Hierarchical models

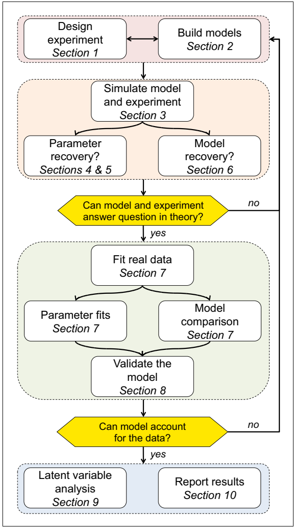
Carsten Murawski

13 October 2021

Overview

In today's lecture, we will be covering:

- Recap of the modelling workflow
- Recap of MLE/Bayesian estimation fundamentals
- Hierarchical Bayes models
- Simulating model(s) (second pass)
- Parameter/model recoverability analysis (second pass)
- Analysing 'real' data



Overview

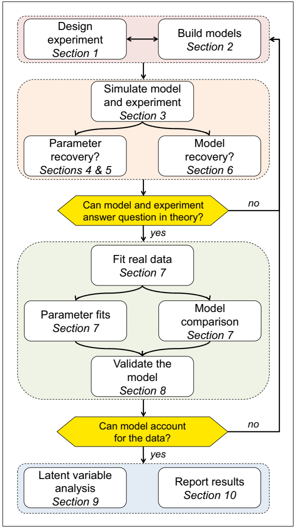
Please refer http://github.com/bmmlab/emds for resources, materials and references.

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

Recap: Modelling workflow (part 1

The main step in the modelling workflow 'pre-data collection' are:

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



Our 'workhorse' experimental paradigm

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

Today, we'll be using another parameterisation of the task based on Kris N. Kirby, Nancy M. Petry, and Warren K. Bickel. "Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls.". In: *Journal of Experimental Psychology: General* 128.1 (1999), pp. 78–87. In this version, the smaller, sooner amounts vary (but always with 0 delay). Delays are expressed in number of days. This task is often used in clinical, psychopathology and other individual differences studies.

The task will present 30 trials. Each trial will last 5 seconds plus inter-trial interval. Trials will be presented in random order (to make them statistically independent).

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 θ, for each model m under consideration. We will denote the parameters associated with a model m by θ_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

$$p(\theta_m|\mathbf{d}, m) = \frac{p(\mathbf{d}|\theta_m, m)p(\theta_m|m)}{p(\mathbf{d}|m)}, \tag{1}$$

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. 1 as

$$\log p(\theta_m|\mathbf{d}, m) = \log p(\mathbf{d}|\theta_m, m) + \log p(\theta_m|m) - \log p(\mathbf{d}|m), \tag{2}$$

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

$$\log p(\mathbf{d}|\theta_m, m) = \log \left(\prod_{i=1}^{I} p(c_i|d_i, \theta_m, m) \right) = \sum_{i=1}^{I} \log p(c_i|d_i, \theta_m, m), \tag{3}$$

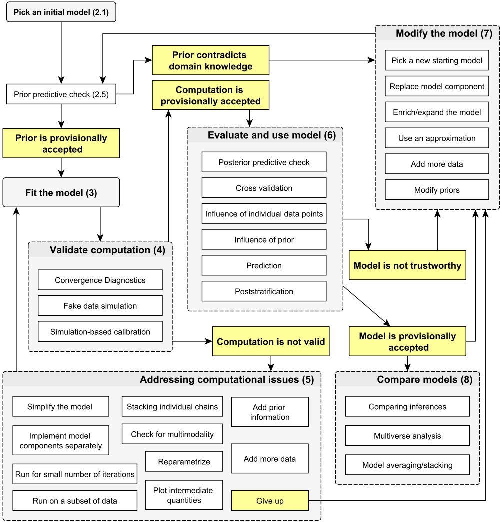
where $p(c_i|d_i, \theta_m, m)$ is the probability of each individual choice given the parameters of the model.

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).

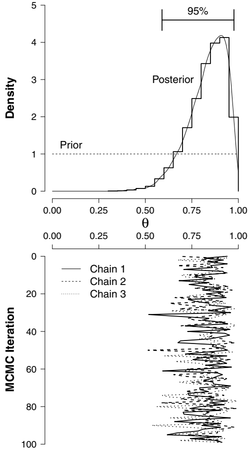


Parameter estimation using Bayesian approaches (cont'd)

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 θ_1 and θ_2, we can use the Bayes factor

$$BF = \frac{p(\mathbf{d}|m_1)}{p(\mathbf{d}|m_2)} = \frac{\int p(\theta_1|m_1)p(\mathbf{d}|\theta_1, m_1)d\theta_1}{\int p(\theta_2|m_2)p(\mathbf{d}|\theta_2, m_2)d\theta_2} = \frac{p(m_1|\mathbf{d})p(m_1)}{p(m_2|\mathbf{d})p(m_2)}. \tag{4}$$

Hierarchical models

In our experiments, we typically assess more than one participant. We might also have other ‘groupings’ of trials, e.g., treatment groups. This introduces structure into our data.

If we combined all trials in one big data set and treated them all as coming from a single (representative) participant, we would potentially end up with misestimation.

We will call a model “non-hierarchical” if is assumed that all data **d** were generated by a set of parameters θ through a likelihood function $p(\mathbf{d}|\theta)$.

Hierarchical models deviate from this structure. In the perhaps simplest form, we might assume that the parameters θ are themselves generated by a function $g(\cdot)$, parameterised by ψ. The latter is usually referred as hyperparameter(s). We now have a model in which the data are generated by $f(\theta)$, but θ is generated by $g(\psi)$, that is, $\mathbf{d} \sim f(\theta)$ and $\theta \sim g(\psi)$.

Hierarchical models (cont'd)

This simple hierarchical structure could be used to account for individual differences in the data. If a non-hierarchical model was used, one would typically estimate parameters for individual participants separately and the analyse individual differences between them through post-hoc analyse.

With a hierarchical model, in contract, the structure in individual differences can be captured by the function g and its parameters ψ.

We can use this approach to model data from an experiment with multiple participants: we choose a set of hyperparameters that determine the distributions from which individual (participant) parameters are drawn.

Hierarchical models (cont'd)

A good starting point for reading up on the use of hierarchical models in cognitive modelling is Michael D. Lee. "How cognitive modeling can benefit from hierarchical Bayesian models". In: *Journal of Mathematical Psychology* 55.1 (2011), pp. 1–7.

And if you would like to see a hierarchical model 'in action', have a look at Christopher A Hill et al. "A causal account of the brain network computations underlying strategic social behavior". In: *Nature Neuroscience* 20.8 (2017), pp. 1142–1149. (Visited on 10/12/2021).

We will do parameter and model recoverability with hierarchical models and compare them to pooled (non-hierarchical) models. See R notebooks for details.

Fitting ‘real’ data

Fitting real data

See R notebooks in Github repo.

Wrap up

The four stages of the modelling workflow

- Experiment design and model building
- Parameter/model recoverability analysis
- Fitting experiment data and robustness tests
- Reporting of results