Bayesian Statistics

Lecture 04 - Hierarchical models

Radboud University
Artificial Intelligence

For today: generative, graphical and hierarchical models

- Tools for reasoning with (multiple!) random variables.
- MCMC does the computations 'under the hood'.
- Note: order of topics is a bit different than in the book.

Generative models

Conceptual models for statistics

- Often, we save the technical details for later and focus on a conceptual level.
- Important: which parameter depends on which other parameter?
- Through which distribution is this dependency / what are the properties of the variable?

The '$\sim$' symbol

- The symbol $\sim$ means 'follows the distribution'. It is used to quickly define a model:

$$\begin{aligned} y_i &\sim \mathrm{Bernoulli}(\theta) \\ \theta &\sim \mathrm{beta}(a, b) \end{aligned} \quad ,$$

and means that $p(y_i|\theta)$ is a Bernoulli distribution and $p(\theta|a, b)$ a beta distribution.

Plate notation

Generative model

$$y \sim \mathrm{Bernoulli}(\theta)$$

$$\theta \sim \mathrm{beta}(a, b)$$

- Shaded variables are *observed*.
- Unshaded variables are *latent* (not observed, but we're interested in them).
- Small circles indicate *hyperparameters*: these are fixed and *not learned*.
- Also known as *plate notation*.

Graphical model

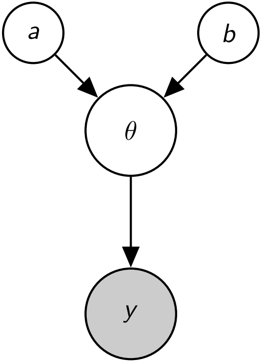
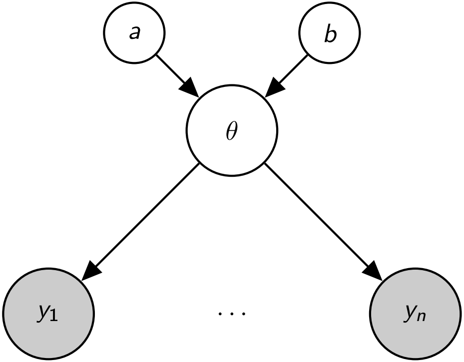


Plate notation



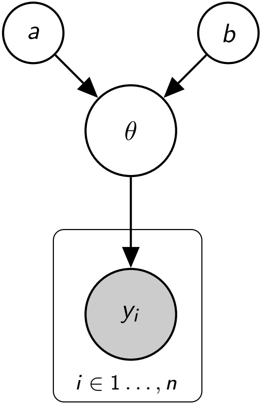


Plate notation and graphical models

Bayesian thinking

- The arrows indicate the dependency: $x \rightarrow y$ implies we specify $p(y|x)$.
- Graphical models and generative models allow you to 'conceptually' think about Bayesian models.
- Implementation is 'just' a technicality (see also next week).

Generative view

- The (graphical+generative) model specifies how you would *generate* **random** data.
- Learning the posterior means that this random data is as close to the real data as it can be (given the model!).
- For next week (and all big exercises): in a script language you simply write down the generative model.

Hierarchical models

We'll get back to generating data in a second.

- So far, we considered models of the form $p(\theta|D)$, with for example $p(\theta|a, b)$.
- Where do a and b come from? Can we learn them too? What priors do we need?
- In realistic problems, variables form a *hierarchy*.

Hierarchical models

Example: modeling text (e.g. news articles)

- Words are not distributed equally over all documents.
- The word ‘football’ occurs often in sports articles, but rarely in weather reports.
- Word counts in a document have parameter θ_{topic} (how frequent is a word given a topic) and
- topic distributions have a parameter θ_{doc} (how frequent each is a topic given a document).

Hierarchical models

- Additional parameters simply extend the parameter space, so using Bayes' Rule:

$$
\begin{aligned}
p(\theta, \phi|D) &= p(D|\theta, \phi)p(\theta, \phi)/p(D) \\
&= p(D|\theta)p(\theta|\phi)p(\phi)/p(D) \ ,
\end{aligned}
$$

  using the chain rule of probabilities.
- We have now factored the posterior in such a way that *if we know* θ, the data D are independent from 'higher level' parameters ϕ.
- If the model can be factored this way, we call it an *hierarchical model*.
- N.B.: for example the beta distribution $p(\theta|a, b)$ itself is *not* hierarchical; a and b are not independent from each other.
- Hierarchical models are useful for our understanding; equivalent mathematical models may not be hierarchical.

Simple example of hierarchy

- In $p(x|\theta)p(\theta|\phi)p(\phi)/p(x)$, let x be your measured height, θ the average height in your country and ϕ your country.
- If the average height in your country θ is known, your height is *independent* of ϕ!
- Meaning we there is no influence of ϕ on x, *over* what is transferred through θ.

Running example: coins from different factories

Extending the beta-Bernoulli model

- Recall (again. . .) the Bernoulli likelihood for coin flips:

$$y_i \sim \mathrm{Bernoulli}(\theta)$$

- and the beta prior:

$$\theta \sim \mathrm{beta}(a, b)$$

- Or, with different interpretation:

$$\theta \sim \mathrm{beta}(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1) \ ,$$

where $\omega = (a - 1)/(a + b - 2)$ (mode) and $\kappa = a + b$ (prior certainty).

Priors on priors

- Parameters *a* and *b* were *pseudocounts*: imagined prior #heads and #tails.

$$\theta \sim \mathrm{beta}(\omega(\kappa-2)+1, (1-\omega)(\kappa-2)+1)$$

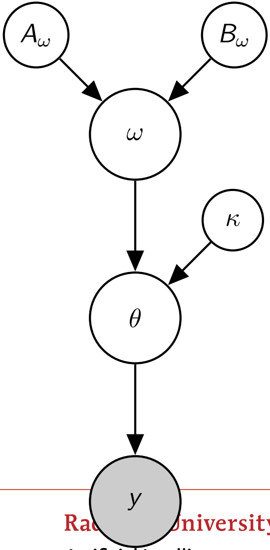
- Parameters ω and κ specify mean and precision;
  - ω is an expectation; θ will likely be near ω.
  - κ tells us how certain we are about that; low κ means θ can vary a lot, large κ mean $\theta \approx \omega$.
- **Introduction of hierarchy:** ω is a value that we can estimate; we estimate **both** the coin flip probability θ, but also its tendency ω.
- To infer a posterior over ω as well, we need a prior for it too, e.g.:

$$p(\omega) = \mathrm{beta}(A_\omega, B_\omega) \quad ,$$

where A_ω, B_ω are again hyperparameters.

Hierarchical model for coin flips

$$
\begin{aligned}
\omega &\sim \mathrm{beta}(A_\omega, B_\omega) \\
\theta &\sim \mathrm{beta}(\omega(\kappa-2)+1, \\
&\qquad (1-\omega)(\kappa-2)+1) \\
y_i &\sim \mathrm{Bernoulli}(\theta)
\end{aligned}
$$



Applying Bayes' Rule to an hierarchical model

- We'll do what we always do, taking into account which variables are independent:

$$p(\theta, \omega | y) = \frac{p(y|\theta, \omega)p(\theta, \omega)}{p(y)} \quad .$$

- But note: $p(y|\theta, \omega) = \mathrm{Bernoulli}(\theta)$, which does not depend on ω! So:

$$p(\theta, \omega | y) = \frac{p(y|\theta)p(\theta, \omega)}{p(y)} \quad . \tag{1}$$

- And recall that $p(\theta, \omega) = p(\theta|\omega)p(\omega)$, so:

$$p(\theta, \omega | y) = \frac{p(y|\theta)p(\theta|\omega)p(\omega)}{p(y)} \quad . \tag{2}$$

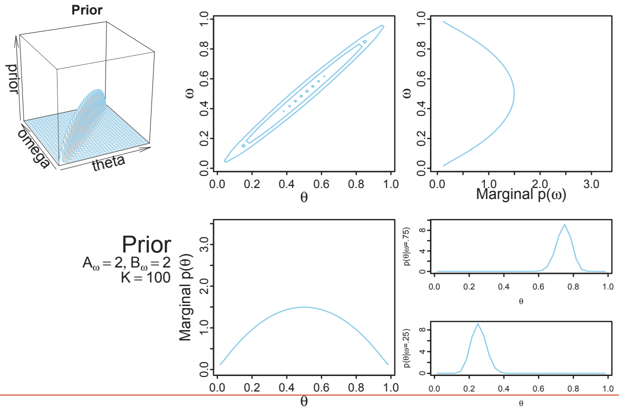
- Now all the numerator terms have conceptual meaning!

Advantages of hierarchical models

- In hierarchical models, independence between particular parameters is crucial.
- This enables more intuitive interpretation for us.
- ...as well as more efficient approximate inference.
- Larger models are often not (entirely) conjugate.

Prior using grid approximation

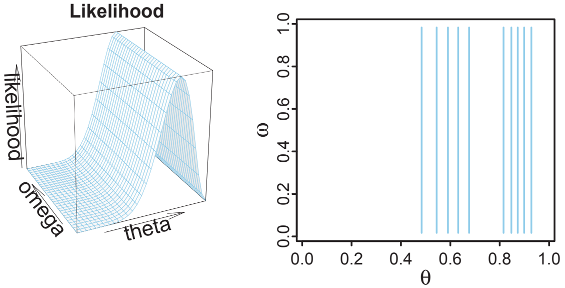
Compute $p(\theta|\omega)p(\omega)$ at discrete points (grid) and normalize by their sum.
Note: $A_\omega = 2, B_\omega = 2, K = 100$ means we are uncertain about ω, but very certain that θ is close to ω.



Radboud University
Artificial Intelligence

Likelihood using grid approximation

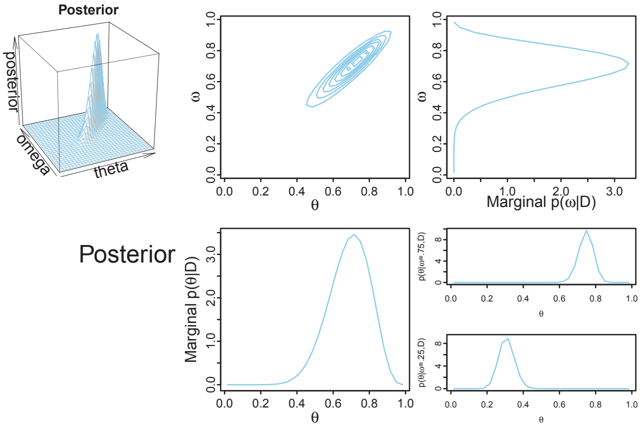
Compute $p(D|\theta)$ at discrete points (grid) and normalize by their sum.



Likelihood
D = 9 heads, 3 tails

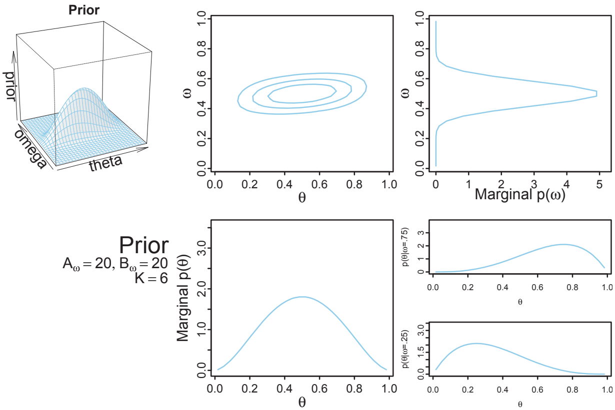
Posterior using grid approximation

Compute $p(D|\theta)p(\theta|\omega)p(\omega)$ at discrete points (grid) and normalize by their sum.



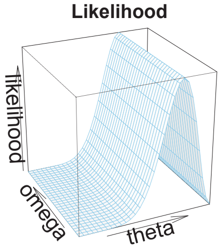
Prior using grid approximation

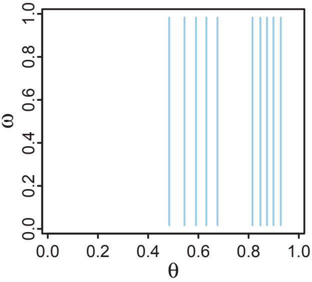
Compute $p(\theta|\omega)p(\omega)$ at discrete points (grid) and normalize by their sum.
Note: $A_\omega = 20, B_\omega = 20, K = 6$ means $\omega \approx 0.5$, but θ may vary.



Likelihood using grid approximation

Compute $p(D|\theta)$ at discrete points (grid) and normalize by their sum.

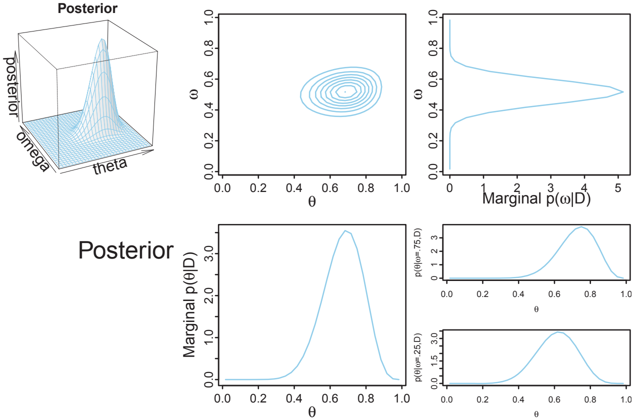




Likelihood
D = 9 heads, 3 tails

Posterior using grid approximation

Compute $p(D|\theta)p(\theta|\omega)p(\omega)$ at discrete points (grid) and normalize by their sum.

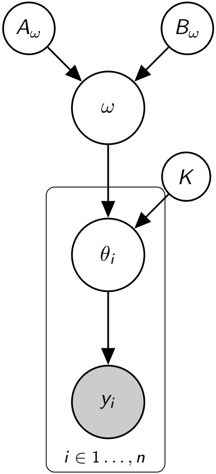


Group- and individual level parameters

- It gets interesting if *multiple* (e.g. subject specific) θ_i depend on *shared* (group-level) ω.
- Allows for modeling of group effects using measurements from individuals (= the realistic case).

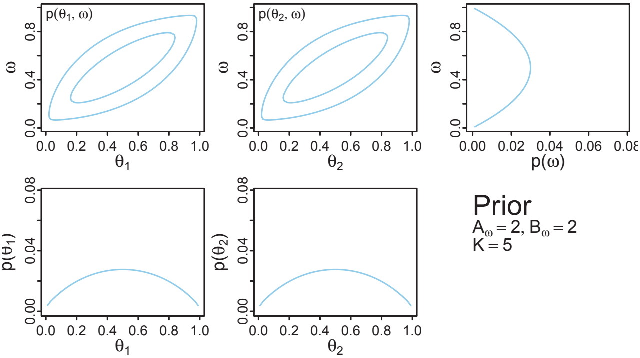
Coins, coins, coins

- Each coin i has $p(\mathrm{heads}|\omega)$.
- ω gives the group-level tendency.



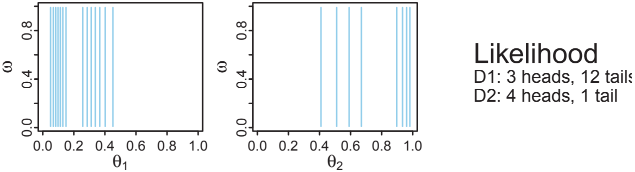
An hierarchical model for two coins

Prior: $p(\theta_1|\omega)p(\theta_2|\omega)p(\omega)$



An hierarchical model for two coins

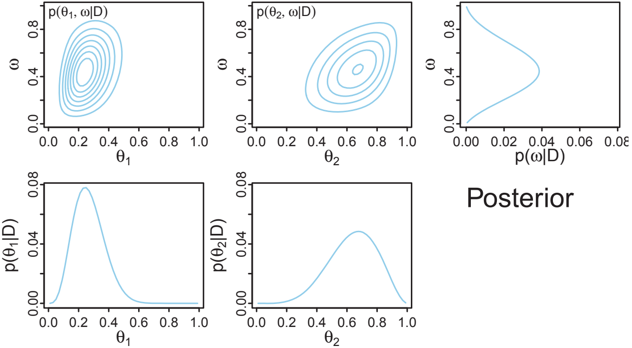
Likelihood: $p(D_1|\theta_1)p(D_2|\theta_2)$



An hierarchical model for two coins

Posterior:

$$p(\theta_1, \theta_2, \omega | D_1, D_2) = \frac{p(D_1|\theta_1)p(D_2|\theta_2)p(\theta_1|\omega)p(\theta_2|\omega)p(\omega)}{p(D_1, D_2)} \qquad (3)$$

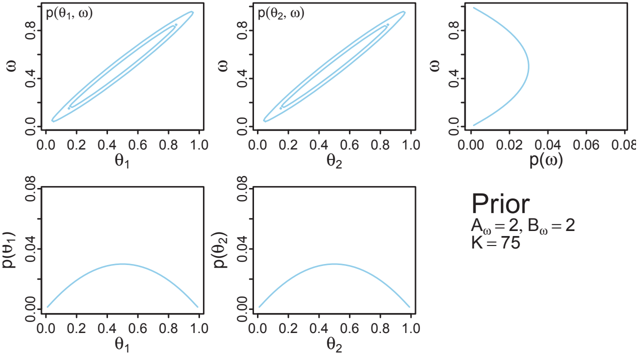


Computation for hierarchical models

- Nothing changes: The Bayesian machinery remains the same, we just have a few more terms to multiply.
- However: Additional observations can be *coupled* in such a model, giving more statistical power.

An hierarchical model for two coins

Prior: $p(\theta_1|\omega)p(\theta_2|\omega)p(\omega)$

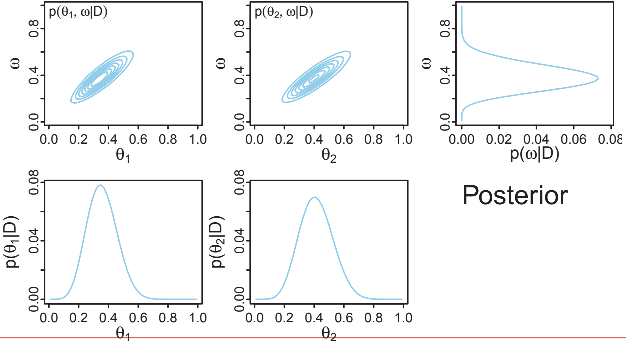


An hierarchical model for two coins

Posterior:

$$p(\theta_1, \theta_2, \omega | D_1, D_2) = \frac{p(D_1|\theta_1)p(D_2|\theta_2)p(\theta_1|\omega)p(\theta_2|\omega)p(\omega)}{p(D_1, D_2)} \qquad (4)$$

Recall D_2: 4 heads, 1 tail. D_1 has a strong influence on θ_2 through ω.



Radbound University
Artificial Intelligence

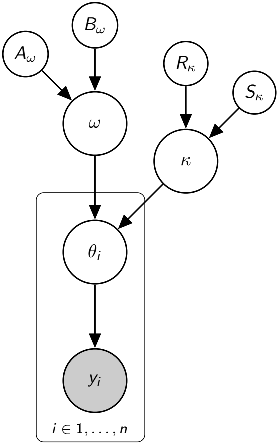
A less-trivial example

How much does θ_i depend on ω?

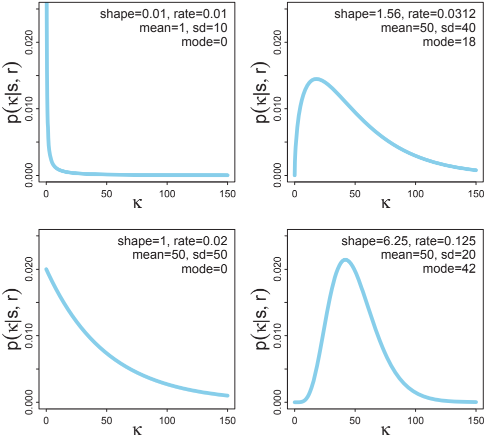
- Recall the prior on
  $\theta_i \sim \mathrm{beta}(\omega(\kappa-2)+1, (1-\omega)(\kappa-2)+1))$.
  κ was fixed!
- If all θ_i are similar, ω has a strong influence.
- If all θ_i are different, ω has a weak influence.
- So can you guess what we need..?

Introducing the Gamma distribution

- $\kappa \sim \mathrm{Gamma}(s, r)$ with
  $\mathrm{Gamma}(\kappa|r, s) = (r^s/\Gamma(s))\kappa^{s-1}e^{-r\kappa}$.
- **Continuous** distribution over $[0, \infty)$.
- Note: $\Gamma(x) = (x-1)!$



Examples of the Gamma distribution for different rate r and shape s



The parametrization of the Gamma distribution

- We have $\kappa \sim \mathrm{Gamma}(s, r)$ with $\mathrm{Gamma}(\kappa|r, s) = (r^s/\Gamma(s))\kappa^{s-1}e^{-r\kappa}$.
- Easier to understand through mean, mode and standard deviation:
  - $\mu = s/r$,
  - $\omega = (s-1)/r$ and
  - $\sigma = \sqrt{s}/r$.
- Rearranging terms gives $s = \mu^2/\sigma^2$ and $r = \mu/\sigma^2$ and
- $\ldots s = 1 + \omega r$ where $r = \frac{\omega + \sqrt{\omega^2 + 4\sigma^2}}{2\sigma^2}$.
- Useful if you know the (a priori) mean/mode/standard deviation.

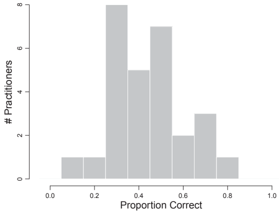
Bigger example: therapeutic touch

Claim: practitioners of ‘therapeutic touch’ can ‘sense a body’s energy field’ (claim investigated by Rosa et al., 1998).

- Experimenter holds out hand above either left or right hand of participant (shielded by a screen), at random.
- Goal for participant: determine above which hand another hand was held.
- Trivia: Experimenter (and co-author) was 9 years old.

Experimental setup

- 10 trials per participant, 21 participants, 7 re-tests a year apart, so 28 measurement sessions.
- Chance performance: 50%.



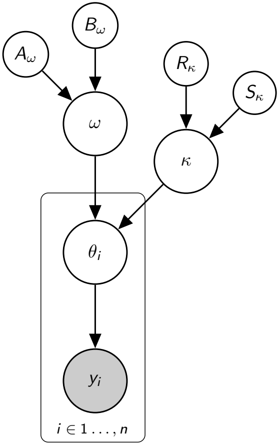
Bigger example: therapeutic touch

Research questions

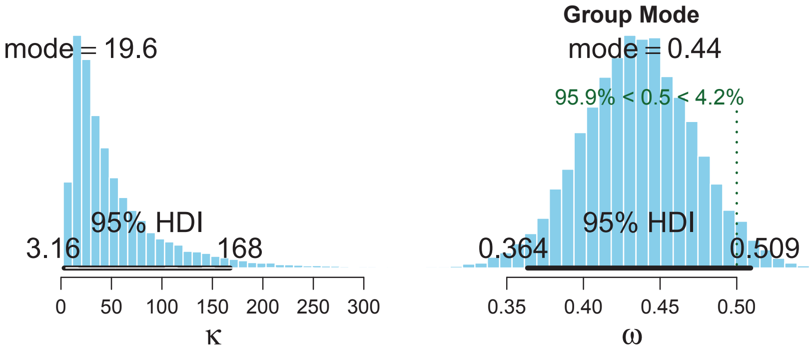
- How much did the group differ from chance performance?
- How much did any individual differ from chance performance?

Parameter estimation

- 30 parameters to estimate: θ_i for 28 participants and (shared) ω and κ.



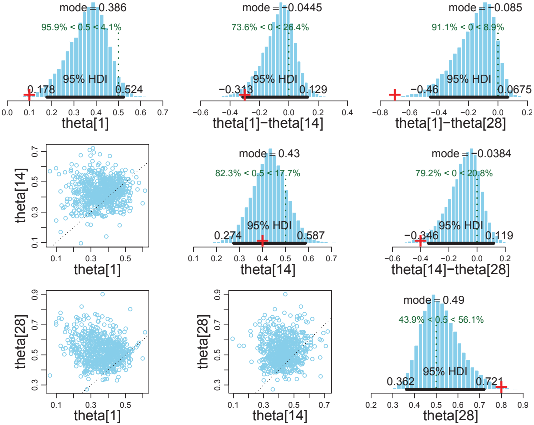
Parameter estimation (using Gibbs) for therapeutic touch data



Note: the 95% HDI **includes** chance level performance!

Parameter estimation (using Gibbs) for therapeutic touch data

Performance of participants 1, 14 and 28 (sorted by % correct).



Shrinkage

- In hierarchical models, lower-level parameters are *pulled closer together*.
- Parameters θ_i are 'averaged' between likelihood contribution z_i/N_i and group-level (prior) parameter ω.
- This effect is known as **shrinkage** and is usually desired (but be aware of it).
- The fewer data are available for a parameter, the more effect shrinkage will have, as the prior has more influence.
- Shrinkage follows from *hierarchical modelling*, not from Bayesian statistics itself!

Taking a step back

- So what model should I use? How deep does the ~~rabbit hole~~ hierarchy go?
- A model is only a *choice* and is dictated by context. The parameters have meaning only within that context.
- At some point, additional depth of the model doesn't explain additional variance.
- In a later lecture we'll look at *comparing* different models in detail.

Using (hierarchical) graphical models to simulate data

Generating data is a good way to confirm that the model makes sense.
Example

- Start with hyperparameters.
- Using the hyperparameters, draw values from the prior, e.g. $\theta \sim \mathrm{beta}(a = 1, b = 1)$.
- With *that* θ, generate random data using $y \sim \mathrm{Bernoulli}(\theta)$.
- For more complex data/models, the generated result may be very different than the observations!

To-read & to-do

- Hierarchical models: Kruschke, sections 9.1, 9.2, 9.3 and 9.5.
- Exercise 04 on Blackboard.