Bayesian Hierarchical Models

Research Module - Econometrics and Statistics - 2019/2020

Tim Mensinger*

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.

*tim.mensinger[at]uni-bonn.de

Contents

1 Introduction

The advent of extensive data collection and data storage has led practitioners of applied econometrics to consider bigger and more complex models. However, with a lack of simultaneous bursts of growth in (economic) theory, users often find themselves employing purely data-driven approaches. This is not bad per se, nonetheless, without imposing structure it is hard to derive non-trivial inferences from such analyses. In this paper we consider a common type of data which is inherently equipped with certain structure and present how to utilize this additional level of information. Moreover we focus on so called Bayesian approaches, that allow for extra knowledge to be readily added to the model.

In section 2 we introduce the formal notion of Bayesian thinking and estimation, which we expose by solving a simple normal model. We end the section by presenting the ideas behind Markov chain Monte Carlo, as it is a fundamental concept in modern Bayesian statistics. In section 3 we present hierachical data —which contains aforementioned additional structure— and ways of modeling it. Our theoretic considerations end with a presentation of hierachical linear models, the class of models we use in the remaining parts of the paper.

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.

We 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 may 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 thought.

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 $Y = \mathcal{M}(X; \epsilon, \theta)$, where θ denotes the model parameters and ϵ an explicitly modeled error term which corrects for uncertainty in the model, e.g. in the case of non-observables.

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 learn about 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 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. $\hat{\theta} \in \operatorname{argmax}_{\theta\in\Theta} \mathcal{L}(\theta; \mathcal{D})$. 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 $\sqrt{n}\widehat{V}^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{I})$. This result we can use to quantify uncertainty in $\hat{\theta}$ by computing standard errors and confidence intervals, as well as 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 coin an infinite number of times, the relative frequency of heads converges to the probability of heads. We do expect the outcome of an experiment (e.g. coin flip) to vary, however, we do not assume the true parameter to vary. For instance, we may interpret the probability of a coin landing on heads as the true model parameter. In this sense, it would be absurd to let this object vary for different experiments. Therefore any hypothesis on θ_0 is either true or false. And it is this binarity which makes hypothesis testing (interpretation of confidence intervals) 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, nevertheless, in the Bayesian paradigm we build 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 on our model

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

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 may be used to incorporate knowledge about the parameter into the analysis that 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 from which 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: $\mathbb{P}(\theta \in \Theta_0 \mid \mathcal{D}) = \int_{\Theta_0} p(\theta \mid \mathcal{D})\mathrm{d}\theta$.

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 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 $y = (y_1, ..., y_n)$ 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 **?**. 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 $p(y \mid \mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\sum_i (y_i - \mu)^2 / 2\sigma^2\right)$, where we dropped all proportionality constants. Application of Bayes theorem yields $p(\mu, \sigma^2 \mid y) \propto p(y \mid \mu, \sigma^2)p(\mu, \sigma^2) \propto (\sigma^2)^{-(n+2)/2} \exp\left(-\sum_i (y_i - \mu)^2 / 2\sigma^2\right)$. 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 $\mu \mid y \sim t_{n-1}(\bar{y}, s^2/n)$ and $\sigma^2 \mid y \sim$ scaled-Inv-$\chi^2\left(n-1, s^2\right)$, where $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$ denotes the (unbiased) sample variance and $\bar{y} = \frac{1}{n}\sum_i y_i$ the sample mean, respectively.*

Proof. See appendix. □

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

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

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

Conjugate Prior. 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* (see definition 2 in the appendix) plays an important part in Bayesian statistics as they are able to provide such assurance.

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, as this assures the posterior to be of known form. 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 $\sigma^2 \sim$ 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 **?** 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 $p(\mu, \sigma^2) = p(\mu \mid \sigma^2)p(\sigma^2) \propto (\sigma^2)^{\frac{3+\nu_0}{2}} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2\right]\right)$. 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* $\nu_n = \nu_0 + n$*,* $\kappa_n = \kappa_0 + n$*,* $\mu_n = \frac{\kappa_0}{\kappa_n}\mu_0 + \frac{n}{\kappa_n}\bar{y}$ *and* $\sigma_n^2 = \left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right]/\nu_n$*.*

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* $\mu \mid y \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$ *and* $\sigma^2 \mid y \sim$ 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. □

Table 2: Comparison of Bayesian estimates using a conjugate prior to ML estimates. See appendix for derivation.

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

Next we consider the results of Proposition 3, 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 deduce 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 —we can use κ_0 then to express our degree of believe in the prior. 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 $\sigma_n^2/\kappa_n = \frac{n}{(\nu_0+n)(\kappa_0+n)}s^2 + \mathcal{O}(1/n^2) = \mathcal{O}(1/n)$, 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 they are also reasonable

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. Closed form solutions allow us to fully investigate the influence of the prior on our results. However, we have also seen that Bayesian analyses are far from trivial and depend 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 may 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 likelihood 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 $1/n \sum_i h\left(\theta^{(i)}\right) \xrightarrow{a.s.} \mathbb{E}\left[h(\theta) \mid \mathcal{D}\right]$, under some regularity conditions on h and $p(\theta \mid \mathcal{D})$, with similar results holding for sample quantiles. Hence, to learn something about $p(\theta \mid \mathcal{D})$ we can formulate questions using quantiles or general expectations and rely on the statement above.

In the subsequent paragraphs we discuss methods to sample from the posterior that work under the general assumption that we can evaluate the posterior at arbitrary points up to an integration constant. 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 motivated above for independent samples. We end this section by introducing an algorithm that accomplishes the above.

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. Assume also that the distribution of the chain at time n converges to this invariant distribution no matter where we initialize the chain. Then, in principle, we could run the chain *long enough* until it converged to the invariant distribution and then consider all subsequent realizations as draws from the stationary distribution. This is the core idea of Markov chain Monte Carlo (MCMC). In practice, however, we do not know when a chain is run *long enough*. In part 3 we present some measures that can be of help with this problem during the application.

Under some regularity conditions, similar but not as strict as in Theorem 1, we get a law of large numbers for such Markov chains (see e.g. **?**, 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 usually too long. That is why we focus on assumptions which admit a central limit theorem with the usual $\sqrt{n}$ convergence rate, as it allows for more rigorous statements about our confidence in the whereabouts of the estimator for large samples.

Remark. (i) Having a central limit theorem in the background does *not* imply that the asymptotic distribution provides a good approximation for finite samples. We still do not know when the asymptotics 'kicks in' (**?**). But under assumptions that allow for a CLT we can be more confident in our results than under assumption that only allow of a LLN. *(ii)* 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 properties when also considering the creational process of the Markov chain. We remark that we will *not* formally introduce all concepts and will provide only a heuristic explaination. 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. We refer to **?** for a survey on recent advances with application to MCMC and **?** 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* $\sigma^2(h) := \mathrm{Var}_\pi(h)\,\tau := \mathrm{Var}_\pi(h)\sum_{k\in\mathbb{Z}}\mathrm{Corr}\left(h(X_0), h(X_k)\right)$. *Assume the Markov chain is* ϕ*-irreducible, aperiodic, reversible with respect to* π *and that* $\sigma^2(h) < \infty$. *Then* π *is stationary for the chain and*[1]

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

Proof. See **?** for a complete proof of the second claim; see **?** Proposition 1 for the first claim and Theorem 27 for the second. □

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 such 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 this condition can be satisfied by construction of the chain.

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 from 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, in the next paragraph we will see that this condition can be satisfied by construction with π being the posterior distribution from which we want to sample.

Finite variance ($\sigma^2(h) < \infty$) implies that the integrated correlation time τ must be finite, as we assumes square integrability of h. The integrated correlation time is finite if the correlation function decreases fast enough to zero. Heuristically speaking, for a CLT to work we need more information as would be available in a sample for which the integrated autocorrelation time is infinite. If it is finite we get the usual large sample variance approximation $\sigma^2(h)/n = \mathrm{Var}_\pi(h)/(n/\tau)$. In this sense we might say that n/τ 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. Here we consider one method which implicitly defines a Markov chain with the desired properties, the Metropolis-Hastings algorithm (**?**, **?**). For other approaches and

[1] In the original paper by **?** 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 **?** 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 $\sigma^2(h) = \sum_{k\in\mathbb{Z}}\mathrm{Cov}\left(h(X_0), h(X_k)\right) = \mathrm{Var}\left(h(X_0)\right) + \sum_{k\neq 0}\mathrm{Cov}\left(h(X_0), h(X_k)\right) = \mathrm{Var}_\pi(h)\left(1 + \sum_{k\neq 0}\mathrm{Cov}\left(h(X_0), h(X_k)\right)\right)/\mathrm{Var}\left(h(X_0)\right) = \mathrm{Var}_\pi(h)\left(1 + \sum_{k\neq 0}\mathrm{Corr}(h(X_0), h(X_k))\right) = \mathrm{Var}_\pi(h)\sum_{k\in\mathbb{Z}}\mathrm{Corr}(h(X_0), h(X_k))$.

more involved algorithms see for example **?**.

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 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. The main idea is that we start with an initial value and iteratively propose new values of the chain, however, we do not accept every proposal, but we do so only with a certain probability —this probability can be thought of as high if the proposal has a high relative density compared to the last chain link. Here we also see why we do not care about any integration constant, as the target density only appears as a ratio (line 4).

Clearly the results depend on the choice of the proposal density. A common pick 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 $y = x_t + \mathcal{N}\left(0, \sigma^2\right)$. See **?** for a recent survey on random walk proposals.

The simplicity of the algorithm is remarkable, but the main question is if the resulting Markov chain inherits favorable properties. And indeed this is the case. By construction the algorithm creates Markov chains which are reversible with respect to π and aperiodic, and if additionally the proposal density is positive and continuous and π is finite then the chain is π-irreducible; see for example **?**. This tells us that under these regularity conditions a CLT holds for chains created using the Metropolis-Hastings algorithm. To avoid using samples that do not come from the stationary distribution, in practice we choose T very large and discard the first few (*burn-in*) samples.

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, **?** show that the classical Metropolis-Hastings algorithm is highly dependent on the proposal density and fails in higher dimensions; **?** provide a geometric intution. Numerous novel methods which deal with dimensionality problems have been published. A particular promising route seems to be *Hamiltonian Monte Carlo* (**?**), which is one of the algorithms used in the probabilistc programming language STAN (**?**) with ongoing research on theoretical properties, see for example **?**.

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 phenomenon 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 $\mathrm{vol}(B_d)/\mathrm{vol}(C_d)$ change as d increases. And second, how does the ratio of probability mass distributed by a standard gaussian on these regions change as d increases. Since closed form expressions of volumina of geometrical objects exist the first questions needs little work. Similarly we can easily compute $\mathbb{P}(X \in C_d) = [\Phi(1) - \Phi(-1)]^d$, where Φ denotes the one-dimensional gaussian cumulative distribution function. However, to compute $\mathbb{P}(X \in B_d)$ 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 $\overline{\mathbb{P}(X \in B_d)} := \sup_{\boldsymbol{x} \in B_d} \phi(\boldsymbol{x}) \text{vol}(B_d) = \sup_{\boldsymbol{x} \in B_d} \phi(\boldsymbol{x}) \int_{B_d} 1 \mathrm{d}\boldsymbol{x} \geqslant \int_{B_d} \phi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{P}(X \in B_d)$. 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 phenomenon 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) which states, inter alia, that most probability mass lies within an annulus centered at the origin with an average distance to the origin of $\sqrt{d}$.

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.

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
$\overline{\mathbb{P}(X \in B_d)}/\mathbb{P}(X \in C_d)$	1.16874	1.07281	0.83589	0.35870	0.10995	0.01184	0.00012

We have seen some unintuitive behavior in higher dimensions which might explain why regular methods do not work or only work very slowly. The next paragraph presents one method which utilizes this behavior to efficiently produce samples.

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

We note that it is impossible to provide a rigorous introduction to HMC here, which is why we will focus on the general intution and refer to a series of papers by Michael Betancourt and several coauthors on, the geometric foundations of HMC (**?**); geometric ergodicity of HMC (**?**) and HMC for hiearchical models (**?**).[2]

One reason why ordinary MCMC methods might converge only very slowly in higher dimensions is that the proposal distribution used in the Metropolis-Hastings algorithm does not properly capture the geometry of the high-dimensional space which leads to many rejected proposals and therefore an inefficient exploration of the parameter space. The main idea of HMC is to extend the parameter space by constructing a specific vector field on it, which moves the chain from one proposed point to another in a way so that we consider points that lie in regions with high probability mass and we regulary jump to faw away points as to explore the space as quickly as possible. But how do we construct this vector field? Note that when considering differentiable posterior densities the gradient defines a vector field. However, this vector field points to the modes of the posterior and as we saw in the last paragraph, in higher dimensions we will find little to no probability mass near the modes. This is where Hamiltonian mechanics comes into play by providing a set of equations (Hamilton's equations) that describe the time-evolution of the interplay of kinetic and potential energy of a system. What this means for our case is best explained by imagining the mode as the center of gravity, with gravity pulling harder as we get closer to the mode —this can be thought of as the gradient vector field. But as most probability mass is spread around an annulus around the mode we do not want to move closer to the center of gravity, we want to move around an orbit around the mode. The distance of the orbit to the mode and exact shape depend of course on the dimensionality and posterior distribution of the problem. Hamilton's equations

[2]Besides doing theoretical research on HMC, Michael Betancourt worked for STAN on integrating the HMC algorithm and runs an educational blog where he presents his research using modern tools, see `https://betanalpha.github.io/`.

provide us with a way to construct a vector field so that moving along the field drifts the Markov chain into this orbit. Once reached the chain explores the relevant space quickly.

Why is the above important? Bayesian statistics and its application to real world problems has gained immense popularity with the invention of Markov chain Monte Carlo methods. It's naive application to general complex high-dimensional problems is not computationally feasible however. These problems are under active development and new approaches on the algorithmic and theoretical side, as the one presented above, prove fruitfull.

3 Hierarchical Models

In the following we consider hierarchical models and when they are applicable. We begin by presenting the idea of hiearchical data and modeling. We then show the general structure of hierarchical models and end this section by introducing the most widely used subclass, hierarchical linear models.

3.1 Hierarchical Data and Modeling

Here we consider what makes data *hierarchical* and how we can use this further 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 in 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 why some authors nowadays prefer the more general terms *multi-level data* and *multi-level model*, see for instance **?**.

We categorize hierarchical models by the number of levels they incorporate and their use of *nested* or *non-nested* data. In this paper we consider two-level models for nested data and refer (again) to **?** and **?** for a treatment of more general settings.

Let us continue with the first example of children in schools. Assume we observe, say, test results, parental income and number of teachers per child per school. Figure 3.1 portrays how multi-level data in this case may be stored.

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 *(test results, parental income, school id). Right: Data on the* school-level *(number of teachers per child).*

To introduce formal notation, assume we observe data on $i = 1, ..., n$ individuals which are clustered among $j = 1, ..., J$ groups. Naturally we consider outcomes y_i on the individual-level. Following the idea of different relations for different levels, we write x_i for covariates that vary by individual and u_j for covariates that only vary by group. 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 $(x_i, u_{j[i]})$. How can we utilize this hierarchical structure?

The main idea of hierarchical modeling is to build (simple) models on different levels, where the outcomes are modeled on the individual-level and we connect the levels by modeling on all higher levels the parameters used in the previous level. The level structure is given by the hierarchical structure of the data and we use features from level ℓ only in the modeling process of level ℓ.

For a general two-level case we may write a model as

$$y_i \mid \theta_{j[i]} \sim p(y_i \mid x_i, \theta_{j[i]}) \,, \qquad \text{(Individual Level)}$$

$$\theta_j \mid \phi \sim p(\theta_j \mid u_j, \phi) \,, \qquad \text{(Group Level)}$$

$$\phi \sim p(\phi \mid \zeta) \text{ with } \zeta \text{ fixed} \;, \qquad \text{(Prior)}$$

where we suppress the dependence on x_i and u_j. On the individual-level the outcomes y_i are modeled depending on the unit-level features x_i and parameters $\theta_{j[i]}$. As in a classical Bayesian model we continue by modeling the parameters θ_j, however, in contrast to section 2 we do not assume a prior distribution but we explicitly model the parameter on the group-level using the group-level features u_j. From a Bayesian point of view this can be seen as a generalization of 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 produce a Bayesian model we assign a prior distribution on all (hyper)parameters that are not explicitly modeled (here ϕ).

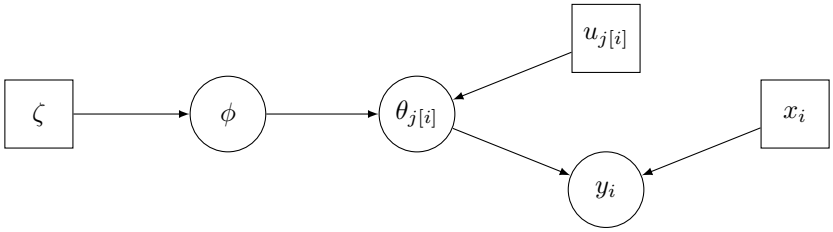

Figure 3.2: A generic two-level Bayesian hierarchical model depicted as a directed acyclical graph modeling a single generic observation. Circled parameters denote random quantities while parameters contained in squares denote fixed quantities.

Figure 3.2 illustrates the structure of modeling a generic observation y_i using the general model from above. 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.

How do we solve for the posterior in these models? Owing to the hierarchical nature we can derive the joint posterior as $p(\phi, \theta \mid y) \propto p(y \mid \phi, \theta)p(\phi, \theta) = p(y \mid \theta)p(\theta \mid \phi)p(\phi)$ where we use that y is independent of ϕ conditional on θ —this can be seen immediatly from figure 3.2. We note that all factors on the right-hand side are known, hence, we can directly apply the methods discussed in the previous section.

3.2 Hierarchical Linear Models

We end our segment on theory with the introduction of hierachical linear models, the class of models we will be using exclusively in our Monte Carlo and application part.

As in classical statistics, linear models are usually simpler to estimate and easier to interpret, which can explain their widespread use in practice. Standard critiques on linear models, e.g. model misspecification issues, also apply here, however, we will see that due to the hiearchical nature we are able to model complex structure, even under a linearity assumption. Being more precise, linearity in a hierarchical context means that on each level parameters enter the *level-model* linearly. Let us formalize this.

General Definition. As before, we consider outcomes y_i for individuals $i = 1, ..., n$ in groups $j = 1, ..., J$, with individual-level characteristics x_i and group-level characteristics $u_{j[i]}$. A general two-level

hierarchical linear model can then be written as

$$y_i = \alpha_{j[i]} + x_i^\intercal \beta_{j[i]} + \epsilon_i \,, \qquad \text{(Individual Level)}$$

$$\beta_j = \gamma_0 + u_j^\intercal \gamma + \eta_j \,, \qquad \text{(Group Level)}$$

where ϵ_i and η_j denote error terms on the respective level. Commonly the group-level errors are modeled as independent and identical, and the individual-level errors are modeled as independent and identical in their respective group. Note that this can be extended for the heteroscedastic or autocorrelated case. Just like in the previous section, this model does not define a Bayesian model per se. We make the model Bayesian by imposing a prior distribution on all paramters that are not explicitly modeled themself. In the above case this would be $\gamma_0, \gamma, \epsilon$ and η.

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

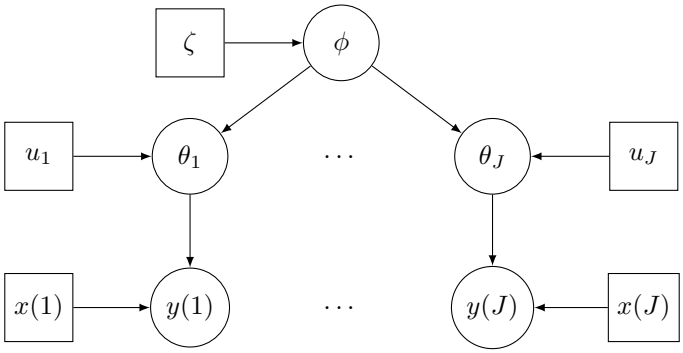

Figure A.1: A generic two-level Bayesian hierarchical model depicted as a directed acyclical graph modeling generic observations $y(j)$ in groups $j = 1, ..., J$. Circled parameters denote random quantities while parameters contained in squares denote fixed quantities.

A.3 Proofs

Derivation of Results in Table 1. For the normal data problem at hand the derivation of the maximum likelihood estimators, $\bar{y}$ for μ and $(n-1)s^2/n$ for σ^2, is well known. We note that the "ML Variance" denotes the theoretical variance of the estimators. Further, the moments of the scaled inverse χ^2 distributions are available in closed form. In particular, if $X \sim \text{scaled-Inv-}\chi^2\left(\nu, \tau^2\right)$, then $\mathbb{E}\left[X\right] = \nu\tau^2/(\nu-2)$ and $\text{Var}\left(X\right) = 2\nu^2\tau^4/((\nu-2)^2(\nu-4))$.

Let us first consider the parameter μ. As the t-distribution is parameterized over its mean and variance we can simply read off these values. Further, as the t-distribution is symmetric the MAP is equal to its mean.

Let us now consider the parameter σ^2. Plugging in the respective parameters, we get

$$\mathbb{E}\left[\,\sigma^2 \mid y\,\right] = \frac{n-1}{n-3}\sigma^2$$

directly, and

$$\operatorname{Var}\left(\sigma^2 \mid y\right) = \frac{2(n-1)^2}{(n-3)^2(n-5)}\sigma^4\,.$$

To get the MAP for σ^2 we need to maximize the posterior of σ^2. Note that we can drop any integration constants, i.e. we need to solve

$$\underset{x>0}{\text{maximize}} \left\{ x^{-(1+\frac{n-1}{2})} \exp\left(\frac{-(n-1)s^2}{2x}\right)\right\},$$

where the term in curly brackets is just the posterior density of σ^2 evaluated at x. Using that the constant of integration is positive we can readily see that the posterior is concave. Hence, differentiating with respect to x and setting this to zero yields the desired result. □

Derivation of Results in Table 2. The proof from above applies here by mutatis mutandis. □

Remark. The subsequent proofs presented here follow **?**; 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 $\mu \mid y \sim t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$. □