Hierarchical Bayesian Models

Research Module - Econometrics and Statistics - 2019/2020

Linda Maokomatanda, Tim Mensinger, Markus Schick

Contents

1 Introduction

2 Bayesian Thinking and Estimation

In this section we will introduce the core topics of Bayesian data analysis, and whenever possible compare proposed methods and results to their frequentist counterpart. As we will see, Bayesian statistics differs from mainstream statistics on a fundamental level; Thus, we have to start there.

2.1 (Probabilistic) Modeling

Before going further, let us first formalize our notion of stochastic modeling. Imagine being interested in some *phenomena*, for example the effect of bigger class sizes on school children performance. Most phenomena cannot be observed directly and only manifest themself through some latent variable system. We can still hope to learn about the phenomena by studying the observational process around it. Clearly the way in which a phenomena reveals itself is dependent on its environment; The observed effects for school-children in sub-saharan africa might look very different to the ones in central europe. To derive sensible results from our analysis we need to postulate the existence of a *true data generating process*, which captures the way in which the observational process adheres to the effects of the phenomena of interest given its environment. We define this object as a probability distribution p_0 on the observational space $\mathbb{Z}$. In this step we move from a model to a probabilistic model, in that we allow our system of interest to be influenced by randomness and not only be characterized by a deterministic process. In practice p_0 is rarely known, therefore the challenge lies in recovering a distribution using the observed data which is as close as possible to p_0. This is usually done by assuming that the true data generating process falls in some class of models, for example a linear model with normal errors. Formally, we restrict our attention to a subset of potential observational processes $\mathbb{M}$ over the whole space of distributions on $\mathbb{Z}$. The beauty of using a model class approach in the construction of $\mathbb{M}$ is that for each distribution $p \in \mathbb{M}$, we can find a parameterization θ in the configuration space Θ; For example, the class of multivariate normal distributions is parameterized by its mean and covariance $(\mu, \Sigma) = \theta \in \Theta$. The goal of all subsequent statistical analysis is then to utilize the observed data to determine the regions in Θ which are most consistent with p_0 and simultaneously to capture our uncertainty about these statements. If $p_0 \in \mathbb{M}$ we can find a parameterization θ_0 which corresponds to the true data generating process; naturally we seek estimators that determine regions close to θ_0. If, however, $p_0 \notin \mathbb{M}$ we enter the world of model misspecification which leads to all sorts of problems. For everything that follows let us therefore make the omnipresent assumption that $p_0 \in \mathbb{M}$.

2.2 Schools of Thought

Next we discuss how the Bayesian and the classical mindset differ. In particular we focus on the predominant way of thinking for most of statistical history, *frequentist statistics*.

Frequentist. The frequentist approach assumes that the true data generating process is completely specified by an unkown but fixed quantity $\theta_0 \in \Theta$. Either implicitly or explicitly we define the *likelihood* $p(z;\theta)$ by modeling the observational process for $z \in \mathbb{Z}$ using a parameterization θ. The main challenges include finding a (point) estimator for θ_0, quantifying the uncertainty of the estimate and testing hypothesis. One fundamental idea which stretches over all these topics is the interpretation of probability as the limit of an infinite sequence of relative frequencies —hence the name. Since θ_0 is fixed all probability statements regarding this object are trivial, that is, either one or zero. This propagates to the problem of hypothesis testing. All hypothesis are either true or false and therefore have probabilities of one or zero. We reject a hypothesis H_0 if conditional on H_0 being true the probability of observing the data in the given sample is lower than some threshold, i.e. $P(\text{data} \mid H_0) < \alpha$. Note that this statement does not tell us anything about H_0 directly but only about the data at hand.

Bayesian. In a Bayesian framework we may believe that there exists a true underlying data generating process which might even be specified by a fixed quantity, but we start to model our initial uncertainty by imposing a probability distribution $p(\theta)$ on Θ. This higher-level uncertainty can also be interpreted as *prior* knowledge about the parameters —we call $p(\theta)$ the *prior distribution*. As we impose probabilistic structure onto Θ....

Bayesian Inference

1. 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.

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

2. 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?)

3. 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.

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

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

5. 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.

6. 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].

7. 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.

8. 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.

9. 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?

10. 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.

2.3 Solving for the posterior analytically

In this subsection we present the analytical estimation of the mean and variance parameters in a univariate normal model using an uninformative and conjugate prior, respectively. We will compare the derived posteriors to the usual maximum likelihood estimate.

In both cases we assume that we observe an iid data sample $y = (y_1, \ldots, y_n)$ with $y_i \sim \mathcal{N}(\theta, \sigma^2)$. Our interest lies in solving for the marginal posterior distributions $p(\theta \mid y)$ and $p(\sigma^2 \mid y)$. Since it will be use frequently, note that by dropping all irrelevant constants we get

$$p(y \mid \theta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - \theta)^2\right) . \tag{1}$$

ADD STUFF TO MAP AND MEDIAN ESTIMATE AND COMPARE TO MAXIMUM LIKELIHOOD!

Uninformative Prior

A reasonable choice for an uninformative prior for (θ, σ^2) stems from defining a uniform prior on the transformed parameters, i.e. $p(\theta, \log \sigma) \propto 1$. The log transformation is necessary since σ is constrained to be positive. This leads to the improper prior[1]

$$p(\theta, \sigma^2) \propto (\sigma^2)^{-1} , \tag{2}$$

which gives the joint posterior as

$$p(\theta, \sigma^2 \mid y) \propto p(y \mid \theta, \sigma^2) p(\theta, \sigma^2) \tag{3}$$

$$\propto (\sigma^2)^{-(n+2)/2} \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - \theta)^2\right) . \tag{4}$$

Thus the marginal posterior of σ^2 can be obtained by integrating the joint posterior over θ. Let $s^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$ denote the sample variance and note that

[1] See appendix for a derivation.

it is easy to show to $\theta \mid \sigma^2, y \sim \mathcal{N}\left(\bar{y}, \sigma^2/n\right)$ (see appendix for a proof). Then,

$$\begin{aligned}
p(\sigma^2 \mid y) &\propto \int p(\theta, \sigma^2 \mid y) \mathrm{d}\theta && (5)\\
&\propto \int \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right) \mathrm{d}\theta && (6)\\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right) \mathrm{d}\theta && (7)\\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \theta)^2]\right) \mathrm{d}\theta && (8)\\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \int \exp\left(\frac{1}{2\sigma^2/n}(\bar{y} - \theta)^2\right) \mathrm{d}\theta && (9)\\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \sqrt{2\pi\sigma^2/n} && (10)\\
&\propto (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right). && (11)
\end{aligned}$$

Hence, $\sigma^2 \mid y \sim \text{scaled-Inv-}\chi^2(n-1, s^2)$.

To finish our analysis we integrate the joint posterior over σ^2 to get the marginal posterior of θ. We evaluate the integral by substitution using $z = \frac{a}{2\sigma^2}$ with $a = (n-1)s + n(\theta - \bar{y})$.[2] Then,

$$\begin{aligned}
p(\theta \mid y) &= \int_{(0,\infty)} p(\theta, \sigma^2 \mid y) \mathrm{d}\sigma^2 && (12)\\
&\propto \sigma^{-(n+2)} \int_{(0,\infty)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \theta)^2]\right) \mathrm{d}\sigma^2 && (13)\\
&\propto a^{-n/2} \int_{(0,\infty)} z^{(n-2)/2} \exp(-z) \, \mathrm{d}z && (14)\\
&\propto a^{-n/2} && (15)\\
&= [(n-1)s + n(\theta - \bar{y})]^{-n/2} && (16)\\
&\propto \left[1 + \frac{(\theta - \bar{y})^2}{(n-1)s^2/n}\right]^{-n/2} && (17)
\end{aligned}$$

which concludes our first analysis implying that $\theta \mid y \sim t_{n-1}(\bar{y}, \sigma^2/n)$.

We know that for the problem at hand the standard maximum likelihood es-

[2] See appendix for explicit derivation

timators and their variances are given by[3]

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_i y_i = \bar{y} \tag{18}$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{n-1}{n}s^2 \tag{19}$$

$$\mathrm{I}(\theta, \sigma)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix} . \tag{20}$$

The Bayesian counterpart to the ML-Estimator is the *maximum a posteriori estimate*, or in short *MAP*. For $\theta \mid y$ we derived a noncentral Student's t-distribution with mean $\bar{y}$ and variance σ^2/n. Since this distribution is unimodal and symmetric the MAP estimate is simply the mean. We note that it is equivalent to the ML estimate on both the point estimate and included variance. For σ^2 the results look slightly different. The mode and the mean of $\sigma^2 \mid y$ are given by $\frac{n-1}{n+1}s^2$ and $\frac{n-1}{n-3}s^2$, respectively. Still we see a very close resemblance of the Bayesian estimator to the ML estimator. One could say that this is a property which is desirable for Bayesian estimators using uninformative priors. One obvious advantage of the Bayesian approach is the ease with which we can compute arbitrary probabilities using the posterior density.

Conjugate Prior

Consider again the likelihood function

$$p(y \mid \theta, \sigma^2) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}a\right) . \tag{21}$$

A conjugate prior for (θ, σ^2) therefore has to be of the form

$$p(\theta, \sigma^2) \propto \sigma^{\alpha} \exp\left(-\frac{1}{2\sigma^2}\beta\right) \tag{22}$$

for some α, β. Using the factorization $p(\theta, \sigma^2) = p(\sigma^2)p(\theta \mid \sigma^2)$ we see that for some $\alpha_0, \alpha_1, \beta_0, \beta_1$ we find

$$p(\sigma^2) \propto \sigma^{\alpha_0} \exp\left(-\frac{1}{2\sigma^2}\beta_0\right) \tag{23}$$

[3] See appendix for proof.

and

$$p(\theta \mid \sigma^2) \propto \sigma^{\alpha_1} \exp\left(-\frac{1}{2\sigma^2}\beta_1\right) . \tag{24}$$

This then tells us that $\theta \mid \sigma^2$ is normal with variance proportional to σ^2, while σ^2 has to be scaled-Inv-χ^2 distributed. We parameterize the distributions as follows

$$\begin{aligned} \theta \mid \sigma^2 &\sim \mathcal{N}\left(\theta_0, \sigma^2/\kappa_0\right) & (25)\\ \sigma^2 &\sim \text{scaled-Inv-}\chi^2(\nu_0, \sigma_0^2)\,, & (26)\end{aligned}$$

for some suitable hyperparameters $\theta_0, \kappa_0, \nu_0, \sigma_0$ and follow Gelman BDA in terming the resulting joint density $p(\theta, \sigma^2)$ by Normal-Inv-$\chi^2(\theta_0, \kappa_0; \nu_0, \sigma_0)$. This in turn ensures that the joint posterior $p(\theta, \sigma^2 \mid y)$ is again Normal-Inv-χ^2. Using an equivalent approach as applied above we can compute the marginal posteriors by intregrating the respective parameters out. This results in

$$\begin{aligned} \sigma^2 \mid y &\sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2) & (27)\\ \theta \mid y &\sim t_{\nu_n}(\theta_n, \sigma_n^2/\kappa_n)\,, & (28)\end{aligned}$$

for

$$\begin{aligned} \nu_n &= \nu_0 + n & (29)\\ \kappa_n &= \kappa_0 + n & (30)\\ \sigma_n^2 &= \left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_n n}{\kappa_0 + n}(\bar{y} - \theta_0)^2\right] / \nu_n & (31)\\ \theta_n &= \frac{\kappa_n}{\kappa_0 + n}\theta_0 + \frac{n}{\kappa_0 + n}\bar{y}\,. & (32)\end{aligned}$$

A detailed derivation is given in the appendix.

2.4 Asymptotics

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

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

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

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 $n \to \infty$.

2.5 Sampling from the posterior

In this section we will consider how to estimate quantities of interest using an unnormalized posterior density (target distribution). These quantities can be very different objects. For instance, we might be interested in visualizing the marginal posterior distribution of some specific parameter, or we want to propagate the posterior distribution through some not necessarily linear transformation function. In Bayesian and frequentist statistics alike, a common approach of summarizing results of this kind is to compute expectations.

In the following we will review methods to draw samples from the posterior distribution which, as we will see, goes hand in hand with approximating expectations. To illustrate potential problems of approximating expectations in higher dimensions, we will first consider the case in which we are equipped with an unimodal normalized posterior density.

Let p be the target density on some smooth space Q. Let f be some function. The quantity of interest is given by the corresponding expectation

$$\mathbb{E}_p [f] = \int_{\mathrm{Q}} f(q)p(q)\mathrm{d}q\,. \tag{33}$$

For complex models we are usually not able to evaluate these integrals analytically, therefore we have to approximate the integral numerically. It is clear that for higher dimensions we run into problems when using naive quadrature methods which is known as the *curse of dimensionality*. In particular when considering high-dimensional spaces most regions will not contribute to the expectation. This can have three causes. One, the region has a negligible density $p(q)$; two, the function has a negligible value $f(q)$; and three, the volume $\mathrm{d}q$ is negligible. Since we care about general methods that can be applied to various functions f we will assume from here on that f has non-negligible values on the whole space.[4] Using the intuition from above we are interested in examining the regions of the space in more detail for which the density has a significant impact and the volume is

[4] Still, for specific applications in which only one expectation is relevant we can make use of the functional form of f.

Dimension	Inner Volume	Outer Volume
1	0.9544997361036416	0.045500261923183016
2	0.9110697462219214	0.08893024983172781
5	0.7922806756813302	0.2077193144527928
10	0.6277086690580651	0.3722913112101811
50	0.0974519722529203	0.9025479290883144
100	0.009496886895983958	0.9905029157864953

Table 1: Comparison of volume close to mode and in the regions around the mode for different dimensions.

non-negligible. We will call this set the typical set.[5] To gain more intuition on how this set looks we will consider the normal density for various dimensions. Divide the space into boxes with side length 4 in such a way that we have one box centered at the mode of the density and one more adjacent box at each side. I.e., in one dimensions we find 3 boxed, in two dimensions we find 9 boxes and so on. Now for each dimension we compute the probability mass falling into the box containing the mode and the boxes surrounding the mode. For the normal case we can compute these values exactly.

```
from scipy.stats import norm
rv = norm(loc=0, scale=1)
def inner_volume(dim):
  return (rv.cdf(2)-rv.cdf(-2))**dim
def outer_volume(dim):
  return (rv.cdf(6)-rv.cdf(-6))**dim-(rv.cdf(2)-rv.cdf(-2))**dim
```

Table 1 shows the probability mass contained in the inner box centered at the mode in comparison to the aggregated probability mass of the boxes surrounding the inner box. As the dimensionality of our problem increases we observe that more and more mass is concentrated not in the regions where the density is highest, around the mode, but in the its adjacent neighborhoods. When constructing estimation techniques we have to account for this counter-intuitive behavior. In particular, we are intersted in exploring some sub-manifold of the relevant space which contains the non-negligible information on the expectation which can be defined to be invariant to the dimensionality of the problem, i.e. the typical set.

[5] REEFFEEERENCE.

2.5.1 Markov Chain Monte Carlo

Next we present a simple class of estimators which produce samples from the posterior distribution by constructing a *Markov Chain* that has the posterior distribution as its stationary distribution.

Let $T(q', q)$ denote the markov transition kernel. If we can find a T s.t.

$$p(q) = \int_Q T(q', q)p(q')\mathrm{d}q' , \tag{34}$$

then the markov chain will have p as its limiting distribution.PROOOOOF. The intuition behind equation 34 is that if we sample from the target distribution and apply the transition kernel, we also want the new ensemble of samples to be distributed accoding the target distribution. But how do we generate actual samples?

Let us assume we can draw samples $\{q_0, ..., q_N\}$ from the posterior distribution through following the markov chain. Then a straightforward estimator is $\hat{f}_N := \frac{1}{N}\sum_i f(q_i)$. Under some conditions WHAT ARE THEEESE??, we get

$$\hat{f}_N \xrightarrow{\mathrm{P}} \mathbb{E}_p(f) . \tag{35}$$

Unfortunately this only tells us something about the limit. To learn more about the finite sample behavior we can make use of a central limit theorem for MCMC estimators. Under ideal circumstances we get

$$\hat{f}_N^{MCMC} \overset{a}{\sim} \mathcal{N}\left(\mathbb{E}_p(f), \text{MCMC-SE}\right) , \tag{36}$$

where MCMC-SE $= \sqrt{\text{Var}_p(f)/\text{ESS}}$ with ESS denoting the effective sample size. ESS can be estimated via $\hat{\text{ESS}} = N/(1 + 2\sum_{l=1}^{\infty} \rho_l)$, where ρ_l represents the autocorrelation of lag l of the simulated markov chain.

Since we start the markov chain with arbitrary starting values it can take some steps until the chain meanders through the relevant regions. Therefore in practice we throw away the first few hundred samples. This is known as *warm up* or *burn-in* in the literature.

What are the *ideal* circumstances so that equation 36 holds? A sufficient condition is given by the assumption of *geometric ergodicity*. WHERE WHERE WHERE??.

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm provides us with a way to jump from on point in a space to another using the established stochastic structure of the density from which we are drawing samples and the transition kernel. The transition kernel can be thought of as a proposal density. Given q we propose a draw q' from $T(q' \mid q)$. The idea of the algorithm is that we accept this new point only with some probability a, where

$$a(q' \mid q) = \min\left(1, \frac{T(q \mid q')p(q')}{T(q' \mid q)p(q)}\right) . \tag{37}$$

Hence with some starting value q_0, the target density p and some transition kernel T we can simulate a markov chain which will at some point produce points that resemble draws from p. However, in higher dimensions this is still not enough, as the estimator does not scale and becomes highly inefficient REFEREREREC NCE.

2.5.2 Foundations of Hamiltonian Monte Carlo

Using random walk metropolis or similar guess and verify algorithms is too costly from a computational perspective in higher dimensions.

Question: How can we use the geometry of the typical set to get information on how to move through it? Answer: For continuous spaces we could have a vector field which is aligned with the typical set? Starting at some point in the typical set we would only have to move in the direction with the given momentum as given by the vector at this point and would again land in the typical set with a new vector.

But this defers the question only to another question: How do we get a vector field which is aligned with the typical set?

A usual starting point for question of this nature is looking at the implied differential structure of the target. This we get via the gradient. In particular, the gradient defines a vector field in the given space which is sensitive to the structure of the target in a way s.t. it points us to the extrema. But as we have seen above, the extrema (modes) are not necessarily where we expect a lot of probability mass. In fact, we've seen that the higher the dimension the more mass is concentrated exactly around some region centered at the mode. A clever analogy from physics can help us out here. Think of the modes as centers of gravity, for example a planet. The typical set floats around the planet. With higher dimensions we've observed that the orbit tends to be farther away from the planet. That is, we want to place an object in the space, such that it does not come to close to the planet but also does not drift away. In particular we need to give the object a certain velocity so that the gravitational (gradient) vector field keeps the object at a steady distance to the center (on the typical set). In our probabilistic setting this means that we have to expand the original probabilistc system with an auxiliary momentum (velocity) parameter.

Phase space and Hamilton's Equations

3 Hierarchical Models

In the following subsections we will introduce the concept of hierarchical data and the models designed to work with that data. We present an analytical derivation for a simple but general case and then illustrate the most common model class, *linear hierarchical models*.

3.1 Hierarchical Data

Hierarchical data is present when there is a natural way to split observations into clusters. For example, we might observe data on children for many different schools. The data could also include schools for different states. So we would observe children in schools and schools in states. This would constitute the case of *nested* groups, which gives meaning to the term hierarchical. There are however many cases which do not feature nested data or an interpretable hierarchical structure but do belong to the type of structured data that can be analyzed using hierarchical models. A prominent example are meta-analysis studies, which can be modeled hierarchically but are non-nested since units might be overlapping. When a clear interpretation of the hierarchy is missing one often encounters the description *multi-leveled data* and *multi-level model*.

Formal Example

Let us assume that we observe test scores y_i for $i = 1, ..., n$ children in $j = 1, ..., J$ different schools. For convenience let us write $j[i]$ for child i's school. We assume that for all children in school j, the outcome y_i follows a common data generating process governed by some parameter θ_j. To make this model hierarchical we assume that these $\theta_1, ..., \theta_J$ also follow a common data generating process governed by some hyperparameter ϕ. This gives

$$p(y_i \mid \theta_{j[i]}, \phi) = p(y_i \mid \theta_{j[i]}) \tag{38}$$

for our model of test scores given all parameters. Note that since ϕ only affects y_i through $\theta_{j[i]}$ we can drop it from the conditioning set. Further we model the common structure of the θ_j's through $p(\theta_j \mid \phi)$. To analyze this problem using Bayesian techniques we must at last assign a prior to ϕ, that is, we define $p(\phi)$.

3.2 Solving for the posterior analytically

As in the previous section, we will first present the analytical derivation of the posterior in a simple normal hierarchical model. We continue using the example of modeling observed test scores of children in different schools.

The objects of our interest are given by the joint posterior $p(\theta_j, \phi \mid y)$ and its two marginal posteriors. In order to derive these analytically we must make convenient distributional assumptions. For the sake of exposition, we choose

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

$$y = \alpha + \beta_j x + \epsilon \, , \tag{39}$$

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

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta \, , \tag{40}$$

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 a *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 equation of a single individual i by

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \, , \tag{41}$$

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

The model defined by the assumptions and equations above (where and what) can of course be made arbitrarily complex. For example we could add higher order polynomial terms or more predictors, as in regular linear regression modeling. Further, the normality assumption of the errors could be relaxed and most importantly, why stop at two levels? Naturally we could model the group-level coefficients using a third level. For the sake of a simpler explaination however, we will stick to the presented model.

4 Monte Carlo Study

5 Application

5.1 MCMC in practice

1. Practioners face mutliple problems when trying to apply Bayesian models. A prominent example is the selection of a right prior.

2. The other important consideration is checking convergence of the mcmc chain. Asymptotic theory tells us that the MCMC will converge with a probability of one to the true density for an unlimited number of steps. Practioners are interested in the performance after only a limited number of steps. Typically we initate our chain with a number of steps we discard later (burn-in) and test convergence based on the rest of the draws.

3. The easiest approach to check convergence is a mere graphical analysis. If the MCMC reached the underlying distribution, new parameters should be drawn around the the mean of the modell. Therefore, the timeseries of the draws should look similar to a stationary process. If the underlying distribution is not reached yet, a slope should be observed.

4. We can quantify this approach by calculating a variety of different convergence criterias. Simply spoken, they measure wether different subsection of a chain describe the same underlying distribution. One of the simplest approaches is based on Geweke(1992) and compares the mean of the draws in one subsection of the chain to an other. Inutitively, both should be the same. One diffulty lies in the correction of means by standard deviations, which need to be adjusted for the autocorrelation as draws are not independent from each other. The underlying test is a t-test $\mathbf{E}\left[g(\theta) \mid Y^T\right], i \in A, C$

   $$\mathbf{CD}_{GWK} = \hat{G}_{S_A}$$

5. Our initial parameter values might habe a sizable effect on our reached distribution. That is why another part of the literature (based on Brooks and Gelman) focuses on starting with different values and comparing the effects on final posterior. If the parameters estimation of the multiple chains align, we can be more convinced that we hit the true distribution of the chain.

6 Conclusion