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 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. fact 5, Roberts and Rosenthatl; General State Space Markov chains and MCMC algorithms). 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 Markov Chains and Stochastic Stability; Meyn and

Tweedie.

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 et al., 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 novel algorithms see Robert and Casella 2004; Monte Carlo Statistical methods or Liang et al. 2010; Advanced Markov Chain Monte Carlo Methods.

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, Fearnhed, Roberts (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.

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. 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* (Theorem 2.9; Foundations of Data Science, Blum et al.) 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

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.

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

$$\mathbb{E}_p[f] = \int_{\mathrm{Q}} f(q)p(q)\mathrm{d}q\,. \tag{16}$$

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.[3] 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 non-negligible. We will call this set the typical set.[4]

2.4.1 Markov Chain Monte Carlo 2

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.

$$p(q) = \int_{\mathrm{Q}} T(q', q)p(q')\mathrm{d}q'\,, \tag{17}$$

then the markov chain will have p as its limiting distribution.PROOOOOF. The intuition behind equation 17 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 $\{q_0, \ldots, q_N\}$ 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

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

conditions WHAT ARE THEEESE??, we get

$$\hat{f}_N \xrightarrow{\text{P}} \mathbb{E}_p(f)\,. \tag{18}$$

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

$$\hat{f}_N^{MCMC} \overset{a}{\sim} \mathcal{N}\left(\mathbb{E}_p(f), \text{MCMC-SE}\right)\,, \tag{19}$$

where $\text{MCMC-SE} = \sqrt{\text{Var}_p(f)/\text{ESS}}$ 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 19 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

$$a(q' \mid q) = \min\left(1, \frac{T(q \mid q')p(q')}{T(q' \mid q)p(q)}\right)\,. \tag{20}$$

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 REFERERERECNCE.

2.5 APPENDIX

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,

$$
\begin{aligned}
p(\sigma^2 \mid y) &\propto \int p(\theta, \sigma^2 \mid y) \mathrm{d}\theta && (21)\\
&\propto \int \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right) \mathrm{d}\theta && (22)\\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right) \mathrm{d}\theta && (23)\\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \theta)^2]\right) \mathrm{d}\theta && (24)\\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \int \exp\left(\frac{1}{2\sigma^2/n}(\bar{y} - \theta)^2\right) \mathrm{d}\theta && (25)\\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \sqrt{2\pi\sigma^2/n} && (26)\\
&\propto (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right). && (27)
\end{aligned}
$$

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 $a = (n-1)s + n(\theta - \bar{y})$.[5] Then,

$$
\begin{aligned}
p(\theta \mid y) &= \int_{(0,\infty)} p(\theta, \sigma^2 \mid y) \mathrm{d}\sigma^2 && (28)\\
&\propto \sigma^{-(n+2)} \int_{(0,\infty)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \theta)^2]\right) \mathrm{d}\sigma^2 && (29)\\
&\propto a^{-n/2} \int_{(0,\infty)} z^{(n-2)/2} \exp(-z)\, \mathrm{d}z && (30)\\
&\propto a^{-n/2} && (31)\\
&= [(n-1)s + n(\theta - \bar{y})]^{-n/2} && (32)\\
&\propto \left[1 + \frac{(\theta - \bar{y})^2}{(n-1)s^2/n}\right]^{-n/2} && (33)
\end{aligned}
$$

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 variances are given by[6]

$$
\begin{aligned}
\hat{\theta}_{ML} &= \frac{1}{n}\sum_i y_i = \bar{y} && (34)\\
\hat{\sigma}^2_{ML} &= \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{n-1}{n}s^2 && (35)\\
\mathrm{I}(\theta, \sigma)^{-1} &= \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix}. && (36)
\end{aligned}
$$

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.

[5] See appendix for explicit derivation

[6] See appendix for proof.

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

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 $i = 1, \ldots, n$ children in $j = 1, \ldots, J$ 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 $\theta_1, \ldots, \theta_J$ also follow a common data generating process governed by some hyperparameter ϕ. This gives

$$p(y_i \mid \theta_{j[i]}, \phi) = p(y_i \mid \theta_{j[i]}) \tag{37}$$

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

$$y = \alpha + \beta_j x + \epsilon \,, \tag{38}$$

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

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta \,, \tag{39}$$

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

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \,, \tag{40}$$

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

6 Conclusion