Bayesian Hierarchical Models

DR. OLIVERA STOJANOVIĆ

PRINCIPAL DATA SCIENTIST @ PYMC LABS

Who am I?

- BSc in Physics and MSc in Electrical Engineering from the University of Novi Sad
- Participated in Petnica’s Summer School of Meteor Astronomy
- PhD in Cognitive Science from Osnabrück University
- Topic areas: Bayesian inference, interpretability in machine learning, causal inference
- Work experience in industry and academia
- Current position: Principal Data Scientist @ PyMC Labs
- Based in San Diego, CA

Outline

- What is Bayesian Reasoning?
- Bayes Theorem (Quick Recap)
- Bayesian Sequential Update
- What does hierarchy in Bayesian model means?
- Example: Bayesian hierarchical models in environmental sciences
- Bayesian model selection
- Applications of Bayesian Hierarchical Models
- Sources for further reading

What is Bayesian Reasoning?

- A way of updating what you believe based on new evidence
- **Is this coin fair?**
- **Frequentist**: Flip the coin many times, calculate proportion of heads, and construct a confidence interval around 0.5
- **Bayesian**: Start with a prior (you believe it's probably fair). After observing flips, you update your belief and get a probability distribution over possible values of the true probability of heads
- Prior -> new information (data) -> posterior

Bayes Theorem (Quick Recap)

likelihood:
probability of observing the data given different parameter values

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

prior:
the initial belief about the distribution of the parameters

posterior:
updated probability distribution after observing the data

marginal probability (evidence):
the total probability of the data, which normalizes the posterior

Bayesian Sequential Update

$$P(\theta \mid \text{data}) \propto P(\text{data} \mid \theta) \cdot P(\theta)$$

- This gives the shape of the posterior, but not a proper probability distribution
- Computing **normalizing constant P(data)** is computationally expensive:

$$P(\text{data}) = \int P(\text{data} \mid \theta) \cdot P(\theta) \, d\theta$$

- P(data) ≅ 1 -> the posterior sums to 1 over all possible parameter values
- Markov Chain Monte Carlo (MCMC) sampling -> generates samples from the posterior without computing the normalizing constant
- *PyMC* package in python, *brms* package in R

What does hierarchy in Bayesian models mean?

- In some cases, we collect data of the **same variable** across **different groups** (e.g. sea temperature at different locations)
- Data from the **same group** often looks more similar to each other than to data from other groups
- Examples:
  - **Air pollution** in cities vs rural areas
  - **Unemployment rates** by county
  - **House prices** in different parts of a city
  - **Disease incidence** (e.g. Lyme borreliosis) across regions

How to model grouped data?

- Option 1: Fit a **separate regression** for each location or group.
- Issues:
  - Requires **lots of parameters**
  - Ignores **shared patterns** across groups

- Option 2: introduce **hierarchies** into a Bayesian model and model group structure directly
- Benefits:
  - Sharing information across similar groups -> helpful when **data is limited**
  - Allows for **group-level differences**

Bayesian hierarchical models

- Also called **multilevel** or **nested models**
- Parameters are organized in levels –> each level has its **own distribution**
- *Hyperparameters* (higher-level parameters) **capture uncertainty** and **improve estimates for lower-lever groups**

Challenges

- Increased complexity in setting up models
- Need to specify appropriate *priors, hyperparameters, and model structures* that capture hierarchical dependencies
- Higher computational costs compared to simpler models
- Accuracy depends on data quality and quantity at each hierarchy level

Bayesian hierarchical models in environmental sciences

- Objective: Predict leaf area index (LAI) from reflectance spectra
- LAI measures the total leaf area per unit ground area -> estimates vegetation density and canopy structure
- Used in models of photosynthesis, evapotranspiration, and climate-vegetation interactions
- Data:
  - 4 datasets of LAI and reflectance spectra of white winter wheat
  - 4 different fields
  - 4 different years
- Challenges:
  - Heterogenous datasets -> large systematic differences
  - Limited dataset (191 measurements)

Siegmann (2015)

Bayesian hierarchical models in environmental sciences

Field A, 2011

Field B, 2012

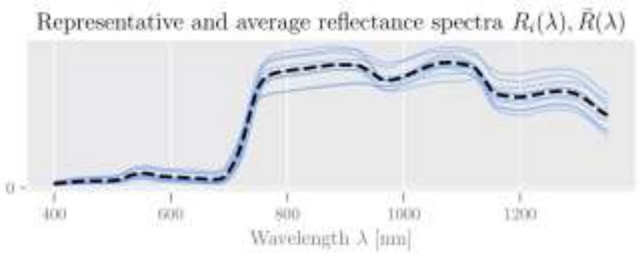
Field C, 2015

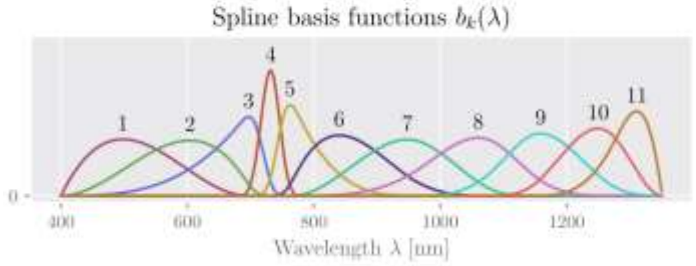
Field D, 2016

data preprocessing: reflectance spectra + LAI measurements

feature extraction: splines with adaptive knot placement

3 Bayesian hierarchical models





Bayesian hierarchical models in environmental sciences

Field A, 2011

Field B, 2012

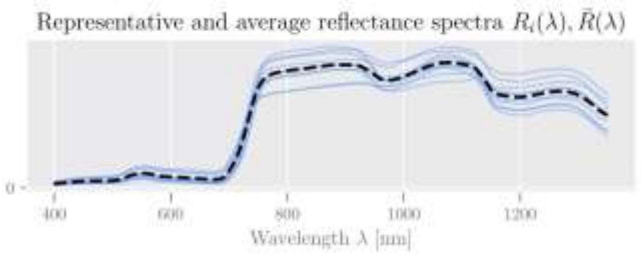
Field C, 2015

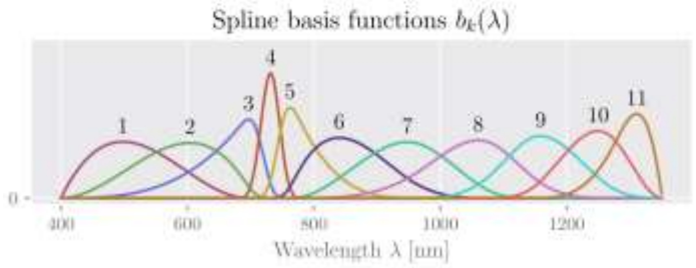
Field D, 2016

data preprocessing:
reflectance spectra
+
LAI measurements

feature extraction:
splines with
adaptive knot
placement

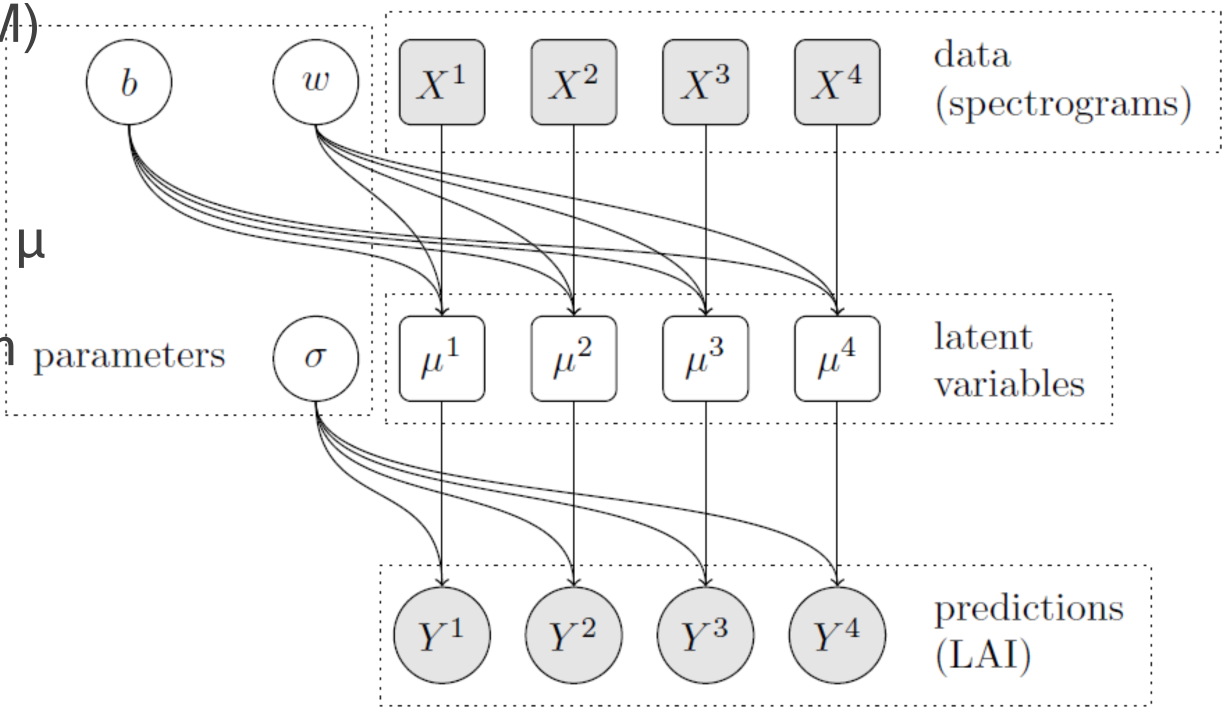
3 Bayesian
hierarchical
models





Model 1: A baseline model

- Simple Generalized linear model (GLM)
- Pool all data together
- log(LAI) is normally distributed around μ
- μ -> feature matrix*weights + bias term
- σ -> deviation parameter

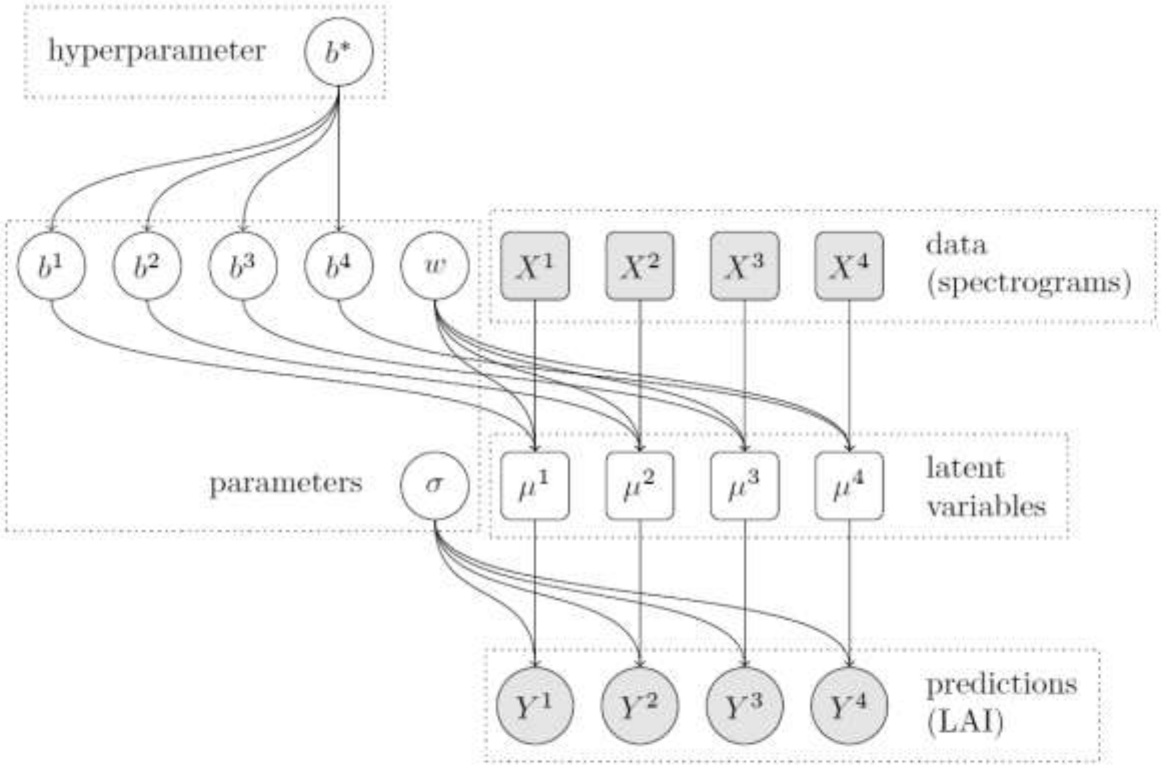


$$
\begin{aligned}
\log(\sigma) &\sim \mathrm{Normal}(0,1) \\
b &\sim \mathrm{Normal}(0,11) \\
w_k &\sim \mathrm{Normal}(0,1) && \forall k \in \{1,\ldots,11\} \\
\mu^j &= X^j w + b, && \forall j \in \{1,\ldots,4\} \\
\log(Y_i^j) &\sim \mathrm{Normal}(\mu_i^j,\sigma), && \forall j \in \{1,\ldots,4\}, i \in I^j
\end{aligned}
$$

Model 2: A model with hierarchical bias

- Additional bias parameter for each dataset
- Bias terms are clustered around a *hyperparameter b**

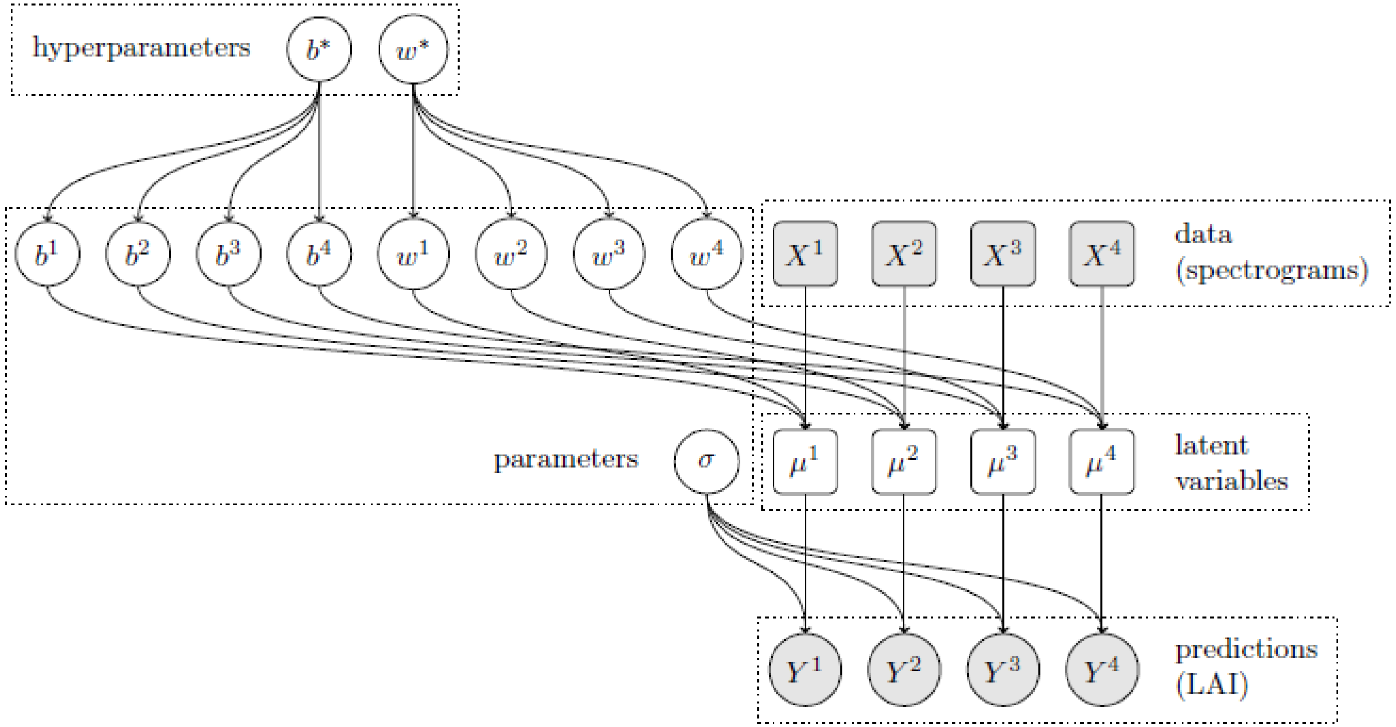
$$
\begin{aligned}
\log(\sigma) &\sim \text{Normal}(0,1) \\
b^* &\sim \text{Normal}(0,11) \\
b^j &\sim \text{Normal}(b^*,1.1) && \forall j \in \{1,\ldots,4\} \\
w_k &\sim \text{Normal}(0,1) && \forall k \in \{1,\ldots,11\} \\
\mu^j &= X^j w + b^j, && \forall j \in \{1,\ldots,4\} \\
\log(Y_i^j) &\sim \text{Normal}(\mu_i^j,\sigma), && \forall j \in \{1,\ldots,4\}, i \in I^j
\end{aligned}
$$



Model 3: Full hierarchical model

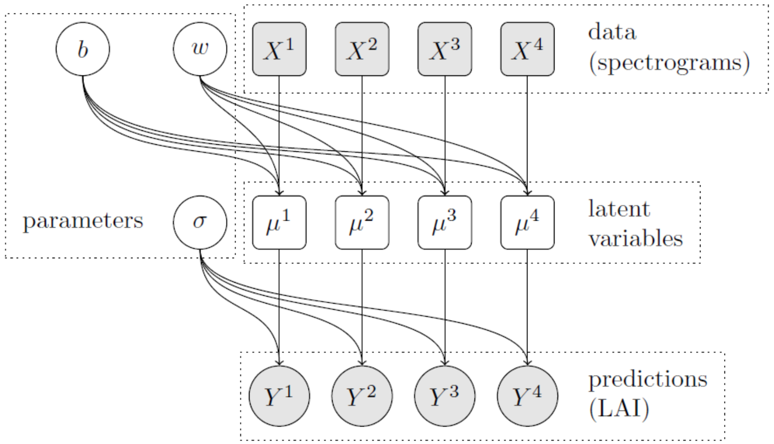
- Weight vector *w* can vary for each dataset
- Weight terms are clustered around around a *hyperparameter w**

$$
\begin{aligned}
\log(\sigma) &\sim \mathrm{Normal}(0,1) \\
b^* &\sim \mathrm{Normal}(0,11) \\
b^j &\sim \mathrm{Normal}(b^*,1.1) && \forall j \in \{1,\ldots,4\} \\
w_k^* &\sim \mathrm{Normal}(0,1) && \forall k \in \{1,\ldots,11\} \\
w_k^j &\sim \mathrm{Normal}(w_k^*,0.1) && \forall k \in \{1,\ldots,11\}, j \in \{1,\ldots \\
\mu^j &= X^j w^j + b^j, && \forall j \in \{1,\ldots,4\} \\
\log(Y_i^j) &\sim \mathrm{Normal}(\mu_i^j,\sigma), && \forall j \in \{1,\ldots,4\}, i \in I^j
\end{aligned}
$$

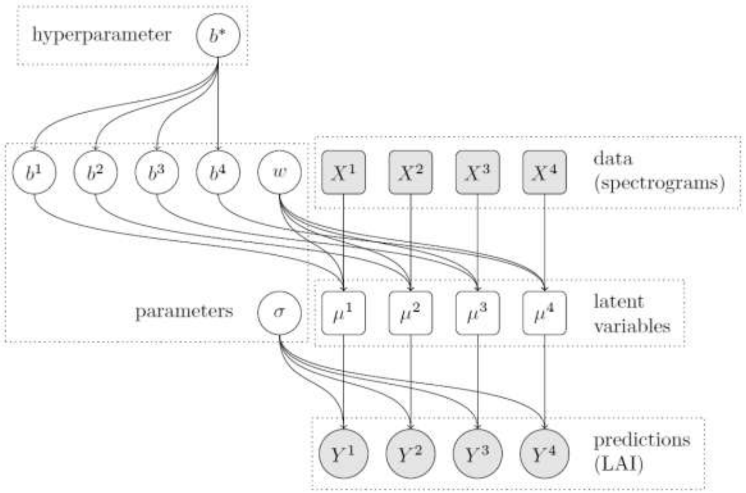


Bayesian hierarchical models in environmental sciences

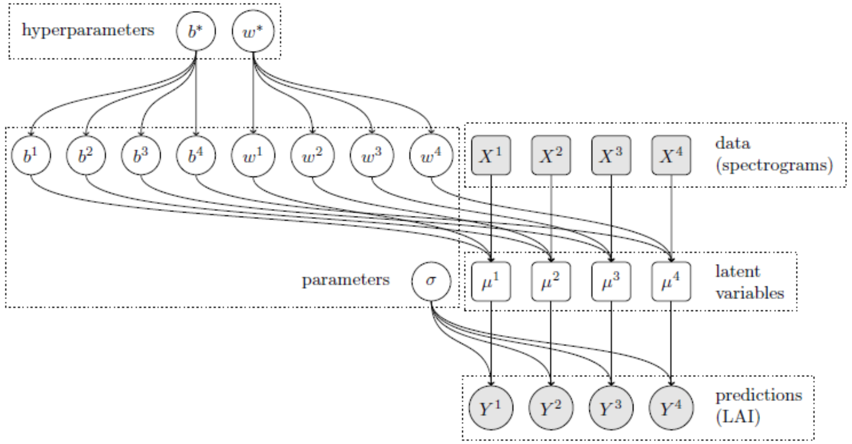
Model 1: A baseline model with pooled data



Model 2: A model with hierarchical bias



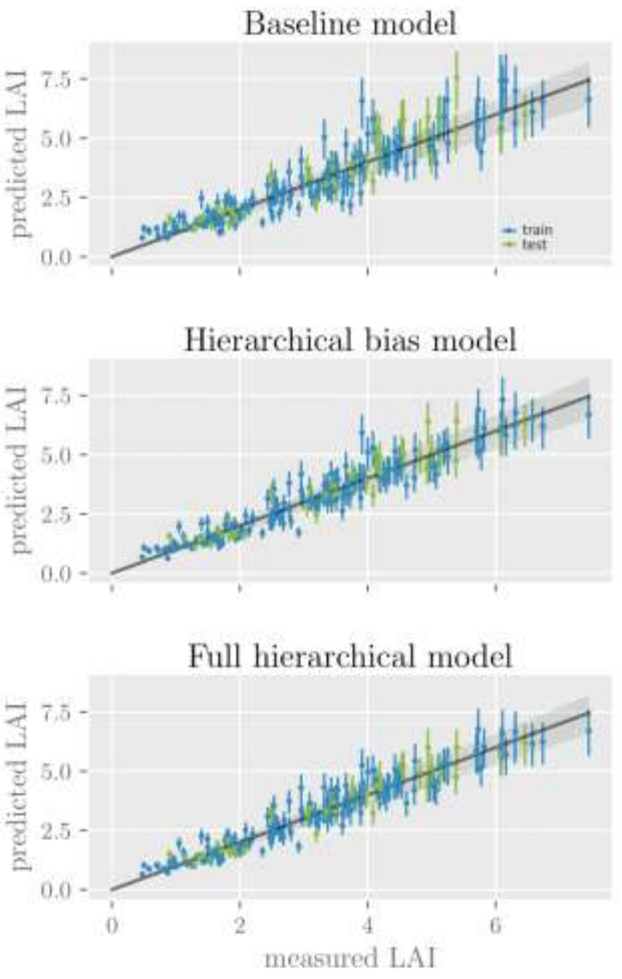
Model 3: Full hierarchical model



Results

- MCMC sampling to infer posterior distribution over parameters
- All models have good (similar) predictions
- Prediction accuracy is similar across models

- How to choose the best model?
  - Can model generalize well on *unseen data*?
  - Model complexity
  - Accuracy



Bayesian model selection

- We want to compare 3 Bayesian models using leave-one-out cross-validation (LOO-CV)
- Refitting the model for each data point with MCMC is computationally expensive

- *Vehtari et al. (2017)* introduced an efficient method for LOO-CV in Bayesian models:
  - **Reweights existing MCMC samples** to approximate the effect of leaving out each data point (importance sampling)
  - **Pareto-smoothed importance sampling (PSIS)** smooths the most extreme weights to reduce variance
  - Produces reliable out-of-sample estimates

Bayesian model selection

- **Predictive density** for a data point -> the probability (for discrete data) or probability density (for continuous data) that the model assigns to that point given the observed data
- Using **leave-one-out (LOO-CV)** cross-validation we compute $p(y_i \mid y_{-i})$ -> the probability of observing data point y_i given the rest of the data
- Probabilities can be very small or large -> we take the logarithm

Bayesian model selection

- **Expected log pointwise predictive density (ELPD)** -> sum over predictive densities
- Calculated as the sum of the log predictive densities -> overall score of model predictive accuracy

$$\mathrm{elpd}_{\mathrm{loo}} = \sum_{i=1}^{n} \log p(y_i \mid y_{-i}),$$

- **Higher ELPD = better fit and better generalization** to unseen data
- **Effective number of parameters (*p_loo*)** -> also based on leave-one-out cross-validation (LOO-CV)
- Estimates how many parameters the model is *effectively* using
- **Higher p_loo = more sensitivity to individual data points**
- Can be lower than the *actual* number of parameters (regularization in hierarchical models)

Bayesian model selection

- We use leave-one-out ELPD to compare 3 models
- We also estimate *effective number of parameters (p_loo)*
- Too high *p_loo* indicates a complex models and can lead to overfitting
- Model 2: hierarchical bias is the best model

Model	ELPD	p_loo
Baseline	-185.5 ± 12.2	13.3
Hierarchical Bias	**-157.0±11.5**	**15**
Hierarchical Full	-157.8 ± 11.5	24.9

Conclusion

- Bayesian hierarchical models work well when data is structured in groups
- Can capture variation across groups without overfitting
- Useful when dealing with limited datasets
- Especially helpful in fields where data is hard to collect
- More complex models don’t necessarily mean better fit

Applications of Bayesian hierarchical models

- Epidemiology - estimating disease rates
- Cognitive Science - understanding perception under uncertainty
- Labor market & economics - predicting unemployment rate
- Finances - calculating loan default chances
- Marketing & e-commerce - measuring marketing channel impact

Further reading/watching/listening

- https://twiecki.io/blog/2017/02/08/bayesian-hierchical-non-centered/
- https://sellforte.com/blog/compared-bayesian-hierarchical-vs-non-hierarchical-modeling
- Video lectures:
  - *Developing Hierarchical Models for Sports Analytics with Chris Fonnesbeck: https://www.youtube.com/watch?v=Fa64ApS0qig*
  - *Hierarchical Bayesian Modeling of Survey Data with Post-stratification (Tarmo Jüristo):* https://www.youtube.com/watch?v=efID35XUQ3I
  - *L3: Hierarchical Modeling (State of Bayes Lecture Series):*

    https://www.youtube.com/watch?v=pnJgDSdgqVg
  - *Chris Fonnesbeck - Probabilistic Python: An Introduction to Bayesian Modeling with PyMC:* https://www.youtube.com/watch?v=911d4A1U0BE

References

- *Stojanovic et al. (2022): Bayesian hierarchical models can infer interpretable predictions of leaf area index from heterogeneous datasets*
  - https://doi.org/10.3389/fenvs.2021.780814
- *Vehtari et al. (2017): Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC*
  - *https://arxiv.org/abs/1507.04544*

Let’s connect!

- GitHub handle: @ostojanovic
- Linkedin profile: https://www.linkedin.com/in/stojanovicolivera/
- Personal website: https://www.stojanovic.science

LinkedIn profile

Personal
website