Bayesian Modeling

Max Kochurov

Moscow State University

Lecture 2

Agenda

Sampling from a distribution

Conjugate models

- Limited set of applications
- Lack of flexibility
- They are scalable

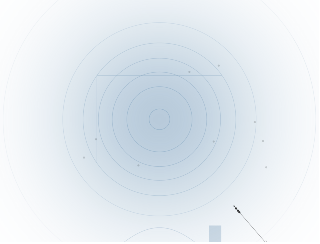


Figure: Easy distribution

Most models

- No closed form solution
- Posterior distributions is complicated
- Less scalable
- Flexible

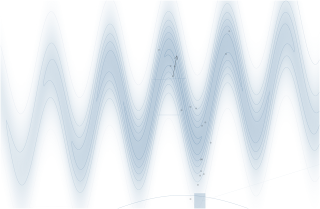


Figure: Complicated distribution

Hamiltonian Monte Carlo Intuition

HMC samples from a complicated distribution

1. Ideas from physics
2. Requires gradient
3. Requires numerical integration

Tuned HMC converges to the target distribution

Warning

I promised a not math heavy course. But this is important for debugging your models.

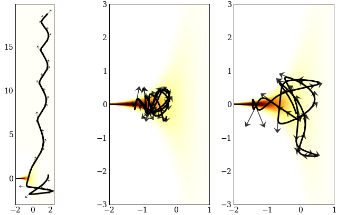


Figure: Leapfrog Integration

HMC Distributions

- $p(\Theta)$ - Target distribution, $\Theta \in \mathbb{R}^d$ (Θ aka **Position**)
- $p(\Delta \mid \Theta)$ - Momentum distribution, $\Delta \in \mathbb{R}^d$ (Δ aka **Velocity**)

Hamiltonian

$$H(\Delta, \Theta) = -\log p(\Delta, \Theta)$$

Notes

- $p(\Delta \mid \Theta) = \text{Normal}(0, M)$, usually a Normal distibution
- Δ and Θ have same dimensions

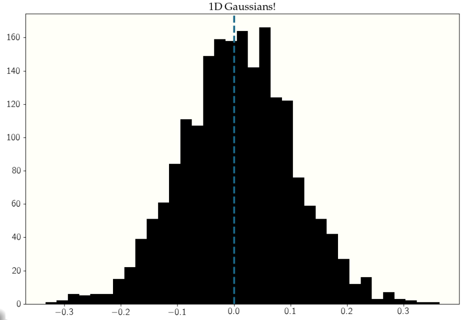


Figure: $p(\Theta) = \text{Normal}(0, 1)$

HMC Differential Equation

$$
\begin{aligned}
H(\Delta, \Theta) &= -\log p(\Delta, \Theta) \\
&= -\log p(\Delta \mid \Theta) \quad -\log p(\Theta) \\
&= \underbrace{K(\Delta, \Theta)}_{\text{Kinetic E}} \quad + \underbrace{V(\Theta)}_{\text{Potential E}}
\end{aligned}
$$

The Physical **motion** equation

$$
\frac{\partial \Theta}{\partial t} = \frac{\partial H}{\partial \Delta}
$$

$$
\frac{\partial \Delta}{\partial t} = -\frac{\partial H}{\partial \Theta}
$$

Motion preserves total energy $H(\Delta, \Theta)$

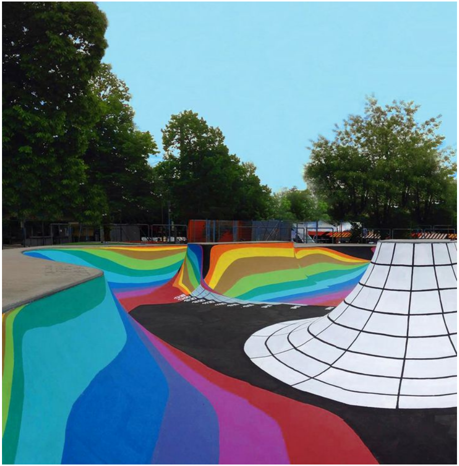


Figure: HMC analogy to skateboarding

HMC Divergences

A divergence is a huge integration error solving the differential equation.

When HMC Fails

Bad geometry for Hamiltonian

Bad geometry comes from a lot of things

1. Strong correlations
2. Narrow funnels in the posterior
3. Strong likelihood
4. Non homogeneous posterior

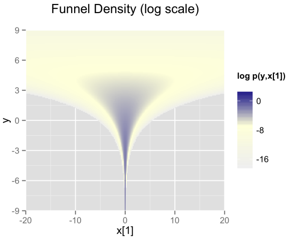


Figure: Neal's Funnel

HMC Reading Materials

Advanced Reading

1. Interactive Demo
2. A tutorial from Colin Carroll
3. A paper from Michael Betancourt
4. NUTS paper from Matthew D. Hoffman, Andrew Gelman

Example

Toy example - Cobb-Douglas

You should all know the Cobb-Douglas function

$$Y \approx A \cdot L^{\beta}$$

In our example:

1. data has 6 groups (hierarchical)
2. We know the groups
3. We know the total factor productivity A is different per group (different equipment)
4. Labour productivity β does not differ much

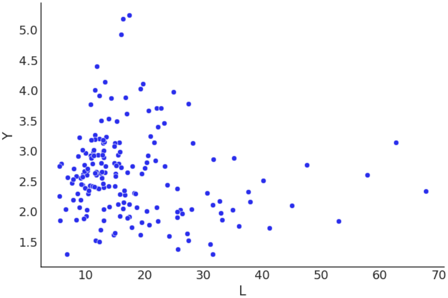


Figure: Example Data (aggregated)

Toy example - Carpet Knitters

Let's put more interpretation in the example

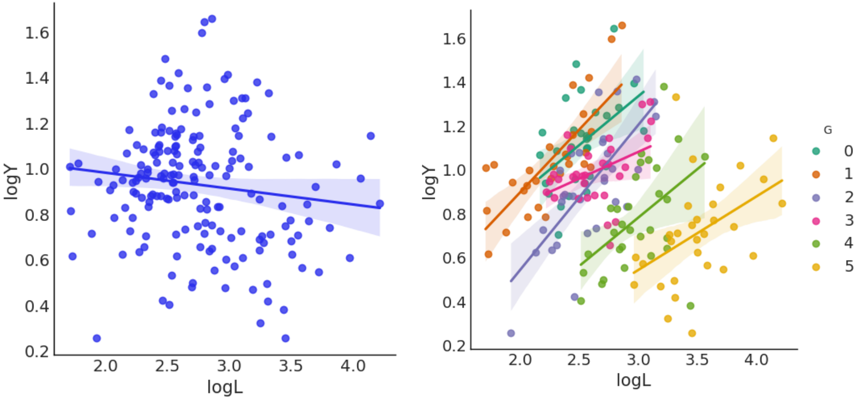
$$Y_g \approx A_g \cdot L^\beta$$

In our example we have a carpet manufacturing plant with 6 workers:

1. Workers make different carpets, thus have total factor productivity A
2. Labour productivity β is like concentration, the more you work the less productive you are
3. Workers produce carpets individually

Figure: Example Y

The Simpson Paradox



One group model

Best practices when you start.

- Start with a most simple model
- Make sure simple model converges well
- Write a more complex model

One group model

Best practices when you start.

- Start with a most simple model
    - You have groups, start with one
    - Make sure priors are well specified, do checks
- Make sure simple model converges well

- Write a more complex model

One group model

Best practices when you start.

- Start with a most simple model
    - You have groups, start with one
    - Make sure priors are well specified, do checks
- Make sure simple model converges well
    - If one group model fails, all fail
    - There are simple checks to verify your model samples well
- Write a more complex model

One group model

Best practices when you start.

- Start with a most simple model
    - You have groups, start with one
    - Make sure priors are well specified, do checks
- Make sure simple model converges well
    - If one group model fails, all fail
    - There are simple checks to verify your model samples well
- Write a more complex model
    - Try several parametrizations
    - Check how model samples
    - Compare models (out of scope for now)

Starting with a simple model

To get an idea why we start simple

$$Y_{g=0} \approx A_{g=0} \cdot L^{\beta}$$

1. What is prior for A?
2. What is prior for β?
3. What is prior predictive for $Y_{g=0}$?

Writing a model

$$Y_{g=0} \approx A_{g=0} \cdot L^{\beta}$$
$$\log Y_{g=0} \approx \log A_{g=0} + \log L \cdot \beta$$

Introducing distributions

$$\begin{aligned}
\log Y_{g=0} &\sim \text{Normal}(\log A_{g=0} + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim ??? \\
\beta &\sim ??? \\
A_{g=0} &\sim ???
\end{aligned}$$

Prior for β

What is a reasonable prior for labour productivity (elasticity) β? Questions to ask yourself

1. Can it be < 0?
2. Can it be large?
3. Can it be > 1?

Prior for β

What is a reasonable prior for labour productivity (elasticity) β? Questions to ask yourself

1. Can it be < 0? No
2. Can it be large?
3. Can it be > 1?

Prior for β

What is a reasonable prior for labour productivity (elasticity) β? Questions to ask yourself

1. Can it be < 0? No
2. Can it be large? No
3. Can it be > 1?

Prior for β

What is a reasonable prior for labour productivity (elasticity) β? Questions to ask yourself

1. Can it be < 0? No
2. Can it be large? No
3. Can it be > 1? No

Prior for β

What is a reasonable prior for labour productivity (elasticity) β? Questions to ask yourself

1. Can it be < 0? No
2. Can it be large? No
3. Can it be > 1? No

Conclusion: It is bounded by $(0, 1)$

The prior is subjective!

Who can argue these bounds do not make sense?

Prior for β

What is a reasonable prior for labour productivity (elasticity) β? Questions to ask yourself

1. Can it be < 0? No
2. Can it be large? No
3. Can it be > 1? No

Conclusion: It is bounded by $(0, 1)$

The prior is subjective!

Who can argue these bounds do not make sense?

Not yet a prior

To get a prior we need a distribution that fits the reasoning

Prior for β

What we know:

- $\beta \in (0, 1)$
- Less probable to be close to the boundary
- Nothing specific about exact value in the range.

In the mind

Enumerate possible distributions that fit the reasoning

Prior for β

What we know:

- $\beta \in (0, 1)$
- Less probable to be close to the boundary
- Nothing specific about exact value in the range.

In the mind

Enumerate possible distributions that fit the reasoning

1. Beta(a, b), $a > 0, b > 0$ with some a, b avoids boundaries

Prior for β

What we know:

- $\beta \in (0, 1)$
- Less probable to be close to the boundary
- Nothing specific about exact value in the range.

In the mind

Enumerate possible distributions that fit the reasoning

1. Beta(a, b), $a > 0, b > 0$ with some a, b avoids boundaries
2. LogitNormal(μ, σ) - always avoids boundaries

Prior for β

What we know:

- $\beta \in (0, 1)$
- Less probable to be close to the boundary
- Nothing specific about exact value in the range.

In the mind

Enumerate possible distributions that fit the reasoning

1. Beta(a, b), $a > 0, b > 0$ with some a, b avoids boundaries
2. LogitNormal(μ, σ) - always avoids boundaries
3. Uniform$(0, 1)$ - a special case of Beta$(1, 1)$

Prior for β

What we know:

- $\beta \in (0, 1)$
- Less probable to be close to the boundary
- Nothing specific about exact value in the range.

In the mind

Enumerate possible distributions that fit the reasoning

1. $\mathrm{Beta}(a, b),\ a > 0, b > 0$ with some a, b avoids boundaries
2. $\mathrm{LogitNormal}(\mu, \sigma)$ - always avoids boundaries
3. $\mathrm{Uniform}(0, 1)$ - a special case of $\mathrm{Beta}(1, 1)$
4. $\mathrm{Kumaraswamy}(a, b),\ a > 0, b > 0$ you do not need to know that

Visualize your prior

Before writing a line of code, visualise your prior. What do you like more?

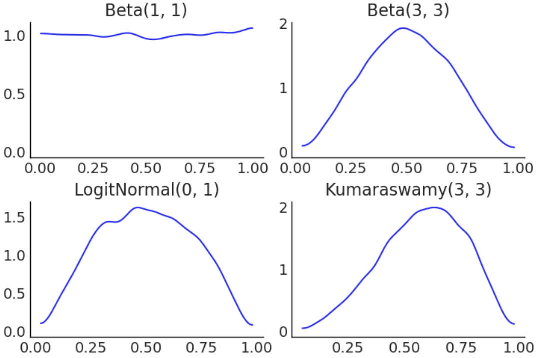


Figure: Visualized Priors

You can choose the form with theory in mind

Setting a prior

I prefer $\text{LogitNormal}(0, 1)$ in this situation. It has a good functional form.

To remember

- Prior is **your** modelling choice
- The choice has to be motivated
- The choice should make sense given practical constraints
- You should always be able to defend your choice
- **Prior is what you do not know, the uncertainty**

The model so far

$$
\begin{aligned}
\log Y_{g=0} &\sim \text{Normal}(\log A_{g=0} + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim ??? \\
\beta &\sim \text{LogitNormal}(0, 1) \\
A_{g=0} &\sim ???
\end{aligned}
$$

Prior for ϵ

Rule of thumb

Error term is something small. Usually avoids zero.

In our case;

- small is "orders of 10-50%"

Let it be

$$\epsilon \sim \text{LogNormal}(-2, 1)$$

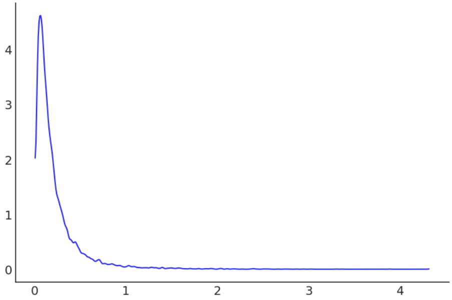


Figure: Prior for ϵ

Prior for ϵ

Rule of thumb

Error term is something small. Usually avoids zero.

In our case;

- small is "orders of 10-50%"

Let it be

$$\epsilon \sim \mathrm{LogNormal}(-2, 1)$$

Useful

In log-log models error term is on the relative scale

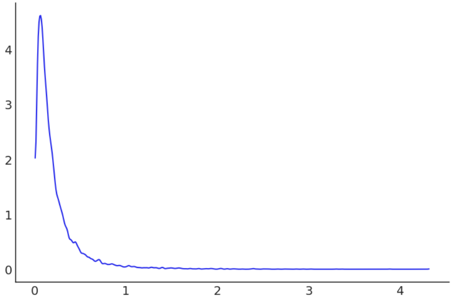
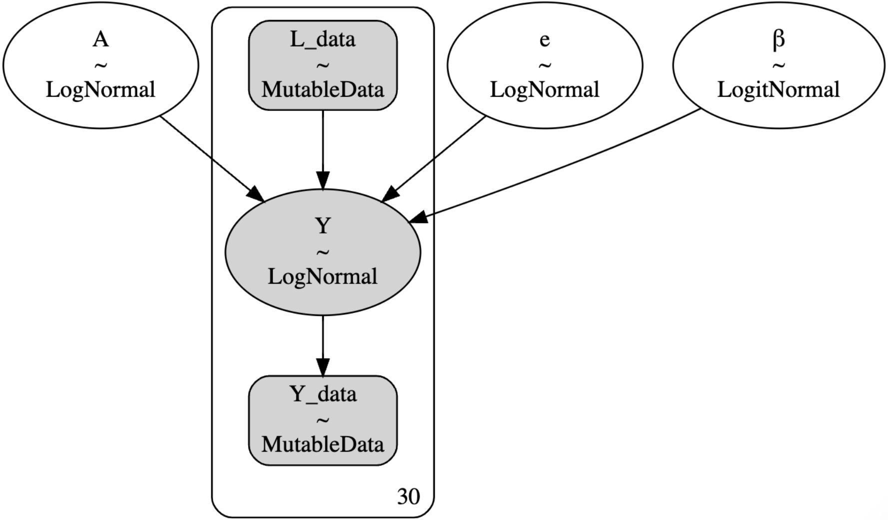


Figure: Prior for ϵ

The model so far

$$
\begin{aligned}
\log Y_{g=0} &\sim \text{Normal}(\log A_{g=0} + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim \text{LogNormal}(-2, 1) \\
\beta &\sim \text{LogitNormal}(0, 1) \\
A_{g=0} &\sim ???
\end{aligned}
$$

Visual Model



Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

- No idea what the prior should be

Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

- No idea what the prior should be
- We have an idea about Y

Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

- No idea what the prior should be
- We have an idea about Y
- Y is positive, thus A is positive

Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

- No idea what the prior should be
- We have an idea about Y
- Y is positive, thus A is positive
- We have practical range for Y, can we infer A at a glance?

Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

- No idea what the prior should be
- We have an idea about Y
- Y is positive, thus A is positive
- We have practical range for Y, can we infer A at a glance?
- Order of 10s for Y makes sense

Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

- No idea what the prior should be
- We have an idea about Y
- Y is positive, thus A is positive
- We have practical range for Y, can we infer A at a glance?
- Order of 10s for Y makes sense
- Prior predictive can help

Prior Predictive

Prior for β was an easy one. We need one for $A_{g=0}$

- No idea what the prior should be
- We have an idea about Y
- Y is positive, thus A is positive
- We have practical range for Y, can we infer A at a glance?
- Order of 10s for Y makes sense
- Prior predictive can help

Definition

Prior predictive is simulated observation model given no data.

The truth

Nobody said setting priors is easy. It is the most work.

Random prior

Why not using e.g.

$$A \sim \mathrm{LogNormal}(0, 1)$$

Random prior

Why not using e.g.

$$A \sim \mathrm{LogNormal}(0, 1)$$

Nonsense

Workers do not produce 800 carpets per week.

That's why

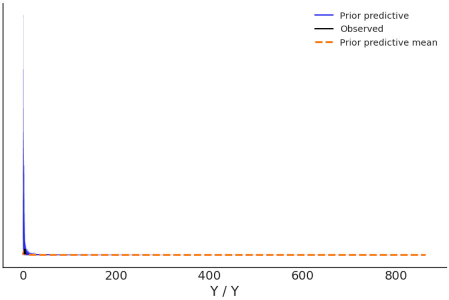


Figure: Prior predictive for Y vs data

Analysing the prior predictive

Getting back to a full model

$$
\begin{aligned}
\log Y_{g=0} &\sim \mathrm{Normal}(\log A_{g=0} + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim \mathrm{LogNormal}(-2, 1) \\
\beta &\sim \mathrm{LogitNormal}(0, 1) \\
A_{g=0} &\sim \mathrm{LogNormal}(0, 1)
\end{aligned}
$$

- We see over dispersion in predictions
- Variance may come from A or ϵ

Actions

1. Try reducing A variance
2. Try reducing ϵ variance

What can we read here?

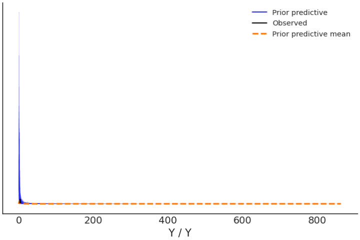


Figure: Prior predictive for Y vs data

Good prior predictive

Seminar

You will play with the example at the seminar.

A good looking prior predictive was with the definition below

$$\begin{aligned}\log Y_{g=0} &\sim \text{Normal}(\log A_{g=0} + \log L \cdot \beta, \epsilon)\\ \epsilon &\sim \text{LogNormal}(-2, 0.1)\\ \beta &\sim \text{LogitNormal}(0, 1)\\ A_{g=0} &\sim \text{LogNormal}(-0.5, 0.1)\end{aligned}$$

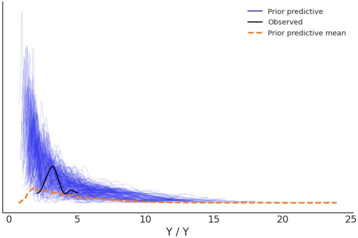


Figure: Prior predictive for Y vs data

What is a good prior predictive?

- Prior predictive covers **reasonable** range for observed data.
- **Data is reference**, not your objective.

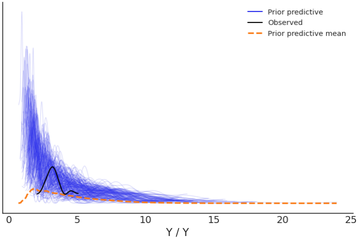


Figure: Prior predictive for Y vs data

What is a good prior predictive?

- Prior predictive covers **reasonable** range for observed data.
    - no astronomic speeds
    - no microscopic distances
    - no black hole densities
    - no superpower workers
- **Data is reference**, not your objective.

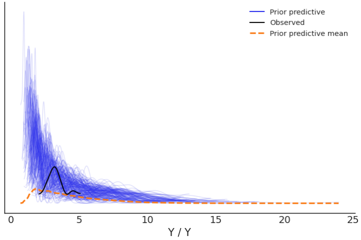


Figure: Prior predictive for Y vs data

What is a good prior predictive?

- Prior predictive covers **reasonable** range for observed data.
  - no astronomic speeds
  - no microscopic distances
  - no black hole densities
  - no superpower workers
- **Data is reference**, not your objective.
  - do not overfit priors on data.
  - in 90% cases you do not need data for prior predictive
  - in 90% cases common sense should work just fine
  - in 10% cases you can ask experts and adjust the priors
  - data is your last resort

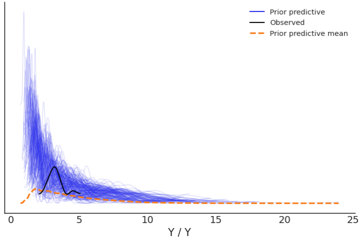


Figure: Prior predictive for Y vs data

HMC in action

Sampling

After we've checked the priors it is time to sample.

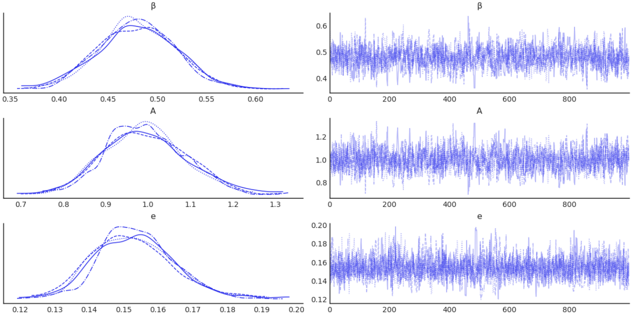


Figure: Posterior MCMC trace

Hierarchies

Hierarchies

Initial data has groups. How to take them in account?

$$\begin{aligned}
\log Y_{\mathbf{g}} &\sim \text{Normal}(\log A_{\mathbf{g}} + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim \text{LogNormal}(-2, 0.1) \\
\beta &\sim \text{LogitNormal}(0, 1) \\
A_{\mathbf{g}} &\sim ???
\end{aligned}$$

What is Hierarchy?

Hierarchy

Once you have similar groups in your data, you have hierarchy.

Examples:

1. Countries, Regions
2. User groups: by age, by profession, etc
3. Treatment groups
4. Time dependent effects
5. Panel Data

Our Example

Workers make different carpets and have total factor productivity A

Treating Hierarchy

Classical Econometrics view:

1. All the groups are independent. **Pooled Model**

$$y_{k,i} = \alpha + \beta x_{k,i} + \varepsilon_{i,k}$$

2. Groups have significant differences. **Fixed Effect Model**

$$y_{k,i} = \alpha_k + \beta x_{k,i} + \varepsilon_{i,k}$$

3. Groups have non significant, random differences. **Random Effects Model**

$$y_{k,i} = \alpha + \beta x_{k,i} + u_k + \varepsilon_{i,k}$$

Where

$$\mathbb{E}u_{k,i} = 0, \quad \mathbb{E}\varepsilon_{k,i} = 0$$

Bayesian Hierarchy

In

$$y_{k,i} = \alpha + \beta x_{k,i} + u_k + \varepsilon_{i,k}$$

Let's rearrange terms

$$y_{k,i} = (\alpha + u_k) + \beta x_{k,i} + \varepsilon_{i,k}$$

- α - population mean
- $\alpha_k = \alpha + u_k$ - group mean

In a Bayesian analysis we need priors. There is more than one way

$$\begin{aligned} \alpha &\sim \text{Normal}(\bar{\mu}, \bar{\sigma}) \\ u_k &\sim \text{Normal}(0, 1) \\ \alpha_k &= \alpha + u_k \cdot \sigma \end{aligned}$$

$$\begin{aligned} \alpha &\sim \text{Normal}(\bar{\mu}, \bar{\sigma}) \\ \alpha_k &\sim \text{Normal}(\alpha, \sigma) \end{aligned}$$

More on priors

Non centered parametrization

$$\alpha \sim \mathrm{Normal}(\bar{\mu}, \bar{\sigma})$$
$$u_k \sim \mathrm{Normal}(0, 1)$$
$$\alpha_k = \alpha + u_k \cdot \sigma$$

Centered parametrization

$$\alpha \sim \mathrm{Normal}(\bar{\mu}, \bar{\sigma})$$
$$\alpha_k \sim \mathrm{Normal}(\alpha, \sigma)$$

Group specific parameter u_k is disentangled

σ is a measure of group differences

1. $\sigma \to 0$: Pooled Model
2. Small σ: Random Effects / Partial Pooling
3. Large σ: Fixed Effects / Unpooled Model

σ interpolates between the models

Degeneracy

Centered parametrization

$$\alpha \sim \text{Normal}(\bar{\mu}, \bar{\sigma})$$
$$\alpha_k \sim \text{Normal}(\alpha, \sigma)$$

Warning

Centered parametrization creates funnel geometry with few data

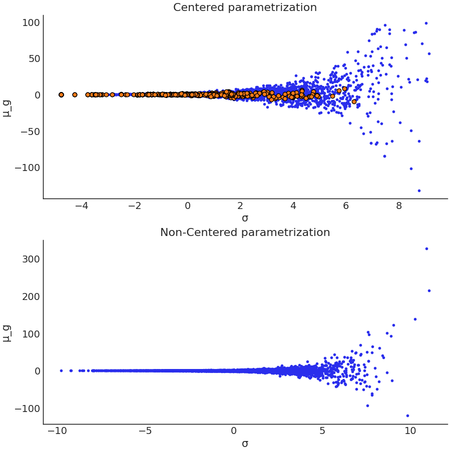


Figure: Divergences appear in the Centered Parametrization

Why Funnel is created?

Geometry is important

1. Sampler has adaptive step size
2. With bad geometry Sampler can't find a good one

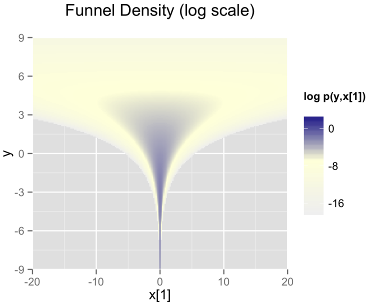


Figure: Funnel Geometry

Suggested reading

Read more on reparametrization in Stan's Guide

Inverted Funnel degeneracy

A "nice" parametrization does have issues as well.

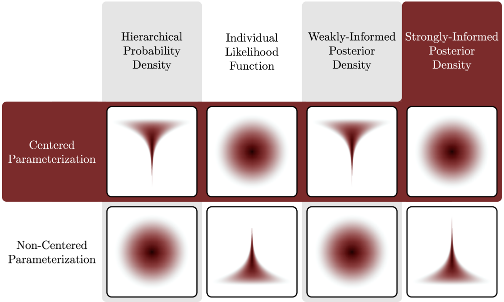


Figure: Inverted Funnel Degeneracy

Advanced Reading

Read more from Michael Betancourt

Setting a Hierarchical Prior

1. Start with a Pooled or Single group model
2. Add Hierarchy

Setting a Hierarchical Prior

1. Start with a Pooled or Single group model
   - You get an idea of prior parameter scales
   - You get a decent model structure
   - Do not care about predictions
2. Add Hierarchy

Setting a Hierarchical Prior

1. Start with a Pooled or Single group model
    - You get an idea of prior parameter scales
    - You get a decent model structure
    - Do not care about predictions
2. Add Hierarchy
    - Decide on which parameters to share
    - Decide on allowed variability for the rest parameters
    - Debug divergences, reparametrize if required

Setting a Hierarchical Prior

1. Start with a Pooled or Single group model
   - You get an idea of prior parameter scales
   - You get a decent model structure
   - Do not care about predictions
2. Add Hierarchy
   - Decide on which parameters to share
   - Decide on allowed variability for the rest parameters
   - Debug divergences, reparametrize if required

Best Practice

Do not hard-code the parametrization, toggle it in the code

The Cobb-Douglas Case

Single group model

$$
\begin{aligned}
\log Y_0 &\sim \text{Normal}(\log A_0 + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim \text{LogNormal}(-2, 0.1) \\
\beta &\sim \text{LogitNormal}(0, 1) \\
A_0 &\sim \text{LogNormal}(-0.5, 0.1)
\end{aligned}
$$

Hierarchical model

$$
\begin{aligned}
\log Y_k &\sim \text{Normal}(\log A_k + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim \text{LogNormal}(-2, 0.1) \\
\beta &\sim \text{LogitNormal}(0, 1) \\
A_k &\sim \text{LogNormal}(\log A_{\text{pop}}, \sigma_A) \\
A_{\text{pop}} &\sim \text{LogNormal}(-0.5, 0.1) \\
\sigma_A &\sim \text{LogNormal}(-2, 0.1)
\end{aligned}
$$

The Cobb-Douglas Case

Single group model

$$\begin{aligned}
\log Y_0 &\sim \text{Normal}(\log A_0 + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim \text{LogNormal}(-2, 0.1) \\
\beta &\sim \text{LogitNormal}(0, 1) \\
A_0 &\sim \text{LogNormal}(-0.5, 0.1)
\end{aligned}$$

Hierarchical model

$$\begin{aligned}
\log Y_k &\sim \text{Normal}(\log A_k + \log L \cdot \beta, \epsilon) \\
\epsilon &\sim \text{LogNormal}(-2, 0.1) \\
\beta &\sim \text{LogitNormal}(0, 1) \\
A_k &\sim \text{LogNormal}(\log A_{\text{pop}}, \sigma_A) \\
A_{\text{pop}} &\sim \text{LogNormal}(-0.5, 0.1) \\
\sigma_A &\sim \text{LogNormal}(-2, 0.1)
\end{aligned}$$

Hint

You can reuse some parameters, just add reasonable variability σ_A

Discussion Time

Setting priors

- Sometimes you do not have expert knowledge
- Sometimes parametrization does not allow you to set a good prior
- Sometimes prior predictive depends on many parameters
- You are limited in time
- Using hyperpriors