Bayesian Hierarchical Models

Research Module - Econometrics and Statistics - 2019/2020

Linda Maokomatanda[*], Tim Mensinger[†], and Markus Schick[‡]

University of Bonn

Abstract

In this paper we introduce the core topics of Bayesian statistics with a focus on modern sampling techniques. We then present hierarchical models and show how they can be seen as a natural extension to the Bayesian prior design. In our simulation study we show how a simple Bayesian hierarchical model behaves under varying sample sizes. We further test in light of the true data generating process how correctly and wrongly specified priors as well as uninformative priors influence the results. At last we provide a small literature review on hierarchical models and fit a standard hierarchical linear model in an educational setting. We supply all code to reproduce our results on our project page.[1]

[*]linda[at]email.de

[†]tim.mensinger[at]uni-bonn.de

[‡]markus[at]email.de

[1]`https://github.com/timmens/bayesian-hierarchical-models`

Contents

1 Introduction

2 Bayesian Thinking and Estimation

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

Before beginning let us shortly introduce our notation. We try to use standard notation wherever possible, nonetheless we make one exception in that we write $p(Z)$ for the probability density function of the random variable Z, where Z might be scalar-valued or vector-valued. If it is clear from the context we will also write $p(z)$ for the density of Z evaluated at z.

2.1 Probabilistic Modeling

We begin by introducing a formal notion of stochastic modeling and continue with a taxonomic description of different schools of thoughts.

Say we observe data $\mathcal{D} = \{(y_i, x_i) : i = 1, ..., n\}$ for which we have some intuition about the relationship between X and Y —this intuition might come from (economic) theory for example. We formalize this by writing the data generating process as a (possibly algorithmic) mathematical model $Y = \mathcal{M}(X; \theta)$, where θ denotes the model parameters. In many settings, however, we are unable to describe the relationship perfectly or the outcome depends on non-observed variables. We deal with this complication by extending the model to account for an explicitly modeled error term, so that $Y = \mathcal{M}(X; \epsilon, \theta)$, where ϵ is the error term.

In the rest of this paper we assume that we know the parametric structure of $\mathcal{M}$ and that our goal lies in learning about the parameters after observing the data $\mathcal{D}$. An important aspect here is to postulate the existence of a *true data generating process*, which we do by assuming that there is some (fixed) θ_0, in the parameter space Θ, so that $Y = \mathcal{M}(X; \epsilon, \theta_0)$ describes reality sufficiently accurate and better than for any other parameter. The subsequent goal is then to recover the value of θ_0. Note that for everything that follows we need to assume that our *true model* actually describes reality accurately; if not we enter the realm of model misspecification which can render any analysis useless.

Next we compare two different schools of thought present in the statistical domain, which consider the estimation of θ_0 and the quantification of our uncertainty in the estimate.

Frequentist. In the literature the so called frequentist methods constitute the most widely used approaches. A particular method —which we choose here as it lends itself nicely to a comparison— is the maximum likelihood approach. There we use the distributional assumptions on our model to construct the likelihood function $\mathcal{L}(\theta; \mathcal{D}) = p(\mathcal{D}; \theta)$, which is simply the joint density of the data evaluated at the observed data points for varying parameter θ. We can then find an estimator $\hat{\theta}$ for θ_0 as the maximizer of this function, i.e. $\hat{\theta} \in \text{argmax}_{\theta \in \Theta} \mathcal{L}(\theta; \mathcal{D})$. There has been published an extensive amount of research on the properties of this estimator, for example on sufficient conditions for the uniqueness of the maximization or large-sample normal approximations. In particular, under some regularity conditions we can find a matrix $\widehat{V}$ so that $\sqrt{n}\widehat{V}^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I)$. This we can use to quantify our uncertainty on $\hat{\theta}$ by computing standard errors and confidence intervals, as well as using the normal result to formulate tests.

One fundamental idea which stretches over all methods in the frequentist world is the interpretation of probability as the limit of an infinite sequence of relative frequencies of events—hence the name. Probability then just counts how many times an event happened or not; for example if we toss a (fair) coin an infinite number of times the relative frequency of heads converges to the probability of heads. We do expect the outcome of such an experiment to vary, however, one thing we do not assume to vary is the true parameter. For instance, we might interpret the probability of a coin landing on heads as the

true model parameter. In this sense it would then be absurd to let this object vary from experiment to experiment. Therefore any hypothesis on θ_0 is either true or false. And it is this binarity which makes hypothesis testing (interpretation of confidence intervals) so awkward in the frequentist setting. By testing some hypothesis $H_0 : \theta_0 = \theta^*$ we do not directly compute the probability of the hypothesis being true –hypothesis are either false or true— but we compute if the observed data $\mathcal{D}$ is more likely to have orginated under the null hypothesis or the alternative.

Bayesian. The main difference of the Bayesian mindset is the understanding of probability as a subjective quantification of uncertainty. We may still believe that θ_0 is fixed, however, in the Bayesian paradigm we build our uncertainty about the true location of the parameter into the model by allowing for probability distributions to be defined on Θ —which, as we saw above, is nonsensical in a frequentist worldview. We can see the direct utility of this liberation by considering Bayes theorem applied to densities

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D} \mid \theta)p(\theta)\,, \qquad \text{(Bayes' theorem)}$$

which reads $$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{evidence}} \propto \textit{likelihood} \times \textit{prior}\,.$$

The posterior distribution is the object of interest for any subsequent Bayesian analysis, it describes the distribution of the parameter of interest given the observed data $\mathcal{D}$. From a naive standpoint this is too good to be true. And in fact it is. To give Bayes' theorem any meaning we have to define the prior $p(\theta)$, a probability distribution of the model parameter on Θ. The prior can be used to incorporate knowledge about the parameter into the analysis which existed prior to observing the data. But this can be highly subjective and can lead to *two* different researchers having *two* different priors which would result in *two* different posteriors. This is where the main criticism of Bayesian statistics is focused on: where does the prior distribution come from? With the scientific goal of objectivity in mind, many feel at unease having results dependent on subjective choices of the 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.

In comparison to the maximum likelihood approach, in a Bayesian analysis there is no need for one specific point estimator or confidence interval. The result of such an analysis is a complete probability distribution so we can compute, in principle, any quantity we like. Being clear on all prior assumptions and giving up (some) *objectivity* we gain the possibility to formulate answers to more natural questions, as for example: $\mathbb{P}\left(\theta \in \Theta_0 \mid \mathcal{D}\right) = \int_{\Theta_0} p(\theta \mid \mathcal{D}) \mathrm{d}\theta$.

2.2 Solving for the posterior analytically

In this subsection we present an 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 maximum likelihood estimator. As it will be of major importance in the subsequent sections, we have included the definition of the scaled inverse χ^2 probability distribution in the appendix (see definition 1).

Let us assume that we observe a sample $y = (y_1, \ldots, y_n)$ with $y_i \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\theta = (\mu, \sigma^2)$. Our interest lies in solving for the marginal posteriors $p(\mu \mid y)$ and $p(\sigma^2 \mid y)$.

Noninformative Prior. We start our analysis with a common prior choice in settings where we have little prior information and sufficent data. In these cases we can use noninformative priors to model complete ignorance of any prior information; in particular, here we use *flat priors*, which assign equal weight to every region in the parameter space. Let us go with the common assumption that μ and σ^2 are independent a priori. Mathematically we can write a flat prior as $p(\mu) \propto 1$. We note that this does

not define a proper proability distribution, which will not matter in this case but can lead to problems in others; see for example section 4.2 in Kass and Wasserman (1996). Since the variance is restricted to be positive we impose a flat prior on the log-transform thereof, i.e. $p(\log\sigma) \propto 1$. Using that $x \mapsto \exp^2(x)$ is one-to-one we get the density of the transformed variable $p(\mu, \sigma^2) = p(\mu)p(\sigma^2) \propto p(\sigma^2) \propto (\sigma^2)^{-1}$.

The likelihood is given by $p(y \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right)$, where we dropped all proportionality constants. Using the above we can apply Bayes theorem to yield $p(\mu, \sigma^2 \mid y) \propto p(y \mid \mu, \sigma^2)p(\mu, \sigma^2) \propto (\sigma^2)^{-(n+2)/2} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right)$. Integrating over the respective parameter yields the marginal posteriors.

Proposition 1. *Under the above setup and a flat prior on μ and $\log\sigma$ we find*

$$\mu \mid y \sim t_{n-1}(\bar{y}, s^2/n), \quad \sigma^2 \mid y \sim \text{scaled-Inv-}\chi^2\left(n-1, s^2\right),$$

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

Proof. See appendix. □

We compare the marginal posteriors to their maximum likelihood counterpart by reporting summary statistics of the distribution. Namely we focus on the mean and variance of the posterior, as well as the *maximum a posteriori* (MAP) estimate $\left(\text{argmax}_{\theta\in\Theta}\, p(\theta \mid y)\right)$. We summarize some results using Proposition 1 in table 1 but withold from a discussion as we consider the results in the next paragraph in more detail.

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	$\frac{n-1}{n}s^2$	$2\sigma^4/n$	$\frac{n-1}{n+1}s^2$	$\frac{n-1}{n-3}s^2$	$\frac{2(n-1)^2}{(n-3)^2(n-5)}s^4$

Table 1: Comparison of Bayesian estimates using a flat prior to ML estimates. See appendix for a derivation.

Conjugate Prior. From table 1 we see that a flat prior (here) leads to similar results as a maximum likelihood approach. In case substantial information on the parameters is available a priori, we can model this information properly to gain more stable results. However, not every product of prior and likelihood results in a sensible posterior. As we are interested in analytical results in this section we seek priors that guarantee posteriors of known form. The class of *conjugate priors* plays an important part in Bayesian statistics as they are able to prodive such assurance (see definition 2 in the appendix).

Consider again the likelihood but written dependent on the sufficient statistics $\bar{y}$ and s^2

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

where s^2 again denotes the (unbiased) sample variance.

We want to construct a two dimensional conjugate prior for (μ, σ^2) such that multiplying the prior by the likelihood does not change its structure. Note that we have $p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2)$. Looking at equation (1) we see that in order to *not* change the inherent structural dependence on the parameters we must have $\mu \mid \sigma^2 \sim \mathcal{N}\left(\mu_0, \sigma^2/\kappa_0\right)$, with *hyperparameters* μ_0 and $\kappa_0 > 0$. Similarly, we observe that we must have $\sigma^2 \sim \text{scaled-Inv-}\chi^2\left(\nu_0, \sigma_0^2\right)$, with hyperparameters ν_0 and $\sigma_0^2 > 0$. This becomes apparent when considering the respective densities. Following Gelman et al. (2004) we write $(\mu, \sigma^2) \sim \text{normal-scaled-Inv-}\chi^2\left(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \sigma_0^2\right)$, with corresponding density function $p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2) \propto (\sigma^2)^{\frac{3+\nu_0}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2\right]\right)$, for the normal scaled inverse χ^2 distribution.

Multiplying the likelihood with our constructed prior we get the joint posterior

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

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

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

Proof. See appendix. □

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

Proposition 3. *The marginal posterior distributions are given by*

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

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

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 prior to ML estimates.

Next we consider the results of Proposition 2, of which some summary statistics are tabulated in table 2. We focus on the analysis of the mean parameter μ.

As the t-distribution is parameterized over its mean and variance (and degrees of freedom) we can directly read off the posterior mean as $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$ and the posterior variance as σ_n^2/κ_n. We see that the posterior mean is simply a convex combination of the prior μ_0 and the sample average $\bar{y}$, with weights determined 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), which can be helpful if we have insufficient data and believe that the parameter should be around μ_0, where we express our degree of believe in the prior using the weight κ_0. As n grows to infinity the information in the data overwhelms all prior information and the posterior mean is dominated by the sample mean.

Similarly we can use the hyperparameters σ_0^2, ν_0 and κ_0 to model our prior knowledge of the variance parameter, which propagates to the posterior variance of the mean parameter. Considering the variance of the posterior mean as a function in n we can use the *Landau notation* to write $\sigma_n^2/\kappa_n = \frac{n}{(\nu_0+n)(\kappa_0+n)}s^2 + \mathcal{O}(1/n^2)$, which resembles the usual $1/n$ convergence rate.

As $v_n = v_0 + n$ tends to infinity the distribution of the posterior mean tends to a normal distribution with parameters behaving (asymptotically) similar to the maximum likelihood estimators. In this sense informative Bayesian priors can be appropriate if the data contains insufficient information *and* we have reasonable knowledge a priori, where we use the prior to stabilize the results, but also if we consider

large samples, where the prior is simply dominated by the likelihood. We refrain from an analogous analysis for σ^2 here and only note that similar results hold, as can be seen from table 2.

Above we considered a simple model, as this allowed us to derive the results analytically. Analytical results allow us to fully investigate the influence of the prior on our results. However, we also seen that Bayesian analyses are far from trivial and depent critically on the complexities of the model structure. If we want to consider more realistic models we have to make ever more restrictive assumptions to yield analytical results. For this reason among others, in the next section we discuss methods which trade off the clarity of an analytical result for the generality of being able to combine near arbitrary priors with complex, possibly high-dimensional likelihoods.

2.3 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 we observed data $\mathcal{D}$ for which we have a (possibly algorithmic) model in mind, which can be represented by the likelihood $p(\mathcal{D} \mid \theta)$. We also assume that a prior distribution $p(\theta)$ has been constructed, so that the posterior is again given by $p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta)$. Unlike before however, we now consider more general settings in which we do not restrict $p(\theta \mid \mathcal{D})$ to be available in analytical form. This can occur in many settings, for example when using priors that do not mix well with the likelihood or more apparent when using computational models which produce evaluations based on algorithms.

To motivate the following, say we are able to draw independent samples $\theta^{(1)}, \ldots, \theta^{(n)}$ from $p(\theta \mid \mathcal{D})$. By the law of large numbers we get $\frac{1}{n}\sum_i h\left(\theta^{(i)}\right) \xrightarrow{a.s.} \mathbb{E}\left[h(\theta) \mid \mathcal{D}\right]$, under some regularity conditions on h and $p(\theta \mid \mathcal{D})$, with similar results holding for sample quantiles. Hence, if we want to learn something about $p(\theta \mid \mathcal{D})$ we can formulate our question using quantiles or general expectations and rely on the statements above.

In the subsequent paragraphs we discuss methods to sample from the posterior that work even if we cannot compute the integration constant $\int p(\mathcal{D} \mid \theta)p(\theta)\mathrm{d}\theta$. We will see that these methods do *not* produce independent samples but instead create Markov chains whose realizations can be seen as autocorrelated samples.

With this in mind, we first consider what properties these chains must fulfill in order to create equivalent results as discussed above for independent samples, before then stating an algorithm that accomplishes this.

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. 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 a particular set of assumptions which will be seen to have favorable prop-

erties when also considering the creation process. 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 Roberts and Rosenthal (2004) for a survey on recent advances with application to MCMC and to Meyn and Tweedie (2009) for a comprehensive treatment of Markov chain theory.

Theorem 1. *(A Central Limit Theorem for Markov Chains). Let $\{X_n\}$ be a (discrete time) Markov chain and π a probability distribution on the same space. Consider some measurable function h with $\mathbb{E}_\pi\left[h^2\right] < \infty$. Define $\sigma^2(h) := Var_\pi(h)\,\tau := Var_\pi(h)\sum_{k\in\mathbb{Z}} Corr\left(h(X_0), h(X_k)\right)$.[2] Assume the Markov chain is ϕ-irreducible, aperiodic, reversible with respect to π and that $\sigma^2(h) < \infty$. Then π is the stationary for the chain and*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathbb{E}_\pi\left[h\right]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(h)\right). \tag{3}$$

Proof. See Roberts and Rosenthal (2004) Proposition 1 for the first claim and Theorem 27 for the second; see Kipnis and Varadhan (1986) for a complete proof of the second claim. □

We end this paragraph by discussing the assumptions of Theorem 1 on an intuitive level.

ϕ-irreducibility assumes that we can find a measure ϕ so that no matter where the chain starts, we eventually reach every region of the state space which has positive measure with respect to ϕ. In the next paragraph we will see that we can make sure that this condition is satisfied by construction such that the chain is π-irreducible.

Aperiodicity assumes that we cannot find disjoint regions on which the chain jumps from one region to another in a cyclical predictable fashion. It seems intuitive that such a behavior will prevent the chain of actually converging to its stationary distribution.

Reversibility with respect to π is a technical assumption which is best explained by its implications. In particular, it implies that the Markov chain has π as its stationary distribution (which is unique by the other assumptions). Again, later we will see that we can construct a chain which fulfills this condition with π being equal to the posterior distribution.

At last, how do we interpret the ***finite variance*** assumption? As we assume square integrability of h we get that $\sigma^2(h)$ is finite if and only if the integrated correlation time τ is finite. This happens if $\mathrm{Corr}(h(X_0), h(X_k))$ goes fast enough to zero. We then get the usual large sample variance approximation of the unnormalized sample mean: $\sigma^2(h)/n = \mathrm{Var}_\pi(h)\,/({}^{n}/_{\tau})$. In this sense we might say that ${}^{n}/_{\tau}$ denotes the *effective sample size*, which corrects for the fact that we are not drawing independent samples and therefore (in most cases) need more samples to yield the same amount of information as in the independent case.

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).

[2] In the original paper by Roberts and Rosenthal (2004) the statement of this theorem differs in that they write $\tau = \sum_{k\in\mathbb{Z}} \mathrm{Corr}\left(X_0, X_k\right)$. We believe that this is an error as Haggstrom and Rosenthal (2007) state in their comparison of different ways of writing the asymptotic variance that $\sigma^2(h) = \sum_{k\in\mathbb{Z}} \mathrm{Cov}\left(h(X_0), h(X_k)\right)$. Now if we use that $X_0 \sim \pi$ we get $\sigma^2(h) = \sum_{k\in\mathbb{Z}} \mathrm{Cov}\left(h(X_0), h(X_k)\right) = \mathrm{Var}\left(h(X_0)\right) + \sum_{k\neq 0} \mathrm{Cov}\left(h(X_0), h(X_k)\right) = \mathrm{Var}_\pi(h)\left(1 + \sum_{k\neq 0} \mathrm{Cov}\left(h(X_0), h(X_k)\right)\right)/\mathrm{Var}\left(h(X_0)\right) = \mathrm{Var}_\pi(h)\left(1 + \sum_{k\neq 0} \mathrm{Corr}(h(X_0), h(X_k))\right) = \mathrm{Var}_\pi(h)\sum_{k\in\mathbb{Z}} \mathrm{Corr}(h(X_0), h(X_k))$.

Algorithm 1 Metropolis-Hastings

Input: $(\pi, q, T) =$ (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 candidate: $y \sim q(\cdot \mid x_t)$
4: compute the acceptance probability: $\mathcal{A} \leftarrow \min\left\{\frac{\pi(y)}{\pi(x_t)}\frac{q(y \mid x_t)}{q(x_t \mid y)}, 1\right\}$
5: update the chain: $x_{t+1} \leftarrow \begin{cases} y & \text{,with probability } \mathcal{A} \\ x_t & \text{,with remaining probability} \end{cases}$
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. And indeed this is the case. The algorithm creates by construction Markov chains which are reversible with respect to π and aperiodic, and if additionally the proposal density is positiv and continuous and π is finite then the chain is π-irreducible; see for example Roberts and Rosenthal (2004).

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.

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. Figure 3.1 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

school	teacher
1	0.5
2	0.7

Figure 3.1: 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

$$y_i \mid \theta_{j[i]} \sim p(y_i \mid x_i, \theta_{j[i]}) \,, \tag{4}$$

$$\theta_j \mid \phi \sim p(\theta_j \mid u_j, \phi) \,, \tag{5}$$

$$\phi \sim p(\phi \mid \zeta) \text{ with } \zeta \text{ fixed} \,, \tag{6}$$

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.

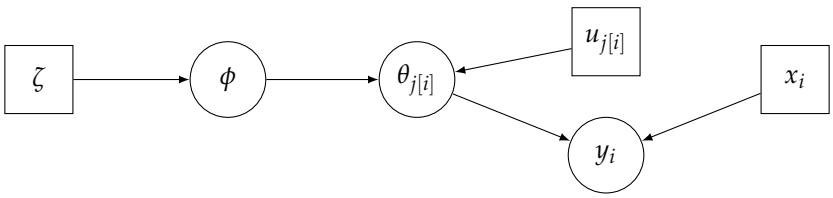

Figure 3.2: A generic two-level Bayesian hierarchical model depicted as a directed acyclical graph modeling a single generic observation. Circled parameters denote random quantities while parameters contained in squares denote fixed quantities.

Figure 3.2 illustrates the conditional dependence structure of modeling a generic observation y_i. In contrast, figure A.1 in the appendix 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 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, i.e. model misspecification, 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.

Being more precise, linearity here means that on each level that is being modeled, parameters enter the *level-model* lineary. We will continue by shortly presenting the general (two-level) hierarchical linear model. In the last subsection we consider the *varying slopes model* which will be further analyzed in the monte carlo study. We also note here that the varying slopes model is closely related to the *varying intercepts model* which will be used in the application part.

General Definition. As before we consider outcomes y_i for $i = 1, ..., n$ in groups $j = 1, ..., J$ with individual-level characteristics x_i and group-level characteristics $u_{j[i]}$. When talking about hierarchical linear models, we consider models that can be written as follows

$$y_i = \alpha_{j[i]} + x_i' \beta_{j[i]} + \epsilon_i \,, \tag{7}$$

$$\beta_j = \gamma_0 + u_j' \gamma + \eta_j \,, \tag{8}$$

where ϵ_i and η_j denote innovation terms on the respective level with distributional assumptions $\epsilon_i \sim p(\epsilon; \theta_\epsilon)$ and $\eta_j \sim p(\eta; \theta_\eta)$. Note that this can be extended for the heteroscedastic or autocorrelated case. The model becomes Bayesian as we impose prior distributions on all parameters that are not explicitly modeled.

Varying Slopes Model with one Predictor in each Level. To finish this section we will showcase the *varying slopes model*, on which we rely heavily in the monte carlo study and which is closely related to the model considered in the application part. In particular we consider the case where each level features one regressor. Written formally we have

$$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i \,, \tag{9}$$

for the individual level, with $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right)$. And for the group-level we get

$$\beta_j = \gamma_0 + \gamma_1 u_j + \eta_j \,, \tag{10}$$

with $\eta_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_\eta^2\right)$. In a classicist interpretation we would call γ_0 and γ_1 the *fixed effects* as they do not vary group and the $\eta_j's$ the *random effects* as they do vary by group. As Gelman and Hill (2007) note, in a Bayesian interpretation this nomenclature is unfortunate as everything in a Bayesian setting is assumed to be random.

Were we to let α vary by group and instead fix β we would get a *varying intercept model* as will be used in the application part. A justified question is if these comparably simple models have any

sensible application. Using the example from Gelman and Hill (2007), consider modeling J different experiments where in each experiment the baseline conditions were the same. In this setting we would like to measure the effect of some treatment (with treatment status T_i) on some outcome y_i. In this case we could model the outcomes as $y_i \mid \theta_{j[i]} \sim \mathcal{N}\left(\alpha + \theta_{j[i]} T_i, \sigma^2\right)$, that is as a varying slopes model. Similar examples can be found to justify the use of varying intercept models. In a real application we usually combine varying intercept and varying slope models; however, this does not mean that every parameter must vary by group. Domain level knowledge can be used to guide the probabilistic modeler in choosing which parameters should vary and how many levels should be build.

A Appendix

A.1 Definitions

Definition 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

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

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

Definition 2. (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 $\mathcal{F}$.

A.2 Figures

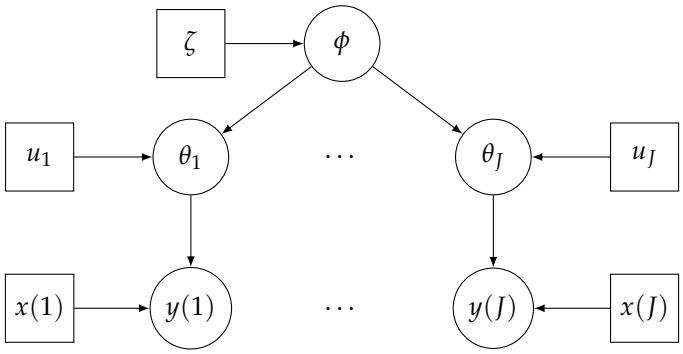

Figure A.1: A generic two-level Bayesian hierarchical model depicted as a directed acyclical graph modeling generic observations $y(j)$ in groups $j = 1, \ldots, J$. Circled parameters denote random quantities while parameters contained in squares denote fixed quantities.

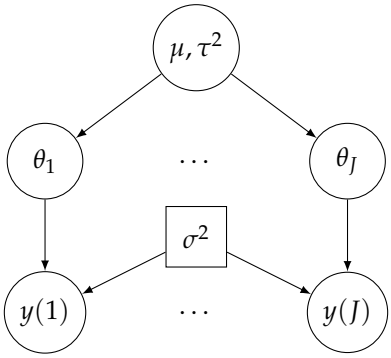

Figure A.2: The directed acyclical graph representation of the model structure of the example model in subsection **??**. *Random and fixed quantities are depicted in circles and squares, respectively.*

A.3 Tables

Derivation of Results in Table 1. We know that for the problem at hand the standard maximum likelihood estimators and their variances are given by

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_i y_i = \bar{y} \tag{11}$$

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{n-1}{n}s^2 \tag{12}$$

$$\mathrm{I}(\mu,\sigma)^{-1} = \begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/(2n) \end{bmatrix}. \tag{13}$$

The Bayesian counterpart to the ML-Estimator is the *maximum a posteriori estimate,* □

Derivation of Results in Table 2. □

Derivation of Results in Table **??**. See file `volume.py` in the online appendix. □

A.4 Proofs

Remark. The proofs presented here follow Gelman et al. (2004); however, we contribute detailed remarks.

Proof of Proposition 1. Consider first the object $\mu \mid \sigma^2, y$. We get

$$p(\mu \mid \sigma^2, y) \propto p(y \mid \mu, \sigma^2) p(\mu \mid \sigma^2) \propto p(y \mid \mu, \sigma^2),$$

where the last step follows as the priors are assumed to be independent. Note then

$$\begin{aligned} p(\mu \mid \sigma^2, y) &\propto \exp\left(-\frac{1}{\sigma^2}\sum_i (y_i - \mu)^2\right) = \exp\left(-\frac{n}{\sigma^2}\frac{1}{n}\sum_i (y_i^2 - 2y_i\mu + \mu^2)\right) \\ &= \exp\left(-\frac{n}{\sigma^2}(\bar{y^2} - 2\bar{y}\mu + \mu^2)\right) \propto \exp\left(-\frac{1}{\sigma^2/n}(\mu - \bar{y})^2\right), \end{aligned}$$

where $\bar{y^2} = \frac{1}{n}\sum_i y_i^2$ and the last step is only proportional as we switch $\bar{y^2}$ for $\bar{y}^2$. Note that proportionality here is with respect to μ. We thus get $\mu \mid \sigma^2, y \sim \mathcal{N}\left(\bar{y}, \sigma^2/n\right)$ as our first intermediate result.

Consider now $\sigma \mid y$. As we already derived the joint posterior we can compute the marginal posterior of σ^2 by integrating out μ. Note that $\sum_i (y_i - \mu)^2 = [(n-1)s^2 + n(\bar{y} - \mu)^2]$, where s^2 denotes the

(unbiased) sample variance. Hence

$$
\begin{aligned}
p(\sigma^2 \mid y) &\propto \int p(\mu, \sigma^2 \mid y) \mathrm{d}\mu \\
&\propto \int \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right) \mathrm{d}\mu \\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right) \mathrm{d}\mu \\
&= \sigma^{-(n+2)} \int \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]\right) \mathrm{d}\mu \\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \int \exp\left(\frac{1}{2\sigma^2/n}(\mu - \bar{y})^2\right) \mathrm{d}\mu \\
&= \sigma^{-(n+2)} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right) \sqrt{2\pi\sigma^2/n} \\
&\propto (\sigma^2)^{-(n+1)/2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2]\right),
\end{aligned}
$$

where the second to last step follows simply by considering the constant of integration of the normal distribution of $\mu \mid \sigma^2, y$. Note that here we consider proportionality with respect to σ^2. By inspection we see that $\sigma^2 \mid y \sim$ scaled-Inv-$\chi^2(n-1, s^2)$, which proves our first claim.

To finish the proof we integrate the joint posterior over σ^2 to get the marginal posterior of μ. We evaluate the integral by substitution using $z = {}^{a}/_{2\sigma^2}$ with $a = (n-1)s^2 + n(\mu - \bar{y})^2$.

Then,

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

where Γ denotes the gamma function (which is finite on the positive real numbers). This concludes the proof by implying that $\mu \mid y \sim t_{n-1}(\bar{y}, s^2/n)$, □

Proof of Proposition 2. Let us first state equation 2 and the premise again. We have to show that

$$
\begin{aligned}
p(\mu, \sigma^2 \mid y) \propto &(\sigma^2)^{-\frac{3+\nu_0+n}{2}} \times \\
&\times \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2 + (n-1)s^2 + n(\bar{y} - \mu)^2\right]\right)
\end{aligned}
$$

is normal-scaled-Inv-$\chi^2\left(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2\right)$ with $\nu_n = \nu_0 + n$, $\kappa_n = \kappa_0 + n$, $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$, $\sigma_n^2 =$

$\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2\right]/\nu_n$. By definition of the normal-scaled-inverse-χ^2 distribution $\nu_n = \nu_0 + n$ follows trivially. Let us therefore consider the term in square brackets in the exponential. We have to show that

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

Plugging in for σ_n^2 we get for the right-hand side

$$\nu_n\sigma_n^2 + \kappa_n(\mu-\mu_n)^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2 + \kappa_n(\mu-\mu_n)^2\,.$$

Therefore we only need to check

$$\kappa_0(\mu-\mu_0)^2 + n(\bar{y}-\mu)^2 = \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2 + \kappa_n(\mu-\mu_n)^2\,.$$

Expanding the right-hand side we get

$$\begin{aligned}
&\frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2 + \kappa_n(\mu-\mu_n)^2 \\
&\quad= \frac{\kappa_0 n}{\kappa_n}\left[\bar{y}^2 - 2\bar{y}\mu_0 + \mu_0^2\right] + \kappa_n\left[\mu^2 - 2\mu\mu_n + \mu_n^2\right] \\
&\quad= \frac{\kappa_0 n}{\kappa_n}\left[\bar{y}^2 - 2\bar{y}\mu_0 + \mu_0^2\right] + \kappa_n\left[\mu^2 - 2\mu\frac{\kappa_0}{\kappa_n}\mu_0 - 2\mu\frac{n}{\kappa_n}\bar{y} + \frac{\kappa_0^2}{\kappa_n^2}\mu_0^2 + \frac{n^2}{\kappa_n^2}\bar{y}^2 + 2\frac{\kappa_0}{\kappa_n}\frac{n}{\kappa_n}\mu_0\bar{y}\right] \\
&\quad= \frac{\kappa_0 n}{\kappa_n}\left[\bar{y}^2 - 2\bar{y}\mu_0 + \mu_0^2\right] + \kappa_n\mu^2 - 2\mu\kappa_0\mu_0 - 2\mu n\bar{y} + \frac{\kappa_0^2}{\kappa_n}\mu_0^2 + \frac{n^2}{\kappa_n}\bar{y}^2 + 2\kappa_0 n\mu_0\bar{y}/\kappa_n \\
&\quad= \left(\kappa_0\mu^2 - 2\mu\kappa_0\mu_0\right) + \left(n\mu^2 - 2n\mu\bar{y}\right) + \frac{\kappa_0 n}{\kappa_n}\bar{y}^2 + \frac{\kappa_0 n}{\kappa_n}\mu_0^2 + \frac{\kappa_0^2}{\kappa_n}\mu_0^2 + \frac{n^2}{\kappa_n}\bar{y}^2 \\
&\quad= \left(\kappa_0\mu^2 - 2\mu\kappa_0\mu_0\right) + \left(n\mu^2 - 2n\mu\bar{y}\right) + \bar{y}^2\left(\frac{\kappa_0 n}{\kappa_n} + \frac{n^2}{\kappa_n}\right) + \mu_0^2\left(\frac{\kappa_0 n}{\kappa_n} + \frac{\kappa_0^2}{\kappa_n}\right) \\
&\quad= \left(\kappa_0\mu^2 - 2\mu\kappa_0\mu_0 + \kappa_0\mu_0^2\right) + \left(n\mu^2 - 2n\mu\bar{y} + n\bar{y}^2\right) \\
&\quad= \kappa_0(\mu-\mu_0)^2 + n(\bar{y}-\mu)^2\,,
\end{aligned}$$

which was what we wanted. □

Proof of Proposition 3. We continue to use the notation of the previous proof. As in the proof of proposition 1 we first compute the distribution of $\mu \mid \sigma^2, y$ and then derive the posterior of σ^2 by integrating μ

out. Note that we actually defined $\mu \mid \sigma^2 \sim \mathcal{N}\left(\mu_0, \sigma^2/\kappa_0\right)$. Hence,

$$
\begin{aligned}
p(\mu \mid \sigma^2, y) &\propto p(y \mid \mu, \sigma^2) p(\mu \mid \sigma^2) \\
&\propto \exp\left(-\frac{1}{2\sigma^2/n}(\mu - \bar{y})^2\right) \exp\left(-\frac{1}{2\sigma^2/\kappa_0}(\mu - \mu_0)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[n(\mu - \bar{y})^2 + \kappa_0(\mu - \mu_0)^2\right]\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left[\mu^2(\kappa_0 + n) - 2\mu(\kappa_0\mu_0 + n\bar{y}) + (\ldots)\right]\right) \\
&= \exp\left(-\frac{1}{2\sigma^2/\kappa_n}\left[\mu^2 - 2\mu(\kappa_0\mu_0 + n\bar{y})/\kappa_n + (\ldots)/\kappa_n\right]\right) \\
&= \exp\left(-\frac{1}{2\sigma^2/\kappa_n}\left(\mu - \mu_n^2\right) + (\ldots)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2/\kappa_n}\left(\mu - \mu_n^2\right)\right),
\end{aligned}
$$

which implies that $\mu \mid \sigma^2, y \sim \mathcal{N}\left(\mu_n, \sigma^2/\kappa_n\right)$, where we used $(\ldots)$ to denote constants independent of μ.

Now we can use this result as

$$
\begin{aligned}
p(\sigma^2 \mid y) &= \int p(y, \sigma^2 \mid y) \mathrm{d}\mu \\
&\propto \int (\sigma^2)^{-\frac{3+\nu_n}{2}} \exp\left(\frac{1}{2\sigma^2}\left[\nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2\right]\right) \mathrm{d}\mu \\
&\propto (\sigma^2)^{-\frac{3+\nu_n}{2}} \int \exp\left(\frac{1}{2\sigma^2}\nu_n\sigma_n^2\right) \exp\left(\frac{1}{2\sigma^2/\kappa_n}(\mu_n - \mu)^2\right) \mathrm{d}\mu \\
&\propto (\sigma^2)^{-\frac{3+\nu_n}{2}} \exp\left(\frac{1}{2\sigma^2}\nu_n\sigma_n^2\right) \int \exp\left(\frac{1}{2\sigma^2/\kappa_n}(\mu_n - \mu)^2\right) \mathrm{d}\mu \\
&\propto (\sigma^2)^{-\frac{3+\nu_n}{2}} \exp\left(\frac{1}{2\sigma^2}\nu_n\sigma_n^2\right) \sqrt{2\pi\sigma^2/\kappa_n} \\
&\propto (\sigma^2)^{-(1+\frac{\nu_n}{2})} \exp\left(-\frac{1}{2\sigma^2}\nu_n\sigma_n^2\right),
\end{aligned}
$$

from which we can conclude that $\sigma^2 \mid y \sim \text{scaled-Inv-}\chi^2\left(\nu_n, \sigma_n^2\right)$.

We end the proof by deriving the marginal posterior of μ using an analogous approach as in the proof of Proposition 1. Define $a := \left[\nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2\right]$. We solve for the posterior by integrating σ^2

out using the substitution $z = \frac{a}{2\sigma^2}$. Then

$$
\begin{aligned}
p(\mu \mid y) &= \int_{(0,\infty)} p(\mu, \sigma^2 \mid y) \mathrm{d}\sigma^2 \\
&\propto \int_{(0,\infty)} (\sigma^2)^{-\frac{3+\nu_n}{2}} \exp\left(\frac{1}{2\sigma^2}\left[\nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2\right]\right) \mathrm{d}\sigma^2 \\
&\propto \int_{(0,\infty)} (\sigma^2)^{-\frac{3+\nu_n}{2}} \exp\left(\frac{a}{2\sigma^2}\right) \mathrm{d}\sigma^2 \\
&\propto \int_{(0,\infty)} (a/2z)^{-\frac{3+\nu_n}{2}} \exp\left(-z\right) \frac{a}{2z^2} \mathrm{d}z \\
&\propto \int_{(0,\infty)} a^{-\frac{3+\nu_n}{2}} a z^{\frac{3+\nu_n}{2}} z^{-2} \exp\left(-z\right) \mathrm{d}z \\
&= a^{-\frac{1+\nu_n}{2}} \int_{(0,\infty)} z^{\frac{\nu_n - 1}{2}} \exp\left(-z\right) \mathrm{d}z \\
&= a^{-\frac{1+\nu_n}{2}} \Gamma\left(\frac{\nu_n + 1}{2}\right) \\
&\propto a^{-\frac{1+\nu_n}{2}} \\
&= \left[\nu_n\sigma_n^2 + \kappa_n(\mu_n - \mu)^2\right]^{-\frac{1+\nu_n}{2}} \\
&= \left[\nu_n\sigma_n^2 \left(1 + \frac{1}{\nu_n} \frac{(\mu_n - \mu)^2}{\sigma_n^2/\kappa_n}\right)\right]^{-\frac{1+\nu_n}{2}} \\
&\propto \left[1 + \frac{1}{\nu_n} \frac{(\mu_n - \mu)^2}{\sigma_n^2/\kappa_n}\right]^{-\frac{1+\nu_n}{2}},
\end{aligned}
$$

which concludes the proof by implying that $\mu \mid y \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$. □

References

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

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.

Haggstrom, O. and J. Rosenthal (2007). On variance conditions for markov chain clts. *Electron. Commun. Probab. 12*, 454–464.

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

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91*(435), 1343–1370.

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

Kipnis, C. and S. R. S. Varadhan (1986). Central limit theorem for additive functionals of reversible markov processes and applications to simple exclusions. *Comm. Math. Phys. 104*(1), 1–19.

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*.

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.

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.