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.

What remains to be introduced:

1. Notation: $p(...)$

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*. We do not aim at an exhaustive overview here nor do we presume that the individual statistician belongs to one and only one of the following categories.

Frequentist. The frequentist approach assumes that the true data generating process is completely specified by an unkown but fixed quantity $\theta_0 \in \Theta$. FOLLOWING IS NOT TRUE, MAKE

IT CLEAR WHAT YOU MEAN. 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. That is, the probability of an event happening is just the limit of the frequency of that event happening over infinitely many independent experiments. What are the implications of this understanding of probability? Many interesting questions do not provide us with a thought experiment in which we can consider an ever increasing sequence of experiments. In these cases using probability is either trivial or lacking an indisputable interpretation. As we use the mathematical rigor of probability theory in our formal derivations, we will obtain (mathematically) correct results; however, the interpretation of these results might be highly unintuitive —for example consider confidence intervals. 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. A hypothesis H_0 is rejected 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. The Bayesian approach also assumes that there may be a true data generating process specified by some (maybe fixed) quantity $\theta_0 \in \Theta$. The main difference is their understanding of probability as a subjective quantification of uncertainty. In this view one is not limited to assigning non trivial probability statements only to objects that appear random in sequential experiments. We can see the direct utility of this liberation by considering a special case of Bayes theorem

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)p(\theta)}{p(\text{data})} \propto p(\text{data} \mid \theta)p(\theta)\,, \tag{1}$$

which reads

$$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}} \propto \textit{likelihood} \times \textit{prior}\,. \tag{2}$$

In the frequentist setting this is of no use, since the statement $p(\theta)$ is nonsensical —remember that probabilistic statements about fixed quantities are meaningless from a frequentist perspective. This already outlines the main criticism of Bayesian analysis: Where does the prior $p(\theta)$ come from? With the scientific goal of objectivity in mind, many feel uneasy with results being dependent on a subjective choice of a prior. In what follows we will embark on the Bayesian idea without providing much more fundamental criticism; nonetheless, when adequate we will consider the influence of different priors on the posterior. To end this comparison, what then are the main tasks associated with a Bayesian analysis? These can be categorized by (i) obtaining the posterior distribution (or something equivalent) and (ii) communicating the information held in the posterior. The first consists of defining the likelihood and (additionally to the frequentist approach) constructing a prior distribution and afterwards combining those to compute the posterior. This computation can sometimes be achieved analytically, but in most cases one has to rely on algorithms to obtain samples of the posterior. The second part consists of plotting the marginal posterior distributions, computing expectations of the form $\mathbb{E}\left[h(\theta) \mid \text{data}\right]$ and testing hypoth-

esis. All of the above can be done independent of the posterior being available analytically or through samples. A clear difference can be seen when considering hypothesis testing. Sacrificing *objectivity* allows us to answer the questions we usually want to ask:

$$P(H_0 : \theta \in S \mid \text{data}) = \int_{\theta \in S} p(\theta \mid \text{data}) \mathrm{d}\theta \,. \tag{3}$$

In the subsequent paragraphs we will be mostly occupied with the first category, computing the posterior, with occasional remarks on the second; in particular, the approximation of expectations.

2.3 Solving for the posterior analytically

In this subsection we present the analytical derivation of the posterior distribution of mean and variance parameters in a univariate normal model for two priors. We will compare the results to the appropriate classical method, namely maximum likelihood.

In both cases, let us assume that we observe an iid sample $y = (y_1, ..., y_n)$ with $y_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Our interest lies in solving for the marginal posteriors $p(\mu \mid y)$ and $p(\sigma^2 \mid y)$.

To be precise, in a full bayesian analysis we assume $\theta = (\mu, \sigma^2)$ to be a random quantity, thus the correct statement should be $y_i \mid \theta \sim \mathcal{N}\left(\mu, \sigma^2\right)$. In situations where this conditional dependence is clear many writers will use the first notation. Here we try to be as pedantic as possible to avoid any confusion and will therefore stick to the second notation.

As it will be of major importance in the subsequent sections we remind the reader of some probability distributions uncommon in the non-Bayesian world.

Definition 2.1. (Scaled inverse χ^2 distribution). Let $\nu > 0$ and $\tau^2 > 0$ be parameters representing degrees of freedom and scale, respectively. The family of *scaled inverse* χ^2 *distributions* is characterized by its probability density function, namely

$$p(x) \propto x^{-(1+\nu/2)} \exp\left(\frac{-\nu\tau^2}{2x}\right) \quad \text{for } x \in (0, \infty)\,, \tag{4}$$

where the constant of integration is ignored for clarity. We write $X \sim$ scaled-Inv-$\chi^2\left(\nu, \tau^2\right)$ to denote that the random variable X follows a scaled inverse χ^2 distribution with parameters ν and τ^2.

Definition 2.2. (Normal scaled inverse χ^2 distribution).

Uninformative Prior

We start our first Bayesian analysis by considering a prior which contains virtually no information. This results in an analysis being mainly, if not completely, driven by the likelihood. A common assumption is independence of the individual priors, that is $p(\theta) = p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$. A natural choice of declaring full ignorance of prior information is to assign a prior over the complete domain of the random parameter. For our case this mean $p(\mu) \propto 1$. We note that this does not define proper proability distribution, which will not matter in this case but can lead to problems in others.[1] Since the variance is restricted to be positive we impose a uniform prior on the

[1] LINK TO PAPER

log-transform thereof: $p(\log \sigma) \propto 1$. This leads to the improper prior[2]

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}. \tag{5}$$

As usual the likelihood is given by

$$p(y \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right), \tag{6}$$

where we dropped all proportionality constants. Using the above we can apply Bayes theorem to yield

$$p(\mu, \sigma^2 \mid y) \propto p(y \mid \mu, \sigma^2) p(\mu, \sigma^2) \tag{7}$$

$$\propto (\sigma^2)^{-(n+2)/2} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right). \tag{8}$$

From here we can derive the marginals by integrating out the respective other parameter. This is formalized in the following two proposition.

Proposition 2.3. *Under the uniform prior from above we find*

$$\mu \mid y \sim t_{n-1}(\bar{y}, s^2/n), \tag{9}$$

$$\sigma^2 \mid y \sim \text{scaled-Inv-}\chi^2\left(n-1, s^2\right), \tag{10}$$

where $s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$ *denotes the sample variance and* $\bar{y} = \frac{1}{n}\sum_i y_i$ *the sample mean.*

Proof. See appendix. □

Having derived the marginal posterior distributions, we can compare the results to their maximum likelihood (ML) counterpart. Since the ML estimates $\left(\text{argmax}_{\theta \in \Theta} p(y \mid \theta)\right)$ are point estimates, we consider the similar *maximum a posteriori* (MAP) estimate $\left(\text{argmax}_{\theta \in \Theta} p(\theta \mid y)\right)$, as well as the posterior mean and variance. All results are summarized in table 1.

Parameter	ML Estimate	ML Variance	MAP	Posterior Mean	Posterior Variance
μ	$\bar{y}$	σ^2/n	$\bar{y}$	$\bar{y}$	s^2/n
σ^2	$\frac{n-1}{n}s^2$	$2\sigma^4/n$	?	$\frac{n-1}{n-3}s^2$	$\frac{2(n-1)^2}{(n-3)^2(n-5)}s^4$

Table 1: Comparison of Bayesian estimates using an uninformative prior and ML estimates. See appendix for a detailed derivation.

Conjugate Prior

We have seen that using an uninformative prior leads to results that are very similar to the ones obtained by a ML approach. In case information on the parameters is available prior to observing the data we can utilize this fact by properly modeling the prior distribution. Since we are interested in analytical results in this section,b we cannot mix any prior with any likelihood, as the product might not be of known form. This leads us to the class of *conjugate priors*.

[2] See appendix for a derivation.

Definition 2.4. (Conjugate prior). Let the likelihood $p(y \mid \theta)$ be given and assume that the prior distribution $p(\theta)$ is a member of some family $\mathcal{F}$ of probability distributions. We say that $p(\theta)$ is a *conjugate prior* if the posterior $p(\theta \mid y)$ is also a member of $\mathcal{F}$.

Conjugate priors were of particular importance in the early stages of Bayesian statistics since these give the practitioner certainty that the posterior follows a distribution which is known and computable. Moreover, nowadays we still see conjugate priors in use as they allow for a full or partial analytical derivation, which increases the accuracy of results or shortens the runtime of programs. For more complex models however conjugate priors can become too restrictive. We discuss solutions to this problem in the next section.

Consider again the likelihood but written to demonstrate its dependence on μ and σ^2

$$p(y \mid \mu, \sigma^2) \propto (\sigma^2)^{n/2} \exp\left(-\frac{1}{2\sigma^2} n \left[(\mu - \bar{y})^2 + (\bar{y^2} - \bar{y})^2\right]\right). \tag{11}$$

We want to construct a two dimensional prior for (μ, σ^2). A theme to which we will be coming back is that modeling higher dimensional parameters by modeling many lower dimensional (sub)parameters using conditioning is often easier than modeling the complete distribution. Here we utilize the equality $p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2)$. By looking at the likelihood (equation 11) we note that in order to *not* change the inherent structual dependence on the parameters, $\mu \mid \sigma^2$ has to be distributed according to $\mathcal{N}\left(\mu_0, \sigma^2/\kappa_0\right)$ with so called *hyperparameters* μ_0 and $\kappa_0 > 0$. Similary, we note that an informative prior for σ^2 has to respect the structure in which σ^2 appears in the likelihood. We achieve this when $\sigma^2 \sim$ scaled-Inv-$\chi^2\left(\nu_0, \sigma_0^2\right)$ with hyperparameters ν_0 and $\sigma_0^2 > 0$. Following Gelman et al. (2004) we write $(\mu, \sigma^2) \sim$ normal-scaled-Inv-$\chi^2\left(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \sigma_0^2\right)$ with corresponding density function

$$p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2) \propto (\sigma^2)^{\frac{3+\nu_0}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2\right]\right). \tag{12}$$

Multiplying the likelihood with our constructed prior we get the joint posterior (up to an integration constant)

$$p(\mu, \sigma^2 \mid y) \propto (\sigma^2)^{-\frac{3+n+\nu_0}{2}} \times \tag{13}$$

$$\times \exp\left(-\frac{1}{2\sigma^2}\left[(\mu - \bar{y})^2 + (\bar{y^2} - \bar{y})^2 + \nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2\right]\right). \tag{14}$$

Proposition 2.5. *The posterior distribution of* $(\mu, \sigma^2) \mid y$*, as given by the conditional density in equation 14, is* normal-scaled-Inv-$\chi^2\left(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2\right)$ *distributed, where*

$$\nu_n = \nu_0 + n\,; \quad \kappa_n = \kappa_0 + n\,; \quad \mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}\,,$$
$$\sigma_n^2 = \left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2\right] / \nu_n\,.$$

Proof. See appendix. □

Since the prior and the posterior are both normal scaled inverse χ^2 distributed, we can speak of a conjugate prior. Using the intermediate finding from proposition 2.5 we can derive the main

result of this section.

Proposition 2.6. *The marginal posterior distributions are given by*

$$\begin{aligned}\mu \mid y &\sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n) \\ \sigma^2 \mid y &\sim \text{scaled-Inv-}\chi^2\left(\nu_n, \sigma_n^2\right),\end{aligned}$$

where ν_n, σ_n^2, μ_n and κ_n are as in proposition 2.5.

Proof. See appendix. □

Parameter	ML Estimate	ML Variance	MAP	Posterior Mean	Posterior Variance
μ	$\bar{y}$	σ^2/n	μ_n	μ_n	σ_n^2/κ_n
σ^2	$\frac{n-1}{n}s^2$	$2\sigma^4/n$	$\frac{\nu_n}{\nu_n+2}\sigma_n^2$	$\frac{\nu_n}{\nu_n-2}\sigma_n^2$	$\frac{2\nu_n^2}{(\nu_n-2)^2(\nu_n-4)}\sigma_n^4$

Table 2: Comparison of Bayesian estimates using a conjugate prios and ML estimates.

Let us first consider the parameter μ. We note that the posterior mean (and MAP) is given by $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$, which forms a convex combination of the prior μ_0 and the sample average $\bar{y}$, with weights given by the sample size and κ_0. For any fixed n this pulls our estimate of the posterior mean away from $\bar{y}$ and closer to μ_0 (and vice versa). Further, we can use the hyperparameter κ_0 to express our uncertainty in μ_0 (or $\bar{y}$ for that matter). Rewritting the posterior variance using the *Laundau notation* we get $\sigma_n^2/\kappa_n = \frac{n-1}{\nu_n\kappa_n}s^2 + \mathcal{O}(\frac{1}{\nu_n\kappa_n}) = \frac{n}{(\nu_0+n)(\kappa_0+n)}s^2 + \mathcal{O}(1/n^2)$. As the sample size n grows the information contained in the likelihood should dominate the prior. We observe this phenomena as the approximate asymptotic behavior of the posterior resembles that of the maximum likelihood estimator. First, as n tends to infinity $\nu_n = \nu_0 + n$ tends to infinity and the t distribution becomes indistinguishable from a normal. Second, as n grows the posterior mean is dominated by $\bar{y}$. And at last, for large n the posterior variance is accurately approximated by σ^2/n. We refrain from an analogous analysis for σ^2 and only note that similar results hold, as can be seen in table 2.

What is gained from using an informative prior here? Using the conjugate prior from above we have four hyperparameters at hand to model our prior knowledge about the parameters. These can be used to represent very detailed to very vague information. In any case, we were able to see that as we collect more and more data, the likelihood dominates our results. A clear advantage of an analytical derivation is that we know exactly how the prior influences the posterior. However, we have also seen that even for this *very* simple model, the derivation is far from obvious. As we consider more complex models using more parameters we have to make more restrictive assumptions on the way we model our prior information, if an analytical analysis is even possible. For this reason among others, in the next section we present a method which trades off the clarity of an analytical result for the generality of being able to combine near arbitrary priors with complex, possibly high-dimensional likelihoods.

2.4 Sampling From The Posterior

In this section we consider approaches that allow us to characterize the posterior distribution in complex settings using sampling methods.

For the rest of this section let us assume that we observe data $z \in \mathbb{Z}$ and can compute the likelihood $p(z \mid \theta)$ and prior $p(\theta)$ for $\theta \in \Theta$. As before, our goal lies in analyzing the posterior distribution given by $p(\theta \mid z) \propto p(z \mid \theta)p(\theta)$. Unlike before however, we now consider cases where the posterior is highly complex or even non-existent in analytical form, which happens for example when the likelihood contribution stems from a algorithmic computational model.

Say we are somehow able to draw independent samples $\theta_1, \ldots, \theta_n$ from $p(\theta \mid z)$. By independence we get the well known result $\frac{1}{n}\sum_i h(\theta_i) \xrightarrow{d} \mathcal{N}\left(\mathbb{E}\left[h(\theta) \mid z\right], \mathrm{Var}\left(h(\theta) \mid z\right)/\sqrt{n}\right)$, under mild conditions on h and $p(\theta \mid z)$. As we are able to formulate many quantities of interest using expectations —probability statements can be written as expectations— and as we can approximate percentiles from a (large) sample, we should be able to adequately summarize the posterior distribution if we are able to draw (independent) samples from it.

In the subsequent paragraphs we will discuss efficient methods to sample from the posterior, even if we cannot compute the integration constant $\int p(z \mid \theta)p(\theta)\mathrm{d}\theta$. We will see that these methods do *not* produce independent but autocorrelated samples. With this in mind, we follow the creational process of these methods and first state the assumptions which have to be satisfied by the sampling process in order to yield good properties as for example a central limit theorem for dependent samples. Then we present the *Metropolis-Hastings algorithm*, which creates samples that fulfill the above criteria. At last we talk about cases in which the Metropolis-Hastings algorithm fails and what can be done instead.

Markov Chain Monte Carlo

Say we are able to construct a *Markov chain* with unique invariant distribution equal to the posterior distribution we want to sample from. Given we know the transition kernel, Markov chains are very easy to simulate. Hence, we could start a chain, let it run *long enough* and at some point consider all subsequent realizations as draws from the posterior; this is the core idea of MCMC —we defer questions regarding the creation of transition kernels which result in specific invariant distributions until next paragraph. In practice we never know for sure when a chain is run *long enough*. In part 3 we present some measures that help during the application. Here we do what statisticians do best: obsess over central limit theorems. Under mild conditions we can get something similar to a law of large numbers for Markov chains (see e.g. Roberts and Rosenthal (2004), fact 5). This tells us that if we run the chain forever, our average will eventually converge to the number we seek. However, forever is a very long time. That is why we focus on assumptions which admit a central limit theorem with the usual $\sqrt{n}$ convergence rate, as it allows us to make more rigorous statements about our confidence in the whereabouts of the estimator for large but finite samples.

Remark. As is often the case, there are many different sets of assumptions that allow for a CLT. The following theorem presents two sets of assumptions which allow for the desired result. We remark that we will *not* formally introduce all concepts and will provide only a heuristic explaination of the assumptions. This is due to the fact that Markov chain theory on general state spaces requires a good understanding of measure theory which we do not want to assume as a prerequisite. The interested reader is refered to Meyn and Tweedie (2009).

Theorem 2.7. *(A Central Limit Theorem for Markov Chains). Let* $\{X_t : t \geqslant 0\}$ *be a positive Harris Markov chain with invariant distribution* π*. Let* h *be measurable with* $\int h^2 \mathrm{d}\pi < \infty$*. Assume either of the following holds:*

1. $\{X_t : t \geqslant 0\}$ *is uniformly ergodic,*
2. $\{X_t : t \geqslant 0\}$ *is* π*-reversible and geometrically ergodic.*

Then, there exists a constant $\sigma^2(h) < \infty$ *such that*

$$\sqrt{t}\left(\frac{1}{t}\sum_t h(X_t) - \int h \mathrm{d}\pi\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(h)\right). \tag{15}$$

Proof. See Cogburn (1972) for 1 and Roberts and Rosenthal (1997) for 2. □

Extend paragraph with heuristic explaination of assumption.

Metropolis-Hastings Algorithm

From the above we know that given a Markov chain with invariant distribution equal to the posterior distribution $p(\theta \mid z)$, we can treat the realizations of the chain as samples from the posterior, under some regularity conditions. Here we consider one method which implicitly defines such a chain, namely the Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970)). For other approaches and more involved algorithms see for example Roberts and Rosenthal (2004) or Liang et al. (2010).

Algorithm 1 Metropolis-Hastings

Input: (π, q, T) = (target density, proposal density, number of samples to draw)

1: initialize x_0 with an arbitrary point from the support of q
2: **for** $t = 0, \ldots, T-1$ **do**
3: sample a canditate: $y \sim q(\cdot \mid x_t)$
4: compute the acceptance probability: $\alpha(x_t, y) \leftarrow \min\left\{\frac{\pi(y)}{\pi(x_t)}\frac{q(y \mid x_t)}{q(x_t \mid y)}, 1\right\}$
5: update the chain: $x_{t+1} \leftarrow \begin{cases} y & \text{,with probability } \alpha(x_t, y) \\ x_t & \text{,with remaining probability} \end{cases}$
6: **return** $\{x_t : t = 1, \ldots, T\}$

Algorithm 1 displays the Metropolis-Hastings algorithm. Line 4 shows why we can use this algorithm with the unnormalized posterior, as the integration constant cancels out in the first fraction. For different settings different proposal distributions are appropriate. A common choice are so called *random walk* proposals which add some random number to the current position of the chain; for example a gaussian random walk proposal is given by $q(\cdot \mid x_t) = \mathcal{N}\left(\cdot \mid x_t, \sigma^2\right)$ or equivalently stated $y = x_t + \mathcal{N}\left(0, \sigma^2\right)$. If the resulting chain has an unique invariant distribution we only know that after some time the chain starts behaving accordingly. Therefore in practice we choose $T = B + T^*$ and drop the first B samples, where B, the so called *burn-in* samples, is large and T^* represents the actual size of samples we want to draw. See Sherlock et al. (2010) for a recent survey on random walk proposals.

The simplicity of the algorithm is remarkable, but the main question of concern is if the resulting Markov chain inherits favorable properties. Need to list what properties can be derived in general and what can only be derived for specific proposal densities.

The ability to sample draws in complex settings using the Metropolis-Hastings algorithm (and other Markov chain Monte Carlo methods for that matter) made Bayesian statistics applicable for real problems. Still, Au and Beck (2001) show that the classical Metropolis-Hastings algorithm is highly dependent on the proposal density and fails in higher dimensions; Katafygiotis and Zuev (2008) provide a geometric intution. The following paragraph illustrates one of the problems of working in higher dimensions.

Volume in Higher Dimensions

Classical MCMC methods can have too slow convergence rates; In higher dimensions this might be due to probability mass being distributed very far from where it is expected (Betancourt (2017)). In this paragraph we motivate this phenomena and in the following we present methods which utilize it.

Let B_d denote the unit ball in $\mathbb{R}^d$ and define C_d as the smallest cube containing B_d. We consider two questions. First, how does the ratio $\text{vol}(B_d)/\text{vol}(C_d)$ changes as d increases. And second, how does the ratio of probability mass distributed by a standard gaussian on these regions changes as d increases. Since closed form expressions of volumina of geometrical objects exist the first questions needs little work. Similary we can easily compute $P(X \in C_d) = [\Phi(1) - \Phi(-1)]^d$, where Φ denotes the one-dimensional gaussian cumulative distribution function. However, to compute $P(X \in B_d)$ we need to integrate over the unit ball with respect to a gaussian distribution, which is non-trivial. For this reason we decide to report an upper bound, as this is sufficient for our motivation. In particular we compute $\overline{P(X \in B_d)} := \sup_{\boldsymbol{x} \in B_d} \phi(\boldsymbol{x}) \text{vol}(B_d) = \sup_{\boldsymbol{x} \in B_d} \phi(\boldsymbol{x}) \int_{B_d} 1 \mathrm{d}\boldsymbol{x} \geqslant \int_{B_d} \phi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = P(X \in B_d)$. The results of these computations are depicted in table 3. We note that both ratios tend to zero very fast as d increases. With this phenomena in mind one has to be cautious when working in high-dimensional spaces, since the regions of interest, that is the regions containing non-negligible probability mass, might not be located where our low-dimensional intution says. This idea is formalized by the *Gaussian Annulus Theorem* (Blum et al. (2017); theorem 2.9) which states, inter alia, that most probability mass lies within an annulus centered at the origin with an average distance to the origin of $\sqrt{d}$.

d	1	2	3	5	7	10	15
$\text{vol}(B_d)/\text{vol}(C_d)$	1.00000	0.78540	0.52360	0.16449	0.03691	0.00249	0.00001
$\overline{P(X \in B_d)}/P(X \in C_d)$	1.16874	1.07281	0.83589	0.35870	0.10995	0.01184	0.00012

Table 3: Comparison of volume ratio of unit ball and cube, and probability ratio of gaussian falling in unit ball and cube for varying dimension d. Numbers are rounded to five decimal places.

Hamiltonian Monte Carlo

We conclude our digression on Bayesian thinking by presenting *Hamiltonian Monte Carlo*, a innovative Markov chain Monte Carlo method from the statistical physics literature which works in higher dimensions Duane et al. (1987). There are of course multiple MCMC algorithms which work in higher dimensions with many more being actively developed. Here we focus on Hamiltonian Monte Carlo as it is the main algorithm used in the probabilistic programming language STAN (Stan Development Team (2018)), which we will be using in our Monte Carlo study and application part.

EXPLAIN HMC HERE.

3 Hierarchical Models

In the following we consider hierarchical models, when they are applicable and how to estimate them. We begin by presenting the idea of hiearchical data and modeling. Then we show how to solve a simple model analytically, before ending the section with introducing two main ideas. One, hierarchical linear models, arguably the most important subclass of hiearchical models. And two, a method to sample from the posteriors arising in complex settings.

3.1 Hierarchical Data and Modeling

Here we consider what makes data hierarchical and how we can use this component to model additional structure. Hierarchical data is present if the data can be clustered on some level; e.g. children in schools, survey responses on different years in different states or experiments in multiple labs. From the examples we see that there must not be a clear *hierarchy* defined on the data. This is one of the reasons why some authors nowadays prefer the more general terms *multi-level data* and *multi-level model*, see for instance Gelman and Hill (2007). We categorize models by the number of levels they incorporate and if they use *nested* or *non-nested* data. In this paper we consider two-level models for nested data and refer again to Gelman and Hill (2007) for a treatment of models with more levels and non-nested data.

In practice we can store hierarchical data very efficiently using normalized relational dataframes. Let us continue the example of children in schools and assume we observe their results on a test, the parental income and the number of teachers per child in the school. Table 4 portrays how multi-level data in this case could be stored. Having gained some intuition let us define our formal notation.

child	result	income	school
1	10	500	1
2	9	450	1
3	12	520	2
4	10	490	1

school	teacher
1	0.5
2	0.7

Table 4: Two tables containing fictional hierarchical data. Left: Data on the child-level, that is the test *results*, parental *income* and *school* id. Right: Data on the school-level, in this case the number of *teachers* per child.

Assume we observe data on $i = 1, \dots, n$ units which are clustered among $j = 1, \dots, J$ groups. Naturally we consider some outcome y_i on the unit-level. Following the idea of different tables for different levels from above, we write x_i for the covariates that vary by unit and u_j for the covariates that only vary on the group-level. We link the two by writing $j[i]$ for the index of the group to which individual i belongs, i.e. the full set of covariates from individual i is given by $(x_i, u_{j[i]})$. But how can we utilize this hierarchical structure?

The main idea behind hierarchical modeling is that we build (simple) models on each level while using the dependent variables from higher levels as input parameters on lower levels. For

a general (two-level) case we may write

$$y_i \mid \theta_{j[i]} \sim p(y_i \mid x_i, \theta_{j[i]}) \,, \tag{16}$$

$$\theta_j \mid \phi \sim p(\theta_j \mid u_j, \phi) \,, \tag{17}$$

$$\phi \sim p(\phi \mid \zeta) \text{ with } \zeta \text{ fixed} \,, \tag{18}$$

where we suppress the dependence on x_i and u_j by assuming they are fixed. In the first level we model the observations y_i depending on the covariates x_i and parameters θ_j. As in a classical Bayesian model we continue by modeling the parameters; However, in contrast to section 2 we do not just assume some prior distribution for the parameters but we *explicitly* model the parameter. From a Bayesian viewpoint this can be seen as a generalization to prior modeling. Despite this Bayesian interpretation, the first two equations define a proper non-Bayesian hierarchical model. We will see that in the linear case these models are well known in the frequentist world as *mixed effects models* or *random coefficient models*. To put the Bayesian in Bayesian hiearchical model we have to assign a prior distributions on the parameters ϕ. As ϕ are themselves parameters for the parameter θ_j, one often speaks of priors on θ_j and *hyperpriors* on the *hyperparameter* ϕ.

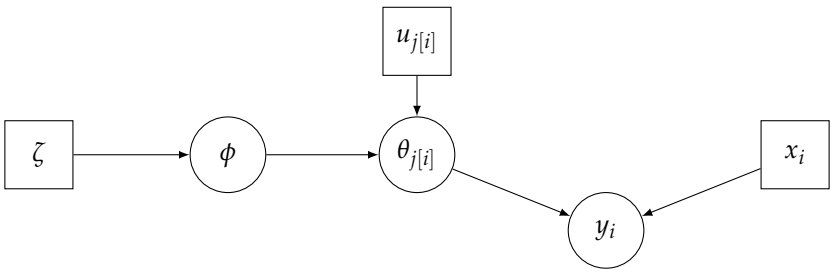

Figure 3.1: A generic two-level Bayesian hierarchical model depicted as a directed acyclical graph modeling a single generic observation. Circled parameters denote random quantities while parameters contained in squares denote fixed quantities.

Figure 3.1 illustrates the conditional dependence structure of modeling a generic observation y_i. In contrast, consider figure 3.1 which illustrates the conditional dependence structure when modeling a generic observation $y(j)$ in group j. We depict random quantities in circles and fixed quantities in squares. Not only do these graphical representations help with understanding the structure of the model but we will see that when solving for the posterior they give an immediate way to check which parameters are conditionally independent.

3.2 Solving for the Posterior Analytically

As in the previous section we consider first an analytical derivation of the posterior distributions using a simple normal model, before dealing with more complex settings.

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

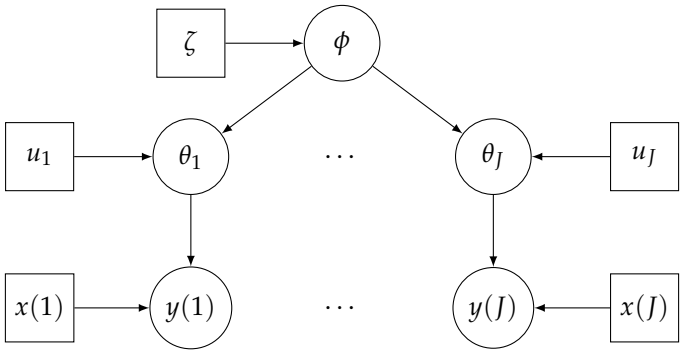

Figure 3.2: A generic two-level Bayesian hierarchical model depicted as a directed acyclical graph modeling generic observations $y(j)$ in groups $j = 1, ..., J$. Circled parameters denote random quantities while parameters contained in squares denote fixed quantities.

3.3 Hierarchical Linear Models

The following subsection presents hierarchical linear models, an important subclass of the general multi-level model. As in classical statistics linear models are usually simpler to estimate and easier to interpret. The standard critique on linear models from classical statistics also applies here; however, we will see that due to the hiearchical nature we are able to model very complex structure even under a linearity assumption.

3.3.1 Varying Slopes Model With One Predictor In Each Level

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

$$y = \alpha + \beta_j x + \epsilon\,, \tag{19}$$

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

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta\,, \tag{20}$$

for $j = 1, ..., J$, 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

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i\,, \tag{21}$$

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

4.1 Convergence

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 take a more quantitative 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. the variance between sequence variance B/N is given by

$$B = \frac{N}{M-1} \sum_{m=1}^{M} (\overline{\theta}_m^{(\bullet)} - \overline{\theta}_\bullet^{(\bullet)})^2$$

7. where

$$\overline{\theta}_m^{(\bullet)} = \sum_{n=1}^{N} \theta_m^{(n)}$$

8. and

$$\overline{\theta}_\bullet^{(\bullet)} = \frac{1}{M} \sum_{m=1}^{M} \theta_m^{(\bullet)}$$

9. The within-chain variance is averaged over the chains,

$$W = \frac{1}{M} \sum_{m=1}^{M} s_m^2$$

10. where

$$s_m^2 = \frac{1}{N-1} \sum_{m=1}^{M} (\theta_m^{(n)} - \overline{\theta}_m^{(\bullet)})^2$$

11. and

$$\overline{\theta}_\bullet^{(\bullet)} = \frac{1}{M} \sum_{m=M}^{N} \theta_m^{(\bullet)}$$

12. The variance estimator is a mixture of the within-chain and cross-chain sample variances,

$$\widehat{var}^+(\theta \mid y) = \frac{N-1}{N} W + \frac{1}{N} B$$

13. Finally, the potential scale reduction statistic is defined by the equation,

$$\widehat{R} = \frac{\widehat{var}^+(\theta \mid y)}{W}$$

14. If the Markov Chain is converged$\widehat{R}$ should be close to 1. Intuitively the variance within a chain should create all the variation of the draws, while the variance between different chains converges to 0.

4.2 Prior selection

15. The selection of a right prior is one critical part of bayesian modelling. And the often the subject to critizism.
16. Following Gelman, we can differentiate between 5 types of priors
17. fllat prior
18. Super-vague but proper prior: normal(0, 1e6);
19. Weakly informative prior, very weak: normal(0, 10);
20. Generic weakly informative prior: normal(0, 1);
21. Specific informative prior: normal(0.4, 0.2) or whatever. Sometimes this can be expressed as a scaling followed by a generic prior: $theta = 0.4 + 0.2 * z; z\ normal(0, 1)$
22. The flat pior (often called uniformative prior) . Consequently, the posterior collapses to the Maximum Likelihood. (from https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations)
23. Another option is a super

4.3 Technical considerations

24. some stuff about bad or good mixing

4.4 Stan

25. For our Bayesian Analysis we use the software Python as well as the software Stan through the interface Pystan

26. Stan compiles the code directly into C and therefore allows the fast analysis need for our monte carlo study.

27. Stan allows a great amount of parametrization. For simplicity we will only focus on a small number of options

28. delta is the metropolis acceptance rate. As shown in above section, mcmc lead to autocorrelated draws. We can therfore set an acceptance rate $\delta \in [0, 1]$. With this probability we accept a new draws with a lower posterior value. Why?

29. A too high acceptance rate will lead to too many draws to be accepted and the chain to wander widly around. As a result the autocorrelation we have a high autocorrelation between each draws.

30. A too low acceptance rate will lead to only values in the middle of the posterior to be accepted. We have a only slowly decaying autocorellation function again.

31. this can be analyzed looking at the autocorrelation plot.(insert some plots here with good or bad mixing)

32. A δ of 0.8 is default. (We change this based on our parametrization)

33. we vary J and N and check the performance of our Bayesian Estimation with the true results

4.5 convergence tests

34. by not setting any starting values stan start automatically with

35. diffuse random initializations automatically satisfying the declared parameter constraints.

4.6 r

esults

36. We cocentrate on 3 different cases in our Simulation study: flat prior, informative true prior and informative wrong prior.

37. we vary J and N and check the performance of our Bayesian Estimation with the true results

38. Based on the package stan-utilty we perform 2 peformance tests in our bayesian analysis

39. First we test wether the empirical percentiles are similar to

5 Application

In this section we discuss the applications of the hierarchical modelling approach in the form of a literature review. We then use real world data to give a practical illustration of the work we have developed in the previous sections of our paper and compare that to the frequentist approach to analyzing multilevel data with the aim of showcasing what differences/similarities there are between the two methods. We then depart form the comparative analysis to look deeper into the Bayesian approach and conduct robustness checks that vary certain aspects (e.g. priors) of it so that we can see how the estimation results change in response to that. The baseline here would be the estimation results from the comparison between Bayesian and frequentist approaches. We close the section off by a discussion of the limitations and challenges encountered in this section.

5.1 Literature review on the application of Hierarchical Models

Bayesian Hierarchical models or multilevel models are a suitable approach to consider social contexts as well as individual respondents or subjects. It becomes attractive to consider hierarchical models in place of the common (or popularized) frequentist approach as soon as there is a need to relax the independence of residuals assumption as a result of similarities in the characteristics of a group of respondents or when the researcher seeks to disentangle variability at various levels of the data. These models have been used in various applications throughout fields in economic research. In education research, Burić and Kim (2020) use these models to examine the relationship between teacher self-efficacy (TSE), instructional quality (i.e., classroom management, cognitive activation, and supportive climate) and student motivational beliefs (i.e., self efficacy and intrinsic motivation) by using responses from both teachers and students and implementing a sophisticated doubly latent multilevel structural equation modelling approach. The results reflect the necessity to disentangle variability at various levels of the data as the researchers find that, at class level, TSE was positively related to the three dimensions of instructional quality but not to students' motivational beliefs. They also find, as expected, that instructional quality was positively related to students' motivational beliefs.

In family economics, Lamnisos et al. (2019) project the total fertility rate and life expectancy at birth probabilistically using Bayesian hierarchical models and United Nations population data for Greece from the period of 1950 to 2015. These are then converted to age-specific mortality rates and combined with a cohort component projection model. This yields probabilistic projections of total population by sex and age groups, total fertility rate (TFR), female and male life expectancies at birth and potential support ratio PSR (persons aged 20-64 per person 65+) by the year 2100. If the forecasts prove in future to be accurate, these models can provide a powerful tool for policy formulation. In agricultural research, Ramsey et al. (2019) develop two econometric models: a hierarchical Bayesian linear model and a hierarchical Bayesian Poisson model to predict exit rates across the towns and prefectures of Japan resulting from off-farm employment opportunities. Off-farm employment opportunities are thought to have an effect on farm exit rates, though evidence on the sign of this effect has been mixed. Examining this issue in the context of Japanese agriculture, the researchers find that farm exits are related to off-farm income as a share of household income, and more specifically to the nature of off-farm work.

In development economics, Meager (2019) jointly estimates the average effect and the heterogeneity in effects across seven studies using Bayesian hierarchical models to answer questions

about external validity that impede consensus on the results from randomized evaluations of microcredit. The researcher finds reasonable external validity: true heterogeneity in effects is moderate, and approximately 60 percent of observed heterogeneity is sampling variation. These paper has the potential to revolutionize the field of development economics as the researcher provides a method to establish external validity using multiple studies form different countries. In health research, Rashid (2019) uses a more advanced application of Bayesian Hierarchical Models in health research. The authors aim to identify the spatial distribution of the three types of misconception factors of HIV transmission (i.e. transmitted by mosquito bite, supernatural means and sharing food with HIV positive person). This study also provides the core socio-economic factors to stop the misconception about HIV/ Aids transmission and helped in reducing its epidemic in Pakistan. Spatial and Non-Spatial Bayesian Hierarchical model were applied to the data and results from them revealed that the Conditional Autoregressive Bayesian Hierarchical Models (Spatial Model) were more appropriate. The results showed that Conditional Autoregressive Bayesian Hierarchical models at level 2 are best fit to the data.

It is evident that Bayesian Hierarchical Models can be useful in the area of microeconomic research. There also have been applications to the Macroeconomics field. It is evident that Bayesian Hierarchical Models can be useful in the area of microeconomic research. There also have been applications to the Macroeconomic research field. Koop et al. (2010) notes that bayesian methods have become increasingly popular as a way of overcoming over-parameterization problems. In this paper, the authors discuss vector autoregressive multivariate time series models (VARs), factor augmented VARs and time-varying parameter extensions and show how Bayesian inference proceeds.

We now demonstrate below, an application to real world data of a comparison between the likelihood (frequentist) and the bayesian inference approaches. For each of these approaches, we will fit three basic multilevel linear models: (a) a varying intercept model with no predictors (Model 1), (b) a varying intercept model with one predictor (Model 2), and (c) a varying intercept and slope model (Model 3). We will use the *lmer* function in the *lme4* package for R to determine maximum likelihood estimates of the parameters in linear mixed-effects models. We then the use *rstanarm* package (also in R) to implement a fully Bayesian approach.

5.2 Frequentist and Bayesian Approaches in Practice: Application to Education Data

A common feature of data structures in education is that units of analysis (e.g., students) are nested in higher organizational clusters (e.g. schools). This kind of structure induces dependence among the responses observed for units within the same cluster. Students in the same school tend to be more alike in their academic and attitudinal characteristics than students chosen at random from the population at large. Multilevel models are designed to model such within-cluster dependence. As mentioned earlier, one advantage of multilevel models is that it allows us to disentangle variability between levels and in our data example, multilevel models recognize the existence of data clustering (at two or more levels) by allowing for residual components at each level in the hierarchy. For example, a two-level model that allows for grouping of student outcomes within schools would include residuals at both the student and school level. The residual variance is thus partitioned into a between-school component (the variance of the school-level residuals) and a within-school component (the variance of the student-level residuals).

5.2.1 The data

We will be analyzing the Gcsemv dataset from Rasbash et al. (2000). The data include the General Certificate of Secondary Education (GCSE) exam scores of 1,905 students from 73 schools in England on a science subject. The Gcsemv dataset consists of the following 5 variables:

- *school*: school identifier
- *student*: student identifier
- *gender*: gender of a student (M: Male, F: Female)
- *written*: total score on written paper
- *course*: total score on coursework paper

Two components of the exam were recorded as outcome variables: written paper and course work. In this application, only the total score on the coursework paper (course) will be analyzed. In our example, we seek to estimate the effect of gender on test scores.

5.2.2 Likelihood inference approach

In this sub-section, we fit the three basic multilevel linear models starting with a varying intercept model with no predictors (Model 1), we then proceed to the varying intercept model with one predictor (Model 2), and the varying intercept and slope model (Model 3) using the *lmer()* functions. Functions such as lmer() are based on a combination of maximum likelihood (ML) estimation of the model parameters, and empirical Bayes (EB) predictions of the varying intercepts and/or slopes resulting in the Best Linear Unbiased Predictions (BLUPs) of the model parameters. We use these functions so that our parameter estimates from both the ML and Bayesian framework are comparable.

Model 1: Varying intercept model with no predictors (Variance components model)

Consider the simplest multilevel model for students $i = 1, ..., n$ nested within schools $j = 1, ..., J$ and for whom we have examination scores as responses. We can write a two-level varying intercept model with no predictors using the usual two-stage formulation as:

$$y_{ij} = \alpha_j + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2) \tag{22}$$

$$\alpha_j = \mu_\alpha + \mu_j \text{ where } \mu_j \sim \mathcal{N}(0, \sigma_\alpha^2) \tag{23}$$

where y_{ij} is the examination score for the ith student in the jth school, α_j is the varying intercept for the jth school, and μ_α is the overall mean across schools. Alternatively, the model can be expressed in reduced form as

$$y_{ij} = \mu_\alpha + \mu_j + \epsilon_{ij}, . \tag{24}$$

If we further assume that the student-level errors ϵ_{ij} are normally distributed with mean 0 and variance σ_y^2, and that the school-level varying intercepts α_j are normally distributed with mean μ_α and variance σ_α^2, then the model can be expressed as

$$y_{ij} =\sim \mathcal{N}(\alpha_j, \sigma_y^2) \tag{25}$$

$$\alpha_j =\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2) \tag{26}$$

We can then fit a linear mixed model by maximum likelihood (ML). We specify an intercept (the predictor "1") and allow it to vary by the level-2 identifier (school). We also specify the *REML = FALSE* option to obtain maximum likelihood (ML) estimates as opposed to the default restricted maximum likelihood (REML) estimates.

[frame=single] Linear mixed model fit by maximum likelihood ['lmerMod'] Formula: course 1 + (1 | school) Data: GCSE

Random effects: Groups Name Variance Std.Dev. school (Intercept) 75.24 8.674 Residual 190.77 13.812 Number of obs: 1725, groups: school, 73

Fixed effects: Estimate Std. Error t value (Intercept) 73.72 1.11 66.4

Under the 'Fixed effects' part of the output, we see that the intercept μ_α, averaged over the population of schools, is estimated as **73.72**. Under the 'Random effects' part of the output, we see that the between-school standard deviation σ_α is estimated as **8.67** and the within-school standard deviation σ_y as **13.81**.

Model 2: Varying intercept model with a single predictor

The varying intercept model[3] with an indicator variable for being female x_{ij} can be written as

$$y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2), \tag{27}$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2). \tag{28}$$

The equation of the average regression line across schools is $\mu_{ij} = \mu_\alpha + \beta x_{ij}$. The regression lines for specific schools will be parallel to the average regression line (having the same slope β), but differ in terms of its intercept α_j. This model can be estimated by adding 'female' to the model, which will allow only the intercept to vary by school, and while keeping the "slope" for being female constant across schools as shown below:

[3]Equivalently, the model can be expressed using a two-stage formulation as

$$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij},$$
$$\alpha_j = \mu_\alpha + u_j,$$

or in a reduced form as

$$y_{ij} = \mu_\alpha + \beta x_{ij} + u_j + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, \sigma_y^2)$ and $u_j \sim N(0, \sigma_\alpha^2)$.

[frame=single] Random effects: Groups Name Variance Std.Dev. school (Intercept) 76.65 8.755 Residual 179.96 13.415 Number of obs: 1725, groups: school, 73

Fixed effects: Estimate Std. Error t value (Intercept) 69.730 1.185 58.87 femaleF 6.739 0.678 9.94

Correlation of Fixed Effects: (Intr) femaleF -0.338

The average regression line across schools is thus estimated as $\hat{\mu}_{ij} = \mathbf{69.73} + \mathbf{6.74}x_{ij}$, with σ_α and σ_y estimated as 8.76 and **13.41** respectively. Treating these estimates of μ_α, β, σ_y^2, and σ_α^2 as the true parameter values, we can then obtain the Best Linear Unbiased Predictions (BLUPs) for the school-level errors $\hat{u}_j = \hat{\alpha}_j - \hat{\mu}_\alpha$.

The BLUPs are equivalent to the so-called Empirical Bayes (EB) prediction, which is the mean of the posterior distribution of u_j given all the estimated parameters, as well as the random variables y_{ij} and x_{ij} for the cluster. These predictions are called "Bayes" because they make use of the pre-specified prior distribution [4] $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$, and by extension $u_j \sim N(0, \sigma_\alpha^2)$, and called "Empirical" because the parameters of this prior, μ_α and σ_α^2, in addition to β and σ_y^2, are estimated from the data.

Compared to the Maximum Likelihood (ML) approach of predicting values for u_j by using only the estimated parameters and data from cluster j, the EB approach additionally consider the prior distribution of u_j, and produces predicted values closer to 0 (a phenomenon described as *shrinkage* or *partial pooling*). To see why this phenomenon is called *shrinkage*, we usually express the estimates for u_j obtained from EB prediction as $\hat{u}_j^{\text{EB}} = \hat{R}_j \hat{u}_j^{\text{ML}}$ where $\hat{u}_j^{\text{ML}}$ are the ML estimates, and $\hat{R}_j = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_y^2}{n_j}}$ is the so-called Shrinkage factor.

By using the *ranef* function, we can also show how much the intercept is shifted up or down in particular schools. For example, in the first school in the dataset, the estimated intercept is about **10.17** lower than average, so that the school-specific regression line is $69.73 - 10.17 + 6.74x_{ij}$.

Model 3: Varying intercept and slope model with a single predictor

We now extend the varying intercept model with a single predictor to allow both the intercept and the slope to vary randomly across schools using the following model[5]:

$$y_{ij} \sim N(\alpha_j + \beta_j x_{ij}, \sigma_y^2),$$
$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right).$$

Note that now we have variation in the α_j's and the β_j's, and also a correlation parameter ρ between α_j and β_j. This model can be fit using *lmer()* as follows:

[4] We elaborate more on prior distributions in Section the full Bayesian approach section

[5] Equivalently, the model can be expressed in a two-stage formulation as

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij},$$
$$\alpha_j = \mu_\alpha + u_j,$$
$$\beta_j = \mu_\beta + v_j,$$

or in a reduced form as

$$y_{ij} = \mu_\alpha + \mu_\beta x_{ij} + u_j + v_j x_{ij} + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, \sigma_y^2)$ and $\begin{pmatrix} u_j \\ v_j \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right)$.

[frame=single] Random effects: Groups Name Variance Std.Dev. Corr school (Intercept) 102.94 10.146 femaleF 47.94 6.924 -0.52 Residual 169.79 13.030 Number of obs: 1725, groups: school, 73

Fixed effects: Estimate Std. Error t value (Intercept) 69.425 1.352 51.336 femaleF 7.128 1.131 6.302

Correlation of Fixed Effects: (Intr) femaleF -0.574 In this model, the residual within-school standard deviation is estimated as $\hat{\sigma}_y = 13.03$. The estimated standard deviations of the school intercepts and the school slopes are $\hat{\sigma}_\alpha = 10.15$ and $\hat{\sigma}_\beta = 6.92$ respectively. The estimated correlation between varying intercepts and slopes is $\hat{\rho} = -0.52$.

5.2.3 Full Bayesian Inference Approach

As previously mentioned, functions such as *lmer()* are based on a combination of maximum likelihood (ML) estimation of the model parameters, and empirical Bayes (EB) predictions of the varying intercepts and/or slopes. However, in some instances, when the number of groups is small or when the model contains many varying coefficients or non-nested components, the ML approach may not work as well in part because there may not be enough information to estimate variance parameters precisely. In such cases, a fully Bayesian approach provides reasonable inferences with the added benefit of accounting for all the uncertainty in the parameter estimates when predicting the varying intercepts and slopes, and their associated uncertainty. This is one of the reasons why a fully Bayesian estimation is particularly interesting. Other reasons are discussed in section 5.3. We now demonstate below, how to fit Models 1, 2 and 3 from above in a fully bayesian framework using the *rstammarm* package. *Rstanarm* is a wrapper for the *rstan* package that enables the most common applied regression models to be estimated using Markov Chain Monte Carlo (MCMC) but still be specified using customary R modeling syntax.

Model 1: Varying intercept model with no predictors (Variance components model)

We can implement a fully Bayesian estimation for multilevel models with only minimal changes to our existing code with *lmer()* from the maximum likelihood application in the previous section. We specify Model 1 with default prior distributions for μ_α, σ_α, and σ_y by prepending *stan_* to the *lmer* call. The *stan_lmer()* function is similar in syntax to *lmer()* but rather than performing maximum likelihood estimation, Bayesian estimation is performed via MCMC. As each step in the MCMC estimation approach involves random draws from the parameter space, we include a seed option to ensure that each time the code is run, *stan_lmer* outputs the same results.

Prior distributions

Model 1 is a varying intercept model with normally distributed student residuals and school-level intercepts: $y_{ij} \sim N(\alpha_j, \sigma_y^2)$, and $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$. The normal distribution for the α_j's can be thought of as a prior distribution for these varying intercepts. The parameters of this prior distribution, μ_α and σ_α, are estimated from the data when using maximum likelihood estimation. In full Bayesian inference, all the hyperparameters (μ_α and σ_α), along with the other unmodeled parameters (in this case, σ_y) also need a prior distribution. For this illustration, we use weakly informative priors that provide moderate regularization [6] and help stabilize computation.

[6] Regularization can be regarded as a technique to ensure that estimates are bounded within an acceptable range of values.

First, before accounting for the scale of the variables, μ_α is given normal prior distribution with mean 0 and standard deviation 10. That is, $\mu_\alpha \sim N(0, 10^2)$. The standard deviation of this prior distribution, 10, is five times as large as the standard deviation of the response if it were standardized. This should be a close approximation to a noninformative prior over the range supported by the likelihood, which should give inferences similar to those obtained by maximum likelihood methods if similarly weak priors are used for the other parameters. *rstanarm* scales the priors in relation to the scale of variables in the estimation proccess.

Second, the (unscaled) prior for σ_y is set to an exponential distribution with rate parameter set to 1.

Third, in order to specify a prior for the variances and covariances of the varying (or "random") effects, *rstanarm* will decompose this matrix into a correlation matrix of the varying effects and a function of their variances. Since there is only one varying effect in this example, the default (unscaled) prior for σ_α that the package uses reduces to an exponential distribution with rate parameter set to 1.

Model 1 results

As we see in the snippet of the summary below, the point estimate of μ_α from the bayesian estimation is 73.75 and this corresponds to the median of the posterior draws. This is similar to the ML estimate obtained previously 73.72. The point estimate for σ_α from the bayesian estimation is 8.92, which is larger than the ML estimate (8.67). This discrepancy may be partly because the ML approach does not take into account the uncertainty in μ_α when estimating σ_α.

[frame=single] Median MAD$_S D(Intercept)$73.751.1

Error terms: Groups Name Std.Dev. school (Intercept) 8.92 Residual 13.8 Num. levels: school 73

When using the bayesian estimation function, standard errors are obtained by considering the median absolute deviation (MAD) of each draw from the median of those draws. It is well known that ML tends to underestimate uncertainties because it relies on point estimates of hyperparameters. Full Bayes, on the other hand, propagates the uncertainty in the hyperparameters throughout all levels of the model and provides more appropriate estimates of uncertainty.

Model 2: Varying intercept model with a single predictor

We can extend the model as we did in the ML framework by including observed explanatory variables at the student level x_{ij}, in this example, an indicator variable for being female. A simple varying intercept model with one predictor at the student level can be written as $y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2)$ and $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$. We use noninformative prior distributions for the hyperparameters (μ_α and σ_α) as specified in Model. Additionally, the regression coefficient β is given normal prior distributions with mean 0 and standard deviation 100. This states, roughly, that we expect this coefficient to be in the range $(-100, 100)$, and if the ML estimate is in this range, the prior distribution is providing very little information for the inference.

[frame=single] Median $\text{MAD}_S D(Intercept) 69.7 1.2 femaleF 6.7 0.7$

Auxiliary parameter(s): Median $\text{MAD}_S Dsigma 13.4 0.2$

Error terms: Groups Name Std.Dev. school (Intercept) 9 Residual 13 As seen from the results above, the point estimates of μ_α, β, and σ_y are almost identical to the ML fitted estimates. However, partly because ML ignores the uncertainty about μ_α when estimating σ_α, the Bayesian estimate for σ_α (9.0) is larger than the ML estimate (8.8), as with Model 1.

Model 3: Varying intercept and slope model with a single predictor

We also fit Model 3 using the baysian estimation framework. Note that here, we use the default priors which are mostly similar to what was done in Model 1. Additionally, we are also required to specify a prior for the covariance matrix Σ for α_j and β_j in this Model. *stan_lmer* decomposes this covariance matrix (up to a factor of σ_y) into (i) a correlation matrix R and (ii) a matrix of variances V, and assigns them separate priors as shown below.

$$\Sigma = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \tag{29}$$

$$= \sigma_y^2 \begin{pmatrix} \sigma_\alpha^2/\sigma_y^2 & \rho\sigma_\alpha\sigma_\beta/\sigma_y^2 \\ \rho\sigma_\alpha\sigma_\beta/\sigma_y^2 & \sigma_\beta^2/\sigma_y^2 \end{pmatrix} \tag{30}$$

$$= \sigma_y^2 \begin{pmatrix} \sigma_\alpha/\sigma_y & 0 \\ 0 & \sigma_\beta/\sigma_y \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \sigma_\alpha/\sigma_y & 0 \\ 0 & \sigma_\beta/\sigma_y \end{pmatrix} \tag{31}$$

$$= \sigma_y^2 VRV. \tag{32}$$

The correlation matrix R is 2 by 2 matrix with 1's on the diagonal and ρ's on the off-diagonal. *stan_lmer* assigns it an LKJ [7] prior (Lewandowski et al. (2009)), with regularization parameter 1. This is equivalent to assigning a uniform prior for ρ. The more the regularization parameter exceeds one, the more peaked the distribution for ρ to take the value 0.

The matrix of (scaled) variances V can first be collapsed into a vector of (scaled) variances, and then decomposed into three parts, J, τ^2 and π as shown below.

$$\begin{pmatrix} \sigma_\alpha^2/\sigma_y^2 \\ \sigma_\beta^2/\sigma_y^2 \end{pmatrix} = 2 \left(\frac{\sigma_\alpha^2/\sigma_y^2 + \sigma_\beta^2/\sigma_y^2}{2} \right) \begin{pmatrix} \frac{\sigma_\alpha^2/\sigma_y^2}{\sigma_\alpha^2/\sigma_y^2 + \sigma_\beta^2/\sigma_y^2} \\ \frac{\sigma_\beta^2/\sigma_y^2}{\sigma_\alpha^2/\sigma_y^2 + \sigma_\beta^2/\sigma_y^2} \end{pmatrix} = J\tau^2\pi.$$

In this formulation, J is the number of varying effects in the model (here, $J = 2$), τ^2 can be viewed as an average (scaled) variance across the varying effects α_j and β_j, and π is a non-negative vector that sums to 1 (called a Simplex/probability vector). A symmetric Dirichlet [8] distribution with concentration parameter set to 1 is then used as the prior for π. By default, this implies a jointly uniform prior over all Simplex vectors of the same size. A scale-invariant Gamma prior with shape and scale parameters both set to 1 is then assigned for τ. This is equivalent to

[7] For more details about the LKJ distribution, see `http://www.psychstatistics.com/2014/12/27/d-lkj-priors/` and `http://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html`

[8] The Dirichlet distribution is a multivariate generalization of the beta distribution with one concentration parameter, which can be interpreted as prior counts of a multinomial random variable (the simplex vector in our context), for details, see `https://cran.r-project.org/web/packages/rstanarm/vignettes/glmer.html#detail`

assigning as a prior the exponential distribution with rate parameter set to 1 which is consistent with the prior assigned to σ_y.

5.3 Looking Deeper into the Baysian Approach

5.3.1 Pooling in Multilevel Models

5.3.2 Convergence

5.3.3 Ranking varying intercepts by school

5.3.4 School to school comparisons

5.3.5 Robustness checks

6 Conclusion

A Appendix

A.1 Tables and Figures

A.2 Code

A.3 Proofs

Uninformative prior 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 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,

$$
\begin{aligned}
p(\sigma^2 \mid y) &\propto \int p(\theta, \sigma^2 \mid y) \mathrm{d}\theta && (33)\\
&\propto \int \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right) \mathrm{d}\theta && (34)\\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right) \mathrm{d}\theta && (35)\\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \theta)^2]\right) \mathrm{d}\theta && (36)\\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \int \exp\left(\frac{1}{2\sigma^2/n}(\bar{y} - \theta)^2\right) \mathrm{d}\theta && (37)\\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \sqrt{2\pi\sigma^2/n} && (38)\\
&\propto (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right). && (39)
\end{aligned}
$$

Hence, $\sigma^2 \mid y \sim$ 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 $a = (n-1)s + n(\theta - \bar{y})$.[9] Then,

$$
\begin{aligned}
p(\theta \mid y) &= \int_{(0,\infty)} p(\theta, \sigma^2 \mid y) \mathrm{d}\sigma^2 && (40)\\
&\propto \sigma^{-(n+2)} \int_{(0,\infty)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \theta)^2]\right) \mathrm{d}\sigma^2 && (41)\\
&\propto a^{-n/2} \int_{(0,\infty)} z^{(n-2)/2} \exp(-z)\, \mathrm{d}z && (42)\\
&\propto a^{-n/2} && (43)\\
&= [(n-1)s + n(\theta - \bar{y})]^{-n/2} && (44)\\
&\propto \left[1 + \frac{(\theta - \bar{y})^2}{(n-1)s^2/n}\right]^{-n/2} && (45)
\end{aligned}
$$

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 estimators and their

[9] See appendix for explicit derivation

variances are given by[10]

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_i y_i = \bar{y} \tag{46}$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{n-1}{n}s^2 \tag{47}$$

$$\mathrm{I}(\theta,\sigma)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix}. \tag{48}$$

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**

[10] See appendix for proof.

References

Au, S.-K. and J. L. Beck (2001). Estimation of small failure probabilities in high dimensions by subset simulation.

Betancourt, M. (2017). The convergence of markov chain monte carlo methods: From the metropolis method to hamiltonian monte carlo.

Blum, A., J. Hopcroft, and R. Kannan (2017, June). *Foundations of Data Science*.

Burić, I. and L. E. Kim (2020). Teacher self-efficacy, instructional quality, and student motivational beliefs: An analysis using multilevel structural equation modeling. *Learning and Instruction 66*, 101302.

Cogburn, R. (1972). The central limit theorem for markov processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, Berkeley, Calif., pp. 485–512. University of California Press.

Duane, S., A. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics Letters B 195*(2), 216 – 222.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis* (2nd ed. ed.). Chapman and Hall/CRC.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*, Volume Analytical methods for social research. New York: Cambridge University Press.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Katafygiotis, L. and K. Zuev (2008, 04). Geometric insight into the challenges of solving high-dimensional reliability problems. *Probabilistic Engineering Mechanics 23*, 208–218.

Koop, G., D. Korobilis, et al. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics 3*(4), 267–358.

Lamnisos, D., K. Giannakou, and T. Siligari (2019). Demographic forecasting of population projection in greece: A bayesian probabilistic study: Demetris lamnisos. *European Journal of Public Health 29*(Supplement_4), ckz186–387.

Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis 100*(9), 1989–2001.

Liang, F., C. Liu, and R. Carroll (2010, 07). Advanced markov chain monte carlo methods: Learning from past samples. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics 11*(1), 57–91.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Meyn, S. and R. L. Tweedie (2009). *Markov Chains and Stochastic Stability* (2nd ed.). USA: Cambridge University Press.

Ramsey, A. F., S. K. Ghosh, and T. Sonoda (2019). Saying sayonara to the farm: Hierarchical bayesian modeling of farm exits in japan. *Journal of Agricultural Economics 70*(2), 372–391.

Rasbash, J., W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford, and T. Lewis (2000). A user's guide to mlwin. *London: Institute of Education 286*.

Rashid, M. (2019). Socio-economic factors of misconception about hiv/aids among ever-married women in punjab: A comparison of non-spatial and spatial hierarchical bayesian poisson model. *Kuwait Journal of Science 46*(4).

Roberts, G. and J. Rosenthal (1997). Geometric ergodicity and hybrid markov chains. *Electron. Commun. Probab. 2*, 13–25.

Roberts, G. O. and J. S. Rosenthal (2004). General state space markov chains and mcmc algorithms. *Probab. Surveys 1*, 20–71.

Sherlock, C., P. Fearnhead, and G. O. Roberts (2010, 05). The random walk metropolis: Linking theory and practice through a case study. *Statist. Sci. 25*(2), 172–190.

Stan Development Team (2018). The Stan Core Library. Version 2.18.0.