Fitting a model: Maximum likelihood

Stefan Frässle

Translational Neuromodeling Unit (TNU)
University of Zurich & ETH Zurich

Computational Psychiatry Course (CPC) 2021
Tuesday, 14.09.2021

Models for the multi-armed bandit

model 1
Random choice

$$p_t^1 = b$$
$$p_t^2 = 1 - b$$

$$0 \leq b \leq 1$$

$$\boldsymbol{\theta} = \{\boldsymbol{b}\}$$

model 2
Noisy win-stay-lose-switch

$$p_t^k = \begin{cases} 1 - \frac{\varepsilon}{2} & \text{if } (c_{t-1} = k \text{ and } r_{t-1} = 1) \text{ OR } (c_{t-1} \neq k \text{ and } r_{t-1} = 0) \\ \frac{\varepsilon}{2} & \text{if } (c_{t-1} \neq k \text{ and } r_{t-1} = 1) \text{ OR } (c_{t-1} = k \text{ and } r_{t-1} = 0) \end{cases}$$

$$\boldsymbol{\theta} = \{\boldsymbol{\varepsilon}\}$$

model 3
Rescorla Wagner

$$Q_{t+1}^k = Q_t^k + \alpha\left(r_t - Q_t^k\right) \quad \text{and} \quad p_t^k = \frac{\exp(\beta Q_t^k)}{\sum_{i=1}^K \exp(\beta Q_t^i)}$$

$$\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$$

Wilson & Collins, 2019, *eLife*

Models for the multi-armed bandit

model 1
Random choice

$$p_t^1 = b \qquad p_t^2 = 1 - b$$

$$0 \le b \le 1$$

$$\boldsymbol{\theta} = \{\boldsymbol{b}\}$$

model 2
Noisy win-stay-lose-switch

$$p_t^k = \begin{cases} 1 - \frac{\varepsilon}{2} & \text{if } (c_{t-1} = k \text{ and } r_{t-1} = 1) \text{ OR } (c_{t-1} \neq k \text{ and } r_{t-1} = 0) \\ \frac{\varepsilon}{2} & \text{if } (c_{t-1} \neq k \text{ and } r_{t-1} = 1) \text{ OR } (c_{t-1} = k \text{ and } r_{t-1} = 0) \end{cases}$$

$$\boldsymbol{\theta} = \{\boldsymbol{\varepsilon}\}$$

model 3
Rescorla Wagner

$$Q_{t+1}^k = Q_t^k + \alpha\left(r_t - Q_t^k\right) \quad \text{and} \quad p_t^k = \frac{\exp(\beta Q_t^k)}{\sum_{i=1}^K \exp(\beta Q_t^i)}$$

$$\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$$

Wilson & Collins, 2019, *eLife*

Maximum likelihood estimator

Maximum likelihood estimator (MLE)

Given: Data: $Y = \{y_1, \ldots, y_T\}$

Assume: Family / set of distributions : $\{p_\theta : \theta \in \Theta\}$ and $Y = \{y_1, \ldots, y_T\}$ is sample from p_θ iid

Goal: Estimate the θ that the data $Y = \{y_1, \ldots, y_T\}$ comes from

Likelihood

$$\boxed{\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\big(p(Y|\theta, m)\big)}$$

where: $$\boxed{p(Y|\theta, m) = p(y_{1..T}|\theta, m) = \prod_{t=1}^{T} p(y_t|\theta, m)}$$

Intuitively: *"Maximum likelihood estimation finds the θ for which the acquired data is most likely"*

Maximum likelihood estimator

Maximum likelihood estimator (MLE)

Given: Data: $Y = \{y_1, \dots, y_T\}$

Assume: Family / set of distributions : $\{p_\theta : \theta \in \Theta\}$ and $Y = \{y_1, \dots, y_T\}$ is sample from p_θ iid

Goal: Estimate the θ that the data $Y = \{y_1, \dots, y_T\}$ comes from

Likelihood

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \big(p(Y|\theta, m)\big)$$

where: $$p(Y|\theta, m) = p(y_{1..T}|\theta, m) = \prod_{t=1}^{T} p(y_t|\theta, m)$$

Notes:

- MLE might not be unique
- MLE might not exist

SPECIFYING THE LIKELIHOOD FUNCTION

model 1
Random choice

$p_t^1 = b$

$p_t^2 = 1 - b$

$0 \leq b \leq 1$

$\boldsymbol{\theta} = \{\boldsymbol{b}\}$

For single trial t: $p(y_t|\theta, m) = \theta^{y_t}(1-\theta)^{(1-y_t)}$

Bernoulli distribution

SPECIFYING THE LIKELIHOOD FUNCTION

model 1
Random choice

$p_t^1 = b$

$p_t^0 = 1 - b$

$0 \leq b \leq 1$

$\boldsymbol{\theta} = \{\boldsymbol{b}\}$

For single trial t: $\boxed{p(y_t|\theta, m) = \theta^{y_t}(1-\theta)^{(1-y_t)}}$

Bernoulli distribution

For all trials $1..T$: $p(Y|\theta, m) = p(y_{1..T}|\theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$

$Y = \{y_1, \ldots, y_T\}$ iid

Maximizing the likelihood

$$p(Y|\theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

Likelihood

Analytical solution

- If $p(Y|\theta, m)$ is differentiable, use the derivative test.
- i.e.: set first derivative to 0
- For simple cases, this yields an explicit (analytical) solution

Numerical solution

- Use numerical routines to find the maximum of $p(Y|\theta, m)$.
- E.g.: fminsearch (MATLAB)

Maximizing the Likelihood

$$p(Y|\theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

Likelihood

Analytical solution

- If $p(Y|\theta, m)$ is differentiable, use the derivative test.
- i.e.: set first derivative to 0
- For simple cases, this yields an explicit (analytical) solution

Numerical solution

- Use numerical routines to find the maximum of $p(Y|\theta, m)$.
- E.g.: fminsearch (MATLAB)

Analytical solution to MLE

model 1
Random choice

$$p(Y|\theta,m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

$$\log\big(p(Y|\theta,m)\big) = \log\left(\prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}\right)$$

$$= \sum_{t=1}^{T} \log\big(\theta^{y_t}(1-\theta)^{(1-y_t)}\big)$$

$$= \sum_{t=1}^{T} \big(y_t \log(\theta) + (1-y_t)\log(1-\theta)\big)$$

$$\frac{d}{d\theta}\log\big(p(Y|\theta,m)\big) = \frac{d}{d\theta}\sum_{t=1}^{T} \big(y_t \log(\theta) + (1-y_t)\log(1-\theta)\big) \overset{!}{=} 0$$

ANALYTICAL SOLUTION TO MLE

model 1
Random choice

$$\frac{d}{d\theta}\sum_{t=1}^{T}(y_t\log(\theta)+(1-y_t)\log(1-\theta))=0$$

$$\left(\frac{d}{d\theta}\log(\theta)\right)\left(\sum_{t=1}^{T}y_t\right)+\left(\frac{d}{d\theta}\log(1-\theta)\right)\left(\sum_{t=1}^{T}(1-y_t)\right)=0$$

$$\frac{1}{\theta(1-\theta)}\left(\sum_{t=1}^{T}y_t-\theta\sum_{t=1}^{T}y_t-\theta T+\theta\sum_{t=1}^{T}y_t\right)=0$$

$$\sum_{t=1}^{T}y_t-\theta T=0$$

Maximum likelihood estimator

$$\boxed{\theta_{MLE}=\frac{1}{T}\sum_{t=1}^{T}y_t}$$

Maximizing the likelihood

$$p(Y|\theta, m) = \prod_{t=1}^{T} \theta^{y_t}(1-\theta)^{(1-y_t)}$$

Likelihood

Analytical solution

- If $p(Y|\theta, m)$ is differentiable, use the derivative test.
- i.e.: set first derivative to 0
- For simple cases, this yields an explicit (analytical) solution

Numerical solution

- Use numerical routines to find the maximum of $p(Y|\theta, m)$.
- E.g.: fminsearch (MATLAB)

Advantages of MLE

Advantages of MLE:

- Easy to compute
- Interpretable
- Desirable asymptotic properties
  - Consistent: $\theta_{MLE} \xrightarrow{p} \theta_0$
  - Normal: $\sqrt{n} \cdot (\theta_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I^{-1})$
  - Statistically efficient: MLE reaches Cramér-Rao bound
- Invariant to reparameterization: if θ_{MLE} is a MLE for θ, then $g(\theta_{MLE})$ is a MLE for $g(\theta)$

Limitations of MLE

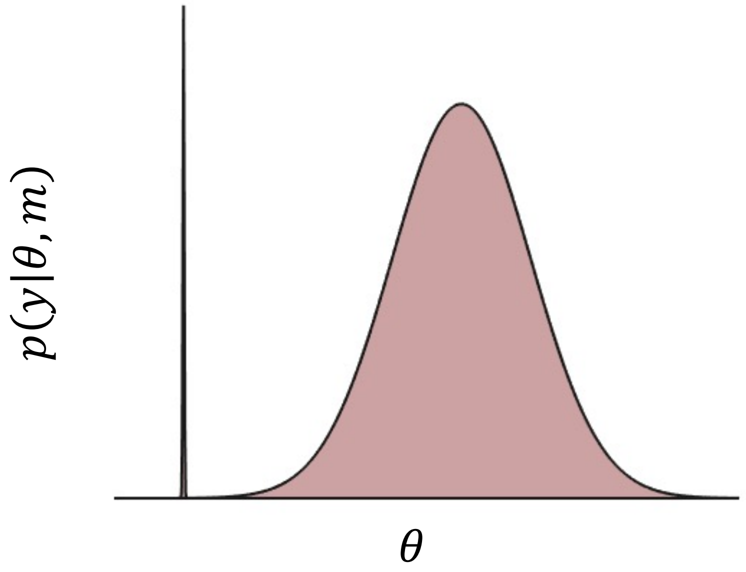
Limitations of MLE:

- Existence & uniqueness of MLE is not guaranteed
- MLE is a point estimate and therefore has no representation of uncertainty

LIMITATIONS OF MLE

Limitations of MLE:

- Existence & uniqueness of MLE is not guaranteed
- MLE is a point estimate and therefore has no representation of uncertainty
- MLE might not be representative of the likelihood function



Limitations of MLE

Limitations of MLE:

- Existence & uniqueness of MLE is not guaranteed
- MLE is a point estimate and therefore has no representation of uncertainty
- MLE might not be representative of the likelihood function
- Overfitting

black swan paradox

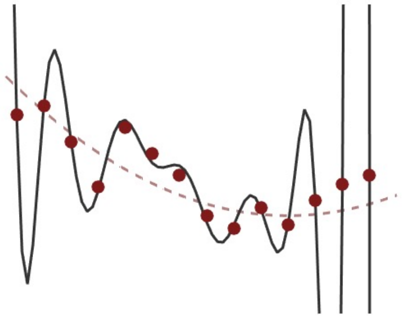
"Maximum likelihood estimation finds the θ for which the acquired data is most likely"

LIMITATIONS OF MLE

Limitations of MLE:

- Existence & uniqueness of MLE is not guaranteed
- MLE is a point estimate and therefore has no representation of uncertainty
- MLE might not be representative of the likelihood function
- Overfitting

black swan paradox



linear regression

Limitations of MLE

Limitations of MLE:

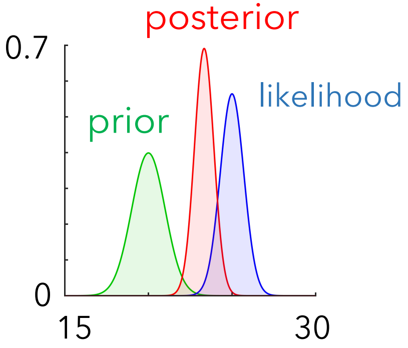
- Existence & uniqueness of MLE is not guaranteed
- MLE is a point estimate and therefore has no representation of uncertainty
- MLE might not be representative of the likelihood function
- Overfitting
- …

ALTERNATIVES

Reverend Thomas Bayes
(1702-1761)

- Alternatives: Bayesian statistics

$$\underbrace{p(\theta|y,m)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta,m)}^{\text{likelihood}}\overbrace{p(\theta|m)}^{\text{prior}}}{\underbrace{p(y|m)}_{\text{model evidence}}}$$

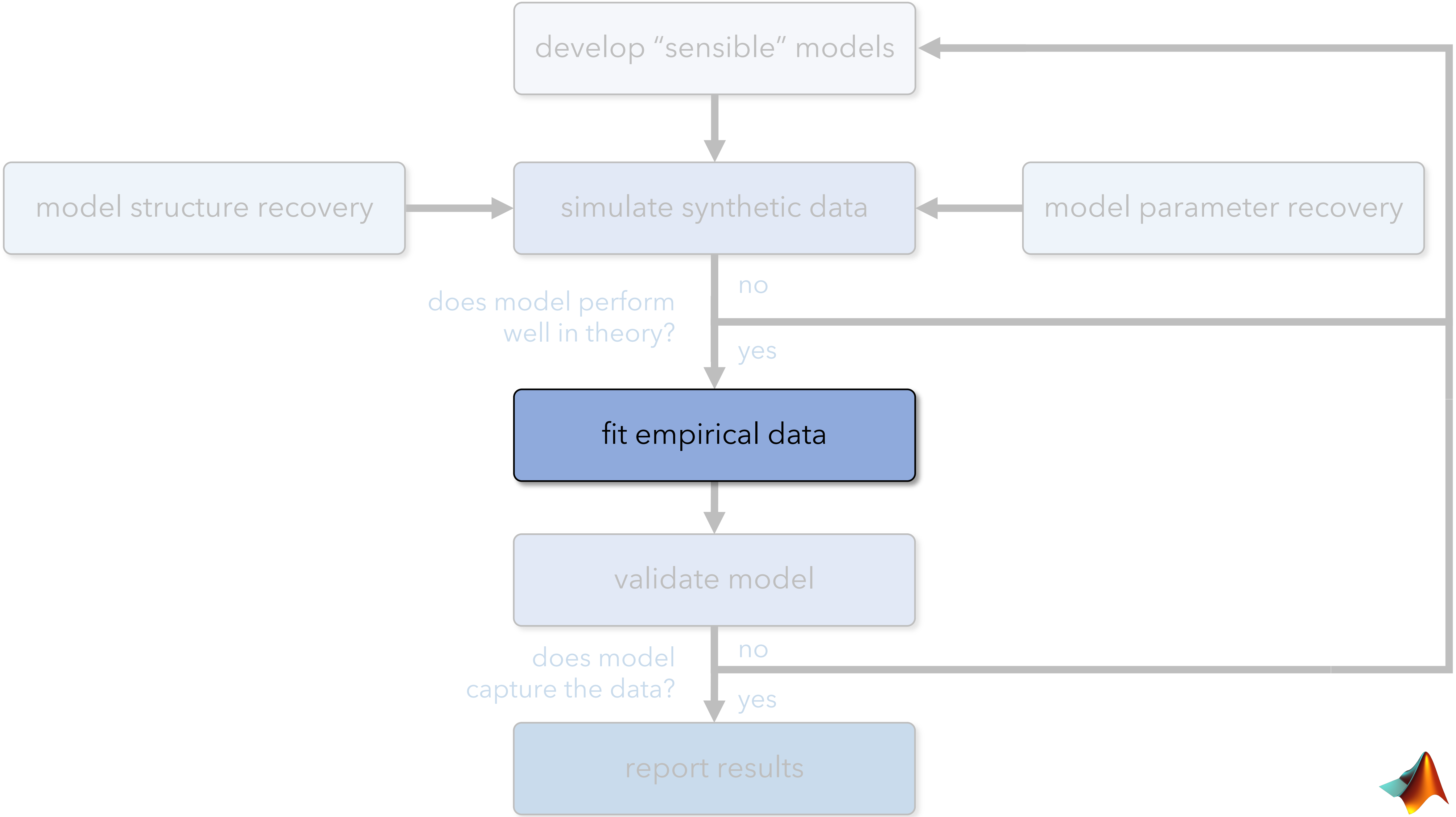


- Point estimates of the posterior can be obtained using maximum-a-posteriori (MAP) estimation

- Note: Under a flat prior, the MAP becomes equal to the MLE

- To obtain full posterior densities, one can resort to Variational Bayesian (VB) or sampling-based (Markov Chain Monte Carlo) techniques

Model inversion: Lecture (*Today, next talk*)



develop "sensible" models
model structure recovery
simulate synthetic data
model parameter recovery
does model perform well in theory?
no
yes
fit empirical data
validate model
does model capture the data?
no
yes
report results

THANK YOU FOR YOUR ATTENTION !

Stefan Frässle

Translational Neuromodeling Unit (TNU)
University of Zurich & ETH Zurich

Email: stefanf@biomed.ee.ethz.ch
Phone: +41 44 634 91 14
@stefan_fraessle