Lecture 4: Intro to MCMC

STATS305C: Applied Statistics III

Scott Linderman

April 6, 2022

Last Time...

- Directed Graphical Models
- Hierarchical Gaussian Model
- Lots of annoying Gaussian integrals

Today...

Outline:

- Intro to MCMC
- Metropolis-Hastings Algorithm
- Gibbs Sampling
- MCMC Diagnostics

Reading:

- Required: Bishop, Ch 11.2-11.3
- Optional: Murphy, Ch 12.1-12.3

Motivating Example: Hierarchical Gaussian Model with Per-School Variance

Let's return to the "8 Schools" example from Lecture 3. We have test scores $\{\{x_{s,n}\}_{n=1}^{N_s}\}_{s=1}^{S}$ where $x_{s,n}$ the score of the n-th student from the s-th school.

We model the collection of scores using a **hierarchical model**. Unlike last time, now we will add a prior on the per-school variance as well.

$$\tau^2 \sim \chi^{-2}(\nu_0, \tau_0^2) \tag{1}$$

$$\mu \sim \mathcal{N}(\mu_0, \tau^2/\kappa_0) \tag{2}$$

$$\theta_s \sim \mathcal{N}(\mu, \tau^2) \qquad \text{for } s = 1, \ldots, S \tag{3}$$

$$\sigma_s^2 \sim \chi^{-2}(\alpha_0, \sigma_0^2) \qquad \text{for } s = 1, \ldots, S \tag{4}$$

$$x_{s,n} \sim \mathcal{N}(\theta_s, \sigma_s^2) \qquad \text{for } n = 1, \ldots, N_s \text{ and } s = 1, \ldots, S \tag{5}$$

Each school has its own mean θ_s and variance σ_s^2. The means are linked via a hierarchical prior: they are conditionally independent given the global mean μ and variance τ^2

When the variance σ_s^2 is unknown, we can no longer get simple closed-form solutions like last time.

Notation

- Let $\boldsymbol{\theta} \in \mathbb{R}^D$ denote all the model parameters.
  - This is an abuse of notation because for the hierarchical Gaussian model $\boldsymbol{\theta} = (\tau^2, \mu, \{\theta_s, \sigma_s^2\}_{s=1}^S)$, but it will be simpler to lump them into one tuple for now.
- Let $\boldsymbol{X}$ denote the observed data.

Posterior expectations

The central object of Bayesian inference is the posterior distribution, $p(\boldsymbol{\theta} \mid \boldsymbol{X})$.

However, we almost always interact with the posterior distribution through *expectations*.

- $\mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{X})}[\boldsymbol{\theta}]$, the posterior mean.
- $\mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{X})}[\mathbb{I}[\boldsymbol{\theta} \in \mathscr{A}]]$, the probability of the parameters being in set $\mathscr{A}$.
- $\mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{X})}[p(\boldsymbol{X}' \mid \boldsymbol{\theta})]$, the posterior predictive density of new data $\boldsymbol{X}'$.

All of these can be written as $\mathbb{E}_{p(\boldsymbol{\theta} \mid \boldsymbol{X})}[f(\boldsymbol{\theta})]$ for some function f.

(One exception is the posterior mode, which is not so easily expressed as an expectation.)

Approximating posterior expectations

Generally, we can't analytically compute posterior expectations. (*Why not?*)

In these cases, we need to resort to approximations.

For example, we could use *quadrature methods* like Simpon's rule or the trapezoid rule to numerically approximate the integral over Θ.

Roughly,

$$\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{X})}[f(\boldsymbol{\theta})] \approx \sum_{m=1}^{M} p(\boldsymbol{\theta}_m \mid \boldsymbol{X}) f(\boldsymbol{\theta}_m) \Delta_m \tag{6}$$

where $\boldsymbol{\theta}_m \subset \Theta$ is a grid of points and Δ_m is a volume around that point.

This works for low dimensional problems (say, up to 5 dimensions), but the number of points (M) needed to get a good estimate grows exponentially with the parameter dimension.

Monte Carlo approximations

Idea: approximate the expectation via sampling,

$$\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{X})}[f(\boldsymbol{\theta})] \approx \frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{\theta}_m) \quad \text{where} \quad \boldsymbol{\theta}_m \sim p(\boldsymbol{\theta} \mid \boldsymbol{X}). \tag{7}$$

Let $\hat{f} = \frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{\theta}_m)$ denote the Monte Carlo estimate. It is a random variable, since it's a function of random samples $\boldsymbol{\theta}_m$.

As such we can reason about its mean and variance. Clearly,

$$\mathbb{E}[\hat{f}] = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{X})}[f(\boldsymbol{\theta})] = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{X})}[f(\boldsymbol{\theta})]. \tag{8}$$

Thus, $\hat{f}$ is an *unbiased* estimate of the desired expectation.

Monte Carlo approximations II

What about its variance?

$$\mathrm{Var}[\hat{f}] = \mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{\theta}_m)\right) \tag{9}$$

$$= \frac{1}{M^2}\left(\sum_{m=1}^{M}\mathrm{Var}[f(\boldsymbol{\theta})] + 2\sum_{1\le m<m'\le M}\mathrm{Cov}[f(\boldsymbol{\theta}_m), f(\boldsymbol{\theta}_{m'})]\right) \tag{10}$$

If the samples are not only identically distributed but also *uncorrelated*, then $\mathrm{Var}[\hat{f}] = \frac{1}{M}\mathrm{Var}[f(\boldsymbol{\theta})]$.

In this case, the *root mean squared error* (RMSE) of the estimate is $\sqrt{\mathrm{Var}[\hat{f}]} = O(M^{-\frac{1}{2}})$.

Compare this to Simpson's rule, which for smooth 1D problems has error rate $O(M^{-4})$. That's roughly 8 times better than Monte Carlo!

However, for multidimensional problems, Simpson's rule is $O(M^{-\frac{4}{D}})$, whereas the **error rate of Monte Carlo does not depend on the dimensionality!**

↑ # derivatives / # dims.

The Catch

So far so good: we'll just draw a lot of samples to drive down our Monte Carlo error.

Here's the catch! How do you draw samples from the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{X})$?

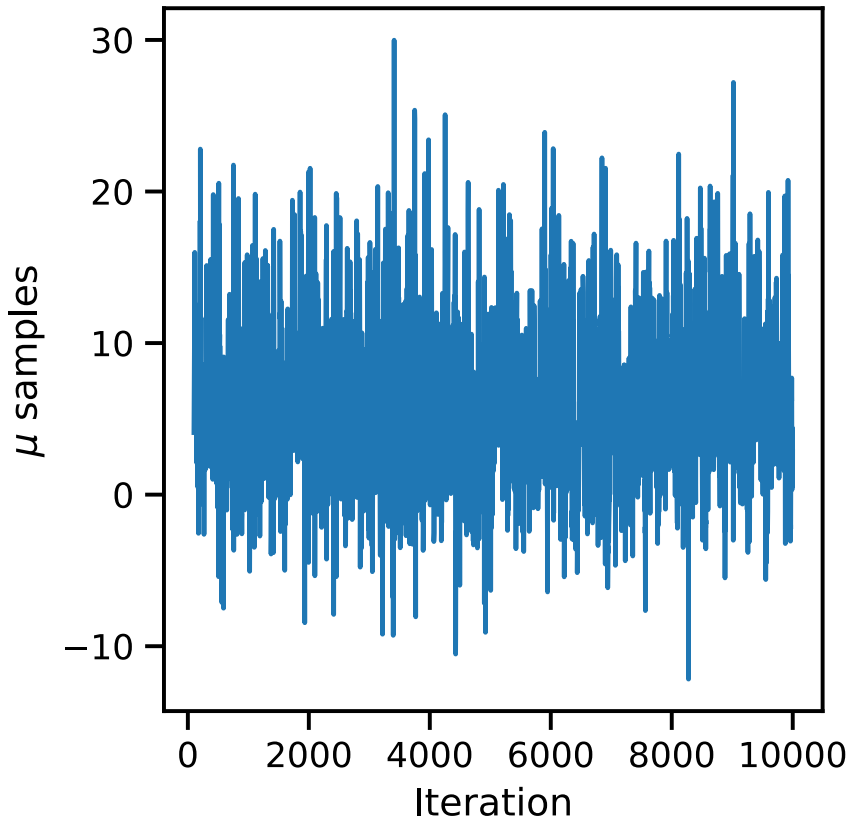
We're interested in Monte Carlo for cases where the posterior does not admit a simple closed form!

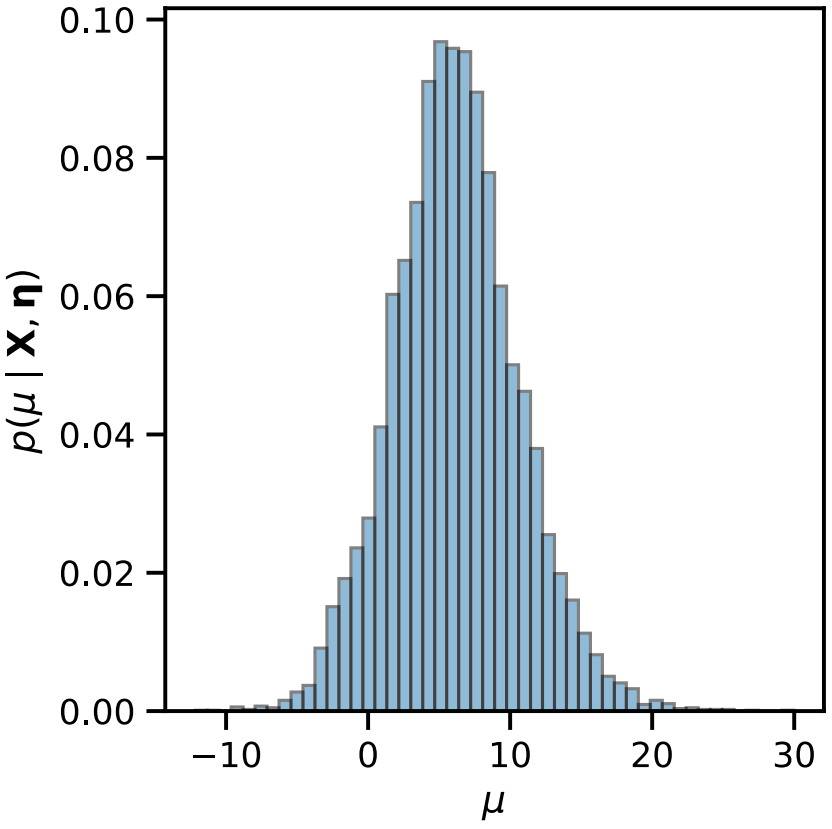
In general, sampling the posterior is as hard as computing the marginal likelihood.

(also, there's presumably cost to sampling so RMSE / unit time might differ)

Markov chain Monte Carlo (MCMC)

Idea: Design a Markov chain whose stationary distribution is the posterior.





Punchline

```
def gibbs(theta, X, num_samples):
    samples = []
    for i in range(num_samples):
        for d in range(len(theta)):
            theta[d] = ... # sample p(theta_d | theta_{\neg d}, X)
        samples.append(theta.copy())
    return samples
```

Markov chains

A *Markov chain* is a joint distribution of a sequence of variables, $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)$.

(To avoid confusion with the model p, we denote the densities associated with the Markov chain by π.)

The Markov chain factorizes so that each variable is drawn conditional on the previous variable,

$$\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M) = \pi_1(\boldsymbol{\theta}_1) \prod_{m=2}^{M} \pi(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_{m-1}). \tag{11}$$

This is called the *Markov property*.

The distribution $\pi_1(\boldsymbol{\theta}_1)$ is called the *initial distribution*.

The distribution $\pi(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_{m-1})$ is called the *transition distribution*. If the transition distribution is the same for each m, the Markov chain is *homogenous*.

Stationary distributions

Let $\pi_m(\boldsymbol{\theta}_m)$ denote the marginal distribution of sample $\boldsymbol{\theta}_m$. It can be obtained recursively as,

$$\pi_m(\boldsymbol{\theta}_m) = \int \pi_{m-1}(\boldsymbol{\theta}_{m-1})\, \pi(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_{m-1})\, \mathrm{d}\boldsymbol{\theta}_{m-1}. \tag{12}$$

We are interested in the asymptotic behavior of the marginal distributions as $m \to \infty$.

A distribution $\pi^*(\boldsymbol{\theta})$ is a **stationary distribution** if,

$$\pi^*(\boldsymbol{\theta}) = \int \pi^*(\boldsymbol{\theta}')\, \pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}')\, \mathrm{d}\boldsymbol{\theta}'. \tag{13}$$

That is, suppose the marginal of sample $\boldsymbol{\theta}'$ is $\pi^*(\boldsymbol{\theta})$. Then the marginal of the next time point is also $\pi^*(\boldsymbol{\theta})$.

Detailed balance

How can we relate transition distributions and stationary distributions?

A sufficient (but not necessary) condition for $\pi^*(\boldsymbol{\theta})$ to be a stationary distribution is that it satisfies *detailed balance*,

$$\pi^*(\boldsymbol{\theta}')\pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = \pi^*(\boldsymbol{\theta})\pi(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) \tag{14}$$

In words, the probability of starting at $\boldsymbol{\theta}'$ and moving to $\boldsymbol{\theta}$ is the same as that of starting at $\boldsymbol{\theta}$ and moving to $\boldsymbol{\theta}'$, if you draw the starting point from the stationary distribution.

To see that detailed balance is sufficient, integrate both sides to get,

$$\int \pi^*(\boldsymbol{\theta}')\pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}' = \int \pi^*(\boldsymbol{\theta})\pi(\boldsymbol{\theta}' \mid \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}' = \pi^*(\boldsymbol{\theta}) \int \pi(\boldsymbol{\theta}' \mid \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}' = \pi^*(\boldsymbol{\theta}). \tag{15}$$

Thus, $\pi^*(\boldsymbol{\theta})$ is a stationary distribution of the Markov chain with transitions $\pi(\theta \mid \theta')$.

Ergodicity

Detailed balance can be used to show that $\pi^*(\boldsymbol{\theta})$ is *a* stationary distribution, but not that it is *the unique* one.

This is where *ergodicity* comes in. A Markov chain is ergodic if $\pi_m(\boldsymbol{\theta}_m) \to \pi^\star(\boldsymbol{\theta})$ regardless of $\pi_1(\boldsymbol{\theta}_1)$.

An ergodic chain has only one stationary distribution, $\pi^*(\boldsymbol{\theta})$.

The easiest way to prove ergodicity is to show that it is possible to reach any $\boldsymbol{\theta}'$ from any other $\boldsymbol{\theta}$. E.g. this is trivially so if $\pi(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) > 0$.

Note: a more technical definition is that all pairs of sets *communicate*, in which case the chain is *irreducible*, and that each state is *aperiodic*. The definitions can be a bit overwhelming.

The Metropolis-Hastings algorithm

Finally we come to our **main objective**: designing a Markov chain for which *the posterior is the unique stationary distribution*.

That is, we want $\pi^*(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{X})$.

Recall our **constraint**: we can only compute the joint probability (the numerator in Bayes' rule), not the marginal likelihood (the denominator).

Fortunately, that still allows us to compute ratios of posterior densities! We have,

$$\frac{p(\boldsymbol{\theta} \mid \boldsymbol{X})}{p(\boldsymbol{\theta}' \mid \boldsymbol{X})} = \frac{p(\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{X})} \frac{p(\boldsymbol{X})}{p(\boldsymbol{\theta}', \boldsymbol{X})} = \frac{p(\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{\theta}', \boldsymbol{X})}. \tag{16}$$

The Metropolis-Hastings algorithm II

Now rearrange the detailed balance condition to relate ratios of transition probabilities to ratios of joint probabilities,

$$\frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}' \mid \boldsymbol{\theta})} = \frac{\pi^*(\boldsymbol{\theta})}{\pi^*(\boldsymbol{\theta}')} = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{X})}{p(\boldsymbol{\theta}' \mid \boldsymbol{X})} = \frac{p(\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{\theta}', \boldsymbol{X})} \tag{17}$$

To construct such a transition distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$, break it down into two steps.

1. Sample a proposal $\boldsymbol{\theta}$ from a *proposal distribution* $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$,
2. Accept the proposal with *acceptance probability* $a(\boldsymbol{\theta}' \to \boldsymbol{\theta})$. (Otherwise, set $\boldsymbol{\theta} = \boldsymbol{\theta}'$.)

Thus,

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = \begin{cases} q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')\, a(\boldsymbol{\theta}' \to \boldsymbol{\theta}) & \text{if } \boldsymbol{\theta}' \neq \boldsymbol{\theta} \\ \int q(\boldsymbol{\theta}'' \mid \boldsymbol{\theta}')\, (1 - a(\boldsymbol{\theta}' \to \boldsymbol{\theta}''))\, \mathrm{d}\boldsymbol{\theta}'' & \text{if } \boldsymbol{\theta}' = \boldsymbol{\theta} \end{cases} \tag{18}$$

The Metropolis-Hastings algorithm III

The constraint in (17) is trivially satisfied when $\boldsymbol{\theta} = \boldsymbol{\theta}'$. When $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, we need

$$\frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}' \mid \boldsymbol{\theta})} = \frac{q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')\, a(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})\, a(\boldsymbol{\theta} \to \boldsymbol{\theta}')} = \frac{p(\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{\theta}', \boldsymbol{X})} \Rightarrow \frac{a(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{a(\boldsymbol{\theta} \to \boldsymbol{\theta}')} = \underbrace{\frac{p(\boldsymbol{\theta}, \boldsymbol{X})\, q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}{p(\boldsymbol{\theta}', \boldsymbol{X})\, q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}}_{\triangleq A(\boldsymbol{\theta}' \to \boldsymbol{\theta})} \tag{19}$$

WLOG, assume $A(\boldsymbol{\theta}' \to \boldsymbol{\theta}) \leq 1$. (If it's not, its inverse $A(\boldsymbol{\theta} \to \boldsymbol{\theta}')$ must be.)

A simple way to ensure detailed balance is to set $a(\boldsymbol{\theta}' \to \boldsymbol{\theta}) = A(\boldsymbol{\theta}' \to \boldsymbol{\theta})$ and $a(\boldsymbol{\theta} \to \boldsymbol{\theta}') = 1$.

We can succinctly capture both cases with,

$$a(\boldsymbol{\theta}' \to \boldsymbol{\theta}) = \min\left\{1, A(\boldsymbol{\theta}' \to \boldsymbol{\theta})\right\} = \min\left\{1, \frac{p(\boldsymbol{\theta}, \boldsymbol{X})\, q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}{p(\boldsymbol{\theta}', \boldsymbol{X})\, q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}\right\}. \tag{20}$$

The Metropolis algorithm

Now consider the special case in which the proposal distribution is symmetric; i.e. $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$.

Then the proposal densities cancel in the acceptance probability and,

$$a(\boldsymbol{\theta}' \to \boldsymbol{\theta}) = \min \left\{ 1, \frac{p(\boldsymbol{\theta}, \boldsymbol{X})}{p(\boldsymbol{\theta}', \boldsymbol{X})} \right\}. \tag{21}$$

In other words, you accept any proposal that moves "uphill," and only accept "downhill" moves with some probability.

This is called the *Metropolis algorithm* and it has close connections to *simulated annealing*.

Gibbs Sampling

Gibbs is a special case of MH with proposals that always accept.

Gibbs sampling updates one "coordinate" of $\boldsymbol{\theta}$ at a time by sampling from its conditional distribution.

Think of this as a proposal distribution. Suppose $\boldsymbol{\theta} \in \mathbb{R}^D$. For each coordinate $d \in 1, \ldots, D$,

$$q_d(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = p(\theta_d \mid \boldsymbol{\theta}'_{\neg d}, \boldsymbol{X})\, \delta_{\boldsymbol{\theta}'_{\neg d}}(\boldsymbol{\theta}_{\neg d}), \tag{22}$$

where $\boldsymbol{\theta}_{\neg d} = (\theta_1, \ldots, \theta_{d-1}, \theta_{d+1}, \ldots, \theta_D)$ denotes all parameters except θ_d.

In other words, the proposal distribution q_d samples θ_d from its conditional distribution and leaves all the other parameters unchanged.

Gibbs Sampling II

What is the probability of accepting this proposal?

$$a_d(\boldsymbol{\theta}' \to \boldsymbol{\theta}) = \min\left\{1,\ \frac{p(\boldsymbol{\theta}, \boldsymbol{X}) q_d(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}{p(\boldsymbol{\theta}', \boldsymbol{X}) q_d(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}\right\} \tag{23}$$

$$= \min\left\{1,\ \frac{p(\boldsymbol{\theta}, \boldsymbol{X}) p(\theta'_d \mid \boldsymbol{\theta}_{\neg d}, \boldsymbol{X}) \delta_{\boldsymbol{\theta}_{\neg d}}(\boldsymbol{\theta}'_{\neg d})}{p(\boldsymbol{\theta}', \boldsymbol{X}) p(\theta_d \mid \boldsymbol{\theta}'_{\neg d}, \boldsymbol{X}) \delta_{\boldsymbol{\theta}'_{\neg d}}(\boldsymbol{\theta}_{\neg d})}\right\} \tag{24}$$

$$= \min\left\{1,\ \frac{p(\boldsymbol{\theta}_{\neg d}, \boldsymbol{X}) p(\theta_d \mid \boldsymbol{\theta}_{\neg d}, \boldsymbol{X}) p(\theta'_d \mid \boldsymbol{\theta}_{\neg d}, \boldsymbol{X}) \delta_{\boldsymbol{\theta}_{\neg d}}(\boldsymbol{\theta}'_{\neg d})}{p(\boldsymbol{\theta}'_{\neg d}, \boldsymbol{X}) p(\theta'_d \mid \boldsymbol{\theta}'_{\neg d}, \boldsymbol{X}) p(\theta_d \mid \boldsymbol{\theta}'_{\neg d}, \boldsymbol{X}) \delta_{\boldsymbol{\theta}'_{\neg d}}(\boldsymbol{\theta}_{\neg d})}\right\} \tag{25}$$

$$= \min\{1, 1\} = 1 \tag{26}$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$ that differ only in their d-th coordinate.

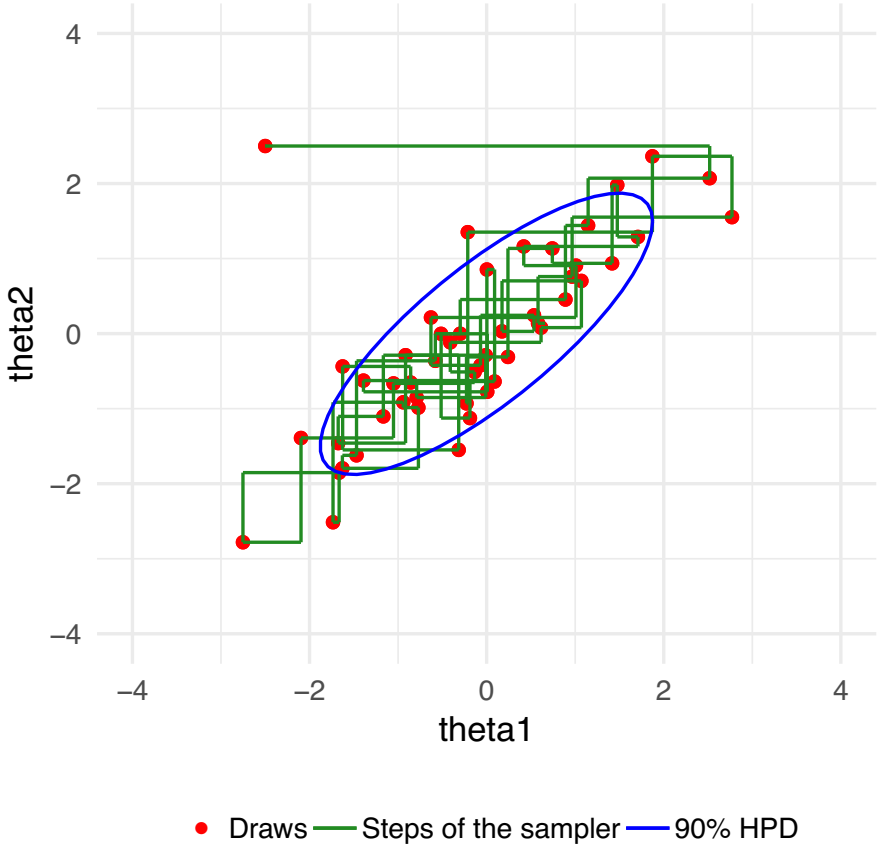
The Gibbs proposal is *an offer you cannot refuse*.

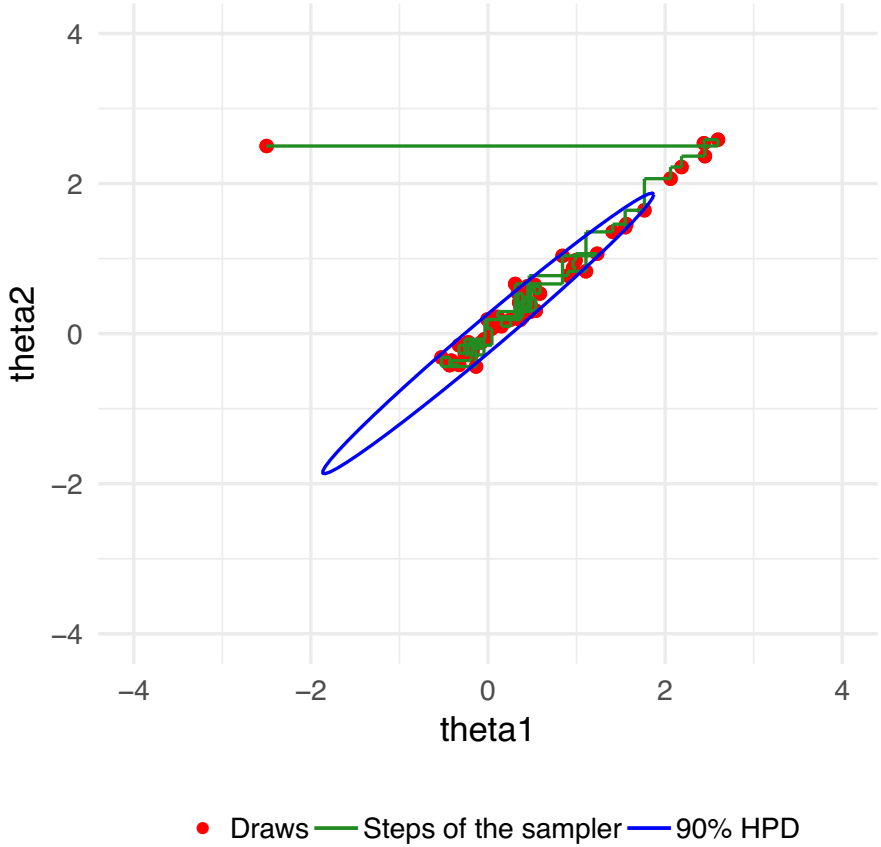
Gibbs Sampling III

Of course, if we only update one coordinate, the chain can't be ergodic. However, if we cycle through coordinates it generally will be.

Question: Does the order in which we update coordinates matter?

Gibbs in a 2D Gaussian example





Gibbs Sampling in the Hierarchical Gaussian Model

Recall the model,

$$\tau^2 \sim \chi^{-2}(\nu_0, \tau_0^2) \tag{27}$$

$$\mu \sim \mathcal{N}(\mu_0, \tau^2/\kappa_0) \tag{28}$$

$$\theta_s \sim \mathcal{N}(\mu, \tau^2) \quad \text{for } s = 1, \ldots, S \tag{29}$$

$$\sigma_s^2 \sim \chi^{-2}(\alpha_0, \sigma_0^2) \quad \text{for } s = 1, \ldots, S \tag{30}$$

$$x_{s,n} \sim \mathcal{N}(\theta_s, \sigma_s^2) \quad \text{for } n = 1, \ldots, N_s \text{ and } s = 1, \ldots, S \tag{31}$$

Gibbs Sampling in the Hierarchical Gaussian Model II

Question: What is the conditional distribution $p(\mu \mid \tau^2, \{\theta_s, \sigma_s^2\}_{s=1}^S, \boldsymbol{X}, \boldsymbol{\eta})$?

Gibbs Sampling in the Hierarchical Gaussian Model III

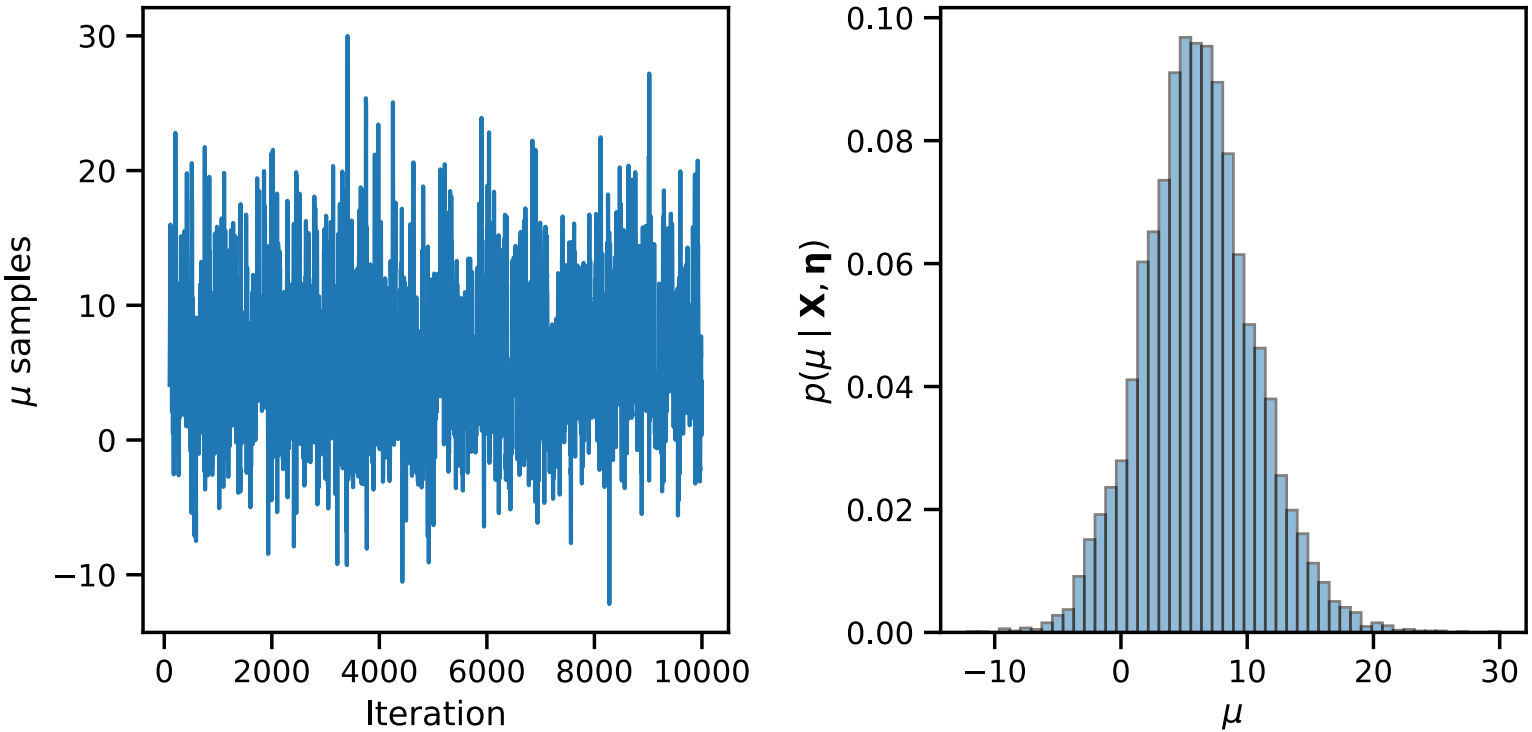


Figure: Trace and histogram of samples of μ

Gibbs Sampling in the Hierarchical Gaussian Model IV

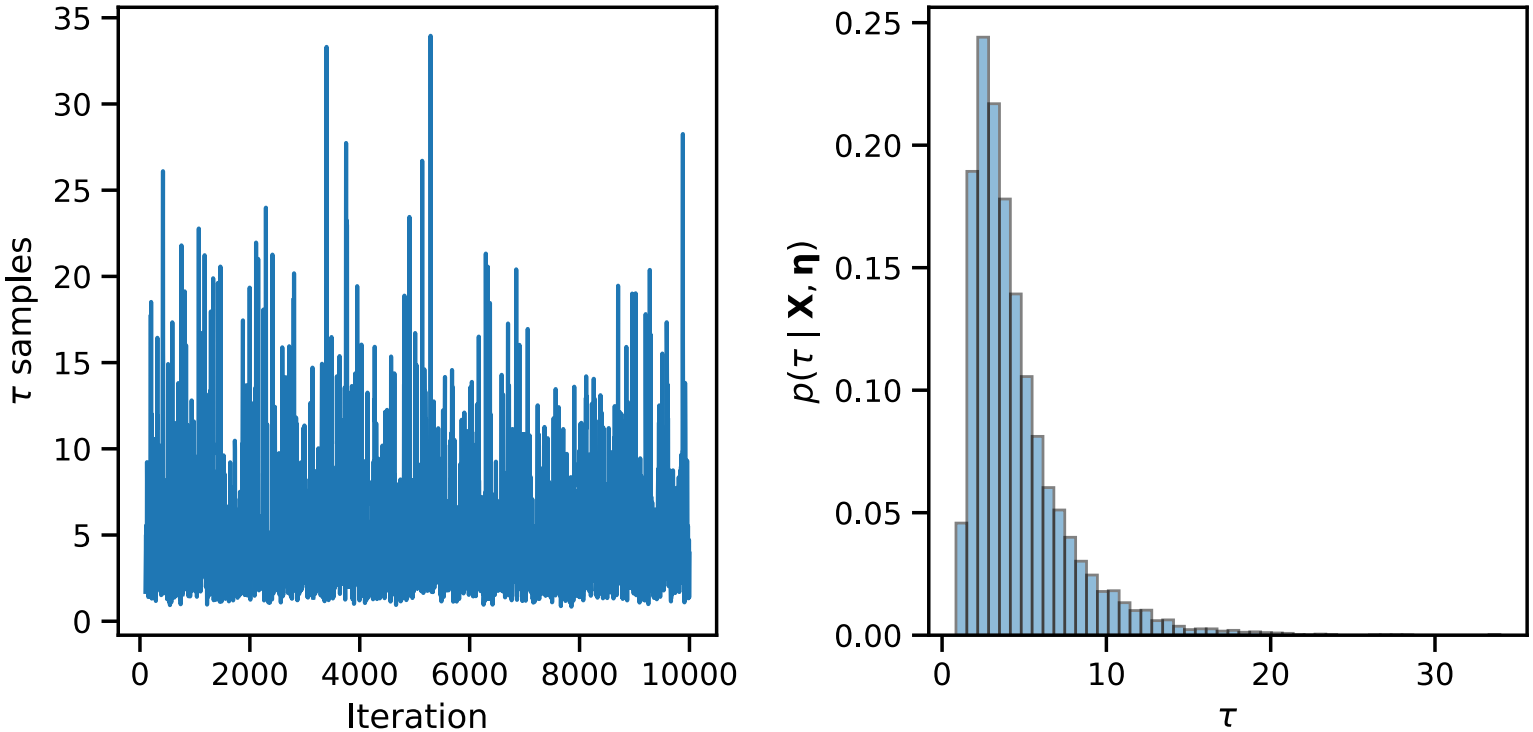


Figure: Trace and histogram of samples of τ

Gibbs Sampling in the Hierarchical Gaussian Model V

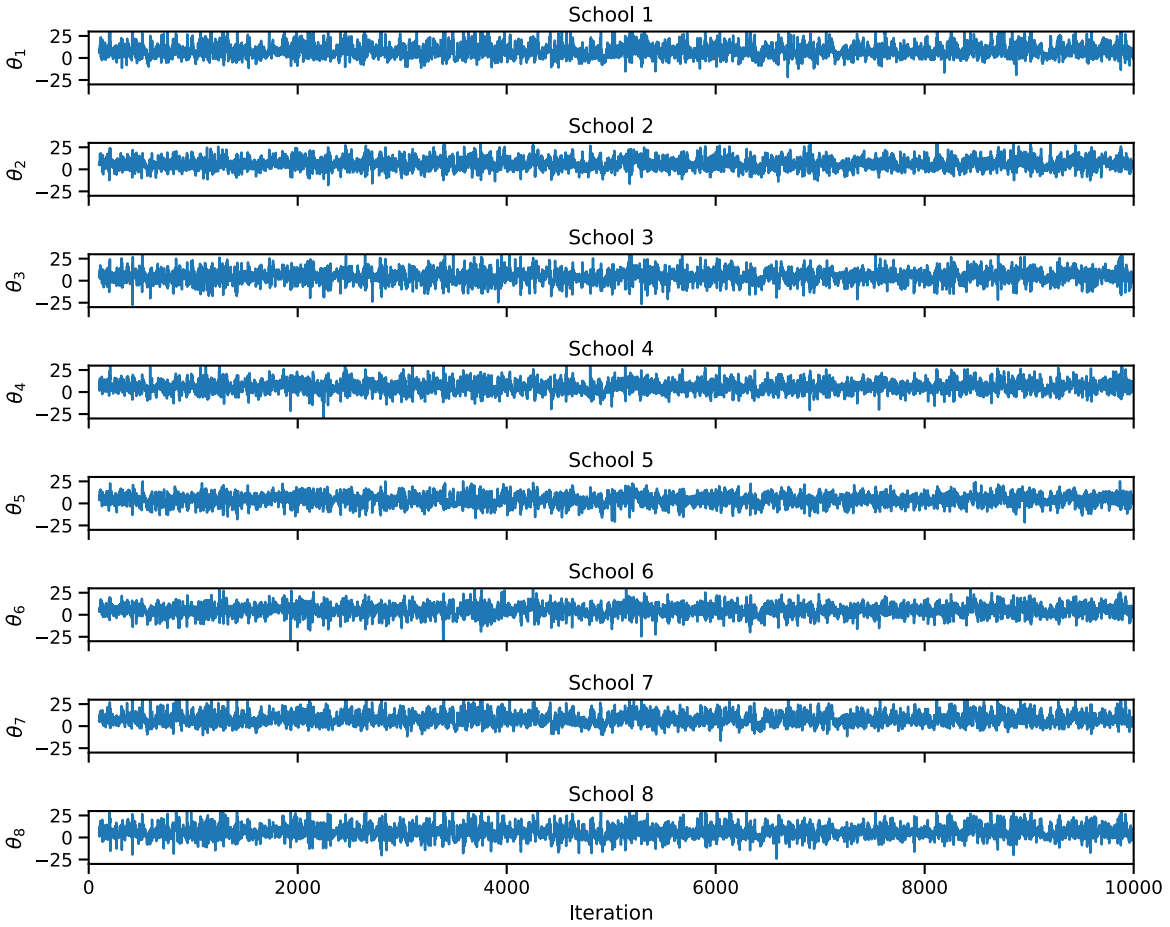


Figure: Traces of samples of θ_s

Gibbs Sampling in the Hierarchical Gaussian Model VI

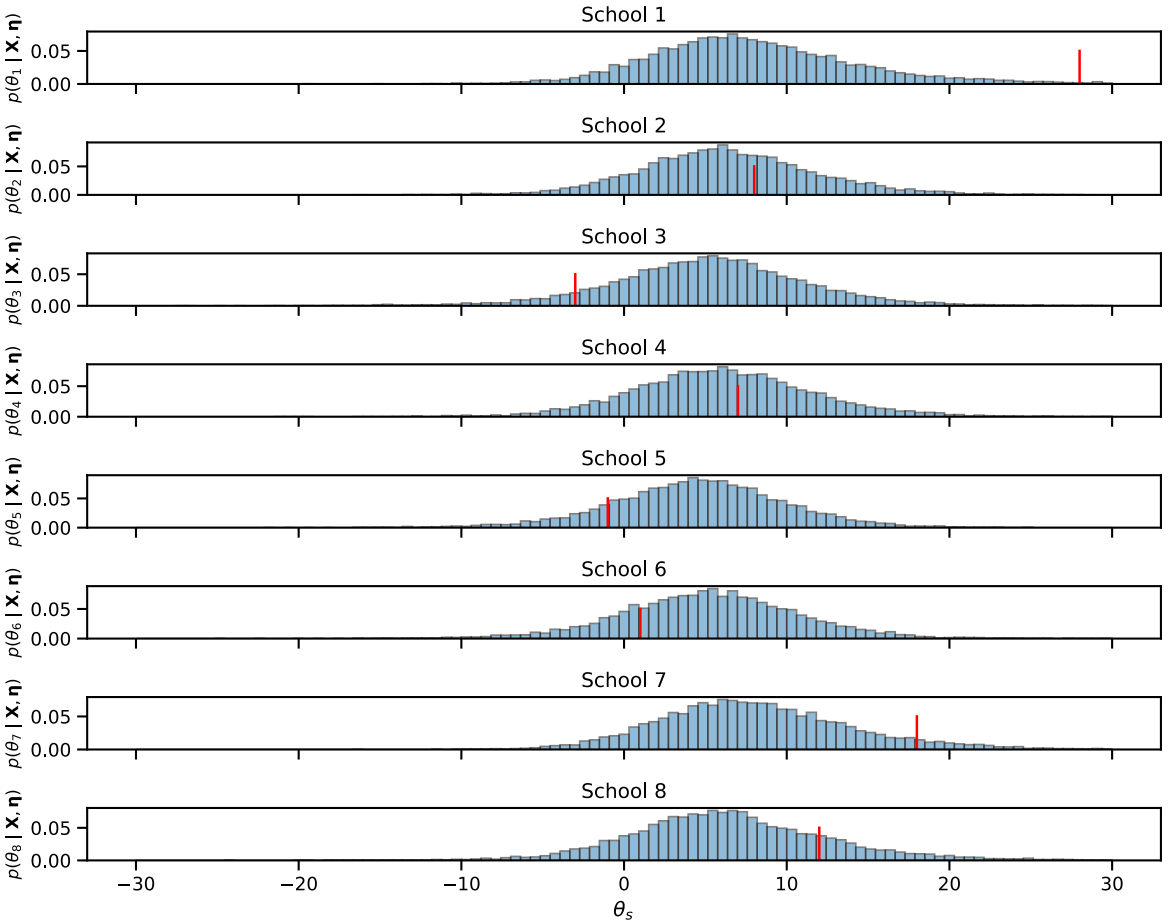


Figure: Histogram of samples of θ_s

Gibbs Sampling in the Hierarchical Gaussian Model VII

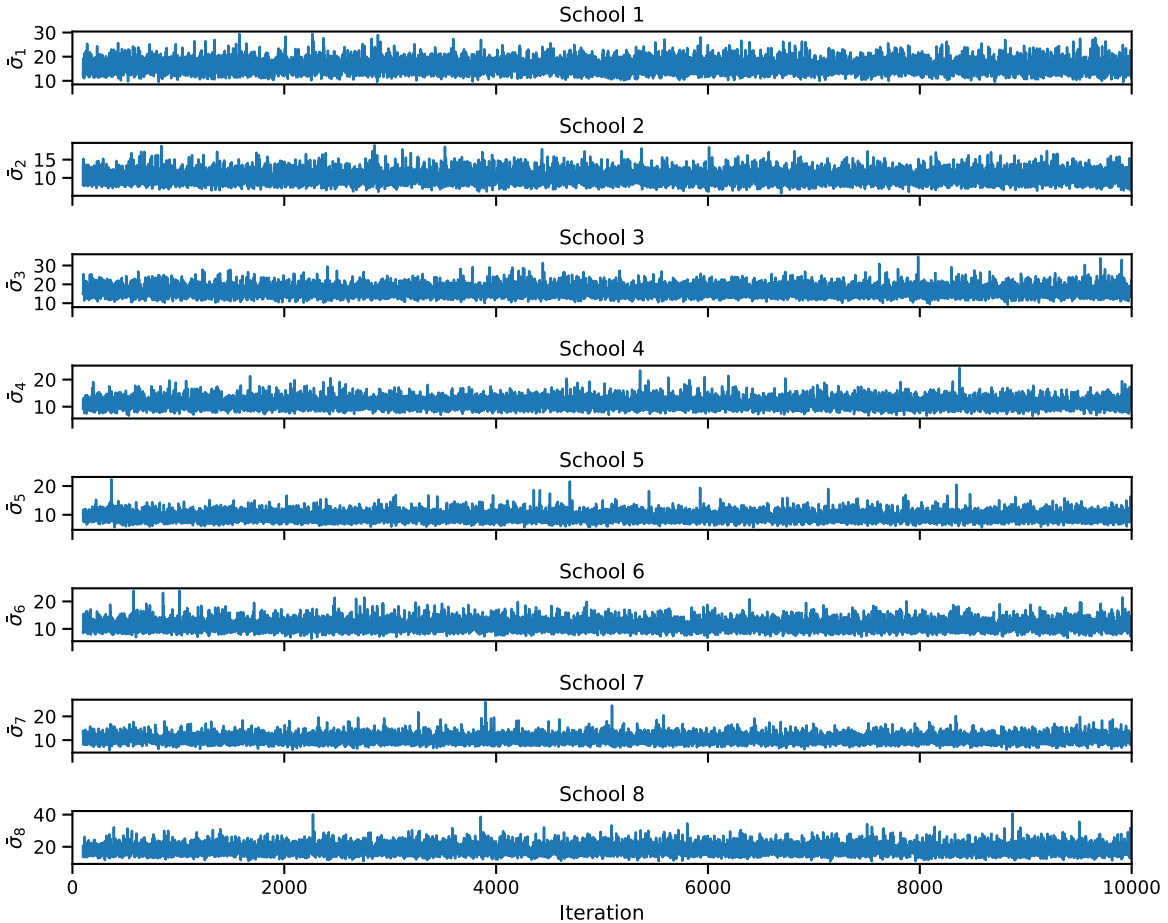


Figure: Traces of samples of σ_s

Gibbs Sampling in the Hierarchical Gaussian Model VIII

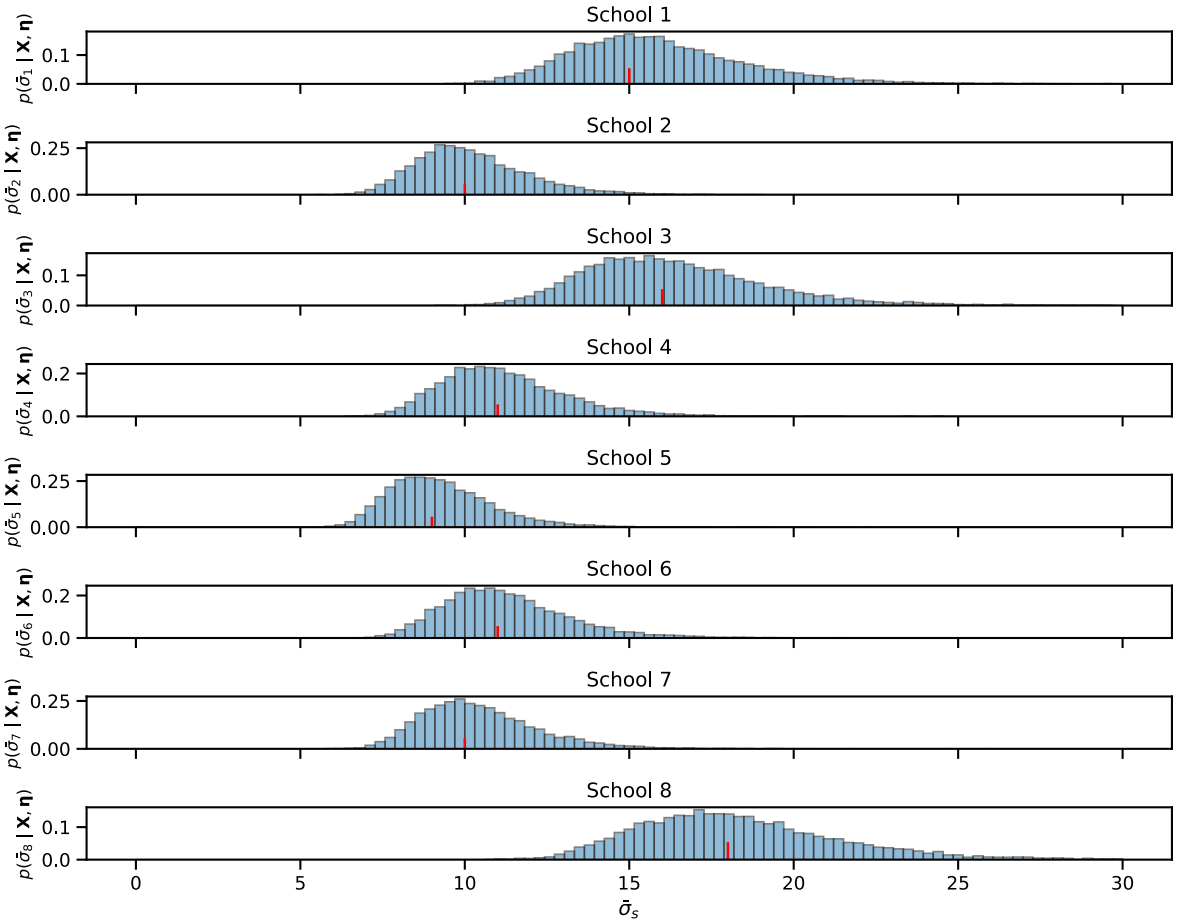


Figure: Histogram of samples of σ_s

Metropolis-Hastings within Gibbs

What if we cannot exactly sample the conditional distribution for some coordinates?

We can mix and match Gibbs and MH updates as long as each update preserves the stationary distribution and the collection of transitions forms an ergodic chain.

Example

Suppose we put a non-conjugate prior on τ^2, something like $\log \tau^2 \sim \mathcal{N}(0, 1)$. Then the conditional distribution, $p(\tau^2 \mid \mu, \{\theta_s, \sigma_s^2\}_{s=1}^S, \boldsymbol{X})$, could not be sampled exactly. However, we could apply an MH update to τ^2 and Gibbs to update other variables.

Part of the "art" of applied Bayesian statistics is designing MCMC transitions to effectively sample the posterior distribution, leveraging model structure (exact conditionals, differentiability, etc.) where possible.

Trace of the Log Joint Probability

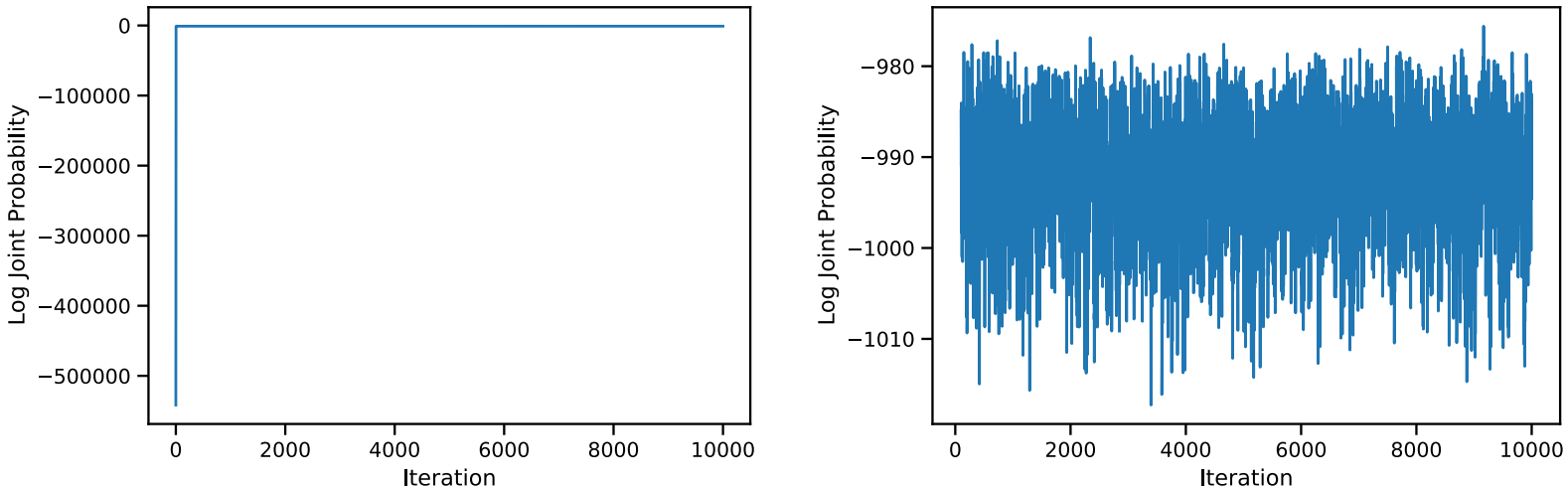


Figure: Log probability of all samples (left) and samples 100+ (right)

Autocovariance and autocorrelation

Recall that,

$$\mathrm{Var}[\hat{f}] = \frac{1}{M^2}\left(\sum_{m=1}^{M}\mathrm{Var}[f(\boldsymbol{\theta})] + 2\sum_{1\le m<m'\le M}\mathrm{Cov}[f(\boldsymbol{\theta}_m), f(\boldsymbol{\theta}_{m'})]\right) \tag{32}$$

$$\approx \frac{1}{M}\left(\mathrm{Var}[f(\boldsymbol{\theta})] + 2\sum_{\ell=1}^{M}\mathrm{Cov}[f(\boldsymbol{\theta}_m), f(\boldsymbol{\theta}_{m+\ell})]\right) \tag{33}$$

since the covariance is only a function of the lag ℓ once the chain has reached stationarity.

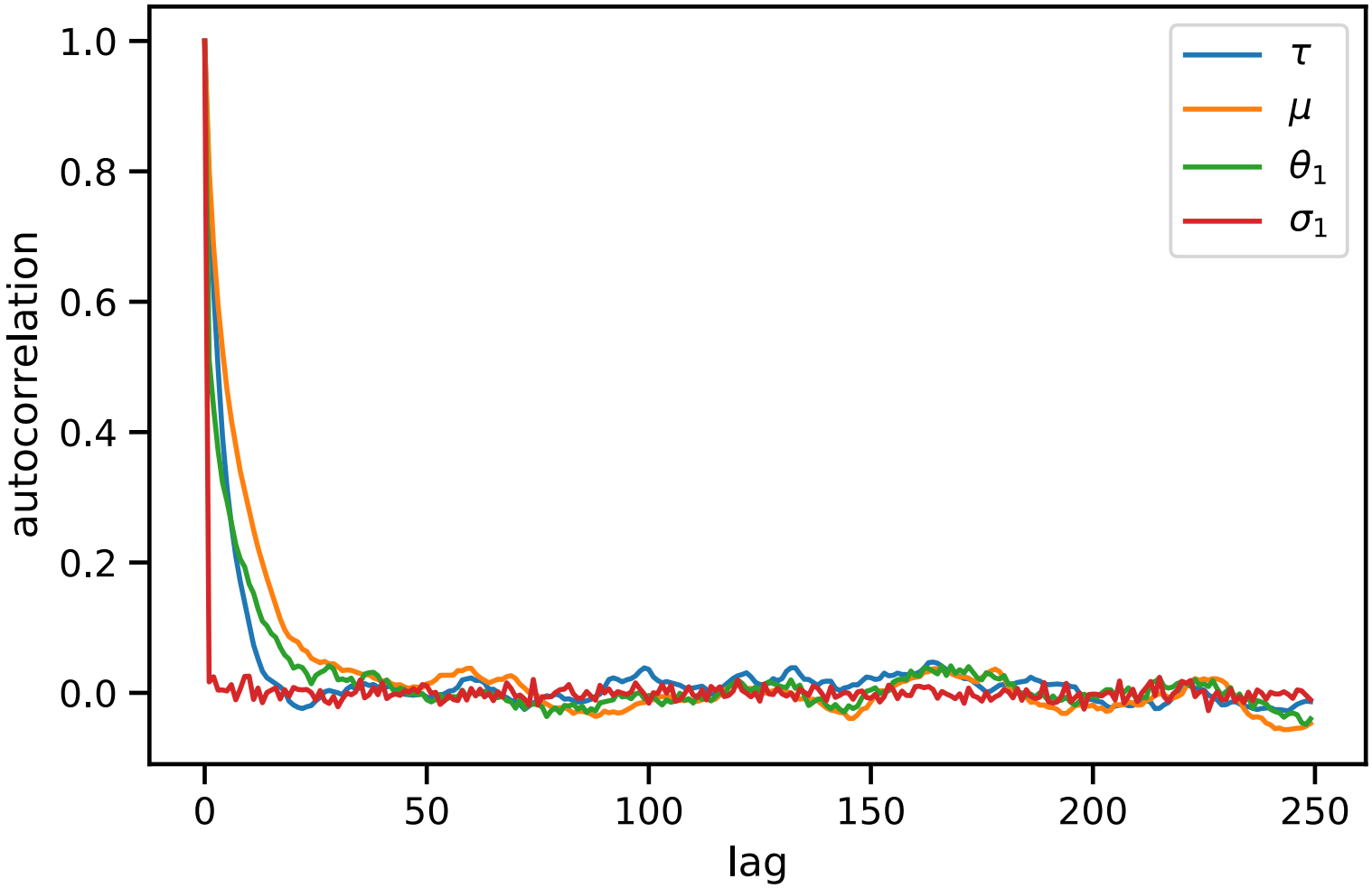
Note: At stationarity, the samples are identically distributed but still correlated!

$\mathrm{Cov}[f(\boldsymbol{\theta}_m), f(\boldsymbol{\theta}_{m+\ell})]$ is called the *autocovariance*. It's a function of the lag ℓ (and the function f).

The *autocorrelation function* (ACF) is defined as, $\mathrm{acf}_f(\ell) = \mathrm{Cov}[f(\boldsymbol{\theta}_m), f(\boldsymbol{\theta}_{m+\ell})]/\mathrm{Var}[f(\theta)]$ so

$$\mathrm{Var}[\hat{f}] \approx M^{-1}\,\mathrm{Var}[f(\boldsymbol{\theta})]\left(1 + 2\sum_{\ell=1}^{M}\mathrm{acf}_f[\ell]\right). \tag{34}$$

Autocorrelation Plots



Effective sample size

The *effective sample size* (ESS) approximates the effective number of independent samples you get from an autocorrelated chain, in terms of the variance of the Monte Carlo estimate,

$$M_{\text{eff},f} = M \frac{\operatorname{Var}[f(\theta)]}{\operatorname{Var}[f(\theta)](1 + 2\sum_{\ell=1}^{\infty} \operatorname{acf}_f[\ell])} = \frac{M}{1 + 2\sum_{\ell=1}^{\infty} \operatorname{acf}_f[\ell]} \tag{35}$$

and we let M_{eff} denote the ESS of the identity functional $f(\theta) = \theta$.

You have to be a bit careful when estimating the ESS—for large values of ℓ the sample correlation is too noisy. Typically, we stop when the sample acf is negative. See Section 11.5 of BDA3 for more details.

In practice, there are already good implementations in Python (c.f. `pyro.ops.stats.effective_sample_size`) and R (c.f. the `coda` package).

Effective Sample Size of the Gibbs Sampler for the Hierarchical Gaussian Model

```
Effective sample sizes (in 10000 Gibbs steps):
tausq:      tensor(1563.3734)
mu:         tensor(625.2454)
theta1:     tensor(1027.9817)
sigamsq1:   tensor(8996.0518)
```

Autocorrelation of a Random Walk

[From Geyer [2011]] To get some intuition, consider the following Markov chain,

$$\theta_m = \rho\,\theta_{m-1} + \epsilon_m \tag{36}$$

where $\epsilon_m \sim \mathcal{N}(0, \tau^2)$.

MH isn't quite a mean-reverting random walk, but that's not a terrible model.

In this toy example, we can calculate the autocovariance of the identity functional $f(\theta) = \theta$,

$$\mathrm{Cov}[\theta_m, \theta_{m+\ell}] = \rho\,\mathrm{Cov}[\theta_m, \theta_{m+\ell-1}] = \rho^{\ell-1}\mathrm{Cov}[\theta_m, \theta_{m+1}] = \rho^{\ell}\mathrm{Var}[\theta_m] \tag{37}$$

so the autocorrelation function decays geometrically as $\mathrm{acf}(\ell) = \rho^{\ell}$.

At stationarity,

$$\mathrm{Var}[\theta_m] = \mathrm{Var}[\theta_{m+1}] = \rho^2\mathrm{Var}[\theta_m] + \tau^2 \Rightarrow \mathrm{Var}[\theta_m] = \frac{\tau^2}{1-\rho^2} \tag{38}$$

For $\rho^2 < 1$, the stationary distribution exists and is $\mathcal{N}(0, \frac{\tau^2}{1-\rho^2})$.

Autocorrelation of a Random Walk II

Letting $f(\theta) = \theta$, we have,

$$\mathrm{Var}[\hat{f}] = \frac{1}{M}\left(\mathrm{Var}[\theta] + 2\sum_{\ell=1}^{M}\mathrm{Cov}[\theta_m, \theta_{m+\ell}]\right) \tag{39}$$

$$= \frac{1}{M}\cdot\mathrm{Var}[\theta]\left(1 + 2\sum_{\ell=1}^{M}\rho^{\ell}\right) \tag{40}$$

$$\approx \frac{1}{M}\cdot\mathrm{Var}[\theta]\left(1 + 2\frac{\rho}{1-\rho}\right) \qquad \text{(for large } M\text{)} \tag{41}$$

$$= \frac{1}{M}\cdot\mathrm{Var}[\theta]\cdot\frac{1+\rho}{1-\rho} \tag{42}$$

and

$$M_{\text{eff}} = M\cdot\frac{1-\rho}{1+\rho}. \tag{43}$$

As $\rho \to 0$, we recover ordinary Monte Carlo. As $\rho \to 1$, the autocorrelation causes the variance of the estimator to blow up and the effective sample size to go to zero.

What about bias?

The mean squared error (MSE) of the estimator is determined by both the bias and the variance.

Ordinary Monte Carlo estimates are unbiased by construction, but MCMC estimates are only *asymptotically unbiased*.

Bias is introduced whenever the initial distribution $\pi_1(\boldsymbol{\theta})$ differs from the stationary distribution; i.e. in all practical cases!

Fortunately, the bias decays as $O(M^{-1})$ whereas the variance decays as $O(M^{-1/2})$, so asymptotically the MSE is dominated by the variance.

If you want to learn more, see Levin and Peres [2017], Meyn and Tweedie [2012], and STATS 318.

References I

Charles J Geyer. Introduction to Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo*, 20116022:45, 2011.

David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.