The Hierarchical Bayesian Inference Model

Sophia Wilson, Kathrine Kuszon (Niels Bohr Institute, UCPH)

Hierarchical Bayesian inference (HBI) is a powerful modelling approach, that can improve the accuracy and precision of parameter estimation by accounting for variability across groups. In this writeup we explore the HBI model and how it differs from 'ordinary' Bayesian methods. We provide an analysis of a hierarchical data set with the aim of comparing the parameter estimates from the hierarchical and non-hierarchical methods. A discussion of the benefits and limitations of the model is also included.

Introduction

Oftentimes, an underlying assumption about the source of data is made: Observations are assumed to be independently and identically distributed (*i.i.d.*) following a single distribution with one or several unknown parameters [1]. It is, however, in many situations not sensible to treat the observations as i.i.d. and as generated from individual but identical distributions since often parameters are non-identical but still related in some way by the structure of the problem. As an example, consider a clinical trial: If patients in one hospital have a certain survival probability, it would be reasonable to expect patients in other hospitals to have a similar survival probability, implying that the survival probabilities at different hospitals can be described by a common population distribution [3].

When observations are somehow grouped hierarchical modelling becomes relevant [2]. Here, observable outcomes are modelled conditionally on certain parameters, which themselves are dependent on higher-level parameters, *hyperparameters*. In the hospital example, this is achieved by using a prior distribution in which the survival probabilities at each hospital are viewed as a sample from a common population distribution dependent on some hyperparameters [3].

To demonstrate how the hierarchical approach improves the parameter estimates, we compare the results when analysing the same hierarchical data set with two other models; one using *completely pooling* and one using *no pooling*.

Theory

When dealing with 'ordinary' Bayesian statistics, we do not consider our parameters as being dependent on other parameters. Instead, we specify the prior distributions using previous knowledge and if no knowledge is available we usually choose our priors to be flat and uninformative. The HBI model differ from 'ordinary' Bayesian statistics by letting the parameters depend on hyperparameters. These hyperparameters have priors of their own called *hyperpriors*.

In order to understand the hierarchical structure of the HBI model, consider a data set consisting of j groups with i observations in each group. For each observation in each group an outcome y_{ji} is sampled from a distribution dependent on parameters θ_j. The parameters θ_j are assumed to be generated exchangeably from a common population with a distribution dependant on hyperparameters ϕ. The key hierarchical part of this model is that the hyperparameters ϕ are *not* known, and thus have their own prior distributions, $P(\phi)$. The joint prior distribution of the data y becomes

$$P(\phi|\theta) = P(\theta|\phi)P(\phi), \tag{1}$$

resulting in the joint posterior distribution

$$P(\theta, \phi|y) = \frac{P(y|\theta, \phi)P(\theta|\phi)}{P(y)} = \frac{P(y|\theta)P(\theta|\phi)P(\phi)}{P(y)} \tag{2}$$

where the latter simplification of the likelihood holds, because the likelihood $P(y|\theta, \phi)$ depends only on θ, meaning that the hyperparameters ϕ only affect y through θ [3].

Methods

We create a hierarchical data set following the approach described by Jørgen Bølstad [2]. This data set consists of 75 groups ($j \in \{1, ...75\}$) with 5 observations ($i \in \{1, ...5\}$) in each group. The outcome data y_{ji} is normally distributed and depends linearly on a covariate x_{ji} according to the following relation

$$y_{ji} = N(\alpha_j + \beta_j x_{ji}, \sigma^2).$$

A key factor is that each group will have its own true intercept α_j as well as its own coefficient β_j generated from the true distributions given by $N(2, 1)$ and $N(-2, 1)$, respectively. The parameters α_j and β_j are themselves normally distributed

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \quad \text{and} \quad \beta_j \sim N(\mu_\beta, \sigma_\beta^2).$$

Their distributions can be described by the four hyperparameters $\mu_\alpha, \sigma_\alpha, \mu_\beta$ and σ_β described by the hyperpriors

$$\mu_\alpha, \mu_\beta \sim N(0, 3^2) \quad \text{and} \quad \sigma_\alpha, \sigma_\beta \sim U(0, 10).$$

σ is furthermore a parameter, of none particular interest, with prior

$$\sigma \sim U(0, 10).$$

The goal is to estimate α_j and β_j. Two other models are used for comparison:

1. Completely pooling In contrast to the HBI model, a model that uses completely pooling assumes that all α_j and β_j are equal. The goal is to only estimate one single value for each of the shared parameters α and β. This corresponds to running a single regression on the entire data set.

2. No pooling Models using no pooling recognize each α_j and β_j but do not assume that these are generated exchangeably from a common population. Instead each α_j and β_j are assigned a wide, flat prior. This model allows no statistical sharing between the different groups. This corresponds to running separate regressions on each data set y_j.

To obtain the best possible estimates of α and β we use a MCMC Metropolis-Hastings sampler for the model using no pooling and the HBI model. The estimates are plotted as a function of the steps for one of the j groups in Figure 1. The

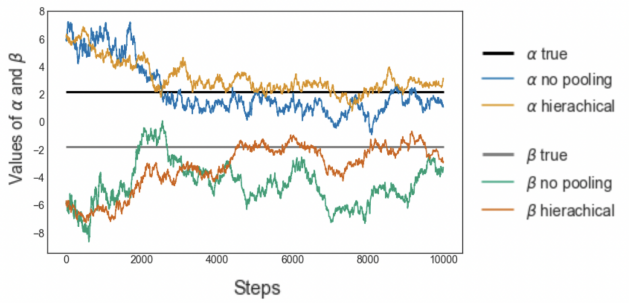

Figure 1: ***A visualisation of the parameter estimations for one of the j groups as a function of the MCMC Metropolis-Hastings sampler steps for the HBI model and the no pooling model.***

initial start guesses are purposely chosen far from the true values to illustrate the convergence towards the true value. In further data analysis, we choose start guesses slightly closer to the true values and remove the first 2000 data points in order for the sampler to reach equilibrium.

Results

The results of the parameter estimation of α and β for all three models can be seen in Figure 2. The estimates for the model using no pooling are clearly more dispersed than the estimates for the HBI model. Considering the root mean square errors (RMSEs) across the sets of α- and β-estimates we observe, that the RMSEs for the models using completely pooling and no pooling are similar and approximately 40% bigger than for the HBI model. This means, that the HBI model on average yields estimates closer to the true parameter values.

The comparison of the two non-hierarchical versus the hierarchical model clearly illustrates the strength of the hierarchical model. The model using completely pooling clearly misunderstands the data set by estimating a single α and a single β for all 75 groups. The α_j- and β_j-estimates in the model using no pooling are pulled away from each other as a consequence of using flat priors. In contrast, the hierarchical model benefits from the prior structure based on data, which pulls the estimates towards the true parameter distribution - an effect known as shrinkage. This means that the HBI model is less sensitive to noise and outliers, making it more robust.

Discussion

The HBI model provides a flexible framework for statistical modelling that can capture variability across groups and improve the accuracy and precision of the parameter estimates. One of the main advantages of the HBI model is its ability to incorporate prior information about the parameters. This can be particularly useful when there is limited data available, as prior information can help to reduce the uncertainty of the parameter estimates. Moreover, the hyperprior distribution can be chosen to reflect prior beliefs about the population-level parameters, which can improve the robustness of the inference. However, the choice of the hyperprior distributions can be challenging and the results can be sensitive to these choices. In most cases it is important to chose hyperprior distributions that are relatively uninformative so the priors are mainly specified by the data itself.
Another advantage of the HBI model is its ability to estimate dispersion within the group observations. This can be important in data sets where there is substantial heterogeneity across the different groups. In the example with hospitals, the HBI model can estimate variations in the treatment effectiveness across the hospitals, which can be useful for examining possible differences in the treatments. One downside of the HBI model is, however, the increased complexity compared with other models which can be more challenging to implement and be more computationally expensive.

Conclusion

In conclusion, HBI is a powerful modelling approach that can improve the accuracy and precision of the parameter estimates. It allows for complex structures of joint priors in order to reduce parameter uncertainties. However, the choice of the hyperprior distributions can be challenging, and the increased complexity of the model can be a barrier to adoption. Nevertheless, the HBI model has great potential for parameter estimation when dealing with hierarchical data structures.

References

[1] Jim Albert and Jingchen Hu. *Probability and Bayesian Modeling*. 2020.

[2] Jørgen Bølstad. The benefits of bayesian hierarchical modeling: Comparing partially pooled and unpooled models in r. Playing with Numbers: Notes on Bayesian Statistics, August 8 2018.

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis Thirds Edition*. 2021.

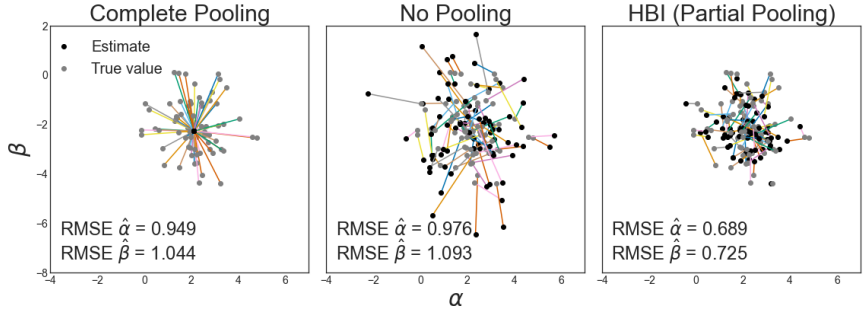

Figure 2: ***True values of α and β visualised as grey dots and the best estimates as black dots. Each pair of estimate and true value are connected with a coloured line.***