Hierarchical Bayesian Models

Research Module - Econometrics and Statistics - 2019/2020

Linda Maokomatanda, Tim Mensinger, Markus Schick

Contents

1 Introduction

2 Bayesian Thinking

2.1 (Probabilistic) Modeling

1. What is Statistics? What is a model? Why are models useful?
   Want to model some *phenomena* which manifests itself in some latent variable system / observational process to learn how it works. What is an observation? What is an environment of the latent system?

2. True Data Generating Process (What even is that)? (Define it as a probability distribution on the observational space: p^1.

3. The Observational Model (Which we assume; Is some subset $\mathbf{S} \subset \mathbf{P}$, where $\mathbf{P}$ is the set of all probability distributions on the observational space $\mathcal{Y}$. (Because $\mathbf{P}$ is too big.)

4. Each $s \in \mathbf{S}$ defines a data generating process, and in doing so provides a (formal) narrative on how the observed data might have been generated.

5. In practice we define the elements of $\mathbf{S}$ via a parameterization. So that we have a one-to-one mapping from $\mathbf{S}$ to some parameter configuration space $\Theta \subset \mathbf{R}^K$.

6. Once we identified $\mathbf{S}$ and the mapping to Θ we want to identify the configuration which is most consistent with the true underlying data generating process p^1. Since we do not know p^1 this comes down to finding the narrative that best explains the observed data and adheres to domain knowledge on the given application. (But what does consistency even mean?)

2.2 Frequentist Inference

1. What is a probability anyways? Define it as the limiting frequency of an event happening in infinitely many repetions of the same experiment. Realistic? Formally use Kolmogorov's axioms. So what is allowed to be stochastic in this framework? What is the source of variability in models? Model configuration parameters are constants. (? said where?)

2.3 Bayesian Inference

1. Can we also put weights on different model configurations prior to observing data? Humans (therefore also scientiest) certainly do; although sometimes subconciously. We want to formalize this by allowing probability distributions on the configurations space, *priors*.

2. What changes? Before we had some space $\mathbf{S}$ and parameterization $\theta \in \Theta$ which defined our observational model through a probability distribution (density) $p_{\mathbf{S}}(y;\theta)$ for $y \in Y$. Since we now allow θ to be stochastic we have to work with the conditional density that is, we model the stochastic relationship of the observations given a certain parameterization, i.e.

$$p_{\mathbf{S}}(y \mid \theta) = p_{\mathbf{S}}(y;\theta)$$

3. Why would we want to do this anyways? Bayes' Theorem! Given a distribution $p_{\mathbf{S}}(\theta)$ on the configuration space we can apply Bayes' Theorem. (State it here or later?)

4. The likelihood function.

$$\ell_y : \Theta \to \mathbf{R}_+, \theta \mapsto p_{\mathbf{S}}(y \mid \theta)$$

   The likelihood function maps model configurations to a numerical quantification which increases for model configurations which are more consistent with the data and decreases with configurations that are less consistent. Hence the likelihood function quantifies the relative consistency of each model configuration with the observed data.

5. The posterior distribution. Applying Bayes' Theorem we get

$$p_{\mathbf{S}}(\theta \mid y) = \frac{p_{\mathbf{S}}(y \mid \theta)}{\int p_{\mathbf{S}}(y \mid \theta) p_{\mathbf{S}}(\theta) \mathrm{d}\theta} p_{\mathbf{S}}(\theta) \propto p_{\mathbf{S}}(y \mid \theta) p_{\mathbf{S}}(\theta)$$

6. The goal of analysis can be inference or prediction When being concerned with the former we would like to understand how the phenomena of interest interacts with the latent variable system and the observational process we measure, as this might give us insights into the phenomena itself.

Having found a parameterization this means that we want to know which parameters are likely to cause the observed data. In the Bayesian setting we answer this question by construction the posterior distribution. That is, a conditional distribution on the configuration space given the observed data. The use of Bayes' Theorem (which makes this possible in the first place) explains the name. But it is not the application of Bayes' Theorem which makes Bayesian statistics different to classical (frequentist) statistics; it is the liberation of the model parameters, which are allowed to vary according to some prior distribution.

7. We can interpret this statement as an updating process: we have beliefs on the model calibration parameter in the form of a prior distribution and we update this belief using the likelihood function. During this updating step three commom patterns can occur. (i) contraction (ii) containment (iii) compromise. [add pictures and gaussian example].

8. Identification of model parameters. If we observe very informative data in the sense that the likelihood is concentrated around a small area then all vague priors will do fine and in a sense we let the data speak. If, however, the observational process was not sensitive to the phenomena of interest, we might observe data with a very low level of information regarding the model parameters. In this case we speak of weakly-identified parameters. This manifests in the likelihood dispersing over large regions of the configuration space. Choosing a prior careless in these situation can result in weak-identifibility of the likelihood propagating to the posterior.

9. Okay now we have $p_{\mathbf{S}}(\theta \mid y)$ so what? Let g be any function on Θ. Compute

$$\mathbf{E}\left[g(\theta) \mid y\right] = \int g(\theta) p_{\mathbf{S}}(\theta \mid y) \mathrm{d}\theta$$

   [Insert analytical gaussian exmaple here:] For very simple models with convenient assumptions we can compute the posterior density in closed-form. Using this we might even be able to compute the above integral for some functions g analytically. For more complicated, i.e. realistic, models this does not work.

10. For more complicated models we utilize the fact that for most questions we

do not need the analytical form of the posterior but we are happy with being able to draw from it. If we are able to draw from the posterior correctly we can approximate quantiles and arbitrary expectations. But how do we draw from a density?

11. Some blabla on how to draw from densities and the normalization constant and this is why there is *Gibbs Sampling* and *Metropolis Hasting Algorithm* and *Monte Carlo Markov Chain* in general.

2.4 Solving for the posterior analytically

In this subsection we will be presenting an example of estimating / solving for the mean and variance parameters when observing data that is assumed to come from a normal distribution.

2.5 Asymptotics

Let $y = \{y_1, ..., y_n\}$.

1. Assume the likelihood function is smooth and consider the maximum likelihood estimator

$$\theta_{ML}(y) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p_{\mathbf{S}}(y; \theta)$$

If there is a $\theta^1 \in \Theta$ such that $p^1 = p_{\mathbf{S}}(;\theta^1)$ then under some minor assumption (what are theeez???) we get the well known result that $\theta_{ML}(y)$ converges in (prob, a.s., ...) to θ^1 as $n \to \infty$.

2.6 Sampling from the posterior

Markov Chain Monte Carlo Methods

1. Metropolis-Hastings Algorithm

3 Hierarchical Models

In the following subsections we will introduce the general idea of Hierarchical models; show how to solve for the posterior analytically in a simplified setting and then present the main model with which we will be working later on, the *Hierarchical Linear model*.

3.1 Hierarchical Data

3.2 Solving for the posterior analytically

As in the previous section, we first will be presenting the analytical derivation of the posterior in a simple normal hierarchical model.

3.3 Hierarchical Linear Models

3.3.1 Varying Slopes Model With One Predictor In Each Level

We assume that units $i = 1, \ldots, n$ can be divided into J distinct groups. We start with a very simple model assuming that intercept is fixed for all groups, that is

$$y = \alpha + \beta_j x + \epsilon\,, \tag{1}$$

with ϵ following a mean zero normal distribution with variance σ_ϵ^2. To incorporate the idea that the groups follow a common structure we also assume

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta\,, \tag{2}$$

for $j = 1, \ldots, J$, with η mean zero normal with variance σ_η^2.

Since γ_0 and γ_1 do not vary by group they are sometimes referred to as *fixed effects*. Similary as η is drawn randomly for each group it is sometimes called *random effect*. Put together this shows the close resemblance of the hierarchical linear model to classical mixed effects models (some reference here would be nice!)

Following the notation of Gelman and Hill (2007) we describe the model equa-

tion of a single individual i by

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \,, \tag{3}$$

where $j[i]$ denotes the group to which individual i belongs.

3.3.2 Varying Intercept and Slope Model with One Predictor in Each Level

We assume that units $i = 1, \ldots, n$ can be divided into J distinct groups. In each group j we model our outcome variable y as a linear function in x, that is

$$y = \alpha_j + \beta_j x + \epsilon \,, \tag{4}$$

with ϵ following a mean zero normal distribution with variance σ_ϵ^2. To incorporate the idea that the groups follow a common structure we also assume

$$\alpha_j = \gamma_0^\alpha + \gamma_1^\alpha u_j + \eta_\alpha \tag{5}$$

$$\beta_j = \gamma_0^\beta + \gamma_1^\beta u_j + \eta_\beta \,, \tag{6}$$

for $j = 1, \ldots, J$, with

$$\begin{bmatrix} \eta_\alpha \\ \eta_\beta \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ & \sigma_\beta^2 \end{bmatrix} \right) \tag{7}$$

Following the notation of Gelman and Hill (2007) we describe the model equation of a single individual i by

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i \,, \tag{8}$$

where $j[i]$ denotes the group to which individual i belongs. This particular model is known as the two-level varying intercept / varying slope model with one unit-level predictor (here x_i) and one group-level predictor (here u_i). The model defined by equations (1) - (4) can of course be made arbitrarily complex by adding higher order polynomial terms or more predictors, as in a regular linear regression models. Further the normality assumption of equation (4) is not mandotory and can be swapped with nearly any other distributional assumption. Also, why stop at two levels? We could naturally model the coefficients in equation 2 and 3

using a third level. All these extensions can in practice be necessary when modelling complex structures; however for the sake of simplicity and clarity we will stick to our clean model.

4 Monte Carlo Study

5 Application

6 Conclusion