Modelling behavioural data

Quentin Huys
MA PhD MBBS MBPsS

Translational Neuromodeling Unit, ETH Zürich
Zurich University Hospital of Psychiatry
Outline

- An example task
- Why build models? What is a model
- Fitting models
- Validating & comparing models
- Model comparison issues in psychiatry
Example task

- **Go**
  - Reward
  - Avoids loss
- **Nogo**
  - Orange square
  - Purple square

Guitart-Masip, Huys et al. Submitted
Example task
Example task

- Go rewarded: Go to win
- Nogo punished: Go to avoid
- Nogo rewarded: Nogo to win
- Go punished: Nogo to avoid
Example task

Think of it as four separate two-armed bandit tasks

Guitart-Masip, Huys et al. Submitted
Analysing behaviour

- **Standard approach:**
  - Decide which feature of the data you care about
  - Run descriptive statistical tests, e.g. ANOVA

- **Many strengths**
- **Weakness**
  - Piecemeal, not holistic / global
  - Descriptive, not generative
  - No internal variables
Analysing behaviour

- **Standard approach:**
  - Decide which feature of the data you care about
  - Run descriptive statistical tests, e.g. ANOVA

- **Many strengths**

- **Weakness**
  - Piecemeal, not holistic / global
  - Descriptive, not generative
  - No internal variables
Models

- Holistic
  - Aim to model the process by which the data came about in its “entirety”

- Generative
  - They can be run on the task to generate data as if a subject had done the task

- Inference process
  - Capture the inference process subjects have to make to perform the task.
  - Do this in sufficient detail to replicate the data.

- Parameters
  - replace test statistics
  - their meaning is explicit in the model
  - their contribution to the data is assessed in a holistic manner
A simple Rescorla-Wagner model

- **Q values**
  
  \[ Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \epsilon (r_t - Q_{t-1}(a_t, s_t)) \]

  - \( a_t \): action on trial \( t \); can be either 'go' or 'logo'
  - \( s_t \): stimulus presented on trial \( t \)
  - \( \epsilon \): learning rate

- **Key points:**
  - Q is the key part of the hypothesis
  - formally states the learning process in quantitative detail
  - formalizes internal quantities that are used in the task
**Actions**

- **Q values**

\[ Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \epsilon (r_t - Q_{t-1}(a_t, s_t)) \]

- **Action probabilities: “softmax” of Q value**

\[
p(a_t | s_t, h_t, \beta) = p(a_t | Q(a_t, s_t), \beta) = \frac{e^{\beta Q(a_t, s_t)}}{\sum_{a'} e^{\beta Q(a', s_t)}}
\]

- **Features:**

\[
p(a_t | s_t) \propto Q(a_t, s_t)
\]

\[
0 \leq p(a) \leq 1
\]

- **links learning process and observations**

  - choices, RTs, or any other data
  - link function in GLMs
  - man other forms
Fitting models I

- Maximum likelihood (ML) parameters

\[ \hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta) \]

- where the likelihood of all choices is:

\[
\mathcal{L}(\theta) = \log p(\{a_t\}_{t=1}^T | \{s_t\}_{t=1}^T, \{r_t\}_{t=1}^T, \{Q(s_t, a_t; \beta, \epsilon)\}_{t=1}^T, \theta) = \log p(\{a_t\}_{t=1}^T | \{Q(s_t, a_t; \epsilon)\}_{t=1}^T, \beta) = \log \prod_{t=1}^T p(a_t | Q(s_t, a_t; \epsilon), \beta) = \sum_{t=1}^T \log p(a_t | Q(s_t, a_t; \epsilon), \beta) \]
Fitting models II

- No closed form
- Use your favourite method
  - gradients
  - fminunc / fmincon...
- Gradients for RW model

\[
\frac{d\mathcal{L}(\theta)}{d\theta} = \frac{d}{d\theta} \sum_t \log p(a_t | Q_t(a_t, s_t; \epsilon), \beta) \\
= \sum_t \frac{d}{d\theta} \beta Q_t(a_t, s_t; \epsilon) - \sum_{a'} p(a' | Q_t(a', s_t; \epsilon), \beta) \frac{d}{d\theta} \beta Q_t(a', s_t; \epsilon)
\]

\[
\frac{dQ_t(a_t, s_t; \epsilon)}{d\epsilon} = (1 - \epsilon) \frac{dQ_{t-1}(a_t, s_t; \epsilon)}{d\epsilon} + (r_t - Q_{t-1}(a_t, s_t; \epsilon))
\]
Little tricks

- Transform your variables

\[
\begin{align*}
\beta &= e^{\beta'} \\
\Rightarrow \beta' &= \log(\beta) \\
\epsilon &= \frac{1}{1 + e^{-\epsilon'}} \\
\Rightarrow \epsilon' &= \log \left( \frac{\epsilon}{1 - \epsilon} \right)
\end{align*}
\]

- Avoid over/underflow

\[
\begin{align*}
y(a) &= \beta Q(a) \\
y_m &= \max_a y(a) \\
p &= \frac{e^{y(a)}}{\sum_b e^{y(b)}} = \frac{e^{y(a)} - y_m}{\sum_b e^{y(b)} - y_m}
\end{align*}
\]
ML characteristics

200 trials, 1 stimulus, 10 actions, learning rate = .05, beta=2

\[ \frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta) \]

Daw 2010
ML characteristics

- ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon

\[ \frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta) \] can be used to derive confidence intervals and identify poorly constrained estimates

Daw 2010
ML characteristics

- ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon

- Hessian $\frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta)$ can be used to derive confidence intervals and identify poorly constrained estimates

- ML can overfit... more later
ML characteristics

- ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon

\[
\mathcal{L}(\beta = 10) \approx \mathcal{L}(\beta = 100)
\]

- Hessian \( \frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta) \) can be used to derive confidence intervals and identify poorly constrained estimates

- ML can overfit... more later
ML characteristics

- ML is asymptotically consistent, but variance high
  - 10-armed bandit, infer beta and epsilon

\[
\mathcal{L}(\beta = 10) \approx \mathcal{L}(\beta = 100)
\]

- Hessian \( \frac{d^2}{d\theta_i d\theta_j} \mathcal{L}(\theta) \) can be used to derive confidence intervals and identify poorly constrained estimates

- ML can overfit... more later
Priors

Not so smooth

Smooth
### Priors

- **Not so smooth**
- **Smooth**

![Graphs showing priors for parameters.](image-url)
Priors

Not so smooth

Smooth

Prior Probability

Prior Probability

Parameter

Parameter

Prior Probability

Prior Probability

Parameter

Parameter

Switches

Dombrovski et al. 2010

Memory

0 20 40 60

0 0.2 0.4 0.6 0.8 1.0

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

0 2 4 6

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

0 5

Am J Psychiatry 167:6, June 2010
Maximum a posteriori estimate

\[ P(\theta) = p(\theta|a_1...T) = \frac{p(a_1...T|\theta)p(\theta)}{\int d\theta p(\theta|a_1...T)p(\theta)} \]

\[ \log P(\theta) = \sum_{t=1}^{T} \log p(a_t|\theta) + \log p(\theta) + \text{const}. \]

\[ \frac{\log P(\theta)}{d\alpha} = \frac{\log L(\theta)}{d\alpha} + \frac{dp(\theta)}{d\theta} \]

- If likelihood is strong, prior will have little effect
  - mainly has influence on poorly constrained parameters
  - if a parameter is strongly constrained to be outside the typical range of the prior, then it will win over the prior
Maximum a posteriori estimate

200 trials, 1 stimulus, 10 actions, learning rate = 0.05, beta=2
m_{beta}=0, m_{eps}=-3, n=1
What prior parameters should I use?
ML characteristics: group data
ML characteristics: group data

\( \theta \)

\( \mathcal{A}_i \mathcal{A}_j \mathcal{A}_k \mathcal{A}_l \)
ML characteristics: group data

- Fixed effect
  - conflates within- and between-subject variability
ML characteristics: group data

- Fixed effect
  - conflates within- and between-subject variability

- Average behaviour
  - disregards between-subject variability
  - need to adapt model
ML characteristics: group data

- Fixed effect
  - conflates within- and between-subject variability
- Average behaviour
  - disregards between-subject variability
  - need to adapt model
- Summary statistic
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy
ML characteristics: group data

- **Fixed effect**
  - conflates within- and between- subject variability

- **Average behaviour**
  - disregards between-subject variability
  - need to adapt model

- **Summary statistic**
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy

- **Random effects**
  - prior mean = group mean
ML characteristics: group data

- **Fixed effect**
  - conflates within- and between-subject variability

- **Average behaviour**
  - disregards between-subject variability
  - need to adapt model

- **Summary statistic**
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy

- **Random effects**
  - prior mean = group mean

\[
p(\mathcal{A}_i | \mu_\theta, \sigma_\theta) = \int d\theta_i p(\mathcal{A}_i | \theta_i) p(\theta_i | \mu_\theta, \sigma_\theta)
\]
ML characteristics: group data

- **Fixed effect**
  - conflates within- and between- subject variability

- **Average behaviour**
  - disregards between-subject variability
  - need to adapt model

- **Summary statistic**
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy

- **Random effects**
  - prior mean = group mean

\[
p(A_i | \mu_\theta, \sigma_\theta) = \int d\theta_i p(A_i | \theta_i) p(\theta_i | \mu_\theta, \sigma_\theta)
\]
**ML characteristics: group data**

- **Fixed effect**
  - conflates within- and between- subject variability

- **Average behaviour**
  - disregards between-subject variability
  - need to adapt model

- **Summary statistic**
  - treat parameters as random variable, one for each subject
  - overestimates group variance as ML estimates noisy

- **Random effects**
  - prior mean = group mean

$$p(A_i | \mu_\theta, \sigma_\theta) = \int d\theta_i p(A_i | \theta_i) p(\theta_i | \mu_\theta, \sigma_\theta)$$
Estimating the hyperparameters

- MAP

\[
\log \mathcal{P}(\theta) = \mathcal{L}(\theta) + \log p(\theta) + \text{const.} \\
= p(\theta | \zeta)
\]

- Empirical Bayes: set them to ML estimate

\[
\hat{\zeta} = \arg\max_\zeta p(A | \zeta)
\]

- where we use all the actions by all the \( k \) subjects

\[
A = \{ a_{k1...T} \}_{k=1}^{K}
\]
ML estimate of top-level parameters

\[ \hat{\zeta} = \arg\max_{\zeta} p(A|\zeta) \]
Estimating the hyperparameters

- Effectively we now want to do gradient ascent on:
  \[ \frac{d}{d\zeta} p(A|\zeta) \]

- But this contains an integral over individual parameters:
  \[ p(A|\zeta) = \int d\theta p(A|\theta) p(\theta|\zeta) \]

- So we need to:
  \[ \hat{\zeta} = \arg\max_{\zeta} p(A|\zeta) \]
  \[ = \arg\max_{\zeta} \int d\theta p(A|\theta) p(\theta|\zeta) \]
Expectation Maximisation

\[
\log p(A|\zeta) = \log \int d\theta \ p(A, \theta|\zeta)
\]
\[
= \log \int d\theta \ q(\theta) \frac{p(A, \theta|\zeta)}{q(\theta)}
\]
\[
\geq \int d\theta \ q(\theta) \log \frac{p(A, \theta|\zeta)}{q(\theta)}
\]

\[k^{th} \text{ E step: } q^{(k+1)}(\theta) \leftrightarrow p(\theta|A, \zeta^{(k)})\]

\[k^{th} \text{ M step: } \zeta^{(k+1)} \leftrightarrow \arg\max_{\zeta} \int d\theta \ q(\theta) \log p(A, \theta|\zeta)\]

- There are other approaches
  - Monte Carlo
  - Analytical conjugate priors
  - Variational Bayes

- Iterate between
  - Estimating MAP parameters given prior parameters
  - Estimating prior parameters from MAP parameters
EM with Laplace approximation

- **E step**: $q^{(k+1)}(\theta) \leftarrow p(\theta|\mathcal{A}, \zeta^{(k)})$
  - only need sufficient statistics to perform M step
  - Approximate $p(\theta|\mathcal{A}, \zeta^{(k)}) \sim \mathcal{N}(m_k, S_k)$
  - and hence:

\[
\text{E step: } q_k(\theta) = \mathcal{N}(m_k, S_k) \nonumber
\]

\[
m_k \leftarrow \arg\max_\theta p(a_k|\theta)p(\theta|\zeta^{(i)}) \nonumber
\]

\[
S^{-1}_k \leftarrow \frac{\partial^2 p(a^k|\theta)p(\theta|\zeta^{(i)})}{\partial \theta^2} \bigg|_{\theta=m_k} \nonumber
\]

Just what we had before: MAP inference given some prior parameters

**matlab**: `[m, L, , S] = fminunc(...)`
EM with Laplace approximation

Next update the prior

M step:

\[
\zeta_{\mu}^{(i+1)} = \frac{1}{K} \sum_k m_k
\]

\[
\zeta_{\nu^2}^{(i+1)} = \frac{1}{N} \sum_i \left[ (m_k)^2 + S_k \right] - (\zeta_{\mu}^{(i+1)})^2
\]

Prior mean = mean of MAP estimates
Prior variance depends on inverse Hessian \( S \) and variance of MAP estimates
Take uncertainty of estimates into account

And now iterate until convergence
Hierarchical / random effects models

- **Advantages**
  - Accurate group-level mean and variance
  - Outliers due to weak likelihood are regularized
  - Strong outliers are not
  - Useful for model selection

- **Disadvantages**
  - Individual estimates $\theta_i$ depend on other data, i.e. on $A_{j\neq i}$ and therefore need to be careful in interpreting these as summary statistics
  - Error bars on group parameters (especially group variance) are difficult to obtain
  - More involved; less transparent
Link functions

- **Sigmoid**
  
  \[ p(a|s) = \frac{e^{\beta Q(a,s)}}{\sum_{a'} e^{\beta Q(a',s)}} \]

- **\( \epsilon \)-greedy**
  
  \[ p(a|s) = \begin{cases} 
  c & \text{if } a = \arg \max_a Q(a,s) \\
  \frac{1-c}{|a|-1} & \text{else}
  \end{cases} \]

- **irreducible noise**
  
  \[ p(a|s) = \frac{1 - 2g}{2} + g \frac{e^{\beta Q(a,s)}}{\sum_{a'} e^{\beta Q(a',s)}} \]

- **critical sanity check 1: reasonable link function?**

- **other link functions for other observations**
Model comparison

- A fit by itself is not meaningful
- Generative test
  - qualitative
- Comparisons
  - vs random
  - vs other model -> test specific hypotheses and isolate particular effects in a generative setting
How well does the model do?

• choice probabilities:
  \[ \mathbb{E} p(\text{correct}) = e^{\mathcal{L}(\hat{\theta})/K/T} \]
  \[ = e^{\log p(A|\theta)/K/T} \]
  \[ = \left( \prod_{k,t=1}^{K,T} p(a_{k,t}|\theta_k) \right)^{1/KT} \]

  “Predictive probabilities”

• typically around 0.65-0.75 for 2-way choice
• for 10-armed bandit example
• pseudo-\(r^2\): I-L/R
• better than chance?

\[ \mathbb{E}[N_k(\text{correct})] = \mathbb{E}[p_k(\text{correct})]T \]
\[ p_{\text{bin}}(\mathbb{E}[N_k(\text{correct})]|N_kd, p_0 = 0.5) < 1 - \alpha \]
Generative test

- Model: probability(\text{actions})
  - simply draw from this distribution, and see what happens

- Critical sanity test: is the model meaningful?
- Caveat: overfitting
Overfitting
Model comparison

Data

$P(x)$

Model 1

Model 2
Model comparison

‣ Averaged over its parameter settings, how well does the model fit the data?

\[ p(A|M) = \int d\theta \, p(A|\theta) \, p(\theta|M) \]

‣ Model comparison: Bayes factors

\[ BF = \frac{p(M_1|A)}{p(M_0|A)} = \frac{p(A|M_1) \, p(M_1)}{p(A|M_2) \, p(M_2)} \]

‣ Problem:
  • integral rarely solvable
  • approximation: Laplace, sampling, variational...
Why integrals? The God Almighty test

![Graph: Powerful model](image1.png)

![Graph: Weak model](image2.png)

---

**Why integrals?**

The God Almighty test is a way to illustrate the difference between powerful and weak models in the context of behavioural data modelling. The graphs show the distribution of observations for two models, with the x-axis representing different values and the y-axis showing the probability density.

**Powerful model:**
- The distribution has high peaks around certain values, indicating that these are more probable outcomes.
- The model captures the nuances of the data effectively.

**Weak model:**
- The distribution is flatter with lower peaks, suggesting that the model is oversimplified and misses important details in the data.

This test is crucial for understanding how our models can either accurately represent the data or fail to capture the underlying patterns, which is fundamental in behavioural data modelling.
Why integrals? The God Almighty test

- Powerful model
- Weak model
Why integrals? The God Almighty test

\[ \frac{1}{N} \left( p(X|\theta_1) + p(X|\theta_2) + \cdots \right) \]

These two factors fight it out
Model complexity vs model fit
Bayesian Information Criterion

- Laplace’s approximation (saddle-point method)
Bayesian Information Criterion

- Laplace’s approximation (saddle-point method)
Bayesian Information Criterion

- Laplace’s approximation (saddle-point method)
Bayesian Information Criterion

- Laplace’s approximation (saddle-point method)

\[ \int dx \ f(x) \approx f^*(x_0) \sqrt{2\pi \sigma^2} \]
Bayesian Information Criterion: one subject

\[ p(A|M) = \int d\theta \, p(A|\theta, M) \, p(\theta|M) \]
Bayesian Information Criterion: one subject

\[ p(A|M) = \int d\theta \, p(A|\theta, M) \, p(\theta|M) \]

\[ p(A|\theta) \, p(\theta|M) \text{ is propto Gaussian} \]
Bayesian Information Criterion: one subject

\[ p(A|\mathcal{M}) = \int d\theta \ p(A|\theta, \mathcal{M}) \ p(\theta|\mathcal{M}) \]
\[ \approx \ p(A|\theta^{ML}, \mathcal{M})p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \]

\( p(\mathcal{A}|\theta) \ p(\theta|\mathcal{M}) \) is propto Gaussian
**Bayesian Information Criterion: one subject**

\[
p(A|M) = \int d\theta \, p(A|\theta, M) \, p(\theta|M) \\
\approx p(A|\theta^{ML}, M)p(\theta^{ML}|M) \times \sqrt{(2\pi)^N|\Sigma|}
\]

- \(p(A|\theta) \, p(\theta|M)\) is proportional to Gaussian
- \(p(\theta|M) = \text{const.}\)
- Model doesn’t prefer particular
**Bayesian Information Criterion: one subject**

\[
p(A|M) = \int d\theta \ p(A|\theta, M) \ p(\theta|M) \\
\approx p(A|\theta^{ML}, M)p(\theta^{ML}|M) \times \sqrt{(2\pi)^N |\Sigma|} \]

\[
\log p(A|M) \approx \log p(A|\theta^{ML}, M) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi)
\]

\[p(A|\theta) \ p(\theta|M)\]

is proportional to Gaussian

\[p(\theta|M) = \text{const.}\]

Model doesn’t prefer particular
Bayesian Information Criterion: one subject

\[
p(A|M) = \int d\theta \, p(A|\theta, M) \, p(\theta|M) \\
\approx p(A|\theta^{ML}, M)p(\theta^{ML}|M) \times \sqrt{(2\pi)^N|\Sigma|} \\
\log p(A|M) \approx \log p(A|\theta^{ML}, M) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi)
\]

\[
p(\theta|M) \text{ is propto Gaussian}
\]

\[
p(\theta|M) = \text{const.}
\]

Model doesn’t prefer particular

\[
p(A|\theta) \, p(\theta|M)
\]

\[
\text{Bayesian Information Criterion: one subject}
\]
Bayesian Information Criterion: one subject

\[ p(A|M) = \int d\theta \, p(A|\theta, M) \, p(\theta|M) \]
\[ \approx p(A|\theta^{ML}, M) p(\theta^{ML}|M) \times \sqrt{(2\pi)^N |\Sigma|} \]

\[ \log p(A|M) \approx \log p(A|\theta^{ML}, M) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \]

\[ \Sigma_{ii} \propto \frac{1}{T} \Rightarrow \frac{1}{2} \log(|\Sigma|) \approx -\frac{N}{2} \log(T) \]

Bayesian Information Criterion (BIC)

\[ \approx -N \]

Akaike Information Criterion (AIC)
Bayesian Information Criterion: one subject

\[ p(A|\mathcal{M}) = \int d\theta \ p(A|\theta, \mathcal{M}) \ p(\theta|\mathcal{M}) \]
\approx p(A|\theta^{ML}, \mathcal{M}) p(\theta^{ML}|\mathcal{M}) \times \sqrt{(2\pi)^N |\Sigma|} \]

\[ \log p(A|\mathcal{M}) \approx \log p(A|\theta^{ML}, \mathcal{M}) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \]

\[ \Sigma_{ii} \propto \frac{1}{T} \Rightarrow \frac{1}{2} \log(|\Sigma|) \approx -\frac{N}{2} \log(T) \quad \text{Bayesian Information Criterion (BIC)} \]
\approx -N \quad \text{Akaike Information Criterion (AIC)} \]

Model fit vs Model complexity

\[ p(A|\theta) \ p(\theta|\mathcal{M}) \quad \text{is proportional to Gaussian} \]
\[ p(\theta|\mathcal{M}) = \text{constant.} \quad \text{Model doesn't prefer particular} \]
Group data

- Multiple subjects
- Multiple models
  - do they use the same model? If not parameters are not comparable
  - which model best accounts for all of them?
- Multiple groups
  - difference in models?
  - difference in parameters?
  - \(2^k\) possible model comparisons
- Multiple parameters
  - \(2^k\) possible correlations with any one psychometric measure
Group data - approaches

- **Summary statistic**
  - Treat individual model comparison measure as summary statistics, do ANOVA or t-test

- **Fixed effect analysis**
  - Subject data independent
    \[
    \log p(A|M) = \sum_i \log p(A_i|M) \\
    = \sum_i \log \int d\theta_i p(A_i|\theta_i)p(\theta_i|M_i) \approx -\frac{1}{2} \sum_i \text{BIC}_i
    \]

- **Random effects analyses**
  - Hierarchical prior on group parameters
    \[
    p(A|M) = \int d\zeta \int d\theta p(A|\theta) p(\theta|\zeta) p(\zeta|M)
    \]
  - Hierarchical prior on models
    \[
    p(A, M_k, r|\alpha) = p(A|M_k) p(M_k|r) p(r|\alpha)
    \]
Group-level likelihood

- Contains two integrals:
  - subject parameters
  - prior parameters

\[
p(A|\mathcal{M}) = \int d\theta \ p(A|\theta, \mathcal{M}) \int d\zeta \ p(\theta|\zeta) \ p(\zeta|\mathcal{M})
\]
Evaluating $p(A|M)$

\[
p(A|M) = \int d\theta \, p(A|\theta, M) \int d\zeta \, p(\theta|\zeta) \, p(\zeta|M)
\]
Evaluating \( p(A|M) \)

- **Two integrals**
  - tricky

\[
p(A|M) = \int d\theta \, p(A|\theta, M) \int d\zeta \, p(\theta|\zeta) \, p(\zeta|M)
\]
Evaluating $p(A|M)$

- **Two integrals**
  - tricky

- **Step by step: approximating levels separately**
  - Top level first:
Evaluating $p(A|M)$

- **Two integrals**
  - tricky

- **Step by step: approximating levels separately**
  - Top level first:

$$p(A|M) = \int d\theta p(A|\theta, M) \int d\zeta p(\theta|\zeta) p(\zeta|M)$$

$$p(A|M) = \int d\zeta p(A|\zeta, M) p(\zeta|M)$$
Evaluating $p(A|M)$

- Two integrals
  - tricky

- Step by step: approximating levels separately
  - Approximate at the top level
  - less action

\[
p(A|M) = \int d\theta \ p(A|\theta, M) \int d\zeta \ p(\theta|\zeta) \ p(\zeta|M)
\]

\[
p(A|M) = \int d\zeta \ p(A|\zeta, M) \ p(\zeta|M)
\]
Evaluating $p(A|M)$

- **Two integrals**
  - tricky
  \[
  p(A|M) = \int d\theta p(A|\theta, M) \int d\zeta p(\theta|\zeta) p(\zeta|M)
  \]

- **Step by step: approximating levels separately**
  - Approximate at the top level
  - less action
  \[
  p(A|\zeta, M) p(\zeta|M)
  \]
  \[
  p(A|M) = \int d\zeta p(A|\zeta, M) p(\zeta|M)
  \]
Evaluating $p(A|M)$

- Two integrals
  - tricky

- Step by step: approximating levels separately
  - Approximate at the top level
  - less action

$$p(A|M) = \int d\theta \ p(A|\theta, M) \int d\zeta \ p(\theta|\zeta) \ p(\zeta|M)$$

$$p(A|\zeta, M) \ p(\zeta|M)$$

is propto Gaussian

$$p(A|\zeta, M) \ p(\zeta|M)$$

$$p(A|\zeta_{ML}, M)p(\zeta_{ML}|M) \times \sqrt{(2\pi)^N|\Sigma|}$$
Evaluating $p(A|M)$

- Two integrals
  - tricky

- Step by step: approximating levels separately
  - Approximate at the top level
  - less action

$$p(A|M) = \int d\theta \ p(A|\theta, M) \int d\zeta \ p(\theta|\zeta) \ p(\zeta|M)$$

$$p(A|\zeta, M) \ p(\zeta|M)$$

is propto Gaussian

Model doesn’t prefer particular $\zeta$

$$p(A|\zeta^ML, M)p(\zeta^ML|M) \times \sqrt{(2\pi)^N|\Sigma|}$$
Evaluating $p(A|M)$

- **Two integrals**
  - tricky

- **Step by step: approximating levels separately**
  - Approximate at the top level
  - less action

\[
p(A|M) = \int d\theta \ p(A|\theta,M) \int d\zeta \ p(\theta|\zeta) \ p(\zeta|M)\]

\[
p(A|\zeta,M) \ p(\zeta|M)\]

\[
p(A|\zeta^{ML},M) p(\zeta^{ML}|M) \times \sqrt{(2\pi)^N |\Sigma|}\]

\[
\log p(A|M) \approx \log p(A|\zeta^{ML},M) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi)\]
Evaluating $p(A|M)$

- **Two integrals**
  - tricky

- **Step by step: approximating levels separately**
  - Approximate at the top level
  - less action

\[ p(A|M) = \int d\theta p(A|\theta, M) \int d\zeta p(\theta|\zeta) p(\zeta|M) \]

\[
p(A|\zeta, M) p(\zeta|M)
\]

\[
p(A|\zeta, M) p(\zeta|M)
\]

\[
\text{Model doesn't prefer particular } \zeta
\]

\[
\text{is proporto Gaussian}
\]

\[
\text{p(A|z, M) p(z|M)}
\]

\[
\text{is proporto Gaussian}
\]

\[
\text{Model doesn't prefer particular } \zeta
\]

\[
\text{p(A|z, M) p(z|M)}
\]

\[
\text{is proporto Gaussian}
\]

\[
\text{Model doesn't prefer particular } \zeta
\]

\[
\text{p(A|z, M) p(z|M)}
\]

\[
\text{is proporto Gaussian}
\]

\[
\text{Model doesn't prefer particular } \zeta
\]

\[
\text{p(A|z, M) p(z|M)}
\]

\[
\text{is proporto Gaussian}
\]

\[
\text{Model doesn't prefer particular } \zeta
\]

\[
\text{p(A|z, M) p(z|M)}
\]

\[
\text{is proporto Gaussian}
\]

\[
\text{Model doesn't prefer particular } \zeta
\]
Evaluating $p(A|M)$

- **Two integrals**
  - tricky
  \[ p(A|M) = \int d\theta \, p(A|\theta, M) \int d\zeta \, p(\theta|\zeta) \, p(\zeta|M) \]

- **Step by step: approximating levels separately**
  - Approximate at the top level
  - less action

\[ p(A|\zeta, M) \, p(\zeta|M) \]

\[ p(A|\zeta^{ML}, M) p(\zeta^{ML}|M) \times \sqrt{(2\pi)^N |\Sigma|} \]

\[ \log p(A|M) \approx \log p(A|\zeta^{ML}, M) + \frac{1}{2} \log(|\Sigma|) + \frac{N}{2} \log(2\pi) \]

\[ \text{just as before, top-level BIC} \]
Approximating level 1

- Still leaves the first level:
  - Approximate integral by sampling, e.g. importance sampling for few dimensions (<10)

\[
\log p(A|\zeta^{ML}, \mathcal{M}) = \log \int d\theta \, p(A|\theta) \, p(\theta|\zeta^{ML})
\]

\[
\approx \log \frac{1}{B} \sum_{b=1}^{B} p(A|\theta^b)
\]

\[
\theta^b \sim p(\theta|\zeta^{ML})
\]
Group-level BIC

$$\log p(A|M) = \int d\zeta p(A|\zeta) p(\zeta|M)$$

$$\simeq -\frac{1}{2} \text{BIC}_{\text{int}}$$

$$= \log \hat{p}(A|\hat{\zeta}^{ML}) - \frac{1}{2}|M| \log(|A|)$$
Example task

Go

Rewarded

Avoids loss

Nogo

Guitart-Masip, Huys et al. Submitted
**Example task**

- **Go rewarded**
  - Go to win

- **Nogo punished**
  - Go to avoid

- **Nogo rewarded**
  - Nogo to win

- **Go punished**
  - Nogo to avoid

![Graph showing probability correct for different conditions](image)

Guitart-Masip, Huys et al. Submitted
Model validation: generating data

Go rewarded
Go to win
Probability(Go)

Nogo punished
Go to avoid

Nogo rewarded
Nogo to win

Go punished
Nogo to avoid

Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. Submitted
Model validation: generating data

\[ p(go | s_t) \propto Q_t(go | s_t) + \text{bias}(go) \]
Model validation: generating data

$p(go|s_t) \propto Q_t(go|s_t) + \text{bias}(go) + V_t(s_t)$

$V_t(s_t) = V_{t-1}(s_t) + \epsilon(r_t - V_{t-1}(s_t))$

Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. Submitted
Model validation: generating data

P(go) \propto \text{value of stimulus}

\begin{align*}
p(go|s_t) & \propto Q_t(go|s_t) + \text{bias}(go) + V_t(s_t) \\
V_t(s_t) &= V_{t-1}(s_t) + \epsilon(r_t - V_{t-1}(s_t))
\end{align*}

Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. Submitted
Model comparison: overfitting?

- RW
- RW + noise
- RW + noise + Q0
- RW + noise + bias
- RW (rew/pun) + noise + bias
- RW + noise + bias + Pav

BIC[\text{int}]

Go rewarded
Go to win

Go punished
Go to avoid

Nogo rewarded
Nogo to win

Nogo punished
Nogo to avoid

Go
Nogo

Rewarded
Avoids loss

Probability(Go)

20 40 60
0
0.5
1

20 40 60
0
0.5
1

20 40 60
0
0.5
1

20 40 60
0
0.5
1
Model comparison: overfitting?

![Graph showing model comparison with BIC values](image)

- RW
- RW + noise
- RW + noise + Q0
- RW + noise + bias
- RW (rew/pun) + noise + bias
- RW + noise + bias + Pav

### Model Options
- Go rewarded
  - Go to win

- Nogo punished
  - Go to avoid

- Nogo rewarded
  - Nogo to win

- Go punished
  - Nogo to avoid
Model comparison: overfitting?

![Graph showing BIC values for different models ranging from RW to RW + noise + bias + Pav.]

- RW
- RW + noise
- RW + noise + Q0
- RW + noise + bias
- RW (rew/pun) + noise + bias
- RW + noise + bias + Pav

BIC_int

- Go rewarded
- Nogo punished
- Nogo rewarded
- Go punished

Probability(Go)

- Go to win
- Go to avoid
- Nogo to win
- Nogo to avoid

Zurich SPM Course, March 13-15 2013

Quentin Huys, TNU/PUK
Model comparison: overfitting?

Note: same number of parameters
How does it do?

\[ \beta \alpha \neq Q \]
\[ \beta \alpha \neq Q \]
\[ 2 \beta \alpha \neq \text{GoOnly} \]
\[ \beta \alpha \neq \text{GoOnly} \]
\[ \beta \alpha \neq \text{GoOnly} \]
\[ \beta \alpha \neq \text{GoOnly} \]
\[ \beta \alpha \neq \text{GoOnly} \]
\[ \beta \alpha \neq \text{GoOnly} \]
How does it do?

Fitted by EM... too nice?
Top-level Laplacian approximation

- Estimating the top-level determinant
  - using 2nd order finite differences

\[
\frac{d^2}{dh_{ij}^2} p(A|\zeta) \bigg|_{\zeta = \hat{\zeta}^{ML}} \approx \frac{1}{\delta^2} \left[ p(A|\hat{\zeta}^{ML} + \delta e_i) - 2p(A|\hat{\zeta}^{ML}) + p(A|\hat{\zeta}^{ML} - \delta e_j) \right]
\]

- the shifted likelihoods can be evaluated by shifting the samples.
Group level errors

**Beta**

- Prior Mean
- Group-level precision

- Increasing Subject # N
- Increasing Trials T

**Learning rate**

- Prior Mean
- True and estimates ± 95% CI

- Prior Variance
- True and estimates ± 95% CI

---

**Total # of observations (N x T)**
Posterior distribution on models

- Generative model for models

\[ \alpha \]
\[ r \sim \text{Dir}(r; \alpha) \]
\[ m_1 \sim \text{Mult}(m;1,r) \]
\[ y_1 \sim p(y_1 | m_1) \]

B

C

\[ \alpha \]
\[ r_k \]
\[ m_{nk} \]
\[ y_n \]

Stephan et al. 2009
Bayesian model selection - equations

- Write down joint distribution of generative model
- Variational approximations lead to set of very simple update equations
  - start with flat prior over model probabilities
    \[ \alpha = \alpha_0 \]
  - then update

\[
\begin{align*}
    u^i_k &= \left( \int d\theta \ p(A_i, \theta | M_k) \right) \exp \left( \Psi(\alpha_k) - \Psi \left( \sum_k \alpha_k \right) \right) \\
    \alpha_k &\leftarrow \alpha_{0,k} + \sum_i \frac{u^i_k}{\sum_k u^i_k}
\end{align*}
\]
Group Model selection

Integrate out your parameters
Questions in psychiatry I: regression

- Parametric relationship with other variables $\psi$
  - do standard second level analyses
  - can use Hessians to determine weights

$$ q_k(\theta) = \mathcal{N}(m_k, S_k) $$
$$ m_k \leftarrow \arg\max_{\theta} p(a_k | \theta) p(\theta | \zeta^{(i)}) $$
$$ S_k^{-1} \leftarrow \frac{\partial^2 p(a_k | \theta) p(\theta | \zeta^{(i)})}{\partial \theta^2} \bigg|_{\theta=m_k} $$

- better: compare two models

Model 1: $\prod_i p(A_i | \theta_i) p(\theta_i | \mu_0, \sigma)$
  - i.e. $\theta_i \sim \mathcal{N}(\mu_0, \sigma)$

Model 2: $\prod_i p(A_i | \theta_i) p(\theta_i | \mu_0, c, \sigma, \psi_i)$
  - i.e. $\theta_i \sim \mathcal{N}(\mu_0 + c\psi_i, \sigma)$
Regression

- Standard regression analysis:

\[ m_i = C r_i + \Sigma^{1/2} \eta \quad \forall i \]

- Including uncertainty about each subject's inferred parameters

\[ m_i = C r_i + (\Sigma^{1/2} + S_i^{1/2}) \eta \quad \forall i \]

- Careful: Finite difference estimates $S$ can be noisy!
  - regularize...
Questions in psychiatry II: group differences

- Do groups differ in terms of parameter(s)?
- **Cannot** compare parameters across different models
  - even very similar parameters can account for different effects
- For models with $k$ parameters, there are $2^k$ possible comparisons
  - multiple comparisons?
  - posterior over models (Stephan et al. 2009)
Group differences in parameters

- Are two groups similar in parameter $x$?
- ANOVA: compare likelihood of two means to likelihood of one global mean. Take degrees of freedom into account.
- But: this tries to account for the parameters with one or two groups, not for the data
- Compare models with separate or joint parameter & prior:

<table>
<thead>
<tr>
<th>Model 1</th>
<th>$\varepsilon$</th>
<th>$\beta_1, \beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 2</td>
<td>$\varepsilon$</td>
<td>$\beta$</td>
</tr>
</tbody>
</table>
Questions in psychiatry III: Classification

- Who belongs to which of two groups?
- How many groups are there?
Model comparison again

- What is ‘significant’?

\[
BF = \frac{p(A|M_1)}{p(A|M_2)} \frac{p(\Lambda < \eta)}{p(\Lambda < \eta)}
\]

<table>
<thead>
<tr>
<th>(\log_{10}(B_{10}))</th>
<th>(B_{10})</th>
<th>Evidence against (H_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 1/2</td>
<td>1 to 3.2</td>
<td>Not worth more than a bare mention</td>
</tr>
<tr>
<td>1/2 to 1</td>
<td>3.2 to 10</td>
<td>Substantial</td>
</tr>
<tr>
<td>1 to 2</td>
<td>10 to 100</td>
<td>Strong</td>
</tr>
<tr>
<td>&gt;2</td>
<td>&gt;100</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

Kaas and Raftery 95

- “Spread of effect” in group comparisons
  - Better model does not mean a behavioural effect is concentrated in one parameter
  - Obvious raw differences spread between parameters
Behavioural data modelling

- Are no panaceas
  - statistics about specific aspects of decision machinery
  - only account for part of the variance

- Model needs to match experiment
  - ensure subjects actually do the task the way you wrote it in the model
  - model comparison

- Model = Quantitative hypothesis
  - strong test
  - need to compare models, not parameters
  - includes all consequences of a hypothesis for choice
Modelling in psychiatry

› Hypothesis testing
  • otherwise untestable hypotheses
  • internal processes

› Limited by data quality
  • Look for strong behaviours, not noisy

› “Holistic” testing of hypotheses

› Marr’s levels
  • physical
  • algorithm
  • computational