monetary outcome $M_t$ for delivery at time $t$ and background consumption $\omega$, $\delta$ is the discount rate, $\tau$ is the delay until receipt of reward at time $t + \tau$.

# Theoretical background: Intertemporal choice (cont'd)

We assume that $u$ is separable and stationary over time. (We will also assume they are completely liquidity constrained, consume monetary rewards from the task at the time stated in the task, and do not smooth consumption over time.)[1]

[1] Cf. Steffen Andersen et al. "Eliciting Risk and Time Preferences". In: *Econometrica* 76.3 (May 2008), pp. 583–618. ISSN: 0012-9682, 1468-0262.

# Theoretical background: Intertemporal choice (cont'd)

Let $U(\omega + M_{t+\tau})$ denote the present value of the utility of a monetary amount received after delay $t + \tau$, that is,

$$U(\omega + M_{t+\tau}) \equiv \frac{1}{(1+\delta)^{\tau}} u(\omega + M_{t+\tau}). \tag{2}$$

We also assume that observation of choices is noisy and that observed utility is random:

$$V(\omega + M_{t+\tau}) \equiv U(\omega + M_{t+\tau}) + \epsilon_{M_{t+\tau}}. \tag{3}$$

In economics, $\epsilon$ is typically referred to as "preference shock". We can think of it as observation error (that is, all the things that affect choice but that we cannot directly observe/measure).

# Theoretical background: Intertemporal choice (cont'd)

It follows that a participant is expected to choose

$$y = \begin{cases} M_t & \text{if } V(\omega + M_t) \geq V(\omega + M_{t+\tau}), \\ M_{t+\tau} & \text{otherwise.} \end{cases} \tag{4}$$

Given Eq. 3, we also have

$$\begin{aligned} V(\omega + M_{t+\tau}) > V(\omega + M_t) \iff \\ U(\omega + M_{t+\tau}) + \epsilon_{M_{t+\tau}} > U(\omega + M_t) + \epsilon_{M_t} \iff \\ U(\omega + M_{t+\tau}) - U(\omega + M_t) > \epsilon_{M_t} - \epsilon_{M_{t+\tau}}. \end{aligned} \tag{5}$$

# Theoretical background: Intertemporal choice (cont'd)

If $\epsilon$ (preference shocks) has a a logistic distribution, then the probability of a participant choosing $M_{t+\tau}$ over $M_t$ (!) is given by

$$F\left(U(\omega + M_{t+\tau})\right) = \frac{1}{1 + e^{-\frac{U(\omega+M_{t+\tau}) - U(\omega+M_t)}{s}}} = \frac{e^{\frac{U(\omega+M_{t+\tau})}{s}}}{e^{\frac{U(\omega+M_t)}{s}} + e^{\frac{U(\omega+M_{t+\tau})}{s}}}, \tag{6}$$

where $s$ is the variance of the logistic distribution.

We will often use $\beta = 1/s$ instead of $s$.

# Theoretical background: Intertemporal choice (cont'd)

The likelihood function (of parameters $\theta$ given choices $\mathbf{y}$) is given by

$$L(\theta \mid \mathbf{y}) = \prod_{i=1}^{I} \left[F\left(U(\omega + M_{t+\tau,i})\right)\right]^{y_i} \left[1 - F\left(U(\omega + M_{t+\tau,i})\right)\right]^{1-y_i}, \tag{7}$$

where $I$ is the number of trials and $y = y_{i \in I}$ is the set of $\mathbf{y}$ binary choices made by the participant. In trial $i$, dummy variable $y_i = 1$ if the participant chose the delayed reward $M_{t+\tau,i}$ over the immediate reward $M_{t,i}$ ($y_i = 0$ otherwise).

# Theoretical background: Intertemporal choice (cont'd)

For the purpose of our 'toy experiment', we will also assume that participants are risk-neutral.

In this case, Eq. 1 simplifies to

$$M_t = \frac{1}{1+\delta)^{\tau}} M_{t+\tau}. \tag{8}$$

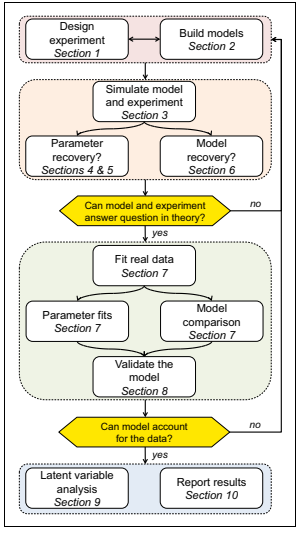
Many other specifications are possible, of course.[2]

[2] Cf. Shane Frederick and George Loewenstein. "Time Discounting and Time Preference: A Critical Review". In: *Journal of Economic Literature* 40.2 (2002), pp. 351–401.

# Let's recall our goal(s)

We were going to do:

- Experiment design
- Building computational models
- Parameter estimation
- Simulating model(s) (first pass)
- Parameter/model recoverability analysis (first pass)



# Recap: Experiment design

Some issues to consider:

- Which scientific (research) question are you asking?
- What is the theoretical foundation of your question?
- Does your experiment engage the targeted cognitive (or neural) processes?
- Will ‘signatures’ of the targeted cognitive (or neural) processes be evident from simple statistics of your data?

# Experimental paradigm

Binary choice task in which participants are given a number of choices between a “smaller, sooner” and a “larger, later” reward.

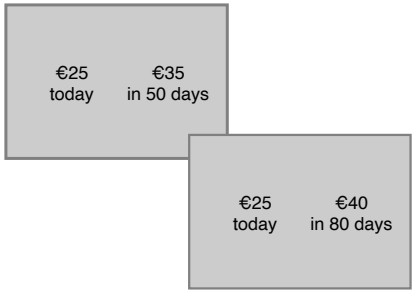
We will keep the smaller, sooner reward constant ($20 today) and modulate the larger, later rewards, for a range of delays (say, 1, 3, 6, 12, 18, 24 months).

We will need to choose the larger, later amounts such that their they ‘capture’ people with different discount rates, yet keep some ‘close’ enough to allow a precise estimate of the discount rate.

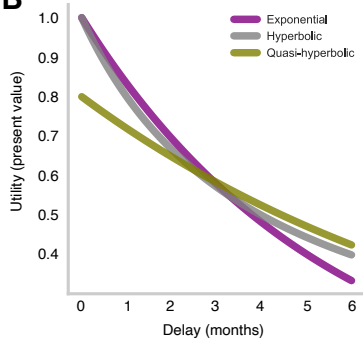
We will use the parameterisation from Steffen Andersen et al. “Eliciting Risk and Time Preferences”. In: *Econometrica* 76.3 (May 2008), pp. 583–618. ISSN: 0012-9682, 1468-0262. The task will present 6 x 10 trials (10 trials with varying amounts per delay). Each trial will last 6 seconds plus inter-trial interval. Trials will be presented in random order (to make them statistically independent).

# Experimental paradigm

A




B



# Side note: Relevance of temporal discounting tasks

Discount rates elicited with temporal discounting tasks have been shown to be correlated with a large number of 'real-life' outcomes, including educational, labour market and health outcomes. For a review, see Kristof Keidel et al. "Individual Differences in Intertemporal Choice". In: *Frontiers in Psychology* 12 (Apr. 2021), p. 643670.

Discount rates are elevated in patients suffering from mental illness, in almost all mental disorders (see Michael Amlung et al. "Delay Discounting as a Transdiagnostic Process in Psychiatric Disorders: A Meta-analysis". In: *JAMA Psychiatry* 76.11 (2019), p. 1176 for a review). Therefore, temporal discounting has been suggested as a candidate endophenotype.

# Parameter estimation

# Parameter estimation

A critical part of our study is the estimation of parameters $\theta$, for each model $m$ under consideration. We will denote the parameters associated with a model $m$ by $\theta_m$.

In our study, we will estimate model parameters based on observed choices $\mathbf{d}$. To do so, we will take a Bayesian approach, with the aim to compute the posterior distribution over the parameters given the data, which we denote by $p(\theta_m|\mathbf{d}, m)$.

Using Bayes rule, we can write the posterior as

$$p(\theta_m|\mathbf{d}, m) = \frac{p(\mathbf{d}|\theta_m, m)p(\theta_m|m)}{p(\mathbf{d}|m)}, \tag{9}$$

where $p(\theta_m|m)$ is the prior on the parameters for model $m$, $p(\mathbf{d}|\theta_m, m)$ is the likelihood of the observations (data) given the parameters, and $p(\mathbf{d}|m)$ is the probability of the data given the model (also called the normalisation constant).

# Parameter estimation (cont'd)

We can rewrite Eq. 9 as

$$\log p(\theta_m|\mathbf{d}, m) = \log p(\mathbf{d}|\theta_m, m) + \log p(\theta_m|m) - \log p(\mathbf{d}|m), \tag{10}$$

The log of the likelihood, $\log p(\mathbf{d}|\theta_m, m)$, is typically referred to as log-likelihood. It can be rewritten as (cf. also Eq. 7)

$$\log p(\mathbf{d}|\theta_m, m) = \log \left( \prod_{i=1}^{I} p(c_i|d_i, \theta_m, m) \right) = \sum_{i=1}^{I} \log p(c_i|d_i, \theta_m, m), \tag{11}$$

where $p(c_i|d_i, \theta_m, m)$ is the probability of each individual choice given the parameters of the model.

# Parameter estimation (cont'd)

Ideally, we could compute the log posterior, $\log p(\theta_m|\mathbf{d}, m)$, directly. However, in practice, this is usually infeasible. Therefore, we typically approximate it.

The most common approach for approximating the posterior are Markov Chain Monte Carlo approaches, which approximate the full posterior distribution (of $\theta_m$).

An alternative approach to estimate the parameters is maximisation of the log-likelihood, which yields point estimates for the parameters:

$$\hat{\theta}_m^{MLE} = \underset{\theta_m}{\operatorname{argmax}} \log p(\mathbf{d}|\theta_m, m). \tag{12}$$

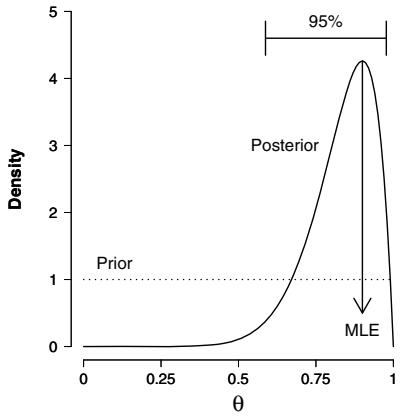
This is called maximum likelihood estimation (MLE). It can be done with standard optimisation functions.

# Properties of the maximum likelihood estimator

In the limit (sample size goes to infinity), the maximum likelihood estimator is

- Consistent (convergence in probability)
- Functionally invariant
- Efficient

A maximum likelihood estimator converges with the most probable Bayesian estimator given a uniform prior distribution on the parameters (cf. Eq. 9, if $p(\theta_m|m)$ has a uniform distribution, $p(\theta_m|\mathbf{d}, m)$, the posterior, is maximised by maximising $p(\mathbf{d}|\theta_m, m)$, the likelihood).




Source: Lee and Wagenmakers (2014)

# Model comparison

If there is more than one candidate model (which is usually the case), an important question is which model is most likely to have generated the data.
We can address this question by computing the probability that the model $m$ generated the data $\mathbf{d}$, $p(m|\mathbf{d})$. Recalling Eq. 9, we have

$$p(m|\mathbf{d}) \propto p(\mathbf{d}|m)p(m) = \int d\theta_m p(\mathbf{d}|\theta_m, m)p(\theta_m|m)p(m), \tag{13}$$

where $p(m)$ is the prior probability that model $m$ is the correct model. Usually, $p(m)$ is assumed to be constant. We can therefore focus on the (log) likelihood

$$E_m = \log p(\mathbf{d}|m) = \int d\theta_m p(\mathbf{d}|\theta_m, m)p(\theta_m|m). \tag{14}$$

$E_m$ is called the Bayesian evidence.

# Model comparison (cont'd)

Computing the Bayesian evidence is often impossible. Therefore, it is typically approximated or replaced with an approximation around the MLE estimates of the parameters.

The latter approach includes the Akaike Information Criterion (AIC), Bayes Information Criterion (BIC) and the Laplace approximation.

We will focus on BIC, which is given by

$$BIC = -2 \log \hat{\mathcal{L}} + k_m \log(T) \approx -2 \log E_m, \tag{15}$$

where $k_m$ is the number of parameters in model $m$ and $\hat{\mathcal{L}}$ is the value of the log-likelihood at $\hat{\theta}_m^{MLE}$ (a lower BIC value is 'better').

# Parameter estimation using Bayesian approaches

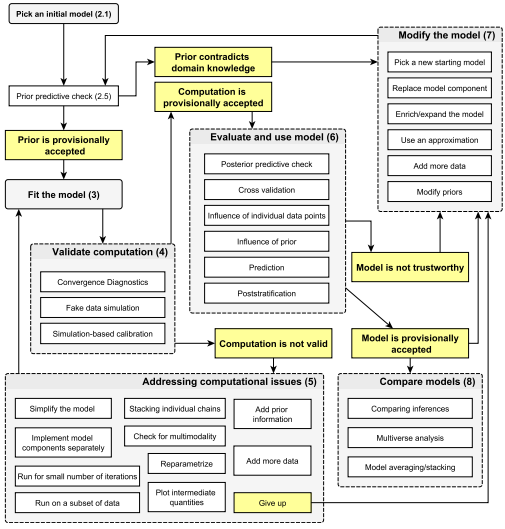
The aim of Bayesian statistics is to represent prior uncertainty about model parameters with a probability distribution and to update this prior uncertainty with new data to produce a posterior probability distribution for the parameter, which contains less uncertainty.

# Bayesian workflow

The three steps of

- model building
- inference
- model checking/improvement

can thought of as the main steps of the Bayesian workflow (cf. Andrew Gelman et al. “Bayesian Workflow”. In: (Nov. 2020). arXiv: 2011.01808).

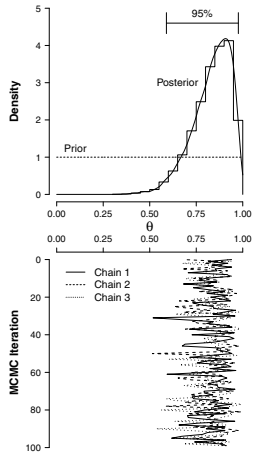


# Parameter estimation using Bayesian approaches (cont'd)

Bayesian approaches compute the posterior distribution over the parameters given the data, $p(\theta_m|\mathbf{d}, m)$. The posterior is typically approximated using techniques based on Markov Chain Monte Carlo (HMC) simulation.

We will illustrate parameter estimation with Bayesian approaches using Stan.

Good introductions to Bayesian estimation are Bayesian Data Analysis by Gelman et al., and Bayesian Cognitive Modeling: A Practical Course by Lee and Wagenmakers (the latter is available online via the University's Library website).




Source: Lee and Wagenmakers (2014)

# Bayesian model comparison

When using a Bayesian approach to parameter estimation, we can use the Bayes factor as a criterion for model comparison.

Remember that the likelihood $p(\mathbf{d}|m)$ represents the probability that the observed data were produced by model $m$. To choose between two models $m_1$ and $m_2$ on the basis of the observed data $\mathbf{d}$, with model parameters $\theta_1$ and $\theta_2$, we can use the Bayes factor

$$BF = \frac{p(\mathbf{d}|m_1)}{p(\mathbf{d}|m_2)} = \frac{\int p(\theta_1|m_1)p(\mathbf{d}|\theta_1, m_1)d\theta_1}{\int p(\theta_2|m_2)p(\mathbf{d}|\theta_2, m_2)d\theta_2} = \frac{p(m_1|\mathbf{d})p(m_1)}{p(m_2|\mathbf{d})p(m_2)}. \tag{16}$$